JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Calculation of Free Energy Landscape in Multi-Dimensions with Hamiltonian-Exchange Umbrella Sampling on Petascale Supercomputer

Wei Jiang,[†,∥] Yun Luo,[†,∥] Luca Maragliano,[§] and Benoît Roux*[,†,§]

[†]Argonne Leadership Computing Facility, Argonne National Laboratory, 9700 South Cass Avenue, Building 240, Argonne, Illinois 60439, United States

[‡]Biosciences Division, Argonne National Laboratory, 9700 South Cass Avenue, Building 202, Argonne, Illinois 60439, United States

[§]Department of Biochemistry and Molecular Biology, Gordon Center for Integrative Science, University of Chicago, 929 57th Street, Chicago, Illinois 60637, United States

ABSTRACT: An extremely scalable computational strategy is described for calculations of the potential of mean force (PMF) in multidimensions on massively distributed supercomputers. The approach involves coupling thousands of umbrella sampling (US) simulation windows distributed to cover the space of order parameters with a Hamiltonian molecular dynamics replica-exchange (H-REMD) algorithm to enhance the sampling of each simulation. In the present application, US/H-REMD is carried out in a two-dimensional (2D) space and exchanges are attempted alternatively along the two axes corresponding to the two order parameters. The US/H-REMD strategy is implemented on the basis of parallel/parallel multiple copy protocol at the MPI level, and therefore can fully exploit computing power of large-scale supercomputers. Here the novel technique is illustrated using the leadership supercomputer IBM Blue Gene/P with an application to a typical biomolecular calculation of general interest, namely the binding of calcium ions to the small protein Calbindin D_{9k}. The free energy landscape associated with two order parameters, the distance between the ion and its binding pocket and the root-mean-square deviation (rmsd) of the binding pocket relative the crystal structure, was calculated using the US/H-REMD method. The results are then used to estimate the absolute binding free energy of calcium ion to Calbindin D_{9k}. The tests demonstrate that the 2D US/H-REMD scheme greatly accelerates the configurational sampling of the binding pocket, thereby improving the convergence of the potential of mean force calculation.

1. INTRODUCTION

Molecular dynamics (MD) simulations based on atomic models can play an increasingly important role in understanding biological macromolecular systems. However, while the efficiency of the MD engine to propagate the classical equation of motions is clearly of prime importance for large-scale simulations, it is often necessary to go beyond simple brute-force trajectories to achieve a quantitative characterization of a complex system. Further compounding the current challenges, it has become essential to develop extremely scalable computational strategies able to make full use of leadership supercomputers evolving toward multimillions of cores. One promising avenue is to exploit computational techniques such as umbrella sampling (US) simulations to calculate the potential of mean force (PMF) or free energy landscape within a subspace spanned by a small set of order parameters.[1−3] In US, the simulations are carried out in the presence of biasing potentials that are introduced to focus the configuration of the system around specific values of chosen order parameters. The distributions of the system along the order parameters from all all the biased MD simulations ("windows") are then combined by postprocessing techniques like the weighted histogram analysis method (WHAM)[4,5] to calculate the unbiased free energy landscape within the relevant subspace.

In US calculations, a large number of window simulations may be needed to cover the relevant part of the subspace. The implication is that US calculations often require a considerable, albeit distributed, computing power. To reduce the requirement of a large number of windows, many efforts have been devoted to develop self-adaptive accelerated sampling strategies, such as adaptive biasing force (ABF),[6,7] metadynamics,[8] adaptive reaction coordinate forces,[9] and temperature-accelerated MD.[10] However, despite the attractive design of these methods, practical experience shows that they can suffer from unanticipated convergence problems in applications to complex systems. Moreover, self-adaptive accelerated sampling strategies along a single MD trajectory renounce the opportunity of using the full scale of modern supercomputers with massive numbers of computer nodes. In contrast, the nature of the US algorithm is well adapted to modern supercomputers, which supports a huge number of windows running concurrently in parallel. Similarly, the postprocessing of umbrella sampling data can be processed very efficiently on multiprocessors using parallelized version of WHAM. Of particular interest for applications on modern supercomputers, various replica-exchange algorithms can be exploited to carry out concurrent window simulations with a minimized communication overhead.

It has been previously shown that classical simulation propagation can be combined with a replica-exchange algorithm

Received: June 7, 2012
Published: September 20, 2012

   dx.doi.org/10.1021/ct300468g | J. Chem. Theory Comput. 2012, 8, 4672−4680

to enhance the sampling achieved by conventional MD.[11−22] In the replica-exchange MD (REMD) approach, several copies of the molecular system are simulated concurrently under slightly different conditions, e.g., different temperatures or Hamiltonians. Attempts are periodically made to exchange (swap) parameters or configurations between different replicas using a Metropolis Monte Carlo acceptance criterion, thus ensuring Boltzmann-weighted statistics. By allowing replicas to exchange their configurations frequently, REMD simulations have a better chance to escape from kinetic bottlenecks and efficiently explore all relevant regions of configurational space. Because the exchanges do not represent actual dynamical propagation during the classical trajectories, complex transitions can occur as the replicas are not required to physically climb over free energy barriers along orthogonal degrees of freedom separating relevant regions of configurational space. More generally, such multiple copy algorithms (MCAs) represent a broad family of extremely scalable strategies aimed at enhancing the sampling efficiency of conventional classical molecular MD. Recently, a systematic parallel/parallel alchemical free energy perturbation FEP/λ-REMD method[17] was developed using the REPDSTR module of the code CHARMM.[23,24] CHARMM/REPDSTR adopts a parallel/parallel mode (MPI level) to launch a large number of replicas of a complex system. Such implementation is considerably more efficient and scalable than script-driven replica-exchange wrappers that launch and control independent simulations. Each replica occupies multiple MPI ranks and has its own I/O. The I/O requirements for CHARMM/REPDSTR methods are similar to those of a standard single-replica parallel MD run, multiplied by the number of used replicas. An extension of the alchemical FEP/λ-REMD scheme allowed exchanges to occur alternatively along the axis corresponding to the thermodynamic coupling parameter λ, and the amplitude of boosting potentials used to cancel the effective dihedral energy barriers opposing the interconversion among different rotameric states of the side chains in the neighborhood of the binding site, in an extended dual array of coupled λ- and H-REMD simulations.[18] It has been shown that such a scheme allowing random moves within a two-dimensional extended replica ensemble significantly improved the statistical convergence in calculations of absolute binding free energy of ligands to proteins.[18] In the present communication, we describe a multidimensional US/H-REMD method for the PMF calculation on large-scale supercomputers, coupling replicas restrained by different biasing potentials. The method was recently implemented in the REPDSTR module of the code CHARMM. To compute the free energy landscape along key order parameters governing large conformational changes in the protein, the US simulations is extended with Hamiltonian-exchange on thousands of windows. The US/H-REMD method described here is similar to the 1D window exchange umbrella sampling MD (WEUSMD) recently introduced by Park et al.[22]

