Combining the Biased and Unbiased Sampling Strategy into One Convenient Free Energy Calculation Method

Haomiao Zhang, Qiankun Gong, Haozhe Zhang, and Changjun Chen *

Constructing a free energy landscape for a large molecule is difficult. One has to use either a high temperature or a strong driving force to enhance the sampling on the free energy barriers. In this work, we propose a mixed method that combines these two kinds of acceleration strategies into one simulation. First, it applies an adaptive biasing potential to some replicas of the molecule. These replicas are particularly accelerated in a collective variable space. Second, it places some unbiased and exchangeable replicas at various temperature levels. These replicas generate unbiased sampling data in the canonical ensemble. To improve the sampling efficiency, biased replicas transfer their state variables to the unbiased replicas after equilibrium by Monte Carlo trial moves. In comparison to previous integrated methods, it is more convenient for users. It does not need an initial reference biasing potential to guide the sampling of the molecule. And it is also unnecessary to insert many replicas for the requirement of passing the free energy barriers. The free energy calculation is accomplished in a single stage. It samples the data as fast as a biased simulation and it processes the data as simple as an unbiased simulation. The method provides a minimalist approach to the construction of the free energy landscape. © 2019 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.25834

Introduction

State transition could be quite complicated for a biomolecule. To find out the related mechanism, one must construct its complete free energy landscape in a space spanned by a few predefined collective variables. This landscape shows the stable states, transition states, optimal pathways, and other important information about the molecule. Free energy calculation costs an enormous amount of simulation time. Up to now, lots of methods have been developed for this purpose, like adaptively biased molecular dynamics (ABMD),[1–3] metadynamics,[4–6] adaptive biasing force (ABF),[7,8] adaptive weighted sampling,[9] standard umbrella sampling method,[10–12] string method,[13,14] and self-healing umbrella sampling method.[15] All these methods can be referred to as the biasing method. As a function of a set of collective variables, the biasing potential flattens the basins and barriers on the original free energy landscape gradually in the simulation. This greatly increases the sampling efficiency of a molecule in the collective variable space. And moreover, there are some other methods that accelerate the simulation by coupling the collective variables to a high temperature, like adiabatic free energy dynamics,[16,17] amplified collective motion,[18] temperature-accelerated molecular dynamics (TAMD),[19–22] integrated tempering sampling (ITS),[23] and the combined ITS–TAMD method.[24]

However, the problem is still unresolved. Even if one has obtained a correct free energy landscape in a collective variable space, the information is still "incomplete." The collective variables in use may not catch all the features on the state transition process of a molecule. There could be lots of minima and barriers overlapped on the landscape (hidden minima and barriers). To get a well-rounded understanding of a transition process, one has to choose many different collective variables and reperform the free energy calculations repeatedly. Of course, it is a very time-consuming process.

An alternative way is to combine the biased simulation with an unbiased simulation by some kinds of temperature- or Hamiltonian-based replica exchange molecular dynamics (REMD) methods.[25–39] Lots of replicas of a molecule are added to the simulation system. The unbiased replicas of a molecule collect the data in the canonical ensemble for the free energy calculation. And the biased replicas modify their original Hamiltonians or environmental temperatures for the fast sampling purpose. Different replicas exchange their Hamiltonians or temperatures at a fixed time step. With the sampling data from the unbiased replicas, one can calculate the free energy landscape in any collective variable space. However, we want to note that the calculated free energy could not be so accurate if the new selected collective variables are completely independent of that in the biasing potential. In such case, the form of the biasing potential must make a corresponding adjustment.

According to the variables in exchange, REMD methods are mainly divided into three categories: Temperature Replica Exchange Molecular Dynamics (T-REMD),[25,26,30] Hamiltonian Replica Exchange Molecular Dynamics (H-REMD),[3,32,33] and Hamiltonian-Temperature Replica Exchange Molecular Dynamics (HT-REMD).[37,38,40] Among them, T-REMD is the most convenient one. All the replicas are simply placed at different temperatures in the simulation. High

[a] H. Zhang, Q. Gong, H. Zhang, C. Chen
Biomolecular Physics and Modeling Group, School of Physics, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China
E-mail: cjchen@hust.edu.cn

Contract Grant sponsor: National Natural Science Foundation of China;
Contract Grant numbers: 31370848, 31770773

temperature allows the replica to move fast in the conformational space. As a comparison, H-REMD is more complicated but efficient at the same time. The Hamiltonians of the biased replicas are well tuned to smooth the original free energy landscape before the formal simulation. Actually, T-REMD and H-REMD can be used simultaneously in the simulation, which is called HT-REMD.[37,38,40] It accelerates the sampling of the molecule by both the modified Hamiltonians and temperatures.

