THE JOURNAL OF
PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

Leveraging the Information from Markov State Models To Improve the Convergence of Umbrella Sampling Simulations

Sunhwan Jo,[†] Donghyuk Suh,[‡] Ziwei He,[‡] Christophe Chipot,[§,∥] and Benoît Roux*[,⊥,#]

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

[‡]Department of Chemistry, University of Chicago, Chicago, Illinois 60637, United States

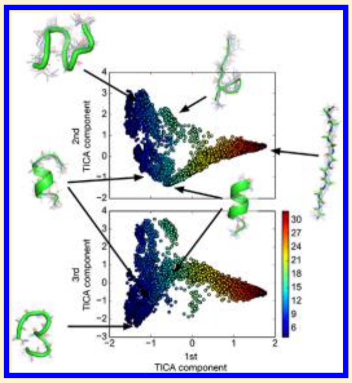
[§]Laboratoire International Associé Centre National de la Recherche Scientifique and University of Illinois at Urbana-Champaign, UMR 7565, Université de Lorraine, BP 70239, 54506 Vandœuvre-lès-Nancy, France

[∥]Department of Physics and Beckman Institute for Advanced Research and Technology, University of Illinois at Urbana-Champaign, 1110 West Green Street, 405 North Mathews, Urbana, Illinois 61801, United States

[⊥]Department of Biochemistry and Molecular Biology, Gordon Center for Integrative Science, University of Chicago, Chicago, Illinois 60637, United States

[#]Center for Nanomaterials, Argonne National Laboratory, Argonne, Illinois 60439, United States

ABSTRACT: Umbrella sampling (US) simulation is a highly effective method for sampling the conformations of a complex system within a small subspace of predefined coordinates. In a typical US stratification strategy, biasing "window" potentials spanning the subspace of interest are introduced to narrow down the range of accessible conformations and accelerate the sampling. The speed of convergence in each biased window simulation may, however, differ. For example, windows that coincide with a large energetic barrier along a coordinate that is orthogonal to the predefined subspace are often plagued by slow relaxation timescales. Here, we design a method that can quantitatively detect this type of issue and gain further insight into the origin of the slow relaxation timescale. Once the problematic windows affected by slow convergence are identified, additional simulations limited to only these windows can be carried out, thereby reducing the overall computational effort. Several possible approaches aimed at performing US simulations adaptively are discussed, and their respective performance is illustrated using a simple model system. Last, simulations of an atomic deca-alanine system are used to demonstrate the efficacy of analyzing US simulation trajectories using the proposed method.



■ INTRODUCTION

Umbrella sampling (US)[1] is an importance-sampling[2] technique that introduces biasing potentials to accelerate the configurational exploration of a complex system. The technique is particularly useful to calculate the free energy along a predefined order parameter, ξ, which is often associated with a putative reaction coordinate in molecular processes. In the commonly used US-stratification strategy, the system of interest is simulated in the presence of a biasing potential, which is introduced to focus the statistical sampling in the vicinity of a small region of configurational space. Because sampling is confined to a small region, only a small piece of the estimated potential of mean force (PMF) is sufficiently accurate to be useful. To obtain the PMF over the whole range of interest in ξ, it is necessary to combine the results from several strata (biased window simulations), each focusing the configurational sampling around a different region of ξ. The weighted histogram analysis method (WHAM)[3−5] or multistate Bennett acceptance ratio (MBAR)[6] estimator can be used to unbias the trajectory and determine the complete PMF along ξ.

US enhances sampling by confining the sampling around the center of the bias; only a reduced portion of configurational space needs to be explored, thus enhancing the overall sampling along ξ.[7,8] However, US does not accelerate sampling along degrees of freedom orthogonal to the chosen order parameter. If a window is located in the vicinity of an energetic barrier along coordinates that are orthogonal to ξ, the gain in sampling efficiency from a stratification of the configurational space rapidly diminishes, leading, in turn, to poor convergence of the free-energy profile.

A simple way to estimate the efficiency of US qualitatively is to monitor the autocorrelation function of $\xi(t)$.[9,10] In a well-designed US simulation, correlation is expected to decay quickly. However, if the free-energy surface is rugged and features hidden barriers in a direction normal to ξ, correlation time can be underestimated because the system may remain

Special Issue: J. Andrew McCammon Festschrift

Received: May 20, 2016
Revised: July 8, 2016
Published: July 13, 2016

DOI: 10.1021/acs.jpcb.6b05125

trapped in a local minimum for a significant amount of time. A more reliable way to diagnose the sampling inefficiency would be comparing the observed distribution, $p^{\mathrm{obs}}(\xi)$, from a biased simulation and the consensus distribution, $p^{\mathrm{cons}}(\xi)$, obtained by taking into account the sampling in the neighboring windows. Zhu and Hummer proposed to compute the consensus distribution by multiplying $\exp[-f(\xi)]$, where $f(\xi)$ is the biasing potential, by the unbiased distribution $p(\xi)$ computed using WHAM.[11] A similar metric has been proposed by Roux and co-workers,[12] which used consensus mean force between neighboring windows rather than probability distribution. These methods can pinpoint windows that are plagued by sampling inefficiencies, but they do not provide any additional insights into the physical origin of these deficiencies.

Recently, transition-based reweighting analysis methods (TRAMs)[13] have been proposed to analyze the results of multiple biased simulations. In particular, the discrete transition-based reweighting analysis method (dTRAM)[14] and dynamic histogram analysis method (DHAM)[15] have been used to infer the unbiased distribution along the predefined order parameters from US trajectories. Traditional analysis methods, such as WHAM or MBAR, assume that the distribution has reached an equilibrium but can produce erroneous free-energy profiles, if, for any reason, an equilibrium has not been attained.[11,15,16] The TRAM estimator requires only locally equilibrated distributions for a given lag time, which makes it more robust.[13,15]

In the present effort, we propose a simple diagnostic, namely, divergence analysis, for identifying problematic windows in US simulations. The basic idea behind our diagnostic is to discretize conformational states, compute the consensus distribution using TRAM estimators, and compare it against the observed distribution from the simulation. TRAM estimators were previously applied only to the chosen order parameter for US as discretized conformational states.[14,15] In our approach, we include other order parameters, for example, dihedral angle or distances, along with the chosen order parameter for US. This scheme not only allows us to compute the divergence between the consensus and the observed distributions obtained within each window but also provides additional insights into the underlying dynamics as well.

The divergence analysis can be used in an adaptive setting where only those windows resulting in a large discrepancy are simulated longer to improve convergence instead of increasing the sampling in all of the windows. We have implemented several basic adaptive simulation protocols and tested their performance with a toy model. The divergence analysis reliably identified windows that have not reached convergence and a protocol that simply extends those windows successfully accelerated the overall convergence. We have applied the divergence analysis and the adaptive simulation protocol to a more biologically relevant system, namely, deca-alanine in vacuum.[16,17] Whereas the divergence analysis correctly identified windows having sampling issues, a simple adaptive protocol did not enhance the convergence in this case, possibly due to stronger electrostatic interactions in a low-dielectric environment. Ultimately, more sophisticated accelerated sampling methods were applied to these pathological windows, which resulted in a free-energy profile that agrees well with the reference profile. This suggests that the divergence analysis could be used to identify windows that have sampling issues. Deriving a single, well-behaved adaptive protocol that can be applied regardless of the simulation system, however, remains challenging.