The proposed US/H-REMD scheme is extremely scalable to tens of thousands of CPUs on leadership computers, making it a very effective tool to characterize complex biomolecular membrane protein systems. The strength of the US/H-REMD method is illustrated by calculating the PMF governing the binding of calcium ion to calbindin D_{9k} as a function of two order parameters (Figure 1). Calbindin D_{9k}[25−28] is a small globular protein of the calmodulin superfamily. It possesses a pair of helix−loop−helix structure called EF-hand, which is a typical functional unit responsible for calcium binding. Because

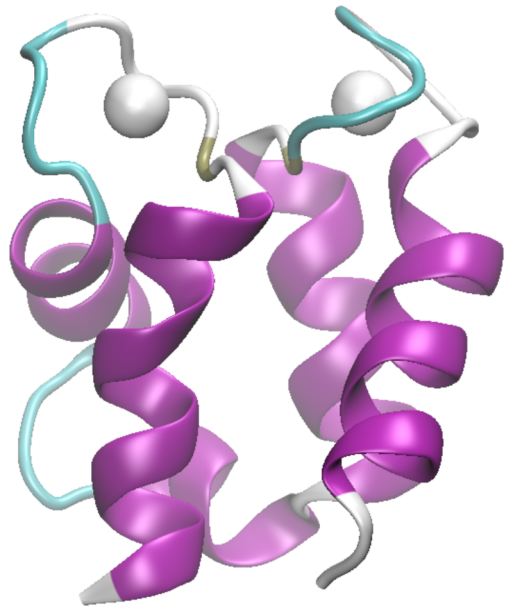

Figure 1. Calbindin D_{9k}. The two helix−loop−helix structures are the functional units for calcium ion binding. The carbonyl groups on the loops form binding pockets for calcium ions. The ions can bind to one or two loops forming singly loaded or doubly loaded states. The conformation shown is a doubly loaded state.

of the very strong interactions formed by the divalent calcium with its coordinating ligands and the conformational complexity of the EF-hand, it is extremely difficult to sample the system from the *apo* to *holo* state using standard US based on a large number of uncorrelated window simulations. These difficulties are overcome with US/H-REMD and further analysis explicitly shows that the coupling of thousands of windows simulations covering the relevant region of a 2D order parameter space with US/H-REMD considerably helps to enhance the sampling.

2. COMPUTATIONAL DETAILS

2.1. REPDSTR Implementation of US/H-REMD. Considering N copies of a system that are identical except for some differences regarding a small number of parameters, it is possible to make ordered lists of these systems such that the difference in the parameters is smallest for the nearest neighbors in the list. In the current application of US in 2 dimensions (2D), different biasing windows potentials are required for each of the simulated systems. Assuming quadratic potentials, there are 2 parameters $\{p_1, p_2\}$ and the associated US biasing window potential is as follows:

$$w(p_1, p_2) = k_1(x - p_1)^2 + k_2(y - p_2)^2 \quad (1)$$

The only variation is with the actual numerical values of the parameters p_1, k_1, p_2, and k_2. Therefore, an ordered list of systems is defined by $\{p_1^j, p_2^j\}$, with j going from 1 to N. Assuming N is even for simplicity, the rule for attempted exchanges during the replica-exchange MD simulation can be defined. First, there is an attempt to exchange between the members of the list and their nearest neighbors according to the odd ↔ even rule, i.e., 1 ↔ 2, 3 ↔ 4, ... $(N - 1) \leftrightarrow N$; then, there is an attempt to exchange their nearest neighbors according to the even ↔ odd rule, i.e., 2 ↔ 3, 4 ↔ 5, ... $(N - 2) \leftrightarrow (N - 1)$. Each neighboring exchange means that all the numerical value of the parameters of the members, or

dx.doi.org/10.1021/ct300468g | J. Chem. Theory Comput. 2012, 8, 4672−4680

equivalently, instant configurations are simply swapped. In principle, a single list would allow swapping systems only with 2 nearest neighbors. However, this is too restrictive in the case of US in 2D. By constructing two different lists, the number of possible attempted exchange for the parameters $\{p_1^i, p_2^i\}$ and $\{p_1^j, p_2^j\}$ during the REMD is doubled and therefore more chances are created for the replicas to transit through the accessible configurational space. For US in 2D, this turns out to be necessary to enable exchange of nearest neighbors within the set of the biasing potentials along the first coordinate, then exchange of nearest enable exchange of nearest neighbors within the set of biasing potentials along the second coordinate. As illustrated in Figure 2, Hamiltonian exchanges are

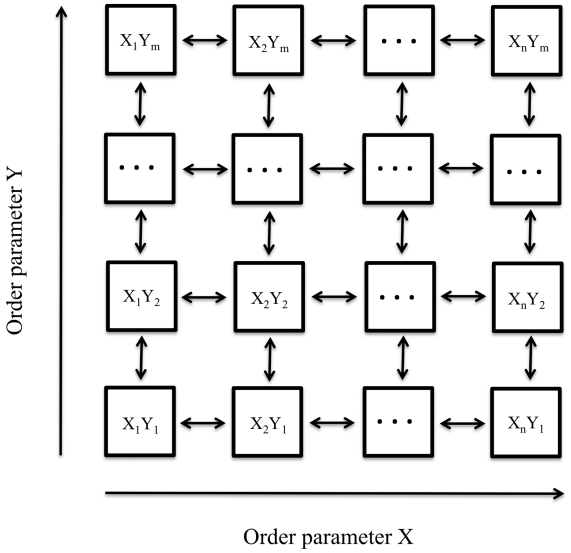

Figure 2. Two-dimensional Hamiltonian replica-exchange MD scheme. Each replica (window) performs periodic exchange attempts with its four neighboring replicas. In each exchange cycle, exchange attempts are performed alternatively along the two dimensions to accelerate the travel through the extended ensemble.