In practice, the efficiency of an REMD simulation is mainly limited by two issues.[37] One is the exchange rate between the biased and the unbiased replicas. This can be resolved by adding more replicas to the system so that the potentials of neighboring replicas are close enough to each other. The other issue is the sampling ability of the hottest (mostly biased) replica on the free energy landscape. To enhance the sampling, one can increase the kinetic energy[25,26] or the potential energy,[41–43] weaken the atom-to-atom interactions,[38] or apply a biasing potential or force.[1–8] The first two ways are particularly effective for the sampling in the conformational space. However, instead of a whole landscape, one might be only interested in a small scope on it, like the region between two functional states of a protein. In such situation, adding a biasing force on some collective variables is an economic technique. One first defines a set of collective variables to indicate the transition process from the reactant to the product and then applies a biasing force to accelerate the sampling in the direction. The biasing force or potential can be prepared in advance[3,37,40] or updated on the fly.[1,44] Generally, preparation of a proper biasing potential may demand expensive computational costs.

In this work, we propose a variant method of ABMD[1–3] and T-REMD.[25,26] It applies an adaptive biasing potential[1–3] to some replicas for fast sampling in the collective variable space. The state variables of the biased replicas are transferred conditionally to the unbiased replicas for producing the widespread unbiased sampling data.[25,26] Unlike previous H-REMD[3,32,33] and HT-REMD,[37,40] it does not need a reference biasing potential prior to the formal simulation.

Here, we want to discuss more about the biasing potential. In many methods, the biasing potential (or biasing force) is involved in both the conformational sampling process and the free energy calculation process.[1–8] It works well. However, the biasing potential is a function of a few predefined collective variables. Its negative value only shows us a projection of the real free energy landscape in a low-dimensional collective variable space. The landscape on other collective variables is still unavailable. To circumvent this issue, we add additional unbiased replicas in the simulation to collect the unbiased sampling data for the construction of a more comprehensive free energy landscape.

Our method is a hybrid simulation method. It is related to some existing methods in the literatures. For example, reservoir REMD (R-REMD)[45] also uses hybrid systems. One is a database that collects lots of the candidate states in the canonical ensemble. It is prepared by a normal molecular dynamics (MD) simulation at a very high temperature. The other is a standard REMD system, whose hottest replica tries to exchange with a randomly selected state in the previous database periodically. In R-REMD, both of the two systems are not particularly accelerated by a biasing potential on the collective variables. And moreover, storage of the candidate states in the database before the formal REMD simulation will consume more memories on the computer for large molecules.

Another related method is named replica exchange with collective-variable tempering (RECT).[46] The method combines well-tempered metadynamics[47,48] with REMD.[25,26,49,50] A set of biasing potentials are exerted on the replicas except the only unbiased one. Each biasing potential has a boosting temperature ΔT from 0 to ∞. $\Delta T = 0$ means that the potential does not grow in the simulation (as a standard MD simulation) and $\Delta T \to \infty$ means that the potential grows at a constant rate (as a standard metadynamics simulation[4–6]). All the neighboring replicas exchange their biasing potentials with each other periodically in the simulation. All the sampling data can be well reweighted to produce the original free energy landscape by the well-known weighted histogram analysis method (WHAM),[51,52] binless WHAM method Multistate Bennett Acceptance Ratio (MBAR),[53] dynamic histogram analysis method (DHAM),[54,55] or other accurate free energy estimator.[56,57] Applications show that RECT performs well with multiple biasing potentials on multiple collective variables. However, two points must be noted here. First, all the replicas in RECT[46] are kept at the same environment temperature. It uses a boosting temperature ΔT to represent the sampling ability. "Cool" replicas (small ΔT) must exchange frequently with "hot" replicas (large ΔT) to obtain a high sampling speed. For a large system, this fast exchange requirement demands a large number of replicas in the simulation. Second, the biasing potentials of all the replicas are growing continuously during the simulation, except the only unbiased one at the ground level. This increases the gap from the unbiased replica to the rests in the potential energy space and disfavors the corresponding exchange gradually in the simulation.

Different from RECT,[46] we separate the biased replicas from the unbiased replicas completely. There is no exchange between them. Biased replicas sample in the collective variable space with a fast-growing biasing potential. They pass the free energy barriers and produce the widespread candidate states. And the unbiased replicas sample in the canonical ensembles by REMD[25,26] at different temperatures. Their sampling data are available for WHAM analysis.[51,52] And moreover, the presented method does not rely on the "hottest" REMD replica to pass the explicit barriers in the collective variable space. This allows an arbitrary number of replicas in the system. Setup for the simulation becomes more flexible to users.