■ METHODS

Toy Model. To illustrate our analysis method, we constructed a toy model system evolving on a simple potential energy surface featuring four wells.

$$\begin{aligned} U(x, y) = &-10 \exp\left[-\frac{1}{5}((x+5)^2 + y^2)\right] \\ &- 5 \exp\left[-\frac{1}{5}((x+2.5)^2 + (y+5)^2)\right] \\ &- 5 \exp\left[-\frac{1}{5}((x+1.25)^2 + (y-5)^2)\right] \\ &- 5 \exp\left[-\frac{1}{5}((x-5)^2 + (y-5)^2)\right] \\ &+ 1 \exp\left[-\frac{1}{5}(y^2)\right] \end{aligned} \tag{1}$$

Figure 1 shows the two-dimensional potential energy surface as well as the numerically derived PMF along the x-axis. There are barriers along the y-axis in the vicinity of $x = -4$ and $x = 0$. Intuitively, sampling around these regions will be slow due to barriers in the orthogonal space because US simulations accelerate the sampling only along the chosen order parameter. Monte Carlo (MC) simulations with a step size of 0.02 for both x and y directions were performed. Thirty-one windows stratify the x-axis from $x = -7.5$ to 7.5 for every 0.5 Å. The initial positions for particles were

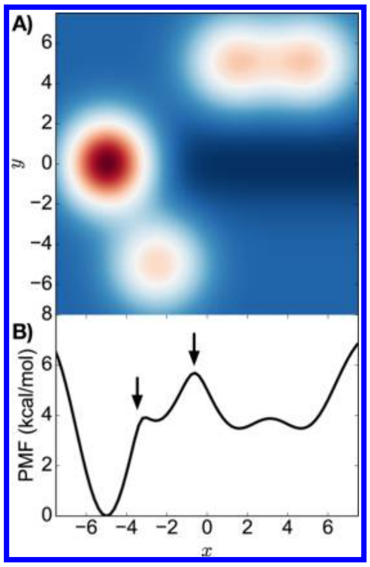


Figure 1. (A) Two-dimensional potential energy surface used for the toy model. (B) Numerically derived PMF along the x-axis. Two arrows indicate the regions where the barriers along the orthogonal space are located.

$$\begin{cases} (x_c, 0.0) & \text{if } x_c < -4 \\ (x_c, -5.0) & \text{if } -4 \le x_c < 0 \\ (x_c, 5.0) & \text{if } 0 \le x_c \end{cases} \tag{2}$$

where x_c is the center of the biasing potential. A harmonic restraint with a 5 kcal/(mol·Å^2) force constant was applied to enforce the bias.

Adaptive Sampling Protocol. To compare the performance of various adaptive sampling strategies, we performed MC simulations with the simple toy model described above. The adaptive sampling is carried out in three steps. First, an initial simulation is performed. After the initial simulation, the divergence analysis is performed (see the Divergence Analysis section for more detail). Lastly, the windows exhibiting large divergence values are subjected to more simulations. This process is iterated until acceptable convergence is reached.

For the divergence analysis, the order parameter (x-axis values) used in umbrella sampling was discretized using a bin size of 0.25, and another parameter (y-axis values) was discretized using k-means clustering with 10 cluster centers, resulting in two discretized time series. Unique states are designated using the Cartesian product of the two discretized time series. DHAM method[15] was used to obtain the unbiased distribution with a lag time of 1.

The initial simulation was performed for 100 000 steps, and trajectories were recorded every 10 steps (10 000 data points per window). The analysis was carried out in every iteration, and the windows that have divergence values larger than a certain threshold were selected and subjected to more simulations. Simulations following the initial simulation were performed for 50 000 steps, and the trajectories were recorded every 10 steps.

Deca-alanine Simulation System. The reference PMF was generated from a 2 μs long simulation, employing the multiple-walker adaptive biasing force (ABF) algorithm.[18] As the simulation proceeds, the algorithm locally estimates the required biasing force to yield a Hamiltonian in which no average force acts along the chosen transition coordinate. It follows that all values of the latter are sampled with an equal probability (in the limit that the transition coordinate is fully decoupled from other slow degrees of freedom), which, in turn, improves the reliability of the computed free-energy changes.[19−21] Among the hosts of available importance-sampling[2] schemes, the ABF algorithm[19,20] has proven to be an effective approach for variance reduction. To further improve the sampling, the multiple-walker ABF algorithm adopts a strategy that consists of collecting force samples from different walkers, concomitantly exploring the free-energy landscape and using a shared buffer.[20,22] The improvement can be spectacular in corrugated free-energy landscapes with a model reaction coordinate, for which standard (single-walker) ABF has proven unable to recover the global minimum in a heap of highly degenerated configurational states.

Two individual 320 ns trajectories were generated (32 windows centered from an end-to-end distance of 3−34 Å, 10 ns long each) with normal US using different initial conformations (extended and helical conformations) for all windows. A subset of windows from the normal US simulation was selected based on divergence analysis, and the trajectories were extended only for these windows. For each individual normal US simulation, 160 ns trajectories were extended (20 umbrella windows and four temperature windows, 2 ns long each) with bias and temperature replica exchange molecular dynamics (REMD). The simulation of this system was carried out using an in-house Python code with the molecular dynamics engine NAMD.[23] The temperature was set to 300 K, and a time-step of 1 fs was used to integrate the equations of motion. Both electrostatic and Lennard-Jones (LJ) cutoff values were set to 9 Å, and LJ interactions were treated with a switching function effective from 8 Å. Langevin dynamics was used with a damping coefficient of 1 ps^{-1}.

Cluster Analysis Method. The clustering analysis was carried out on the deca-alanine system in a reduced dimension as a result of time-lagged independent component analysis (TICA) using C_α pair distances as input order parameters. TICA enables dimensionality reduction by removing those degrees of freedom that decorrelate quickly and separating components that are independent of each other.[24−26] This method is advantageous when working with many order parameters and can also improve the accuracy of the results compared with traditional clustering methods. The TICA algorithm[27] solves a generalized eigenvalue problem from time-lagged covariance matrices $\mathbf{C}(\tau)$ with elements

$$c_{ij}(0) = \langle r_i(t) r_j(t) \rangle$$
$$c_{ij}(\tau) = \langle r_i(t) r_j(t+\tau) \rangle \qquad (3)$$

where τ is the lag time and $r(t)$ are the input order parameters, for example, distance between two atoms. Then, the generalized eigenvalue problem can be solved

$$\mathbf{C}(\tau)\mathbf{U} = \mathbf{C}(0)\mathbf{U}\boldsymbol{\Lambda} \qquad (4)$$

where $\mathbf{U}$ is an eigenvector matrix with the independent TICA components and $\boldsymbol{\Lambda}$ is a diagonal eigenvalue matrix. The data can be projected onto the TICA space to yield the transformed coordinates,

$$\mathbf{z}^{\mathrm{T}}(t) = \mathbf{r}^{\mathrm{T}}(t)\mathbf{U} \qquad (5)$$

A reduced dimension is achieved here by selecting only the first few columns of $\mathbf{U}$, which correspond to the slowest transition modes. For the deca-alanine system, using PyEMMA software package,[28] we computed the first five independent components which correspond to the five slowest modes with a lag time of 2 ps. Mini batch k-means clustering was carried out in this reduced dimension with 300 clusters that map the end-to-end distances.[29]

RESULTS AND DISCUSSION

Divergence Analysis. Let us assume a system described by collective variables, ξ_1, ξ_2, ..., ξ_n. One of the collective variables, ξ_1, is chosen, and US simulation is performed along ξ_1. The divergence analysis consists of three parts: discrete state definition, computing unbiased distribution, and analyzing the divergence of each window.

