JCTC Journal of Chemical Theory and Computation

Letter

pubs.acs.org/JCTC

Improved Statistical Sampling and Accuracy with Accelerated Molecular Dynamics on Rotatable Torsions

Urmi Doshi and Donald Hamelberg*

Department of Chemistry and the Center for Biotechnology and Drug Design, Georgia State University, Atlanta, Georgia 30302-4098, United States

ABSTRACT: In enhanced sampling techniques, the precision of the reweighted ensemble properties is often decreased due to large variation in statistical weights and reduction in the effective sampling size. To abate this reweighting problem, here, we propose a general accelerated molecular dynamics (aMD) approach in which only the rotatable dihedrals are subjected to aMD (RaMD), unlike the typical implementation wherein all dihedrals are boosted (all-aMD). Nonrotatable and improper dihedrals are marginally important to conformational changes or the different rotameric states. Not accelerating them avoids the sharp increases in the potential energies due to small deviations from their minimum energy conformations and leads to improvement in the precision of RaMD. We present benchmark studies on two model dipeptides, Ace-Ala-Nme and Ace-Trp-Nme, simulated with normal MD, all-aMD, and RaMD. We carry out a systematic comparison between the performances of both forms of aMD using a theory that allows quantitative estimation of the effective number of sampled points and the associated uncertainty. Our results indicate that, for the same level of acceleration and simulation length, as used in all-aMD, RaMD results in significantly less loss in the effective sample size and, hence, increased accuracy in the sampling of $\varphi-\psi$ space. RaMD yields an accuracy comparable to that of all-aMD, from simulation lengths 5 to 1000 times shorter, depending on the peptide and the acceleration level. Such improvement in speed and accuracy over all-aMD is highly remarkable, suggesting RaMD as a promising method for sampling larger biomolecules.

INTRODUCTION

Molecular dynamics (MD) is now commonly used as a complementary tool to experiments in understanding the dynamical behavior of biomolecules. Among advances in force fields[1−4] and computational power,[5−7] this has been made possible due to developments in sampling techniques. Advanced sampling techniques, such as those discussed in several references,[8−24] allow representation of relevant conformational space and access to long time scale motions of biological interest, not observed in all-atom conventional MD. Although the method of choice depends largely on the particular question in hand, there remain general problems while studying biomolecules, which typically have a complex and hyper-dimensional energy landscape. While some methods may require prior knowledge of the landscape, others may lose the kinetic information about transitions between conformational states. Techniques that involve modification of the Hamiltonian or simulation at elevated temperatures for quicker sampling often run into reweighting issues to obtain the correct canonical Boltzmann distribution of the original Hamiltonian or at the desired ambient temperature.[25] The more aggressive the alteration of the Hamiltonian, the faster is the conformational sampling but the larger the error on recovered properties of the original landscape. Therefore, one has to make a trade-off between the level of desired accuracy and the speed of conformational sampling. The reason for larger inaccuracies in retrieving true statistics is the reduction in the effective number of sampled points due to dominance of some points with very large weights. To explain this further, we will utilize the context of accelerated MD.[18]

In accelerated MD, a continuous and non-negative bias potential is added to the original potential only when it is lower than the preset boost energy such that the potential energy wells are elevated whereas the transition state regions remain unaltered. Simulating on such a modified potential significantly increases the rate of escape from potential energy wells, resulting in better and faster sampling of the conformational space. Each configuration is reweighted by the factor $\exp[\beta\Delta V(\mathbf{r})]$, where $\Delta V(\mathbf{r})$ is the value of the bias potential at position $\mathbf{r}$ and given by $\Delta V(\mathbf{r}) = (E - V(\mathbf{r}))^2/(\alpha + (E - V(\mathbf{r})))$, where E, the boost energy, and α, the shape parameter, determine the magnitude of the bias and, thereby, the extent of acceleration. Accurate calculation of probability distributions and free energy profiles (or any other thermodynamic or kinetic properties) corresponding to the unmodified potential depends on the distribution of weights. For example, let us suppose that there are 10 points sampled in a bin of conformational space defined by $\mathbf{r}$, and all points have equal weights, then the effective number of sampled points will always be 10, as is the case in unbiased simulations. However, in a scenario typical of biased simulations, if each point has a different weight such that there are six data points with negligible weights and four data points with $\exp[\beta\Delta V]$ values that make up more than 99% of the total weight, the effective number of points sampled in this well would then decrease to roughly four and thus reduce the precision of the estimated probability distributions and free energies. It is, therefore, important to gauge the errors introduced in enhanced sampling techniques due to the added bias.