alternatively performed along the first coordinate and the second one, forming a 2D lattice pattern. The whole set of algorithm is implemented in the CHARMM/REPDSTR module, and coordinates instead of biasing parameters are exchanged during one swap. Programming of coordinate swap is more complicated than parameter swap, however the former can achieve a "universal" exchange that is independent of specific energy term at the cost of slightly larger communication volume when moderate exchange frequency say 1/50 steps is used. The Hamiltonian replica-exchange molecular dynamics (H-REMD) algorithm follows the conventional Metropolis Monte Carlo exchange criterion:

$$P(p_1^i \leftrightarrow p_1^j; p_2^l) = \min\left\{1, e^{-[U(p_1^i,p_2^l,\mathbf{r}_{i,l}) + U(p_1^j,p_2^l,\mathbf{r}_{j,l}) - U(p_1^i,p_2^l,\mathbf{r}_{l,l}) - U(p_1^i,p_2^l,\mathbf{r}_{j,l})]/k_BT}\right\} \quad (2)$$

$$P(p_1^k; p_2^m \leftrightarrow p_2^n) = \min\left\{1, e^{-[U(p_1^k,p_2^m,\mathbf{r}_{k,m}) + U(p_1^k,p_2^n,\mathbf{r}_{k,n}) - U(p_1^k,p_2^m,\mathbf{r}_{k,n}) - U(p_1^k,p_2^n,\mathbf{r}_{k,m})]/k_BT}\right\} \quad (3)$$

where U denotes the potential energy of the underlying replica, and $(p_1^i, p_2^j\ p_1^j, p_2^m, p_2^n)$ denote the biasing parameters.

2.2. Simulations with US/H-REMD. All of the US/H-REMD simulations for calbindin D_{9k} were carried out on the IBM Blue Gene/P cluster Intrepid of the Argonne Leadership Computing Facility (ALCF) at Argonne National Laboratory. The simulations were carried out in a high performance mode using the version c36a2 of the CHARMM program,[29] which was modified and extended for the present study. The all-atom potential energy function PARAM27 is used for protein and ions. The TIP3P water potential[30] as modified for the CHARMM force field[31] was used. The Lennard–Jones parameters of the Ca^{2+} ion were taken from previous work.[25] Following the formulation introduced by Woo and Roux,[32] the absolute binding free energy is expressed in terms of a PMF function of two variables, $W(R,\chi)$, where R is the distance between the ion and the binding pocket along a prescribed axis, and χ is the root-mean-square deviation (rmsd) of the conformation of the EF-hand relative to its ion-bound crystallographic structure. An axial restraining potential $U_{axis}(\phi, \theta)$ is introduced to maintain the Ca^{2+} ion aligned relative to the protein and the EF-hand binding site. The axis is specified by the angles (ϕ, θ) in spherical coordinates using the center-of-mass (COM) of three groups of atoms on the protein. They are the COM of the EF-hand binding loop, the COM of two helicies in one EF-hand, and the COM of the whole protein. The contribution of the restraining potential in the bound state is calculated from a standard FEP/MD simulation of 0.4 ns per window with 10 intermediate values of the restraining force constant. The value is 0.20 ± 0.01 kcal/mol. The standard error is calculated from 3 independent FEP/MD simulations. The doubly loaded structure of calbindin D_{9k}, determined by X-ray crystallography at 1.6 Å resolution,[28] was solvated with 0.15 M KCl solution in a rectangular simulation box under periodic boundary conditions (PBC). The model includes a total of 12 251 atoms.

All MD simulations were carried out under constant temperature and pressure (NPT) condition at 300 K and 1 atm using Langevin thermostat and Andersen-Hoover barostat. The replicas were propagated with a 2 fs time step using Langevin dynamics. The particle-mesh Ewald (PME) method was used to evaluate the Coulombic interactions. After a 200 ps equilibrium run of the solvated structure, 512 initial configurations along the two reaction coordinate axis were generated using strong biasing potential and short equilibrium run (200 ps). These configurations were expanded to 1024, then 2048 and finally 4096 initial configurations for the replicas. Exchange attempts started after a 200 ps independent equilibrium simulation of each replica. The time interval of attempted exchange was set as 100 steps (0.2 ps), corresponding roughly to the decorrelation time scale in the system.

The two order parameters must be chosen to properly describe the progress of the calcium binding process, from the holo to the apo state of calbindin D_{9k}. In the current study, they are the doubly loaded (two calcium ions) and singly loaded

(one ion) calbindin D_{9k}. To obtain an acceptable acceptance ratio, empirically set at about 20%, the distribution of discrete order parameter values along each dimension was carefully chosen by monitoring short US/H-REMD test runs. In the current calbindin D_{9k} system, the distance R between the ion and the binding loop is chosen to go from 1.5 Å to 23 Å, with a window spacing of 0.25 Å for distances smaller than 11.5 Å, and a window spacing of 0.5 Å for distances larger than 11.5 Å. The rmsd of the EF-hand binding loop 1 relative to the X-ray crystal structure is chosen to cover variations going from 0.6 Å to 6.9 Å rmsd, with a window spacing of 0.1 Å rmsd. In total, 4096 replicas (64 replicas along each dimension) were used to guarantee that the successful exchange is evenly distributed and the average acceptance ratio reaches 28% (Figure 3). Finally, the data from the US simulations were unbiased using WHAM,[4] with a bin size of 0.05 Å and a stringent tolerance of 0.00 001 kcal/mol on every point in the PMF.

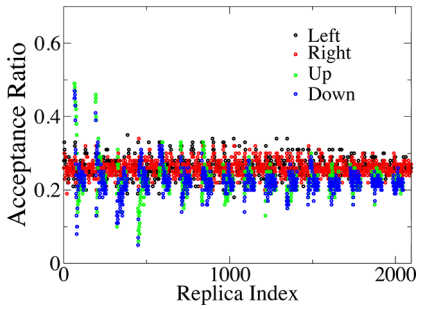

Figure 3. Acceptance ratios of Hamiltonian-exchange of inner 2100 replicas (that are closer to binding sites). For a replica, the right and left denote the exchange with its two neighbors along the first reaction coordinate; the up and down denote the exchange along the second reaction coordinate.