First, discrete states are defined using multiple collective variables. The order parameter employed in US simulation, ξ_1, is discretized using binning. Binning is a preferable choice for those order parameters chosen for US simulation to work with a TRAM estimator.[14,15] The rest of the order parameters are then discretized using either clustering or systematic binning. It is also possible to perform dimensionality reduction, such as TICA or principal component analysis, before clustering in case many other parameters are desired.

The discretization procedure results in multiple discrete time series. The unique discrete states are assigned by taking Cartesian product of these discrete time series. For example, let two order parameters, ξ_1 and ξ_2, be used to define discrete states using binning, Ξ_1 and Ξ_2, respectively. The final discrete

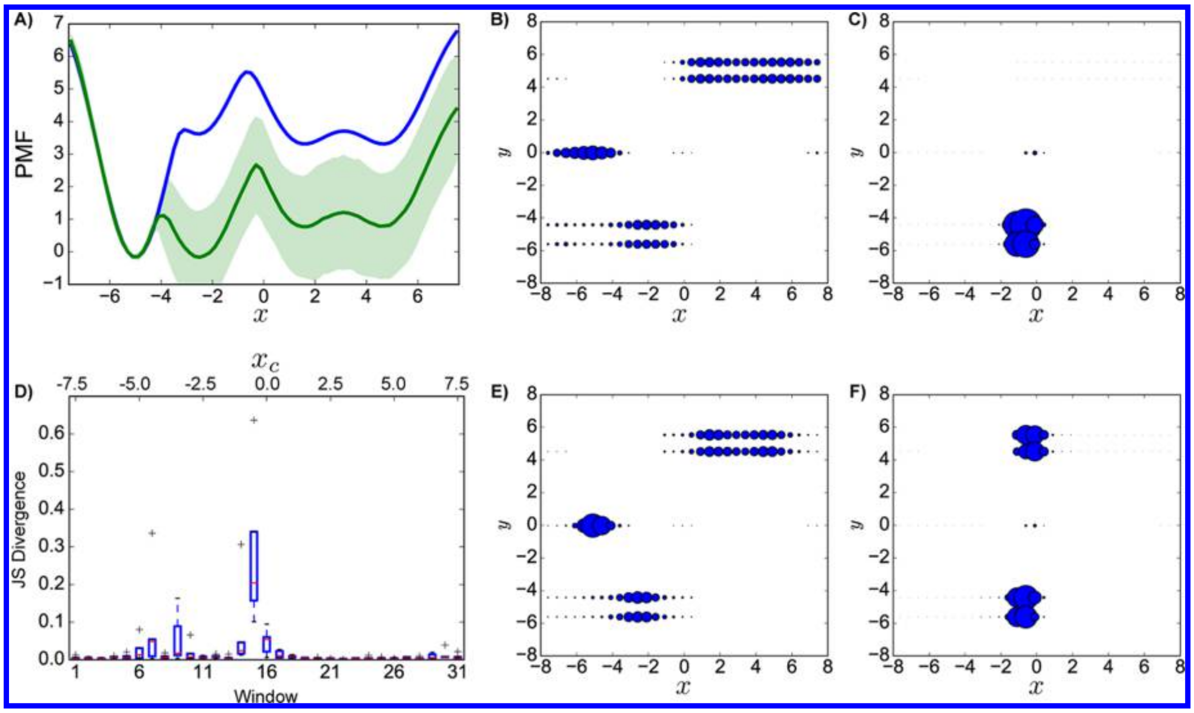


Figure 2. Illustration of the method to quantify discrepancies between the consensus and the observed distributions. (A) PMF calculated after 500 000 steps of simulations. The blue curve is calculated as $W(x) = -1/\beta \log[\int_y \exp[-\beta U(x,y)]\ dy]$, and the green curve is obtained from the simulation. (B) Aggregated biased and (E) unbiased distributions. Each circle and size represents the discretized state and its relative population. (C) Observed and (F) consensus distributions from the window centered at $x_c = 0.0$ as an example. (D) The divergence for each window, which represents the discrepancy between the consensus and observed distributions. The label of the second axis refers to the center of the biasing potential (x_c) for each window. The errors were estimated as the standard deviation from five independent simulations.

states index can be designated by $D = \Xi_1 \times N_{\Xi_2} + \Xi_2$ where N_{Ξ_2} is the number of discrete states for the order parameter ξ_2.

The final discrete time series is used to construct the transition count matrix. Transition-based methods, such as DHAM or TRAM, are used to compute the equilibrium distribution of these discretized states. Only TRAM formalism will be briefly discussed here. Detailed derivation of TRAM and DHAM can be found in refs 13−15. The biased and unbiased distributions of the discretized states are related as

$$\pi_i^{(k)} = f^{(k)}\gamma_i^{(k)}\pi_i \tag{6}$$

where $\pi_i^{(k)}$ represents the biased population of state i in the US window k, and π_i represents the unbiased population of state i. $\gamma_i^{(k)}$ is the bias applied to the corresponding state i, and f is a normalization constant. To satisfy the detailed balance, the following relation also holds

$$\pi_i^{(k)}\gamma_i^{(k)}T_{ij}^{(k)} = \pi_j^{(k)}\gamma_j^{(k)}T_{ji}^{(k)} \tag{7}$$

where $T_{ij}^{(k)}$ is the probability of transition from state i to state j observed in the US window k. Because $\mathbf{T}^{(k)}$ is a transition probability matrix and π is the probability vector, they should be normalized to 1, that is, $\sum_j T_{ij}^{(k)} = 1$ and $\sum_j \pi_j = 1$. The likelihood of observing the transition count $c_{ij}^{(k)}$ from state i to state j in a simulation is

$$L = \Pi_k \Pi_i \Pi_j (T_{ij}^{(k)})^{c_{ij}^{(k)}} \tag{8}$$

The optimal solution that maximizes the likelihood as well as the normalization conditions fulfills the following conditions

$$\sum_k \sum_j \frac{(C_{ij}^{(k)} + C_{ji}^{(k)})\gamma_i^{(k)}\pi_i\upsilon_j^{(k)}}{\gamma_i^{(k)}\pi_i\upsilon_j^{(k)} + \gamma_j^{(k)}\pi_j\upsilon_i^{(k)}} = \sum_k \sum_j C_{ji}^{(k)}$$

$$\sum_j \frac{(C_{ij}^{(k)} + C_{ji}^{(k)})\gamma_i^{(k)}\pi_i}{\gamma_i^{(k)}\pi_i\upsilon_j^{(k)} + \gamma_j^{(k)}\pi_j\upsilon_i^{(k)}} = 1 \tag{9}$$

where $\upsilon_i^{(k)}$ are unknown Lagrange multipliers. The unbiased distribution π_i can be found iteratively in a self-consistent manner. The initial population is set uniformly, and $\upsilon_i^{(k)} = \sum_j c_{ij}^{(k)}$.