Received: May 23, 2012
Published: August 30, 2012

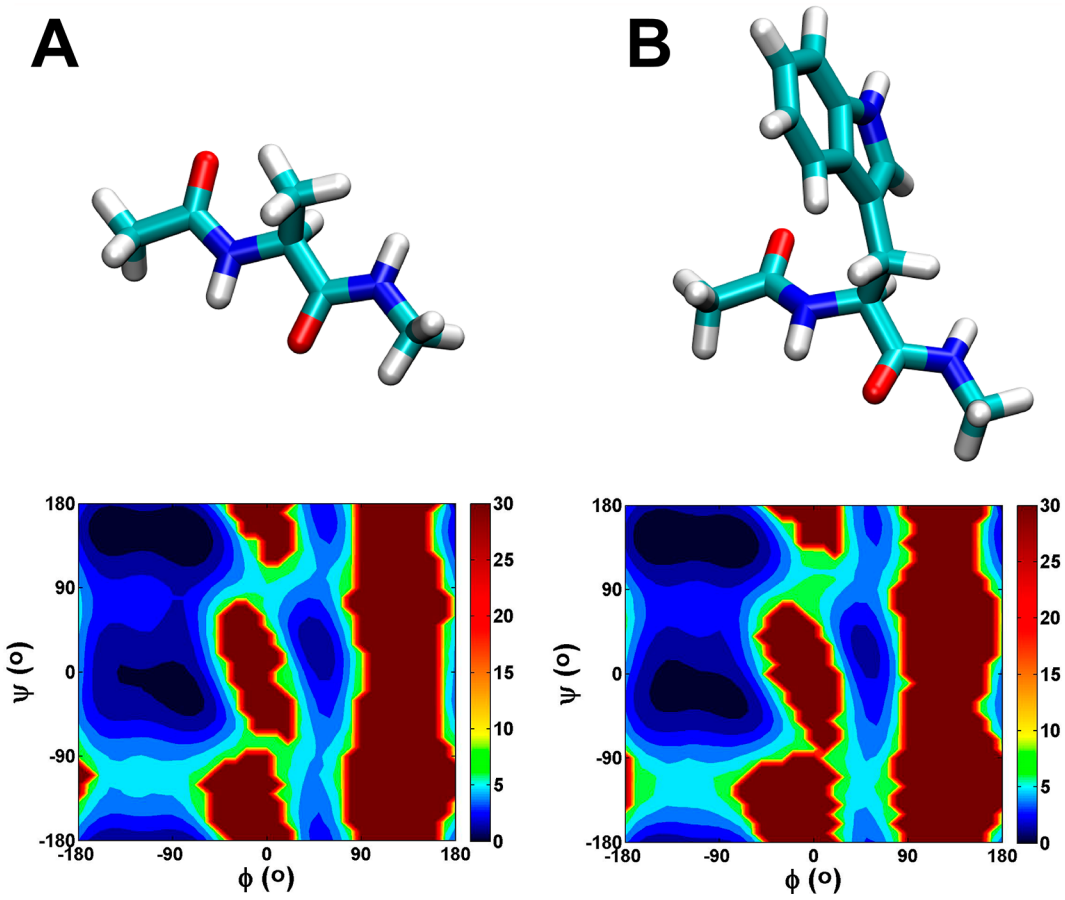

Figure 1. Structures of (A) Ace-Ala-Nme and (B) Ace-Trp-Nme. (Lower panel) Free energy landscapes projected onto ϕ andψ of Ala (left) and Trp (right) obtained from 260 ns normal MD. Contour lines are every 1.2 kcal/mol. Data were collected every 0.01 ps. These 2-D free energy profiles were generated with a resolution of 10° × 10° per bin.

In the original implementation of accelerated MD[18] (aMD), a bias potential was added to the total potential energy such that all the dihedrals as well as 1−4 nonbonded interactions were boosted. However, the total potential energy increases with system size, usually requiring a larger magnitude of bias potential to achieve adequate acceleration, which, in turn, results in significant errors due to exponential reweighting of large varying statistical factors. To obtain better convergence, running larger biomolecular systems for much longer times, therefore, becomes necessary, but is not always feasible. The general approach to address this problem has been to selectively boost the relevant (e.g., all dihedral angles as in all-aMD[26] or torsions of choice belonging to the same or different molecules in a system[27]) or the desired (e.g., part of the system such as only the substrate in an enzyme−substrate complex[28,29] or a particular loop in a protein) degrees of freedom while simulating the rest of the system with conventional MD. Such an approach, which has been previously adopted in many different flavors of aMD,[26,27,30] allows improved statistics and more accurate estimation of the free energy profile projected onto the chosen degree(s) of freedom. However, rather than the strategy of reducing the magnitude of statistical weights, it is important to examine the distribution of statistical factors, which influences the accuracy of retrieving the correct canonical properties, as we will see in the Theory section.

Here, we propose an aMD approach in which a bias potential is added only to the torsional potential of all rotatable dihedrals, which are degrees of freedom most pertinent to conformational changes in biomolecules. The nonrotatable torsions such as the ones in the ring systems and the improper torsions that are mainly responsible for maintaining the planarity do not alter significantly. They seldom contribute to conformational changes and are, therefore, not subjected to acceleration. We implement our approach and test it on two tractable model systems—N-acetyl-N′-methyl alanine amide (alanine dipeptide; Ala-dp) and N-acetyl-N′-methyl tryptophan amide (tryptophan dipeptide; Trp-dp) (Figure 1).

THEORY

Recently, Shen and Hamelberg developed a theory[25] in order to quantitatively assess the relation between the potential loss in the number of sampled points after reweighting and the reduction in accuracy in the retrieved original statistics. To make this text self-explanatory, we detail this theory here.

i. Statistical Uncertainty in Free Energy Associated with Sampling Size. According to the basic statistical theory of ensembles, the variation in a large and independent sampling size, N, is on the order of $\sqrt{N}$.[31] Therefore, the magnitude of the uncertainty, ε, in the free energy at a particular point $\mathbf{r}$, $F(\mathbf{r}) = -k_B T \ln N(\mathbf{r})$, is approximately given by $\varepsilon \approx k_B T(\ln(N(\mathbf{r}) \pm (N(\mathbf{r}))^{1/2}) - \ln(N(\mathbf{r})))$. Rearranging this equation relates the error, ε, associated with a given independent sampling number N at $\mathbf{r}$ as

$$N(\mathbf{r}) = 1/(e^{\varepsilon/k_B T} - 1)^2 \quad (1)$$

dx.doi.org/10.1021/ct3004194 | J. Chem. Theory Comput. 2012, 8, 4004−4012

Generally speaking, increasing the independent sampling number N will decrease the error; however, it will not guarantee the sampling of all relevant areas of conformational space. Increasing N may simply populate the already sampled regions even more and not represent other areas of interest at all. Therefore, depending on the system at hand, one has to decide up to what level of free energy sampling is desired before estimating the associated statistical uncertainty.

ii. Estimating Sampling Size of Regions of Conformational Space with the Highest Desired Level of Free Energy. Let us consider a region or a bin x up to which one desires to sample, i.e., assuming it to be the least sampled region. The idea is to determine how long one should sample so as to obtain the representation of bin x. Once region x is sampled, it is assured that other regions of relatively lower free energy will be sampled better or as well. Also, if we can estimate the uncertainty associated with the least sampled region, we can obtain the upper limit of error for other regions of phase space. Now, let us only consider the least and the most sampled bins such that $f_x = N_x/N_o$, which is the fraction of data points sampled in x relative to the most sampled bin, where $f_x < 1$. N_x is the actual number of points sampled in bin x, whereas N_o refers to the sampling in the reference state, which is usually the free energy minimum and set to zero. The free energy of bin x, relative to the most sampled bin, is therefore $\Delta F_x = -k_B \ln(N_x/N_o) = -k_B T \ln f_x$. Rearranging these terms yields

$$f_x = e^{-\Delta F_x/k_B T} \tag{2}$$

If an MD simulation is run for a total length of τ_{sim} time (or N_{sim} steps) and τ_{corr} is the correlation time of an observable of interest, the total number of independent sampling values for that observable can be estimated as $N^* = (\tau_{sim}/\tau_{corr})$. τ_{corr} is the simulation time for the data values of that observable to become uncorrelated with their previous values. Alternatively, N^* can be expressed in terms of total number of simulation steps, N_{sim}, and the inverse of data collection frequency, n, i.e., $N^* = (N_{sim}/n)$. To ensure that data collected are uncorrelated, the condition $n \geq (\tau_{corr}/(dt))$ should be met. Here, dt is the integration time step. Now, considering that the most populated bin has a sampling size $N_o \approx N^*$, we can determine the uncorrelated sampling size of region x at a particular level of ΔF_x as

$$N_x = N^* \times f_x = N^* \times e^{-\Delta F_x/k_B T} \tag{3}$$

Equation 3 and the above expressions allow us to estimate how long one should run unbiased simulations and how frequently one should collect data to sufficiently sample bin x, i.e. $N_x \gg 1$.