2.3. Analysis of PMF Convergence. To ascertain the quality of the sampling, we examine the consistency of the adjacent pieces of PMF obtained by neighboring US windows in the region where they overlap. Such overlap exists if the product of the biased histograms from two windows is nonzero. Let $\rho_n(\mathbf{x})$and $\rho_m(\mathbf{x})$ be the biased histograms extracted from the n-th and m-th windows, and $W_n(\mathbf{x})$ and $W_m(\mathbf{x})$ be the pieces of PMF extracted from the same windows. Aside from an arbitrary offset constant, the free energy landscape from these two pieces of PMFs should be consistent for all values of $\mathbf{x}$ in the overlapping region if the sampling scheme leads to well-converged results. To quantitatively determine whether this condition is satisfied, we write $W_n(\boldsymbol{x}) = M_{nm}W_m(\boldsymbol{x}) + B_{nm}$, and determine the slope M_{nm} from a linear regression of all data (i.e., using all values of $\mathbf{x}$ from the overlapping region). The slope M_{nm}, which reports on the consistency of the sampling from the n-th and m-th windows, is determined as follows:

$$M_{nm} = \frac{\left(\sum_{\mathbf{x}} p_{nm}(\mathbf{x})W_n(\mathbf{x})W_m(\mathbf{x})\right) - \left(\sum_{\mathbf{x}} p_{nm}(\mathbf{x})W_n(\mathbf{x})\right)\left(\sum_{\mathbf{x}} p_{nm}(\mathbf{x})W_m(\mathbf{x})\right)}{\left(\sum_{\mathbf{x}} p_{nm}(\mathbf{x})W_n(\mathbf{x})W_n(\mathbf{x})\right) - \left(\sum_{\mathbf{x}} p_{nm}(\mathbf{x})W_n(\mathbf{x})\right)\left(\sum_{\mathbf{x}} p_{nm}(\mathbf{x})W_n(\mathbf{x})\right)} \tag{4}$$

where the weighting factor $p_{nm}(\mathbf{x})$ is given by,

$$p_{nm}(\mathbf{x}) = \frac{\rho_n(\mathbf{x})\rho_m(\mathbf{x})}{\sum_{\mathbf{x}} \rho_n(\mathbf{x})\rho_m(\mathbf{x})} \tag{5}$$

If the slope M_{nm} is close to 1, then the data for all values of $\mathbf{x}$ from the two pieces of PMFs obtained from the n-th and m-th windows display maximum consistency. However, it is possible that the slope departs from 1 without damaging statistical consistency if the overall PMF is nearly flat in the overlap region. To detect these situations, it is useful to consider also the total overall magnitude of the free energy variations within the overlap region using this expression,

$$\sigma_{nm}^2 = \left(\sum_{\mathbf{x}} p_{nm}(\mathbf{x})W_n(\mathbf{x})W_n(\mathbf{x})\right) - \left(\sum_{\mathbf{x}} p_{nm}(\mathbf{x})W_n(\mathbf{x})\right)^2 + \left(\sum_{\mathbf{x}} p_{nm}(\mathbf{x})W_m(\mathbf{x})W_m(\mathbf{x})\right) - \left(\sum_{\mathbf{x}} p_{nm}(\mathbf{x})W_m(\mathbf{x})\right)^2 \tag{6}$$

with these two measures, it is possible to quantitatively compare the convergence and consistency of standard US with US/H-REMD.

3. RESULTS AND DISCUSSION

3.1. Illustrating the US/H-REMD with a Simplified Model. One of the main advantages of US/H-REMD is that it enables one to achieve a better conformational sampling of slow degrees of freedom that are "orthogonal" to the tagged order parameters chosen for the PMF calculation. Because different values of these orthogonal degrees of freedom map onto the same value of the coordinates used to compute the PMF, the source of the problem might be described as "degeneracy". It is useful to try illustrating this point with a simple model. For this purpose, we construct a simple 2D toy system,

$$U(x, y) = -\frac{332}{4} \times \left\{ \frac{1}{\left[x^2 + y^2 + (z-3)^2\right]^{1/2}} + \frac{1}{\left[(x-10)^2 + (y-10)^2 + (z-4)^2\right]^{1/2}} \right\} + A\exp\left[\frac{(y-5)^2}{2.5^2}\right] \tag{7}$$

The entire 2D potential map is shown in Figure 4 (top). Essentially, the 2D toy model corresponds to a positively charged particle moving within the xy plane and interacting with two negative charges fixed at (0, 0, 3) and (10, 10, 4) respectively (with a dielectric constant of 4), and restricted to a square region via flat-bottom harmonic potentials (0 Å < x < 10 Å and 0 Å < y < 10 Å). In addition, a Gaussian potential energy barrier of amplitude A = 3 kcal/mol is imposed along the line y = 5 to increase the sampling difficulties. The goal is to compute the 1D PMF along the x coordinates using US or US/H-REMD and compare their respective performance (in both cases, harmonic potentials evenly separated by 0.5 Å along the x-axis with a force constant of 5 kcal/mol/Å^2 were used for the biasing window potentials). While this 2D model is extremely simple, it nevertheless displays some pathological sampling problems because there is a slow orthogonal degree of freedom along the y-axis relative to the x-axis. When trying to calculate the PMF along the x-axis using standard US with uncorrelated MD simulations, the slow orthogonal degree of freedom along the y-axis gives rise to hysteresis problems and poor sampling. As shown in Figure 4 (bottom), these problems are reflected in the PMF calculated with standard US, which is poorly converged (green lines). In contrast, US/H-REMD (calculated