$$\upsilon_i^{(k),\text{new}} = \upsilon_i^{(k)} \sum_j \frac{(C_{ij}^{(k)} + C_{ji}^{(k)})\gamma_i^{(k)}\pi_i}{\gamma_i^{(k)}\pi_i\upsilon_j^{(k)} + \gamma_j^{(k)}\pi_j\upsilon_i^{(k)}}$$

$$\pi_i^{\text{new}} = \frac{\sum_{k,j} C_{ji}^{(k)}}{\sum_{k,j} \frac{(C_{ij}^{(k)} + C_{ji}^{(k)})\gamma_i^{(k)}\upsilon_j^{(k)}}{\gamma_i^{(k)}\pi_i\upsilon_j^{(k)} + \gamma_j^{(k)}\pi_j\upsilon_i^{(k)}}} \tag{10}$$

Once a converged π is obtained, the discrepancy between the consensus and observed distributions can be calculated.

$$\pi_i^{(k),\text{consensus}} = f^{(k)}\gamma_i^{(k)}\pi_i$$

$$\pi_i^{(k),\text{observed}} = \sum_j c_{ij}^{(k)} / \sum c_{ij}^{(k)} \tag{11}$$

To quantify the discrepancy between the consensus and observed distributions within each window, we chose Jensen−Shannon (JS) divergence[30] because it has several remarkable mathematical properties, for example, it is symmetric and has finite bounds.[31] The divergence of the US window k is defined as

$$\mathrm{Div}(k) = \frac{1}{2}D(\pi^{(k),\mathrm{consensus}}||M) + \frac{1}{2}D(\pi^{(k),\mathrm{observed}}||M) \tag{12}$$

where $D(\pi||M)$ is the Kullback−Leibler divergence defined as

$$D(\pi||M) = \sum_i \pi_i \log \frac{\pi_i}{M_i} \tag{13}$$

and M_i is defined as

$$M_i = \frac{1}{2}(\pi_i^{\mathrm{consensus}} + \pi_i^{\mathrm{observed}}) \tag{14}$$

Divergence Analysis of a Simple 2D Toy Model. For illustrative purposes, we carried out five independent US simulations using a simple toy model (see Figure 1). Each simulation was carried out for 500 000 steps, and trajectories were recorded every 10 steps (i.e., 50 000 data points per window). The PMF was calculated using WHAM algorithm.[4,5] As shown in Figure 2A, the difference between numerically derived and calculated PMFs is significant, and the convergence of the independent simulations was not reached. The potential energy function has a large energetic barrier in orthogonal space; therefore, it is not surprising to have a larger error in the PMF.

The order parameter used for the US simulation, that is, x position, was discretized using a bin size of 0.5, and y position was used as an extra order parameter for the divergence analysis. The y positions were discretized using k-means clustering with five cluster centers. Larger grid spacing and a small number of cluster centers were employed here for clarity of the illustration. The unique states were designated by the Cartesian product of the two discretized time series.

Figure 2B shows the aggregated biased distribution of these unique states, which would be close to a flat distribution along the order parameter employed in the US simulation. Figure 2E shows the unbiased distribution using DHAM algorithm,[15] which reproduces the general features of the major free-energy basins quite well. The description, however, remains coarse at this stage and probably inaccurate due to limited sampling. We have also tested the dTRAM algorithm[14] using the PyEMMA program,[28] and the results were similar.

For each window, the corresponding biasing potential was re-applied to obtain the consensus distribution in each window. For instance, Figure 2C,F shows the observed and consensus biased distributions from the window centered at $x_c = 0.0$ as an example. This particular window is chosen because the window coincides with a large energetic barrier in the orthogonal space. In this case, the consensus biased distribution is calculated as

$$\pi_i^{\mathrm{consensus}} = \exp[-\beta f(x_i)]\pi_i / \sum_j \exp[-\beta f(x_j)]\pi_j \tag{15}$$

where π_i and x_i are the unbiased probability distribution and the center of the discrete state i, respectively. $f(x)$ is the biasing potential. Clearly, over the present simulation timescale, only the energy minima around $y = -5$ were explored, but on the basis of our analysis, the wells at both $y = -5$ and 5 are accessible and have about the same probability. The TRAM estimator allows states explored by neighboring windows to be included. Moreover, by incorporating additional collective variables that are deemed important, it is possible to gain further insight into the underlying potential energy surface.

The TRAM estimators can be used to obtain the transition probability matrix, $T_{ij}^k(\tau)$ for a particular window k. The transition matrix has to be valid and Markovian for a given lag time, τ, to be used in computing kinetic properties, for example, transition rates or timescale. Although the kinetic properties from US can be inaccurate if the energy barrier in the orthogonal space is sufficiently large, the Markov model can still provide useful information.

The characteristic time of the slowest relaxation for each window can be used as a measure of whether or not the simulation has reached convergence. As shown in Figure 3A, the windows coinciding with the orthogonal barriers correspond to a characteristic time that is 1 order of magnitude larger compared with the other windows, and extending simulations in those windows would be beneficial for enhanced convergence. Another piece of useful information extracted from the Markov model inferred from the US window is the eigenvectors of the transition matrix. The sign of the second eigenvector can be used to partition the microstates into two

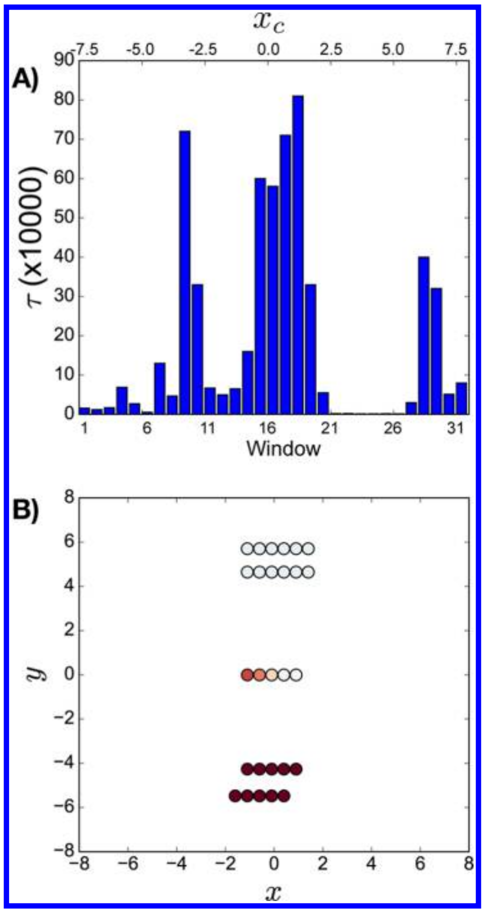


Figure 3. Timescale and mode of transition within the windows using a TRAM estimator. (A) Characteristic time of slowest relaxation, $\tau = \tau_{\mathrm{lag}}/\log \lambda_2$, within the windows. τ_{lag} is the lag time, and λ_2 is the second eigenvalue. The label of the second axis refers to the center of the biasing potential (x_c) for each window. (B) An example of eigenvector corresponds to the slowest relaxation in the window located at $x_c = 0$. Each circle represents an accessible state, for example, $\pi_i > 10^{-6}$, and the colors reflect the values of the second eigenvector from red (−0.1) to blue (0.1).