iii. Estimating Effective Sample Size in Biased Simulations. In general, N_x can be expressed as $\sum_{i=1}^{N_x}(w_i/w_h)$, where weight $w_i = \exp(s_i)$ of the ith sample point is normalized by the highest statistical weight w_h and $s_i = [\beta \Delta V_i(\mathbf{r})]$. In the case of normal MD, since $w_i = w_h = 1$, $\sum_{i=1}^{N_x}(w_i/w_h) = \sum_{i=1}^{N_x} 1 = N_x$. In accelerated MD, $w_i < w_h$ and, therefore, $\sum_{i=1}^{N_x}(w_i/w_h) = N_e$, where N_e is defined as the effective number of sampled points in bin x. Data points with weights closer to the highest weight will have a larger contribution to N_e. The loss in N_x is then given by $(N_e/N_x) = 1/N_x \sum_{i=1}^{N_x}(w_i/w_h)$ or alternatively as $(N_e/N_x) = \langle w/w_h \rangle = \langle e^s/e^{s_h} \rangle$. This can be further rewritten as

$$\frac{N_e}{N_x} = \int_{s_l}^{s_h} e^{s-s_h} p(s)\, ds \text{ or } N_e = N_x \int_{s_l}^{s_h} e^{s-s_h} p(s)\, ds \tag{4}$$

where $p(s)$ is the distribution of statistical reweighting factors $s = \ln w = [\beta \Delta V(\mathbf{r})]$, and s_l and s_h are the smallest and the largest values of these reweighting factors, respectively. $p(s)$ from the whole simulation and not just for bin x will include the largest range of s and yield an upper bound for the loss in sampling size.

iv. Sampling Error in Reweighting-Based Simulations. The effective sampling size for the entire simulation can be approximated from eqs 3 and 4 as $N^* \times f_x \times \int_{s_l}^{s_h} e^{s-s_h} p(s)\, ds$. Substituting eq 2 in this expression and relating it to the uncertainty (eq 1) results in

$$N^* \times e^{-\Delta F_x/k_B T} \times \int_{s_l}^{s_h} e^{s-s_h} p(s)\, ds = \frac{1}{(e^{\varepsilon/k_B T} - 1)^2} \tag{5}$$

The uncertainty associated with the least sampled region x provides the upper limit of the sampling error for the entire conformational space in biased and unbiased simulations.

Equation 5 allows estimation of the uncertainty up to a desired free energy, given the length of the simulation and the distribution of reweighting factors. Or, one can estimate how long one should run the simulation in order to obtain a particular level of accuracy up to a certain free energy. Equation 5 further provides (an additional) general quantitative assessment of convergence (discussed further below), rather than the case-based qualitative measures existing in the field. It should be noted that the above theory is not only limited to MD simulations requiring reweighting. It can be applied to evaluate the accuracy of normal MD simulations in which case the integral in eq 5 would equal unity, and the uncertainty in the free energy values up to a desired free energy level can be related to the independent sampling size. The relative free energy difference between two states is a quantity of wide interest in many experimental and theoretical studies.[32−35] The expected systematic error in free energy differences from finite-length sampling has been presented by Zuckerman and Woolf.[31] According to the above theory, the uncertainty in the free energy difference would be less than 2ε, as ε is the upper bound of error. These estimates refer to the statistical errors from a given length of simulation due to reweighting.

RESULTS AND DISCUSSION

Ala-dp has served as a classic archetypal system to evaluate the performance of sampling techniques in several earlier studies,[18,27,30,37] as it samples very well in routinely carried out simulation lengths (in a vacuum or implicit or explicit solvent) and its free energy landscape can be simply described along $\phi-\psi$ dimensions. In Ala-dp, improper torsions constitute only 6% of the total torsions, whereas nonrotatable dihedrals are totally absent. In contrast to Ala-dp, Trp-dp has a bulky side chain with the nonrotatable torsions of the Trp ring and the improper torsions forming 43% of the total number of dihedral angles. Benchmarking our approach, which we hereafter refer to as RaMD, on these model systems gives us an opportunity to determine the effects of neglecting irrelevant torsions on the accuracy of retrieved equilibrium properties while still simulating the system in its entirety. This has important implications for larger biomolecular systems; i.e., one can accelerate the entire biomolecule instead of its specific regions or parts and yet obtain better statistics. We subjected Ala-dp and Trp-dp to 260 ns of normal MD (nMD) and generated free energy profiles projected onto the $\phi-\psi$ dimensions

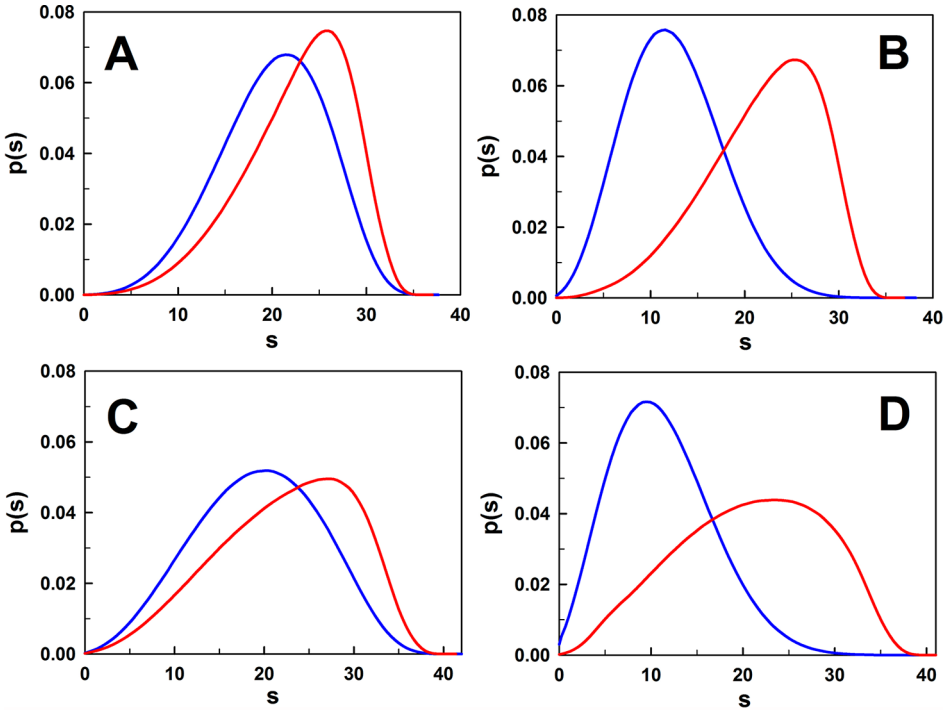

Figure 2. Distributions of reweighting factor s from all-aMD (blue) and RaMD (red) for Ala-dp and Trp-dp using two levels of acceleration. Boost energy E was set at 29 kcal/mol over the average dihedral energy of either all torsions or only rotatable torsions for both Ala-dp and Trp-dp. (Top panel) α = 15 kcal/mol and (lower panel) α = 10 kcal/mol were used for both all-aMD and RaMD. (A) Ala-dp (aMD: E = 40.6 kcal/mol, α = 15 kcal/mol; RaMD: E = 39.95 kcal/mol, α = 15 kcal/mol), (B) Trp-dp (aMD: E = 45 kcal/mol, α = 15 kcal/mol; RaMD: E = 39.7 kcal/mol, α = 15 kcal/mol), (C) Ala-dp (aMD: E = 40.6 kcal/mol, α = 10 kcal/mol; RaMD: E = 39.95 kcal/mol, α = 10 kcal/mol), (D) Trp-dp (aMD: E = 45 kcal/mol, α = 10 kcal/mol; RaMD: E = 39.95 kcal/mol, α = 10 kcal/mol).