dx.doi.org/10.1021/ct300468g | J. Chem. Theory Comput. 2012, 8, 4672−4680

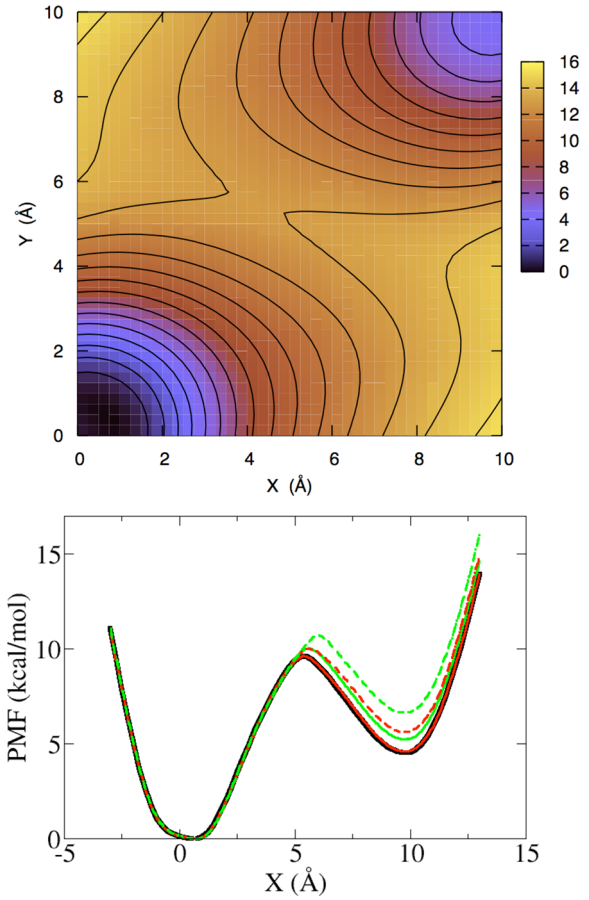

Figure 4. Top: 2D energy surface of the toy model from eq 2 as a function of the position of the particle in the *xy* plane. Bottom: Comparing the 1D PMF along the *x*-axis. Black line is the exact PMF obtained by numerical integration of the Boltzmann factor, $\exp[-U(x,y)/k_BT]$, over the *y*-axis; red lines are PMF profiles from US/H-REMD; green lines are PMF profiles from US. Dash lines are PMF profiles from the trajectories of 1 ns; solid lines are PMF profiles from the last 4 ns trajectories of 5 ns simulations.

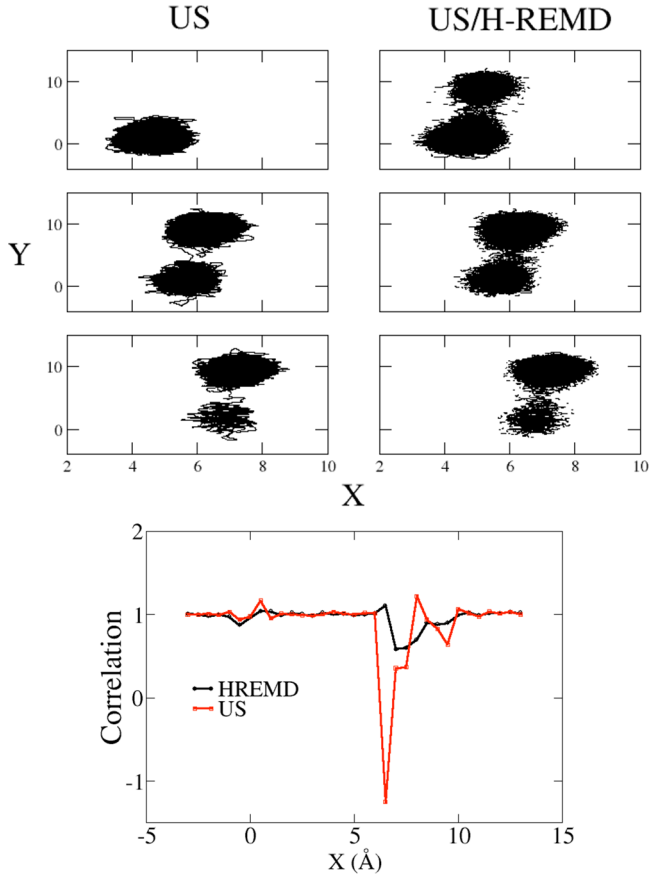

Figure 5. Top: Comparing 1 ns trajectories of 3 windows along *x*- and *y*-axis from US (left) and US/H-REMD (right) simulations. Window positions from top to bottom: x = 5, 6, 7 Å. US/H-REMD enhances sampling around orthogonal energy barrier y = 5 Å. Bottom: Consistency of the sampling from neighboring windows along the *x*-axis calculated from as the slope $M_{n,n+1}$ from eqs 4 and 5; the amplitude of the Gaussian barrier was increased to 6 kcal/mol to enhance the differences. Note the big peak in the case of standard US near the center of the *x*-axis due to poor sampling, while US/REMD retains a consistent correlation (∼1.0) for the whole range. In both cases, the data were collected from 1 ns trajectory. The exchange frequency for the US/H-REMD simulation is 0.02 ps, yielding an average acceptance ratio of 35%.

with an exchange frequency of 0.02 ps) generates a faster and well-converged PMF (red lines). The reason is that both sides of the orthogonal energy barrier along the *y*-axis need to be sampled properly to accurately evaluate the free energy landscape along the *x*-axis. These difficulties are revealed in Figure 5 (top) by comparing the trajectories of three biased windows generated near the midpoint x = 5 Å from standard US (left) and US/H-REMD (right). It is clear that the sampling along the orthogonal slow degree of freedom (*y*-axis) is significantly improved by the replica-exchange scheme, even within the first nanosecond of simulations. Nevertheless, it is important to realize that the exchange process does not alter the actual distribution of systems with respect to the *y*-axis. Thus, exchanges can occur even if the amplitude of the energy barrier at y = 5 is set to infinity, but the relative number of replica locked above and below the line y = 5 will then remain the same throughout the entire US/H-REMD simulations. Therefore, while the US/H-REMD scheme helps accelerate the configurational exploration by allowing the possibility of nondynamical transitions across a slow orthogonal degree of freedom to avoid kinetic bottlenecks, US/H-REMD still requires the occurrence of some spontaneous dynamical ("physical") transitions across this slow orthogonal degree of freedom to yield a proper Boltzmann distribution.

One way to characterize the quality of sampling is to verify whether segments of PMF extracted from adjacent windows, $W_n(x)$ and $W_{n+1}(x)$ are consistent with one another. If so, the value of $M_{n,n+1}$ calculated from eq 4 should be close to unity. Conversely, a significant departure from unity is indicative of statistical inconsistencies between adjacent windows. The value of $M_{n,n+1}$ evaluated along the *x*-axis from 4 ns trajectories is shown in Figure 5 (bottom). Here, the amplitude A of the Gaussian barrier was increased to 6 kcal/mol to enhance the constrast between US and US/H-REMD. It is observed that poor sampling of the slow degrees of freedom (*y*-axis) causes a lack of correlation between segments of the PMF from adjacent US windows near x = 5 Å. In contrast, the PMF segments from US/H-REMD display an excellent correlation between neighboring windows along the *x*-axis, resulting in a well-converged PMF (Figure 4, bottom).

dx.doi.org/10.1021/ct300468g | J. Chem. Theory Comput. 2012, 8, 4672−4680

3.2. Binding Free Energy of Ca^{2+} to Calbindin D_{9k}.

Sampling difficulties in the case of Ca^{2+} binding to calbindin D_{9k} arise from the very strong electrostatic interactions between the divalent cation and the negatively charged carboxylate groups of the EF-hand binding site. Presumably, there are numerous configurations near the bound state(s) that have overlapping values of the two order parameters R and χ, but are otherwise kinetically trapped with respect to some other (orthogonal) coordinate. The phenomenon of degeneracy is associated with the fact that the chosen order parameters account imperfectly for the progress of all the slow processes in the system. Some degrees of freedom, orthogonal to the two order parameters R and χ used for the 2D PMF, are slowly varying on the time scale of the MD simulations. Except for the simplest systems, some amount of degeneracy is unavoidable. In such situations, standard US simulations are expected to encounter considerable difficulties to sample the configurational space appropriately, and yield a correct Boltzmann-weighted PMF in the subspace of the chosen order parameters.