DOI: 10.1021/acs.jpcb.6b05125

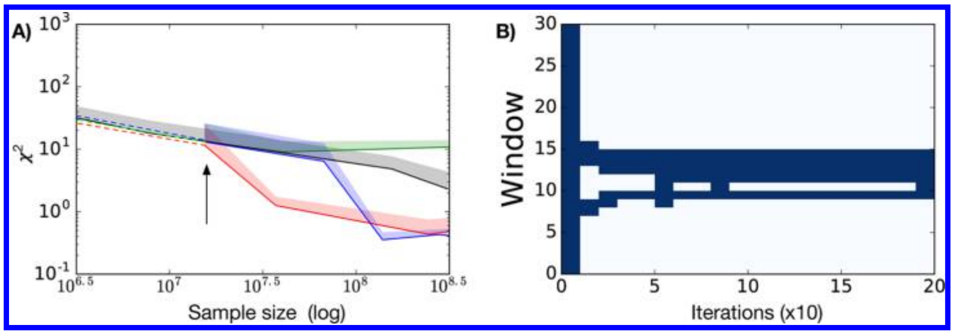


Figure 4. Performance of adaptive simulation protocols. (A) The relative error in the PMF is measured with respect to the total number of data collected from the simulation. The error bars are calculated by taking the standard deviation from five independent simulations, and only upper error bars are displayed for clarity. The red, blue, and green curves are obtained from an adaptive simulation protocol with continued simulation, independent sampling, and RE, respectively. The black curve is obtained from simulations without adaptive protocol. (B) Windows selected for each iteration in one of the adaptive simulation protocols. All windows were used for the initial simulation and only three or four windows were selected in the following iterations.

metastable states.[32,33] Visualizing the eigenvector of the accessible microstates of those windows exhibiting large divergence could provide insight into the underlying dynamics of the system. For example, Figure 3B shows the accessible microstates, for example, $\pi_i > 10^{-6}$ for the window centered at $x_c = 0.0$. Each microstate is colored from red (negative) to blue (positive) representing the value of the second eigenvector of the rate matrix, which shows the existence of potential wells at $y = 5$ and $y = -5$.

Adaptive Sampling Protocol and Performance Comparison. In many instances, US simulations span reaction pathways, which comprise two or more well-defined states separated by barriers. Typically, the sampling of the barrier is more challenging than the sampling of the well-defined states because diffusion along the orthogonal degree of freedom is more accessible at the barrier.[34] Efficient use of computational resources would consist of sampling more at those windows that reflect nonergodicity instead of simulating all of the windows simultaneously.

With the proposed method for quantifying the discrepancy between the consensus and the observed distributions, one can imagine an adaptive sampling strategy. Here, we focus on adaptive strategies that can harness computational resources effectively without alteration of the reaction coordinate model. In our adaptive sampling strategy, divergence analysis is carried out to identify the pathological windows; then, the simulation is resumed only for these windows.

We have used a dynamic threshold, which is determined based on a generalized extreme value (GEV) distribution.[35] Assuming that most windows have similar divergence values and only a few windows exhibit a large divergence, the GEV distribution can be used to model the divergence values. The GEV distribution has a cumulative distribution function

$$F(x;\, \mu,\, \sigma,\, \xi) = \exp\left[-\left[1 + \eta\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right] \tag{16}$$

where μ, σ, and ξ are the parameters that determine the location, scale, and shape of the distribution, respectively. The divergence values from the simulation were fitted to find the parameters of the distribution using the machine-learning Python library scikit-learn.[36] Once the parameters are determined, the threshold is determined as $F(x^{\text{threshold}};\mu,\sigma,\xi) = 0.9$ and the windows that exhibit divergence values higher than the threshold are selected. The adaptive sampling protocol is set to stop when the difference between the maximum and minimum divergence is smaller than 0.05.

The general adaptive sampling framework will be to (1) perform an initial simulation, (2) analyze the trajectory collected so far, (3) pursue the simulation for those windows that exhibit large discrepancies, and repeat steps (2) and (3) until divergence values become small enough or a set amount of resource has been utilized.

Between steps (2) and (3), there could be a variety of ways to restart the simulation to enhance ergodic sampling. We have tested three strategies:

1. Pursue the simulation from the previous run.
2. Independent sampling among the accessible states.
3. Replica exchange (RE) between neighboring windows.

Pursue the Simulation from the Previous Run. This is the simplest adaptive protocol. For each round of simulation, windows that exhibit large deviations are selected as mentioned above, and a total of 20 rounds of simulations are performed. Simulation of the selected windows is simply continued without any further treatment.

Figure 4A shows the relative error in the PMFs obtained from straightforward US simulations (black curve) and from the adaptive protocol (red curve). The relative error is defined as $\chi^2 = \sum(W^{\text{ref}} - W^{\text{calc}})^2/W^{\text{ref}}$, where W^{ref} is the numerically derived PMF and W^{calc} is the PMF obtained from the simulation. The relative error decreased as soon as the adaptive scheme started (arrow). Overall, the adaptive scheme reduced the relative errors by more than 1 order of magnitude compared with straightforward US simulation.

The increased computational efficiency, measured by error/computational cost, could be a result of simply using fewer windows for simulation because the protocol does nothing to the sampling rate of individual windows. As shown in Figure 4B, only about 10% of the windows were pursued after the initial simulation phase.