Table 1. Comparison of Statistical Accuracies, Total Simulation Lengths, and Speedup of β-to-α_R Dynamics for Ala-dp and Trp-dp in all-aMD and RaMD[a]

	simulation method	α (kcal/mol)	% coverage of ϕ–ψ plots	n[b]	$\int_{s_l}^{s_h} e^{s-s_h} p(s)\, ds$;[c] i.e., N_e/N_x	reduction factor[d]: $N_{\text{all-aMD}}$*/N_{RaMD}*	$\tau_{\beta-\alpha}(\psi)$ (ps)	ε[e] (kcal/mol)
Ala-dp	**nMD**		68.1	6.55			81.8	0.02
	all-aMD	15	87.7	5.92	1.56×10^{-4}	4.7	24.5	0.76
	V_{avg} = 11.6 kcal/mol	10	89.4	5.68	3.29×10^{-5}	10.2	28.4	1.12
	66 torsions	5	93.8	6.06	1.42×10^{-5}	4.1	25.4	1.35
	RaMD	15	87.3	6.1	7.55×10^{-4}		26.6	0.46
	V_{avg} = 10.95 kcal/mol	10	89.2	6.33	3.74×10^{-4}		29.8	0.58
	62 torsions	5	93.9	6.28	5.98×10^{-5}		29.7	0.96
Trp-dp	**nMD**		66.1	6.5			289.8	0.02
	all-aMD	15	82.2	5.85	2.46×10^{-6}	296	72.7	1.84
	V_{avg} = 16 kcal/mol	10	81.9	5.85	1.26×10^{-6}	315	71.6	2.04
	119 torsions	5	87.8	5.4	6.05×10^{-8}	1070	62.1	2.94
	RaMD	15	87.4	6.1	7.6×10^{-4}		83.7	0.46
	V_{avg} = 10.7 kcal/mol	10	90.1	5.8	3.94×10^{-4}		79.3	0.57
	68 torsions	5	94.1	6.15	7.36×10^{-5}		75.4	0.94

[a] E is set to 29 kcal/mol above the V_{avg} in each case. [b] $n = \tau_{corr}/dt$ where τ_{corr} is the bin dwell time in fs, longer than that for which the data collected becomes uncorrelated, and dt is the integration time step in fs. Data collected less than every n steps will be correlated. [c] Equation 4, which estimates the loss in the effective number of sampled points. [d] Reduction factor is the ratio of the total number of independent sampling values from all-aMD to that from RaMD. It is obtained by rearranging eq 5 and assuming the same level of accuracy for both all-aMD and RaMD and sufficient sampling of regions below contour lines corresponding to the same free energy level, i.e., 6 kcal/mol. $\tau_{\beta-\alpha}(\psi)$ is the escape time from the β region into the α_R-helical well along the ψ dihedral. [e] ε is the error as defined in eq 5 and was estimated from simulations with $N_{sim} = 1.3 \times 10^8$ steps and n steps that were different for nMD, all-aMD, and RaMD. The error calculated is for free energy values up to 6 kcal/mol. Listed also are the number of torsions boosted in each case.

(Figure 1). For both the dipeptides, all the relevant α-helical and β-strand regions were well-sampled and the smoothness of the contour lines suggested relatively less statistical error up to ~6.0 kcal/mol. Next, we simulated Ala-dp and Trp-dp with all-aMD, boosting all the torsions (excluding the 1−4 interactions), and with RaMD in which only rotatable torsions were accelerated. While assessing the performance of different methods, it is important to compare the quality of sampling