Materials and Methods

In the free energy calculation, large degrees of freedom in a molecule and high free energy barriers between the metastable states may cause severe sampling problems. To accomplish the free energy calculation, one has to use some efficient methods to overcome the sampling problem. These methods might have the following features.

I. Biasing potential on some collective variables is applied to a molecule. The biasing potential cannot only increase the sampling speed but also restrict the sampling area to some important states.

II. No reference parameters or potentials are required before the formal simulation. The implementation is simplified to a single-step process.
III. Free energy landscape is not calculated from the biasing potential. And the final biasing potential is not required to flatten out all the minima and barriers on the landscape. A coarse potential is fine.
IV. Unbiased sampling data are available at different temperatures. When the simulation is over, the data can be processed by WHAM[51,52] for the production of the original free energy surface on different collective variables.

As far as we know, most of the existing methods only satisfy a part of the requirements above. To obtain some features, one might have to give up the others. For a normal user, sometimes, it is hard to make a choice. In this work, inspired by a lot of well-known methods like ABMD[1–3], REMD,[25,26] R-REMD,[45] RECT[46], and Monte Carlo method,[58] we present an integrated method to have all the features in one simulation. For convenience, the method is called mixing REMD method. Main idea of the method is illustrated in Figure 1.

The system is constituted of three kinds of replicas, namely, biased replicas, equilibrium replicas, and unbiased REMD replicas, respectively. Biased replicas are responsible for fast search in the collective variable space. They are simulated with the biasing potential. In this work, the potential is built by the well-tempered ABMD method.[59] Similar to metadynamics,[4–6,47,60,61] ABMD applies a time-dependent biasing potential in the simulation.[1–3] However, instead of recording and summing lots of history-dependent Gaussian functions, ABMD builds the biasing potential by a set of cubic B-splines[1]

$$U(t,\xi) = \sum_{m=1}^{n_{grids}} U_m(t) B_m(\xi) \quad (1)$$

As shown in the expression, this biasing potential $U(t, \xi)$ is a sum of basis functions $B_m(\xi)$ with different weights $U_m(t)$ (t is the time, ξ is the collective variable, and m is the grid index on ξ). The calculation quantity of the biasing potential is constant from the beginning to the end of the simulation.[1] As for the well-tempered version of the ABMD method,[59] it has the same idea as the well-tempered metadynamics method.[47,48] It uses a pseudotemperature T' to control the growing speed of the biasing potential. Large T' makes the potential grow fast in the simulation with large fluctuations (in the case of an infinite T', it turns into a standard ABMD). Small T' reduces the growing speed of the biasing potential as well as the free energy errors in the iterations. A well-tuned biasing potential can remove the distinguishable barriers or minima and let a molecule sample freely on the smoothed free energy landscape.

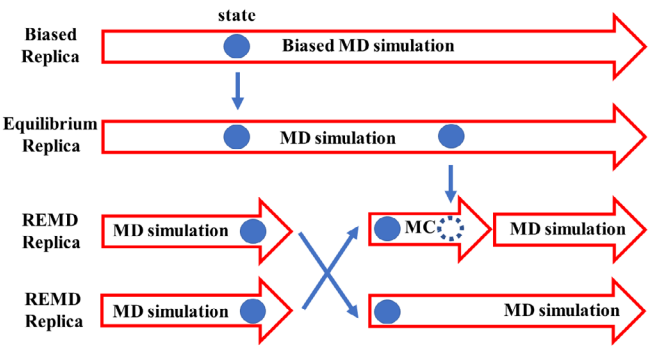

Figure 1. Schematic diagram for the mixing REMD method in this work. [Color figure can be viewed at wileyonlinelibrary.com]

Instead of ABMD, there are many other efficient methods available for the simulation of the biased replicas, like metadynamics,[4–6] ABF,[7,8] and steered MD.[62–64] These methods can all help the replica escape from the local minima on the free energy landscape quickly.