The sampling difficulties are clearly displayed in Figure 6, with a series of calculated 1D PMFs as a function of distance R

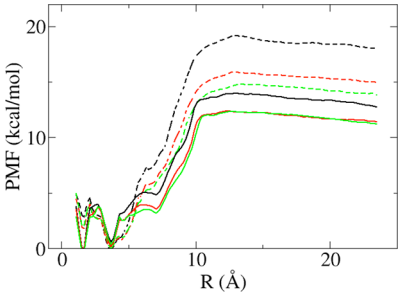

Figure 6. Comparing the convergence of PMF between US (dash lines) and US/REMD (solid lines) of calcium ion binding to the first binding site of D_{9k}. In both cases, the color black, red and green represents the block average from each 400 ps/per replica simulation. Each simulation starts from the last configuration of the previous run. The 1D PMF from US/H-REMD exhibits convergence from the second run (800 ps), while the results from standard US without replica-exchange do not converge even after 1.2 ns. Note that US/H-REMD samples the binding sites much better. The 1D PMF is calculated from 2 D PMF profiles (see Figure 8 for details).

obtained via US/H-REMD and from standard US with uncorrelated MD simulations (note that theses 1D PMFs were obtained by integrating out the rmsd coordinate χ of 2D PMFs). It can be observed that the 1D PMF from US/H-REMD exhibits two deep free energy wells at distances of 1.5 Å and 4.0 Å corresponding to the two most probable binding configurations of the Ca^{2+} ion. In contrast, standard US simulations suffer from slow convergence, even after 1.2 ns per window. Other features, such as an inflection of the PMF at 7.5 Å, are only observed in the PMF obtained from US/H-REMD. Figure 7 shows the histogram of windows at 1.5 Å along the coordinate χ in the region corresponding to the first bound state. It can be seen that the histogram of US/H-REMD exhibits a much higher degree of overlap (top) than standard US simulations (bottom). The calculated 2D PMFs are shown in Figure 8. Two binding wells near the binding site are clearly observed in the 2D PMF calculated from US/H-REMD simulations (top). In contrast, the corresponding region is fuzzy and the binding wells are barely resolved in the 2D PMF calculated from standard US with uncorrelated MD simulations

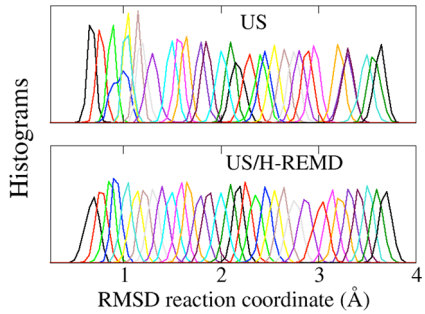

Figure 7. Enhanced overlaps of histograms along the rmsd coordinate χ with standard US (top) and US/H-REMD (bottom). The selected replicas were located at the binding site (distance R = 1.5 Å). Notice the significantly enhanced overlap around χ equal to 1, 1.5, 3.1, and 3.5 Å.

(bottom). Thus, US based on uncorrelated MD simulations fails to adequately sample the two dominant binding configurations that are revealed and sampled correctly by the US/H-REMD simulations.

The free energy landscape shows two free energy binding wells, corresponding to two distinct configurations. The conformation of the first energy well is located at (R 1.5 Å, rmsd 1.0 Å), and the second binding well is located at (R 3.7 Å, rmsd 2.3 Å). The first is very close to the crystal structure, with the Ca^{2+} ion coordinated by the oxygen atoms from five residues (Figure 8: six oxygen atoms shown as red spheres) and one or two water molecules (oxygen atoms shown as magenta spheres), for a total of 7 to 8 ligands. The existence of the second binding well is intriguing. In this configuration, the EF-hand binding loop is slightly open and only two residues provide coordinating ligands (glutamic acid 17 and 27). In addition, four water molecules coordinate the ion, yielding again a total of 7 ligands coordinating the Ca^{2+} ion. Following the statistical mechanical PMF formulation of noncovalent association,[32] the absolute binding free energy of a second Ca^{2+} ion to a singly occupied calbindin calculated from the 2D-PMF is −11.5 kcal/mol, in fairly good agreement with the experimental value −9.4 kcal/mol.[33,34] In contrast, the free energy value estimated from standard US simulations is about −23 kcal/mol, a result that is obviously wrong and inaccurate. The 2D-PMF calculated from the US/H-REMD method described here will be extended to singly loaded state to provide important information about the molecular basis of Ca^{2+} binding cooperativity

The enhanced convergence of US/H-REMD can be attributed to more efficient sampling, especially near these two free energy minima. Because of configuration exchanges between neighboring US windows, the system is not required to undergo actual dynamical transition along the orthogonal degrees of freedom that are at the origin of the sampling difficulties. It is for this reason that US/H-REMD accelerates the convergence of the calculation. The gain in statistical convergence from US/H-REMD over standard US with uncorrelated MD simulations can be ascertained by considering the consistency of the pieces of the 2D-PMF obtained from neighboring US windows in the region where they overlap. This is quantified from the factor M_{nm} obtained from a linear regression fitting as defined by eq 4. The results are shown in Figure 9. The PMFs of neighboring windows from standard US with uncorrelated MD simulations are dramatically inconsistent thought the entire 2D range of the order parameters. In