dx.doi.org/10.1021/ct3004194 | J. Chem. Theory Comput. 2012, 8, 4004−4012

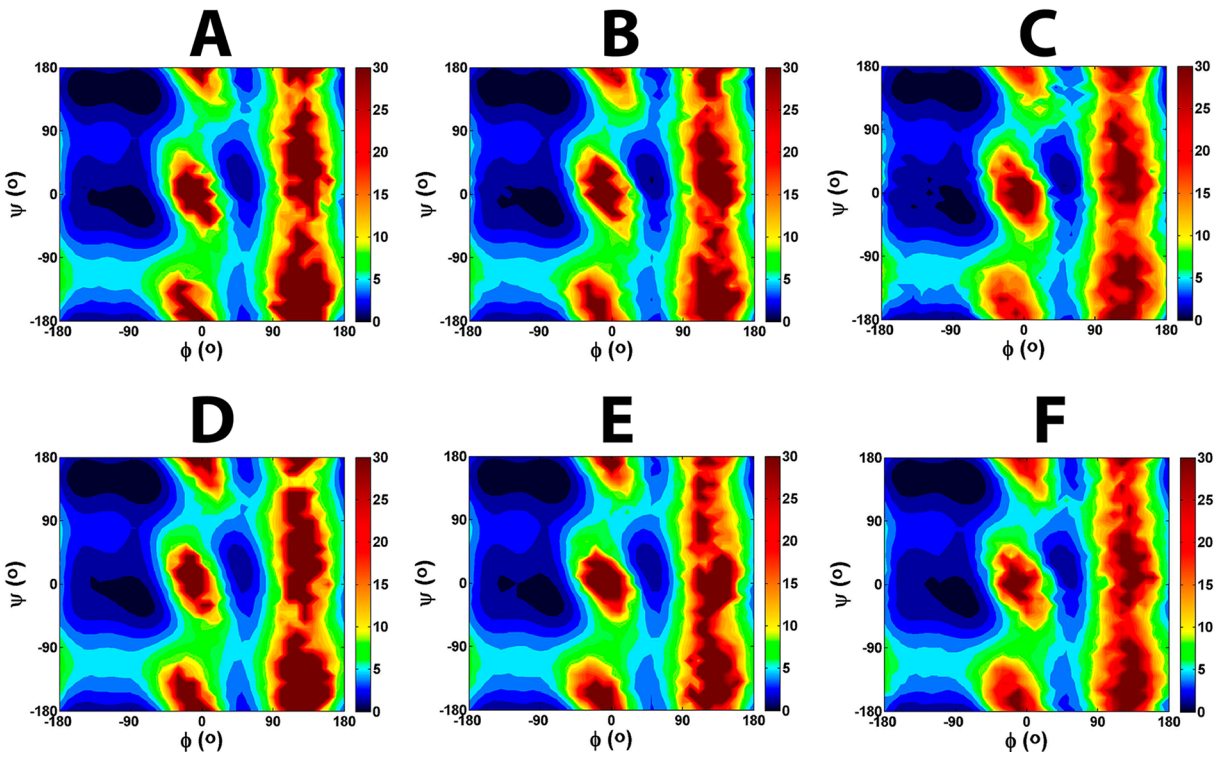

Figure 3. Free energy landscapes of Ala-dp obtained from all-aMD (top panel) and RaMD (lower panel) using three increasing levels of acceleration. (A) all-aMD: α = 15 kcal/mol. (B) all-aMD: α = 10 kcal/mol. (C) all-aMD: α = 5 kcal/mol. (D) RaMD: α = 15 kcal/mol. (E) RaMD: α = 10 kcal/mol. (F) RaMD: α = 5 kcal/mol. For all-aMD, E = 40.6 kcal/mol, and for RaMD, E = 39.95 kcal/mol was set. Contour lines are every 1.2 kcal/mol. These $\phi-\psi$ plots were obtained from a simulation length of 260 ns for each case with data collected every 0.01 ps, yielding 2.6 × 10^7 steps and sorted in a bin size of 10° × 10°.

resulting from each method and the statistical uncertainty in computing the equilibrium properties.[38] In this work, we computed ε, the uncertainty in free energy, from the ensemble average of reweighting factors and the independent sampling size, as given in eq 5. Figure 2 shows the distributions of reweighting factors for each dipeptide obtained from both all-aMD and RaMD, using two different levels of acceleration. Intriguingly, as compared to all-aMD, RaMD resulted in $p(s)$'s that were shifted to higher statistical factors in spite of not boosting all the torsions. The reweighting factors got closely clustered near s_h such that the difference between the mean and s_h reduced. The increase in the proportion of reweighting factors that became relatively more similar to s_h led to their higher contribution to the integral in eq 5, resulting in relatively higher N_e/N_x (i.e., less loss in sampled points) from RaMD than from all-aMD (Table 1). Therefore, it is not only the dominance of few points with higher statistical weights that causes the reweighting problem. The effective number of sampled points and hence the accuracy of a simulation, in fact, depends on how similar the boost factors are to the highest one. During molecular dynamics simulations, high-energy conformations are sometimes visited due to ring puckering and out-of-plane motions of improper torsions, thus giving rise to spikes in the potential energy.

Therefore, the average total torsional potential energy, V_{avg}, will be slightly higher (Table 1) than the true minimum, obligating a relatively larger bias potential in all-aMD. Also, conformations sampled in a well with potential energy very close to the global or a local minimum will experience high boost, and those with higher energy will be minimally accelerated; i.e., ΔV will be close to zero. The sudden, frequent variation in the potential energy thereby results in a distribution of reweighting factors with a larger deviation from s_h in all-aMD, since the improper and nonrotatable torsions are easily boosted to higher energies. Neglecting the improper and the nonrotatable ring torsions as in RaMD helps in avoiding such potential energy spikes and gives rise to $p(s)$ that is more lopsided toward s_h (Figure 2).

There is ambiguity in the meaning of 'sampling'. It refers to 'conformational' sampling when the representation or population of different regions of phase space is concerned. Sampling also means the statistical noise or error introduced in the recovery of equilibrium properties. To assess the sampling error, we next compared the free energy profiles generated from the different setups of all-aMD and RaMD for Ala-dp (Figure 3) and Trp-dp (Figure 4). The smoothness of the free energy contour lines indicates the statistical error. Both, all-aMD and RaMD sampled the relevant conformations, i.e., essentially four wells, as seen in the $\phi-\psi$ plots from nMD of each dipeptide. Despite the fact that all-aMD and RaMD gave more coverage in the $\phi-\psi$ space, the reweighted free energy profiles were relatively rougher.

As the level of acceleration increased, the modified potential deviated more from the original potential, giving rise to rougher free energy profiles with larger statistical error. Consistently for each dipeptide and level of acceleration, we found that RaMD yielded smoother free energy profiles than those from all-aMD. Comparing the performance of all-aMD and RaMD in recovering the free energy profiles, the decrease in errors was most dramatic in the case of Trp-dp, especially for α = 5 kcal/mol (Figure 4). The reason for the outperformance of RaMD can be justified using the expression in eq 5. From the