Independent Sampling among Accessible States. This protocol "reseeds" an initial condition for each iteration based on the consensus biased distribution. The accessible states are defined as follows

$$\Omega^{\text{accessible}} = \{d_i | i = 1 \dots N, \quad \underset{i}{p} = e^{-\beta V_i}\pi_i > 10^{-6}\} \tag{17}$$

where d_i is the discrete state index, V_i is the biasing potential at the center of the corresponding discretized state, and π_i is the unbiased probability. The next initial configuration is proposed

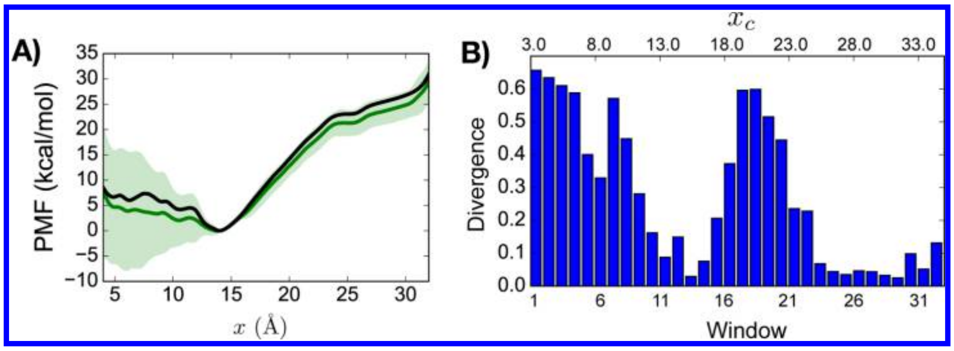


Figure 5. (A) PMF obtained by two US simulations using different starting conformations. The thick green line is the averaged PMF, and the shaded area is the standard deviation of the two PMFs. The black curve is the reference PMF obtained with the multiple-walker ABF algorithm. (B) The divergence for each window after US simulation. The label of the second axis refers to the center of the biasing potential (x_c) for each window.

based on the consensus probability, p_i, and the simulation continues from that configuration.

The relative error in the PMFs obtained from independent sampling and reseeding protocol is shown in Figure 4A (green curve). Surprisingly, the relative error remained poor and did not show perceptible improvement over time. This conspicuous error stems from the fact that the PMF converges to a wrong free-energy profile. There are two factors in play, namely, the inaccurate unbiased probability and slow diffusion in the orthogonal space. The initial unbiased distribution produced by the TRAM estimator is often reasonable, but it may not be accurate given a limited sampling. Hence the independent sampling based on the inaccurate unbiased distribution may introduce error in the PMF.

Even if the inaccurate unbiased distribution was used to initialize a new simulation, a correct equilibrium distribution could be attained in successive simulations if the particle can diffuse quickly. However, if the diffusion is slow for any reason, it is possible that the inaccuracies in the equilibrium distribution may persist. We hypothesize that the diffusion in the orthogonal space can be even slower in US simulations because the diffusion along the reaction coordinate is restrained at a region where the barrier in the orthogonal space is steeper. Put together, combining an enhanced sampling method with US simulations, for example, temperature- or Hamiltonian-RE method, might prove beneficial.

RE between Neighboring Windows. This protocol uses an RE algorithm[37] for all windows using the last configuration. Only those windows exhibiting large divergences will be restarted. We used a neighbor-exchange algorithm to attempt switch configurations between windows using the latest snapshot. Exchange attempts were made for all windows regardless of whether a window was selected for further simulation or not. For each iteration, 500 exchange attempts were made.

The performance of the RE algorithm is shown in Figure 4A (blue curve). The relative error decreased significantly compared with the straightforward simulations. However, the reduction of the relative error was at a similar level as pursuing simulations from the previous run. In addition, the reduction of error appears to be stochastic, that is, the degree of error reduction differs for each run, because the protocol selects different windows for exchange in different runs.

Application of Adaptive Sampling Protocol to Ala$_{10}$. Deca-alanine (Ala$_{10}$) in vacuum is a popular, paradigmatic model for studying peptide conformational equilibria, featuring a high free-energy barrier toward unfolding. The choice of its end-to-end distance has also proven to be a relatively ineffective order parameter to guide the sampling, notably for compact conformations where it is highly degenerated.[17,18] As a result, the one-dimensional PMF determined along the end-to-end distance converges very slowly and features several metastable states. In this section, we illustrate the difficulties of sampling a system with (hidden) high free-energy barriers in the orthogonal space using conventional US and show how to identify and address the sampling issues in a computationally effective manner.

The end-to-end distance between the carbonyl atoms at both N- and C-termini was used for US, and 32 windows were used to cover the transition pathway from 3 to 34 Å with an equal spacing of 1 Å. Figure 5 depicts the PMF obtained by performing two independent US simulations, using the extended and helical conformations as the starting conformations. When compared with the reference PMF obtained with the multiple-walker ABF algorithm,[19,20] US simulations appear biased by the starting conformation, which probably reflects a lack of convergence. We also have performed window-exchange US, which periodically exchanges biasing potential between windows and is often used to enhance convergence.[12] However, the convergence between two simulations that were started using different initial structures did not significantly improve (data not shown). A similar observation has been made by Park and Im that window-exchange US does not improve the sampling when the bias in the orthogonal space is too high.[16]

To use the divergence analysis, a set of discretized states needs to be defined. Root-mean-square deviation (RMSD) or radius of gyration is commonly used for comparing structural similarity. However, for a small peptide like Ala$_{10}$, it may not be ideal because of possible degeneracies.[38] Instead, pairwise distances derived from Euclidean distances between all pairs of C$_\alpha$ atoms are used. The time series of pairwise distances are subjected to a TICA, which yields a small set of independent components.[24−26] Using TICA, the pairwise distance time series, which has a dimension of 28, is reduced to five independent components. The *k*-means clustering algorithm is used to cluster the TICA components into 300 cluster centers. Figure 6 shows the resulting discrete states in TICA component space. Only the first three independent components are shown here for clarity. To examine the correlation between the end-to-end distance and the TICA components, each state is color coded based on the average end-to-end distance. In our analysis, the first TICA component is highly correlated with the end-to-end distance. TICA is designed to decompose the input time series into a set of slow- and fast-

DOI: 10.1021/acs.jpcb.6b05125

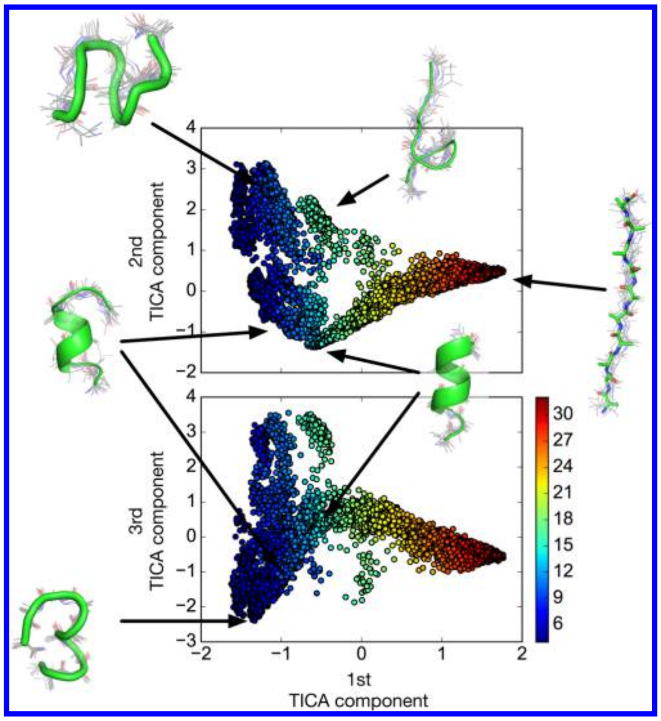


Figure 6. TICA analysis of US simulations. Only the first three TICA components were used for visualization purposes. Each dot represents a trajectory snapshot, and the dots are colored based on the end-to-end distance for that snapshot. After the clustering analysis, 10 snapshots that have the same cluster membership are randomly selected and drawn as a line representation. The centroids are drawn as a cartoon representation.

varying components. The input data are a set of distances, and some of these distances are highly correlated with the end-to-end distance for the US simulation. For example, the distance between two C_α atoms in the N- and C-termini would be highly correlated with the end-to-end distance, that is, distance between carbonyl atoms in the N- and C-termini. Such variables that are highly correlated with the end-to-end distance would appear to diffuse “slowly” to the TICA algorithm because the end-to-end distance is stratified in the US and does not change much within a stratified region, but it ultimately spans a wide range when trajectories from other windows are combined.