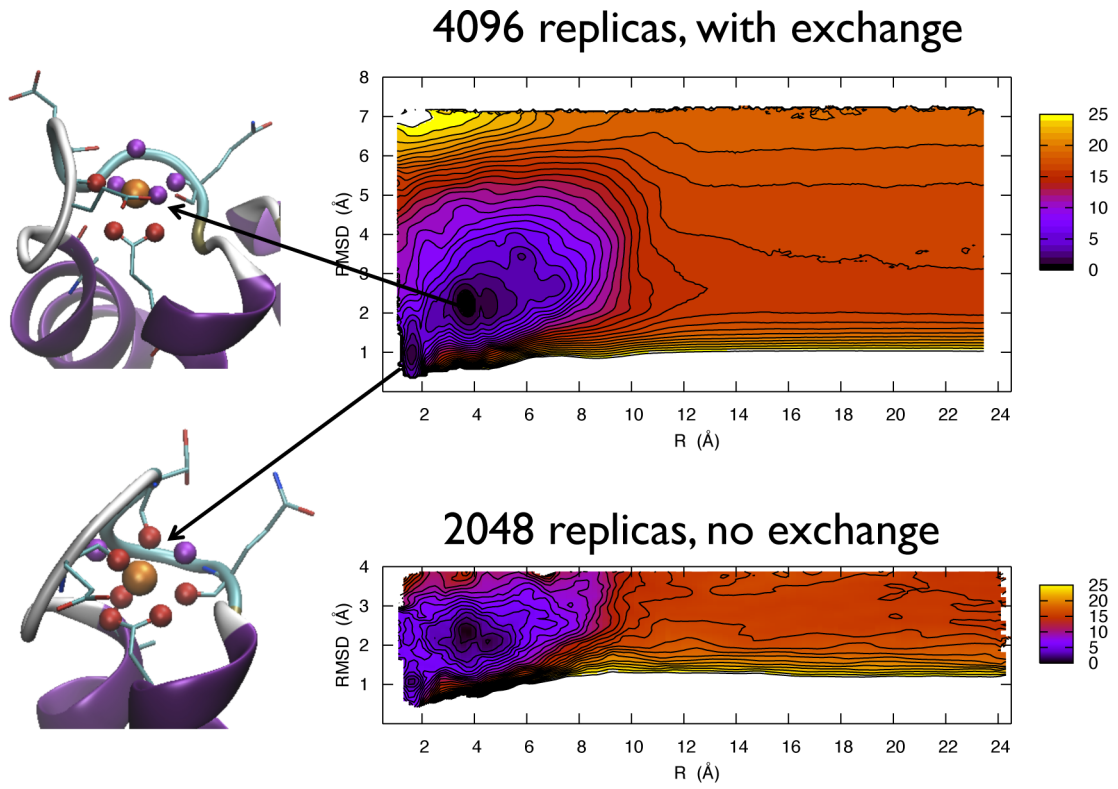

Figure 8. Two-dimensional (2D) free energy landscape from US/H-REMD (top) compared to standard US without exchange (bottom). The result from US/H-REMD exhibits 2 clear potential wells against the large fuzzy region from standard US. The latter fails to sample the (principal) binding site at 1.5 Å. Two typical snapshots of the binding site are shown for the two most probable binding states. Note that the US/H-REMD panel contains 4096 replicas that cover a larger area far from binding region while the 2D PMF from standard US panel only covers the area near to binding region.

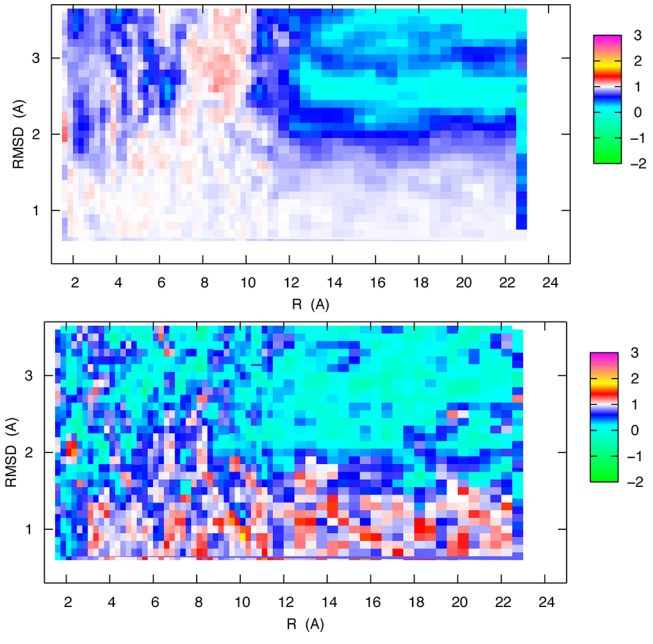

Figure 9. Consistency of the pieces of PMF obtained from neighboring US windows in the region where they overlap from the factor M_{nm} obtained by linear regression from eqs 4 and 5 fitting between neighboring windows from US/H-REMD (top) and from standard US without exchange (bottom). The color scale is defined as follows: −2.0 green, 0.2 cyan, 0.6 blue, 1.0 white, 1.4 red, 1.8 yellow, and 3.0 magenta.

contrast, the PMFs of neighboring windows from US/H-REMD are systematically consistent with a factor M_{nm} close to 1. One exception is the quadrant of the 2D PMF with R larger than 14 Å and χ larger than 4 Å. However, the overall magnitude of the free energy variations within the windows is much smaller in this region than in the rest of the 2D PMF, and the smaller values of M_{nm} merely reflects small statistical fluctuations in the PMFs.

3.3. Parallel Performance on Petascale Supercomputers. Parallel performance benchmark results for the 2D US/H-REMD simulations performed on Calbindin D_{9k} are presented for simulations runs on petascale supercomputers IBM Blue Gene/P Intrepid in Figure 10. It can be seen that the US/H-REMD computation effectively scales up to 32 racks (130K CPUs) on Intrepid, that is, to 80% resource of the entire machine. The strong scaling graph demonstrates the scalability of the parallel/parallel implementation. While CHARMM does not scale most efficiently to a large number of cores for a single replica comprising a large number of atoms, the scaling performance of CHARMM/REPDSTR is dramatically enhanced in the context of a multiple copy algorithm (MCA) methodology. Nevertheless, with the current exchange frequency and simulation time, the present US/H-REMD calculations do not guarantee that any given trajectory will travel through the entire space spanned by the order parameters. Ultimately, despite of the reasonable free energy value and the evidence of well-sampled bound configurations, longer production run and/or higher exchange frequency would be desirable to quantitatively characterize the Ca^{2+} binding cooperativity to Calbindin D_{9k}. Efforts are currently