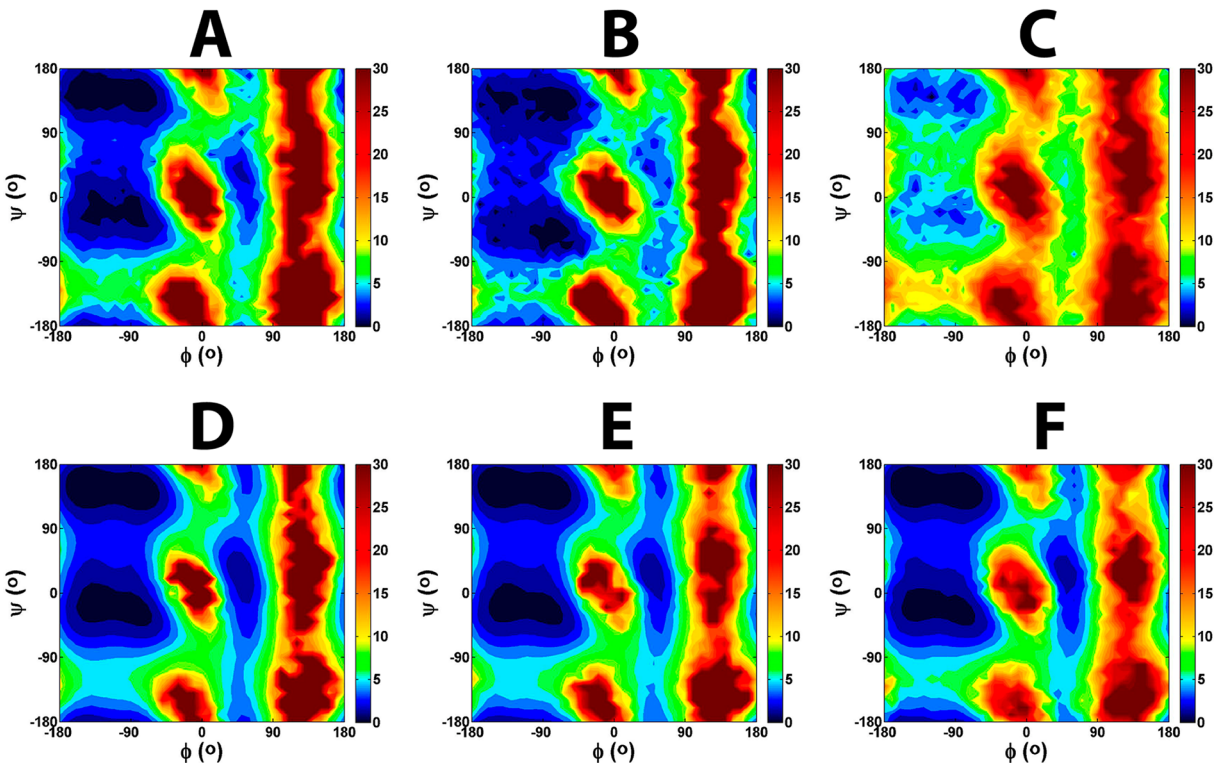

Figure 4. Free energy landscapes of Trp-dp obtained from all-aMD (top panel) and RaMD (lower panel) using three increasing levels of acceleration. (A) all-aMD: α = 15 kcal/mol. (B) all-aMD: α = 10 kcal/mol. (C) all-aMD: α = 5 kcal/mol. (D) RaMD: α = 15 kcal/mol. (E) RaMD: α = 10 kcal/mol. (F) RaMD: α = 5 kcal/mol. E was set to 45 kcal/mol for all-aMD and to 39.7 kcal/mol for RaMD. Contour lines are every 1.2 kcal/mol. These $\phi-\psi$ plots were obtained from a simulation length of 260 ns for each case with data collected every 0.01 ps, yielding 2.6×10^7 steps and sorted in a bin size of $10° \times 10°$.

distributions of reweighting factors, we calculated the loss in the effective number of independent sampled points (Table 1), which was reduced considerably in all setups of RaMD as compared to the corresponding ones of all-aMD for the same level of acceleration. Plugging the values of the integral, total simulation length, and correlation time in terms of the number of steps in eq 5, we estimated ε for each setup for free energy up to 6 kcal/mol. Although we ran RaMD and all-aMD for the same amount of time (corrected for the differences in the correlation time), the higher effective sample sizes in case of RaMD resulted in the reduction in the overall error ε. Alternatively, we determined what the reduction in simulation lengths would be for the same level of accuracy acceptable up to a certain free energy level for both all-aMD and RaMD. We found that for the levels of acceleration used in this study, the simulation lengths reduced by 5−10 times in the case of Ala-dp, while for Trp-dp the simulation lengths decreased by 300−1000 times. This meant that convergence could be reached 5−1000 times faster in RaMD than in all-aMD, depending on the composition of the peptide and extent of acceleration. Since Ala-dp lacks nonrotatable torsions in the side-chain, it serves as a good model system for understanding the separate effects of neglecting only the backbone improper torsions in aMD and, thus, provided the lower limit of RaMD performance. It is amazing that by ignoring only 6% (four improper torsions out of 66 total dihedrals) of the total dihedrals in Ala-dp, such notable improvement is observed in the statistics. Using Trp-dp, on the other hand, provided the additional and much larger speedup from not accelerating nonrotatable torsions of the side-chain and most probably yielded the upper limit. As the omission of nonrotatable dihedrals resulted in better performance than improper torsions, RaMD is expected to provide between a 5 to 1000 times speedup in sampling other amino acids, of which more than half have side chains with improper and/or nonrotatable torsions and of larger peptides and proteins containing a combination of different amino acids. It should also be noted that as compared to nMD, it appeared as though we traded the accuracy to considerably greater coverage of $\phi-\psi$ space in both forms of aMD. The error of 0.02 from nMD seems obviously small as there is no need for reweighting in this case, but it may be a false sense of good statistics since sampling might be limited to regions that are separated by low energy barriers.

There is always a trade-off between the statistical error, which depends on how much the target system is altered and the sampling speed is increased, and the simulation length. Therefore, how much to accelerate and for how long one should run the simulations to sample conformations up to a required level of free energy and up to what accuracy should be decided on the basis of the particular question at hand. As the number of torsions boosted was significantly lower, the acceleration per degree of freedom was effectively higher in the case of RaMD than in all-aMD, despite the use of the same level of acceleration for both forms of aMD. If we were to use acceleration per degree of freedom equal to that in all-aMD, we would obtain more accurate results from RaMD than those reported in Table 1.

Although from Figure 1 it appears that Ala-dp and Trp-dp have sampled the relevant conformational states with less statistical error in nMD, it does not mean that the nMD simulations have reached convergence. While comparing the performance of sampling methods or force fields, it is essential

dx.doi.org/10.1021/ct3004194 | J. Chem. Theory Comput. 2012, 8, 4004−4012

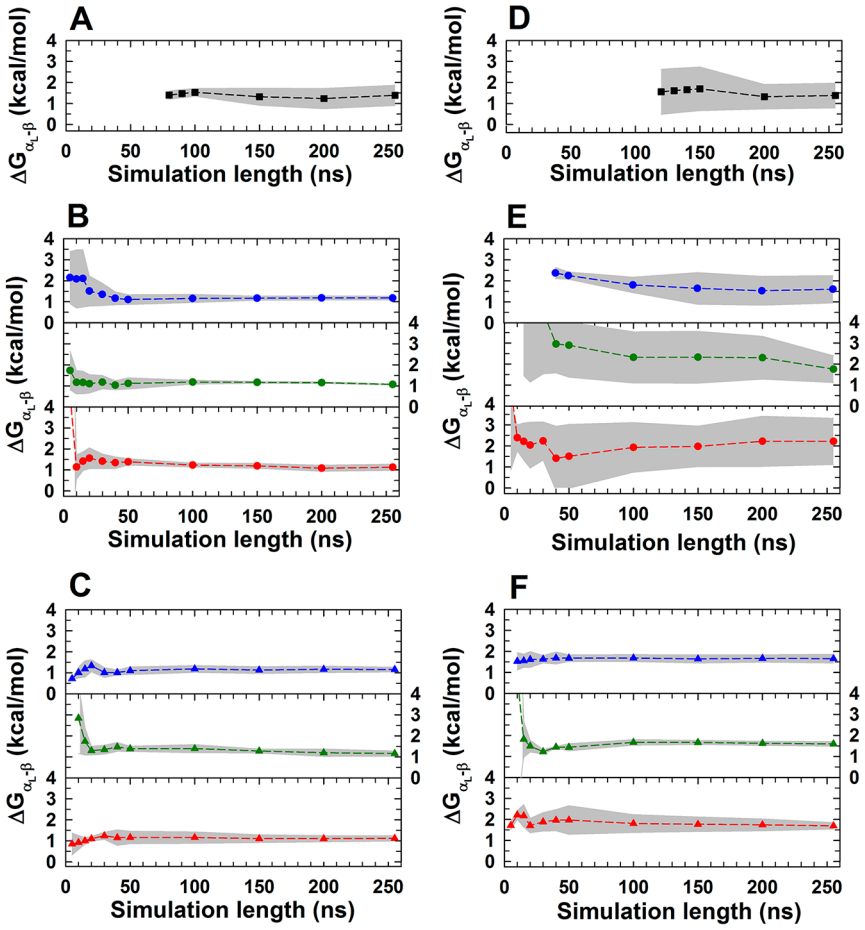

Figure 5. Free energy difference between α_L-helical and extended β-sheet regions as a function of simulation lengths. Shown are the plots for Ala-dp simulated with (A) nMD, (B) all-aMD, and (C) RaMD and Trp-dp simulated with (D) nMD, (E) all-aMD, and (F) RaMD using various levels of acceleration: α = 15 kcal/mol (blue), α = 10 kcal/mol (green), and α = 5 kcal/mol (red). Free energy differences from three independent MD runs of 260 ns each are averaged and shown with their respective standard deviations (gray filled areas).