Biased replicas produce biased sampling data. To obtain unbiased data in the canonical ensemble, the instantaneous state variables of the biased replicas are copied to the equilibrium replicas periodically and further simulated for a short time period (equilibrium time, e.g., 20 ps). In the short equilibrium simulation, biasing potentials are removed. Then, the state variables of the equilibrium replicas are accepted or rejected by the REMD replicas at the same temperatures according to the Metropolis criterion.[65]

$$P = \min\{1, \exp(-(V(\mathbf{r}') - V(\mathbf{r}))/k_B T)\} \quad (2)$$

Here, P is the probability function for acceptance. $V(\mathbf{r}')$ and $V(\mathbf{r})$ are the potential energies of an equilibrium replica and an REMD replica, respectively. T is the environment temperature. If we look at the REMD replica, it just undergoes a state transition process as a standard Monte Carlo move.[58]

The acceptance probability could be small on the first use when the equilibrium replica just obtains the state variables from the biased replica. The potential energy could be much different from the subsequent REMD replica, even if they stay at the same temperature. To promote the transition, the criterion in eq. (2) is checked periodically in the following time. The acceptance probability should increase afterward when the conformation is sufficiently relaxed in the simulation.

After the REMD replicas successfully accept the new states of the equilibrium replicas, they have two ways to handle their own previous state variables (coordinates, velocities). The first way is to dump the data immediately. It lets the equilibrium replicas and the corresponding Central Processing Units (CPUs) stand idle until the next transition moment. The second way is to copy the state variables back to the equilibrium replicas for further simulation. The state variables might or might not return to the REMD replicas according to the same Metropolis criterion[65] [eq. (2)]. More candidate states are generated in the process. This way maximizes the CPU utilization and allows the REMD replicas to sample in more trajectories. Both of the two ways are optional in mixing REMD. In this work, we use the second one.

The REMD replicas are simulated at different temperatures as in a standard REMD simulation.[25,26] They exchange with each other according to the criterion:

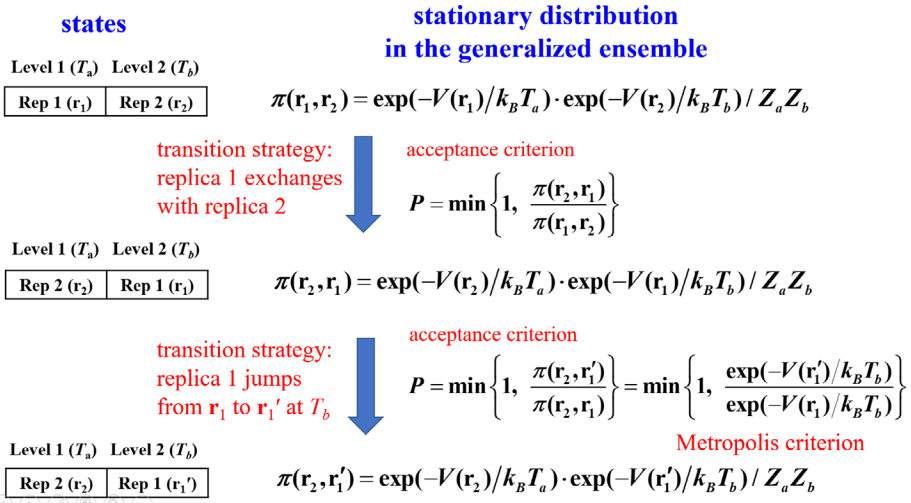

Figure 2. Illustration of the conformational transitions for the REMD replicas in an mixing REMD simulation. In the figure, $\mathbf{r}_1$ and $\mathbf{r}_2$ are the state variables of replica 1 and 2, respectively. $\mathbf{r}_1'$ is the new state of replica 1 that comes from the equilibrium replica (not shown in the figure). V is the potential energy. Z_a and Z_b are the partitional functions at T_a and T_b. π is the stationary distribution function of the states in the generalized ensemble. [Color figure can be viewed at wileyonlinelibrary.com]

$$P_{ij} = \min\left\{1, \frac{\exp\left(-V\left(\mathrm{r}_j^{(i)}\right)/k_BT^{(i)}\right)\cdot\exp\left(-V\left(\mathrm{r}_i^{(j)}\right)/k_BT^{(j)}\right)}{\exp\left(-V\left(\mathrm{r}_i^{(i)}\right)/k_BT^{(i)}\right)\cdot\exp\left(-V\left(\mathrm{r}_j^{(j)}\right)/k_BT^{(j)}\right)}\right\} \quad (3)$$

Here, P_{ij} is the probability function for a successful exchange. $V(\mathbf{r}i^{(j)})$ is the potential energy of the ith replica at jth temperature $T^{(j)}$. Different from the biased replicas, the unbiased REMD replicas are designed for the production of the unbiased sampling data. These data allow us to construct the original free energy landscape on different collective variables by WHAM method[51,52] during or after the simulation.