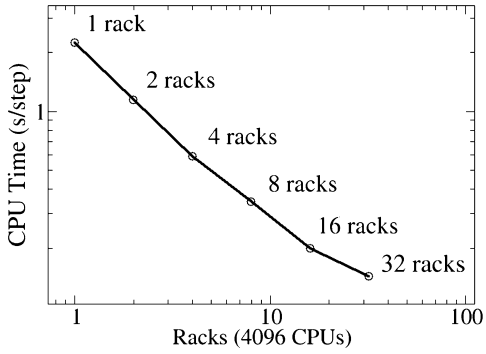

Figure 10. Strong scaling of the US/H-REMD algorithm on IBM Blue Gene/P Intrepid for the 2D PMF calculation. The scalability is proportional to the scalability of a single replica and number of replicas. Note that infrequent point-to-point communications (Hamiltonian exchange) between neighboring replicas do not affect parallel performance.

underway to implement a MPI level Hamiltonian exchange with parameter swap supporting various MCAs in NAMD[35] to further expand the range of applications enabled by extended ensemble sampling methods.

4. CONCLUSIONS

An extremely scalable umbrella sampling (US) Hamiltonian replica-exchange molecular dynamics scheme that is suitable for leadership computing platforms, US/H-REMD, was introduced to enhance the sampling for biological potential of mean force (PMF) calculations. The method was applied to calculate a PMF for a two-dimensional (2D) subspace of order parameters. Hamiltonian exchanges are performed in an extended ensemble of 2D reaction coordinates, with the exchange attempts occurring alternatively along the two dimensions. For US simulation of complex biological processes, a very large number of windows (replicas) along each dimension may be employed to achieve a meaningful acceptance ratio. Correspondingly, a high exchange attempt frequency is necessary to guarantee the samplings of each trajectory through the extended ensemble. Aided by the massively distributed computing power of leadership supercomputers, this US/H-REMD in 2D can be extended to more dimensions to enhance the sampling of a fairly large number of degrees of freedom. A high performance application of US/H-REMD on IBM Blue Gene/P shows that the sampling of the Ca^{2+} binding sites to calbindin D_{9k} is significantly enhanced and that the binding free energy of Ca^{2+} ion in the doubly loaded state is calculated accurately. Further implementations of the proposed strategy to the highly scalable simulation program NAMD for arbitrary number of dimensions are currently in progress.

AUTHOR INFORMATION

Corresponding Author

*E-mail: roux@uchicago.edu.

Author Contributions

[‖]These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful to Dr. Andrew Binkowski for his support. This research is funded by Grant MCB-0920261 from the National Science Foundation. W.J.'s research was supported by the Computational Postdoctoral Fellowship of Argonne Leadership Computing Facility (ALCF). Y.L.'s research is supported by the Early Science Postdoctoral Fellowship for Blue Gene/Q of ALCF. This research used the ALCF resource at ANL, which is supported by the Office of Science of the U.S. Department of Energy under Contract DE-AC02-06CH11357.

REFERENCES

(1) Kirkwood, J. *J. Chem. Phys.* **1935**, *3*, 300.
(2) Torrier, G.; Valleau, J. *Chem. Phys. Lett.* **1974**, *28*, 578.
(3) Kastner, J. *Wiley Interdisc. Rev.: Comput. Mol. Sci.* **2011**, *1*, 932.
(4) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011.
(5) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **1995**, *135*, 40.
(6) Darve, E.; Wilson, M.; Pohorille, A. *Mol. Sim.* **2001**, *28*, 113.
(7) Rodriguez-Gomez, D.; Darve, E.; Pohorille, A. *J. Chem. Phys.* **2004**, *120*, 3563.
(8) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *199*, 12562.
(9) Barnett, C.; Naidoo, K. *Mol. Phys.* **2009**, *107*, 1243.
(10) Maragliano, L.; Vanden-Eijnden, E. *J. Chem. Phys.* **2008**, *128*, 184110.
(11) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13711.
(12) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.
(13) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042.
(14) Rick, S. W. *J. Chem. Theory Comput.* **2006**, *2*, 939.
(15) Fajer, M.; Hamelberg, D.; McCammon, J. A. *J. Chem. Theory Comput.* **2008**, *4*, 1565.
(16) Kannan, S.; Zacharis, M. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 697.
(17) Jiang, W.; Hodoscek, M.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 2583.
(18) Jiang, W.; Roux, B. *J. Chem. Theory Comput.* **2010**, *6*, 2559.
(19) Moors, S.; Michielssens, S.; Ceulemans, A. *J. Chem. Theory Comput.* **2010**, *7*, 231.
(20) Meng, Y.; Sabri, D.; Roitberg, A. *J. Chem. Theory Comput.* **2011**, *7*, 2721.
(21) Zuckerman, D. M. *Annu. Rev. Biophys.* **2011**, *40*, 41.
(22) Park, S.; Kim, T.; Im, W. *Phys. Rev. Lett.* **2012**, *108*, 108102.
(23) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
(24) Brooks, B. R.; Brooks, C. L., 3rd; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.
(25) Marchand, S.; Roux, B. *Proteins: Struct., Funct., Bioinf.* **1998**, *33*, 265.
(26) Maler, L.; Blankenship, J.; Rance, M.; Chazin, W. *Nat. Struct. Biol.* **2000**, *7*, 245.
(27) Julenius, K.; Robblee, J.; Thulin, E.; Finn, B.; Fairman, R.; Linse, S. *Proteins: Struct., Funct., Bioinf.* **2002**, *47*, 323.
(28) Svensson, L.; Thulin, E.; Forsen, S. *J. Mol. Biol.* **1992**, *223*, 601.
(29) Brooks, B. R.; Brooks, C. L., III; Mackerell, J. A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Paster, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.
(30) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. K. *J. Chem. Phys.* **1983**, *79*, 926.

(31) MacKerell, A. D., Jr.; Banavali, N.; Foloppe, N. *Biopolymers* **2000**, *56*, 257.
(32) Woo, H.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *102*, 6825.
(33) Linse, S.; Johansson, C.; Brodin, P.; Grundstrom, T.; Drakenberg, T.; Forsen, S. *Biochemistry* **1991**, *30*, 154.
(34) Fast, J.; Hakansson, M.; Muranyi, A.; Gippert, G. P.; Thulin, E.; Evenas, J.; Svensson, L. A.; Linse, S. *Biochemistry* **2001**, *40*, 9887.
(35) Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kale, L.; Shulten, K. *J. Comput. Chem.* **2005**, *26*, 1781.