to define the criteria for convergence and test for them, as convergence is a relative term with no standard meanings. This issue has been discussed in great length elsewhere.[24,39] Convergence usually refers to adequate conformational sampling such that the correct relative populations of the various states (or some observable of interest) have equilibrated. To assess the quality of nMD, all-aMD, and RaMD, in terms of sampling speed and convergence, we investigated the following properties: (a) the escape rates from the β- to the α_L-helical region, $\tau_{\beta-\alpha}(\psi)^{-1}$ (Table 1), (b) free energy difference, $\Delta G(\alpha_L-\beta)$, between the extended β-strand and α_L-helical regions as a function of simulations lengths (Figure 5), and (c) percent coverage of $\phi-\psi$ space at different simulation lengths (data not shown). Expectedly, both forms of aMD showed significant speedup in the $\beta-\alpha$ transition rates as compared to nMD. However, there was only a slight slow-down in the $\beta-\alpha$ dynamics in RaMD relative to all-aMD. As extended β strand and left-handed α-helical regions are separated by a relatively higher barrier, their relative population serves as a good metric for testing sampling as well as convergence. As compared to Ala-dp, Trp-dp with a bulky side chain was more difficult to sample, and this was evident from the longer times required to reach equilibrated values of $\Delta G(\alpha_L-\beta)$ as shown in Figure 5 and percent coverage of $\phi-\psi$ space (data not shown). On average, nMD did not start sampling the α_L-helical region until 80 ns for Ala-dp and 120 ns for Trp-dp, not allowing the calculation of $\Delta G(\alpha_L-\beta)$ at shorter simulation lengths (Figure 5A,D). While the $\Delta G(\alpha_L-\beta)$ values for Ala-dp from all-aMD equilibrated between 50 and 100 ns (Figure 5B), those from RaMD converged earlier than 50 ns with fluctuations of <0.5 kcal/mol (Figure 5C) calculated from three independent runs. Such deviations from RaMD were comparable to those from all-aMD. However, for Trp-dp, the fluctuations in $\Delta G(\alpha_L-\beta)$ from all-aMD were noticeably much greater as compared to those from RaMD (Figure 5E and F). RaMD exhibited convergence in terms of equilibration of $\Delta G(\alpha_L-\beta)$ values before 50 ns, which was much earlier than all-aMD. In the case of the least acceleration, i.e., α =15 kcal/mol (Figure 5E), all-aMD had not visited the α_L-helical region until 40 ns, thus marking delayed convergence. In the absence of a reference from infinite sampling, we independently tested the credibility of the converged $\Delta G(\alpha_L-\beta)$ values from RaMD by running two-dimensional umbrella sampling simulations along ϕ and ψ directions for Trp-dp. Our estimates of $\Delta G(\alpha_L-\beta)$ from RaMD closely agreed with those from umbrella sampling, i.e., ~1.7 kcal/mol, suggesting that, indeed, RaMD had reached convergence. Figure 5, thus, clearly demonstrated that RaMD can reproduce the relative populations of α_L-helical and extended β-sheet regions in significantly less time and statistical error than all-aMD.

CONCLUDING REMARKS

We propose RaMD, an accelerated MD approach, in which only the total rotatable torsional potential is subjected to acceleration. Rotatable dihedrals are most pertinent to conformational changes, and considering only them for aMD is a valid approach, given the importance of rotamer distribution and its changes in protein structure prediction and designing of proteins and drugs.[40−43] Our all-aMD and RaMD simulations on Ala-dp and Trp-dp showed that RaMD outperformed all-aMD in three aspects: (1) Improved conformational sampling: for a simple dipeptide like Ala-dp, sampling of conformational states, mainly right- and left-α-helical, β-strand, and polyproline II regions in terms of % coverage was comparable for both all-aMD and RaMD. However, for Trp-dp with a bulky side chain, RaMD was able to sample not only the four main regions mentioned above but also the higher energy regions between these states. (2) Sampling with improved statistics: the precision of aMD approaches depends on the reweighting and the level of alteration of the original Hamiltonian. For three different levels of acceleration, consistently, we found that RaMD produced sampling of $\phi-\psi$ space (i.e., two-dimensional free energy profiles) with significantly less statistical error. RaMD resulted in higher effective sampled data, leading to more accuracy than all-aMD. (3) Faster sampling: with RaMD, Ala-dp and Trp-dp sampled conformational space ∼5−10 and ∼300−1000 times faster, respectively, to certain accuracies identical to those obtained from corresponding all-aMD setups. This enhancement in sampling speed and accuracy could have positive repercussions on simulating larger biomolecules with current limited computational resources. For nucleic acids, especially, that typically have several nonrotatable dihedrals, RaMD could prove to be a very promising sampling method.

COMPUTATIONAL METHODS

All molecular dynamics simulations were carried out using the *pmemd* module of the AMBER10[44] with the modified parm99[45] (ff99SB[2]) force field. Using the *xleap* program, Ala-dp and Trp-dp were solvated with 906 and 731 TIP3P[46] water molecules, respectively, in a cubic periodic box. The solvent was initially allowed to relax by carrying out 5000 steps of minimization with a steepest descent algorithm, and the peptide atoms were restrained with a force constant of 50 kcal/mol·Å^2. Subsequently, 100 ps of MD was performed in which the restraining force constant was reduced to 25 kcal/mol·Å^2, after which another 100 ps MD equilibration step was carried out wherein both water and peptide atoms were allowed to relax. After equilibrating the water density, production runs were carried out in the NPT ensemble at 1 bar of pressure and at 300 K. All bonds involving hydrogens were constrained with SHAKE[47] using a tolerance of 0.0001. A cutoff of 9 Å was used for nonbonded interactions while long-range electrostatic interactions were treated with the particle mesh Ewald[48] method. The temperature was regulated with a Langevin thermostat with a collision frequency of 1 ps^{-1}. The systems were maintained at the reference pressure by coupling to an external bath with a coupling constant of 1 ps. Newton's equations of motions were integrated using a time step of 2 fs. For each setup of normal MD, all-aMD, and RaMD, three independent 260-ns-long production runs were performed, giving a total simulation time of (780 ns × 14 setups) 10.92 μs. Each 260-ns production run was carried out in 26 10-ns segments, restarting the simulations in each segment with a distinct random number seed. For all-aMD and RaMD, the boost energy E was set to 29 kcal/mol above the average dihedral energy of all torsions and only rotatable torsions, respectively. α was set to three different values: 5, 10, and 15 kcal/mol for all-aMD as well as RaMD. Out of the 260-ns trajectory for each setup, the first 5 ns of data were discarded as further equilibration, and data were collected every five steps for the rest of the simulation, i.e., every 0.01 ps. ϕ and ψ dihedrals were collected from the production runs and binned into a two-dimensional $\phi-\psi$ plot containing 36 × 36 10° regions. The $\phi-\psi$ plot was converted into free energy profiles by $\Delta G_i = -RT \ln(N_i)$ where $N_i = N_x$, i.e., the actual number of hits in bin i from nMD or $N_i = N_e$, i.e., the sum of the statistical weights of configurations that sampled bin i from aMD approaches. The free energy profiles were normalized with respect to the global minimum (0 kcal mol), which is the most populated bin. The free energy profiles were plotted as contour plots every 1.2 kcal/mol using Matlab 7.11.0 (R2010b). The free energy difference between the left-handed α-helical (ϕ,ψ: +45° ± 20°, +30° ± 20°) and extended β sheet (ϕ,ψ: −150° ± 20°, 150° ± 20°) regions was calculated for various simulation lengths from 5 to 255 ns, starting from the 50 0001st frame.