The state distribution in the second and the third TICA components indicates that many conformational states degenerate in the short end-to-end distance regime. To examine the origin of the degeneracy, several discrete states are selected along different end-to-end distance regimes and the corresponding conformations are visualized (see Figure 6). When the end-to-end distance is large, Ala_{10} must necessarily adopt an extended conformation, which reduces a number of possible ways to arrange the chain, whereas there is a large degeneracy of possible conformations when the end-to-end distance is small.

The final set of discrete states was obtained by taking a Cartesian product of the cluster membership and the gridded end-to-end distance, which resulted in 1813 unique states and the divergence of windows with respect to the biased equilibrium distribution inferred from the DHAM algorithm (see Figure 5B). The windows corresponding to shorter end-to-end distances are burdened by larger divergence values whereas those near the minimum of the PMF (end-to-end distance around 14 Å, i.e., the α-helical state) show only very small divergence. The windows corresponding to larger end-to-end distances, for example, around 16−30 Å, are also burdened by uncertainty, albeit to a lesser degree than windows corresponding to the compact conformations of the peptide chain.

To improve the convergence of the PMFs, we have employed several simple adaptive simulations, that is, pursuing simulation at the end of the previous run, or a window exchange algorithm. However, simply performing adaptive simulations on Ala_{10} simulation did not show improved convergence, particularly for the short end-to-end distance regime (data not shown). It appears that the conformations are tightly locked by a network of polar interactions, hampering ergodicity in a sampling.

To overcome the observed nonergodicity, an enhanced sampling technique would be desirable. To use computational resources effectively, the simulations are divided into two regimes, that is, shorter (3.0−13 Å) and longer (16−22 Å) end-to-end distances, where the windows exhibit large divergences (see Figure 5B). For different regimes, window- and temperature-exchange US simulations were performed separately. For each simulation, four temperatures ranging from 300 to 450 K were used. More specifically, a total of four simulations were performed: extended/short (44 windows = 11 bias (3−13 Å) × 4 temperatures, starting from extended initial conformation), extended/long (36 windows = 9 bias (16−22 Å) × 4 temperatures, starting from extended initial conformation), helix/short (same as extended/short but starting from helical initial conformation), and helix/long (same as extended/long but starting from helical initial conformation). A conventional neighbor-exchange algorithm was used for every 0.4 ps, and the samplings were performed for 2 ns.

Figure 7 shows the PMFs after 2 ns of accelerated sampling for the selected windows. With only a limited sampling, the convergence of PMFs is greatly improved. One downside of using accelerated sampling is that it makes the divergence analysis complicated because of mixed thermodynamic states in the trajectories. A new TRAM estimator, named xTRAM, has been proposed to perform the analysis for such trajectories; however, it is not pursued in this contribution.

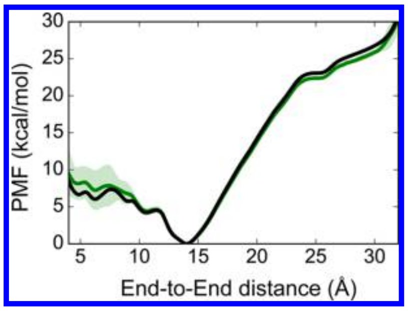


Figure 7. PMF obtained after extending the windows that exhibit large divergence values. The original US trajectories and the new trajectories from the enhanced sampling simulations were combined to calculate the final PMF (green curve). The shaded lines are the standard deviation of PMFs initiated from extended and helical conformations. The black curve is the reference PMF obtained by the multiple-walker ABF simulation.

■ CONCLUSIONS

A method for the analysis of US simulations was designed to help identify those specific window simulations that suffer most from sampling issues. In the divergence method, discrete conformational states are defined by using a set of order parameters, and TRAM estimators are used to estimate the unbiased distribution of the discrete states. For each window, the corresponding biasing potential is reapplied to the global unbiased distribution, resulting in a consensus distribution within a US window. Then, a divergence between the observed and the consensus distributions can be computed to identify windows that are plagued by sampling issues.

Similar measures have been proposed previously.[11,12] For example, the inconsistency between the PMF and the observed distribution along the predefined order parameter was used to identify pathological windows,[11,12] but such methods are limited to only identifying whether a certain window has inconsistent sampling. Our method can include any order parameter that is deemed important, which allows not only identifying problematic windows but also illuminating dynamics of the order parameters in a given window.

Important caveats of the method ought to be underlined. First, it relies on a transition matrix that is obtained from the simulation based on a limited amount of sampling. This transition matrix, therefore, may not have the desired level of accuracy. Alternatively, it is possible that the simulation may fail to capture a particular transition in the orthogonal space. If diffusion along the orthogonal space is so slow and is hidden to all windows in the course of the US simulation, then the method will not be able to detect sampling issues. Generating multiple, independent US trajectories could help circumvent these important shortcomings.

In turn, the quantified discrepancy can be used in an adaptive protocol to harness the available computational resources with increased efficiency. US simulations are embarrassingly parallelizable with minimum communication between the computing units, which makes it especially appealing for high-performance computing resources devoid of fast interconnection, for example, graphics processing unit (GPU) clusters or cloud computing resources. Employing a toy model, a simple restart of the simulation appears sufficient to reduce computational cost significantly. However, for a more complex system, more advanced reseeding approach, such as RE algorithms, would be necessary. Coupling with enhanced sampling methods, such as multiple-walker ABF, to discover accessible states and incorporate in the divergence analysis may also be a good strategy.

Recently, many nonequilibrium reseeding methods have been proposed for adaptive simulations and have demonstrated their effectiveness.[39,40] Unfortunately, our metric assumes equilibrium sampling in the observed distribution, and, as a consequence, it is not compatible with those nonequilibrium adaptive simulation protocols. Instead, equilibrium-accelerated sampling techniques, such as RE or weighted ensemble schemes,[41,42] could be used in conjunction with the present method.

AUTHOR INFORMATION

Corresponding Author

*E-mail: roux@uchicago.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to acknowledge Frank Noé and Christoph Wehmeyer at Freie Universitaet Berlin for helpful discussions and comments. We also thank Wonpil Im and Soohyung Park at the University of Kansas for helpful comments. This research used resources of the Argonne Leadership Computing Facility (ALCF), which is a DOE Oce of Science User Facility supported under Contract DE-AC02-06CH11357, an award of computer time provided by the INCITE program. We also gratefully acknowledge the computing resources provided by Argonne's Laboratory Computing Resource Center (LCRC). S.J. is supported by a Computational Post- doctoral Fellowship from ALCF. This work is supported by grant MCB-1517221 from the National Science Foundation and by the France and Chicago Collaborating in the Sciences (FACCTS) Center of the University of Chicago.

REFERENCES

(1) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte-Carlo Free-Energy Estimation—Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.

(2) Lelièvre, T.; Stoltz, G.; Rousset, M. *Free Energy Computations: A Mathematical Perspective*; Imperial College Press, 2010.

(3) Ferrenberg, A. M.; Swendsen, R. H. Optimized Monte-Carlo Data Analysis. *Phys. Rev. Lett.* **1989**, *63*, 1195−1198.

(4) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The Weighted Histogram Analysis Method for Free-energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13*, 1011−1021.

(5) Souaille, M.; Roux, B. Extension to the Weighted Histogram Analysis Method: Combining Umbrella Sampling with Free Energy Calculations. *Comput. Phys. Commun.* **2001**, *135*, 40−57.

(6) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.

(7) *Free Energy Calculations*; Chipot, C., Pohorille, A., Eds.; Springer Series in Chemical Physics; Springer: Berlin, 2007; Vol. *86*.

(8) Madras, N.; Randall, D. Markov Chain Decomposition for Convergence Rate Analysis. *Ann. Appl. Probab.* **2002**, *12*, 581−606.

(9) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations. *J. Chem. Theory Comput.* **2007**, *3*, 26−41.

(10) Hub, J. S.; de Groot, B. L.; van der Spoel, D. A. Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J. Chem. Theory Comput.* **2010**, *6*, 3713−3720.

(11) Zhu, F.; Hummer, G. Convergence and Error Estimation in Free Energy Calculations Using the Weighted Histogram Analysis Method. *J. Comput. Chem.* **2012**, *33*, 453−465.

(12) Jiang, W.; Luo, Y.; Maragliano, L.; Roux, B. Calculation of Free Energy Landscape in Multi-Dimensions with Hamiltonian-Exchange Umbrella Sampling on Petascale Supercomputer. *J. Chem. Theory Comput.* **2012**, *8*, 4672−4680.

(13) Wu, H.; Noé, F. Optimal Estimation of Free Energies and Stationary Densities from Multiple Biased Simulations. *Multiscale Model. Simul.* **2014**, *12*, 25−54.

(14) Wu, H.; Mey, A. S. J. S.; Rosta, E.; Noé, F. Statistically Optimal Analysis of State-Discretized Trajectory Data from Multiple Thermodynamic States. *J. Chem. Phys.* **2014**, *141*, 214106.

(15) Rosta, E.; Hummer, G. Free Energies from Dynamic Weighted Histogram Analysis Using Unbiased Markov State Model. *J. Chem. Theory Comput.* **2015**, *11*, 276−285.

(16) Park, S.; Im, W. Theory of Adaptive Optimization for Umbrella Sampling. *J. Chem. Theory Comput.* **2014**, *10*, 2719−2728.

DOI: 10.1021/acs.jpcb.6b05125
J. Phys. Chem. B 2016, 120, 8733−8742

(17) Chipot, C.; Hénin, J. Exploring the Free Energy Landscape of a Short Peptide Using an Average Force. *J. Chem. Phys.* **2005**, *123*, 244906.

(18) Comer, J.; Phillips, J. C.; Schulten, K.; Chipot, C. Multiple-walker Strategies for Free-energy Calculations in NAMD: Shared Adaptive Biasing Force and Walker Selection Rules. *J. Chem. Theory Comput.* **2014**, *10*, 5276−5285.

(19) Darve, E.; Pohorille, A. Calculating Free Energies Using Average Force. *J. Chem. Phys.* **2001**, *115*, 9169.

(20) Comer, J.; Gumbart, J. C.; Hénin, J.; Lelièvre, T.; Pohorille, A.; Chipot, C. The Adaptive Biasing Force Method: Everything You Always Wanted to Know, But Were Afraid to Ask. *J. Phys. Chem. B* **2015**, *119*, 1129−1151.

(21) Comer, J.; Roux, B.; Chipot, C. Achieving Ergodic Sampling Using Replica-Exchange Free-Energy Calculations. *Mol. Simul.* **2014**, *40*, 218−228.

(22) Minoukadeh, K.; Chipot, C.; Lelièvre, T. Potential of Mean Force Calculations: A Multiple-Walker Adaptive Biasing Force Approach. *J. Chem. Theory Comput.* **2010**, *6*, 1008−1017.

(23) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−1802.

(24) Molgedey, L.; Schuster, H. G. Separation of a Mixture of Independent Signals Using Time Delayed Correlations. *Phys. Rev. Lett.* **1994**, *72*, 3634−3637.

(25) Hyviirinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis*; Wiley and Sons, 2001.

(26) Naritomi, Y.; Fuchigami, S. Slow Dynamics in Protein Fluctuations Revealed by Time-structure Based Independent Component Analysis: The Case of Domain Motions. *J. Chem. Phys.* **2011**, *134*, 065101−065109.

(27) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139*, 015102.

(28) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525−5542.

(29) Sculley, D. Web-Scale K-Means Clustering, *Proceedings of the 19th International Conference on World Wide Web*; ACM: New York, NY, USA, 2010; pp 1177−1178.

(30) Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145−151.

(31) Crooks, G. E. Inequalities between the Jensen-Shannon and Jeffreys Divergences, Technical Report 004, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA, 2008.

(32) Buchete, N.-V.; Hummer, G. Coarse Master Equations for Peptide Folding Dynamics. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.

(33) Noé, F.; Fischer, S. Transition Networks for Modeling the Kinetics of Conformational Change in Macromolecules. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154−162.

(34) Zheng, L.; Chen, M.; Yang, W. Random Walk in Orthogonal Space to Achieve Efficient Free-energy Simulation of Complex Systems. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20227−20232.

(35) Jenkinson, A. F. The Frequency Distribution of the Annual Maximum (or Minimum) Values of Meteorological Elements. *Q. J. R. Meteorol. Soc.* **1955**, *81*, 158−171.

(36) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(37) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(38) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139*, 015102.

(39) Huang, X.; Bowman, G. R.; Bacallado, S.; Pande, V. S. Rapid Equilibrium Sampling Initiated from Nonequilibrium Data. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 19765−19769.

(40) Zimmerman, M. I.; Bowman, G. R. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J. Chem. Theory Comput.* **2015**, *11*, 5747−5757.

(41) Huber, G. A.; Kim, S. Weighted-Ensemble Brownian Dynamics Simulations for Protein Association Reactions. *Biophys. J.* **1996**, *70*, 97−110.

(42) Bhatt, D.; Zuckerman, D. M. Beyond Microscopic Reversibility: Are Observable Non-equilibrium Processes Precisely Reversible? *J. Chem. Theory Comput.* **2011**, *7*, 2520−2527.

DOI: 10.1021/acs.jpcb.6b05125