AUTHOR INFORMATION

Corresponding Author

*E-mail: dhamelberg@gsu.edu. Tel.: 404-413-5564. Fax: 404-413-5505.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We acknowledge support from the National Science Foundation CAREER Grant MCB-0953061, the Georgia Cancer Coalition (GCC) scholar award, and Georgia State University. The Information Systems and Technology center at Georgia State University is acknowledged for allocating computational time on the IBM System p7 supercomputer, acquired through a partnership of the Southeastern Universities Research Association and IBM's support to the SURAgrid initiative.

REFERENCES

(1) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27−85.
(2) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712−725.
(3) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., 3rd; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817−3829.
(4) Zhu, X.; Lopes, P. E. M.; MacKerell, A. D. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 167−185.
(5) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. *J. Comput. Chem.* **2009**, *30*, 864−872.
(6) Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J. P.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J.; Chow, E.; Eastwood, M. P.; Ierardi, D. J.; Klepeis, J. L.; Kuskin, J. S.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*; ACM: Portland, OR, 2009; pp 1−11.
(7) Stone, J. E.; Phillips, J. C.; Freddolino, P. L.; Hardy, D. J.; Trabuco, L. G.; Schulten, K. *J. Comput. Chem.* **2007**, *28*, 2618−2640.
(8) Zheng, L.; Chen, M.; Yang, W. *J. Chem. Phys.* **2009**, *130*, 234105.
(9) Wu, X.; Wang, S. *J. Chem. Phys.* **1999**, *110*, 9401−9410.

(10) Voter, A. F. *Phys. Rev. Lett.* **1997**, *78*, 3908−3911.
(11) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187−199.
(12) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141−151.
(13) Rahman, J. A.; Tully, J. C. *Chem. Phys.* **2002**, *285*, 277−287.
(14) Kruger, P.; Verheyden, S.; Declerck, P. J.; Engelborghs, Y. *Protein Sci.* **2001**, *10*, 798−808.
(15) Kleinjung, J.; Bayley, P.; Fraternali, F. *FEBS Lett.* **2000**, *470*, 257−262.
(16) Huber, T.; Torda, A. E.; van Gunsteren, W. F. *J. Comput-Aided Mol. Des.* **1994**, *8*, 695−708.
(17) Hansmann, U. H.; Okamoto, Y. *Phys. Rev. E* **1996**, *54*, 5863−5865.
(18) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919−11929.
(19) Grubmuller, H. *Phys. Rev. E* **1995**, *52*, 2893−2906.
(20) Elber, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 151−156.
(21) Bussi, G.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* **2006**, *96*, 090601.
(22) Bruccoleri, R. E.; Karplus, M. *Biopolymers* **1990**, *29*, 1847−1862.
(23) Berne, B. J.; Straub, J. E. *Curr. Opin. Struct. Biol.* **1997**, *7*, 181−189.
(24) Zuckerman, D. M. *Annu. Rev. Biophys.* **2011**, *40*, 41−62.
(25) Shen, T.; Hamelberg, D. *J. Chem. Phys.* **2008**, *129*, 034103.
(26) McQuarrie, D. A. *Statistical Mechanics*; University Science Books: Sausalito, CA, 2000.
(27) Singh, S. B.; Wemmer, D. E.; Kollman, P. A. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 7673−7677.
(28) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690−2693.
(29) Hummer, G.; Szabo, A. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 3658−3661.
(30) Israelewitz, B.; Gao, M.; Schulten, K. *Curr. Opin. Struct. Biol.* **2001**, *11*, 224−230.
(31) Zuckerman, D. M.; Woolf, T. B. *Phys. Rev. Lett.* **2002**, *89*, 180602.
(32) Hamelberg, D.; Shen, T.; Andrew McCammon, J. *J. Chem. Phys.* **2005**, *122*, 241103.
(33) Wereszczynski, J.; McCammon, J. A. *J. Chem. Theory. Comput.* **2010**, *6*, 3285−3292.
(34) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2009**, *131*, 147−152.
(35) Doshi, U.; McGowan, L. C.; Ladani, S. T.; Hamelberg, D. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 5699−5704.
(36) de Oliveira, C. A.; Hamelberg, D.; McCammon, J. A. *J. Chem. Phys.* **2007**, *127*, 175105.
(37) Velez-Vega, C.; Borrero, E. E.; Escobedo, F. A. *J. Chem. Phys.* **2009**, *130*, 225101.
(38) Grossfield, A.; Zuckerman, D. M. *Annu. Rep. Comput. Chem.* **2009**, *5*, 23−48.
(39) Lyman, E.; Zuckerman, D. M. *J. Phys. Chem. B* **2007**, *111*, 12876−12882.
(40) Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. *Science* **2003**, *302*, 1364−1368.
(41) Richardson, J. S.; Bryan, W. A., 3rd; Richardson, D. C. *Methods Enzymol.* **2003**, *374*, 385−412.
(42) Yi-Ching Yang, A.; Källblad, P.; Mancera, R. L. *J. Comput-Aided Mol. Des.* **2004**, *18*, 235−250.
(43) Samish, I.; MacDermaid, C. M.; Perez-Aguilar, J. M.; Saven, J. G. *Annu. Rev. Phys. Chem.* **2011**, *62*, 129−149.
(44) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. In *AMBER 10*; University of California: San Francisco, 2008.
(45) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(46) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(47) Ryckaert, J.; Cicotti, G.; Berendsen, H. *J. Comput. Phys.* **1977**, *23*, 327−341.
(48) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.