In simple terms, the presented mixing REMD method in this work uses the biased replicas to search the collective variable space and uses the unbiased REMD replicas to generate the sampling data. Both of them are indispensable to the simulation. Without biased replicas, REMD replicas cannot efficiently pass the high barriers on the free energy landscape. And without REMD replicas, biased replicas cannot withdraw the biasing potential and sample the canonical ensembles in an effective way.

Compared to a standard REMD, mixing REMD allows the REMD replicas to do additional conformational transitions at the constant temperatures. This can be illustrated by a system of two replicas at two different temperatures (T_a and T_b) in Figure 2. As shown in the figure, mixing REMD has two different transitions in the generalized ensemble. One is a normal replica exchange for all the REMD replicas and the other is a jump from $\mathbf{r}_1$ to $\mathbf{r}_1'$ for the hotter one at T_b (replica 1). Both transitions obey the same acceptance criterion as a standard REMD. It satisfies the detailed balance condition and ensures the stationary distribution of the states in the generalized ensemble.

Mixing REMD is an extension of many existing methods. It does have some advantages. We give a brief description in Table 1. For example, R-REMD[45] must prepare a conformational database at a very high temperature before the formal simulation and mixing REMD Does not have to do this. H-REMD[3,32,33] and HT-REMD[37,40] require an initial biasing potential for fast sampling purpose at the beginning. This is not necessary for mixing REMD too. Mixing REMD uses an adaptive biasing potential as well as ABMD[1,2] or metadynamics.[4–6] The shape of the potential evolves from a flat profile. It is easy to use in the simulation.

Moreover, adaptive biasing potential produces ABF. It helps a molecule jump out of any free energy minima efficiently in the simulation. And in H-REMD and HT-REMD, due to a static biasing potential, the sampling can only be efficient when the potential is well defined.

Finally, for the methods that have the adaptive biasing potentials, like ABMD[1,2] and metadynamics,[4–6] they construct the free energy landscape from the biased sampling data. As comparison, mixing REMD generates purely unbiased sampling data by its REMD part. The free energy calculation is easy to do in any collective variable space without an additional reweighting process. Moreover, with the REMD part, mixing REMD can further enhance the sampling on the other degrees of freedom that are not accelerated in the biased simulation.

To present the performance of the method, we use two models. The first is a simple Alanine (ALA) dipeptide.[14,44,66–69]

Table 1. Comparison of mixing REMD with some existing methods.

Method	Easy to run? (no initial dataset)	Efficient for fast sampling? (use adaptive potential)	Flexible in free energy calculation? (have unbiased data)
T-REMD[a]	Yes	No	Yes
R-REMD[b]	No	No	Yes
ABMD[c]	Yes	Yes	No
Metadynamics[d]	Yes	Yes	No
H-REMD[e]	No	No	Yes
HT-REMD[f]	No	No	Yes
Mixing REMD[g]	Yes	Yes	Yes

[a] T-REMD[25,26,30] enhances sampling by a set of high temperatures.
[b] R-REMD[45] enhances sampling by a widespread conformational database.
[c] ABMD[1,2] enhances sampling by an adaptive biasing potential of cubic B-splines.
[d] Metadynamics[4–6] enhances sampling by an adaptive biasing potential of Gaussian functions.
[e] H-REMD[3,32,33] enhances sampling by a set of static biasing potentials.
[f] HT-REMD[37,40] enhances sampling by a set of static biasing potentials and high temperatures.
[g] Mixing REMD enhances sampling by a set of adaptive biasing potentials and high temperatures.

The sequence is acetyl (ACE)-ALA-N-methyl (NME). Five replicas of the molecule are added to the simulation. Three of them are the unbiased REMD replicas at 400, 529.150, and 700 K, respectively. The rests are one equilibrium replica and one biased replica. For convenience, they are both placed at the same temperature as the hottest REMD replica (700 K). The small system allows us to do a sufficient search in its conformational space for the calculation of the potential energy distribution and the free energy surface. The data are compared with a standard REMD simulation at equivalent temperatures.

The second model is the N-terminal domain of ribosomal protein L9 (NTL9, pdb id: 2HBB[70]). It is an RNA binding protein with 51 residues and 813 atoms. Its native structure contains two α-helices and three β-sheets (Fig. 3). The formations of the secondary and tertiary structures in the protein are quite complicated. In the mixing REMD simulation of NTL9, we also use five replicas: three REMD replicas at 300, 314.643, and 330 K, one equilibrium replica at 330 K and one biased replica at 330 K. The simulation of the protein starts from an extremely extended conformation.

All the simulations are performed by the AMBER 16 software[71] in implicit solvent[72] (igb = 8). The force field is AMBER FF14SB[73] and the time step is 1 fs. The temperature is controlled by Langevin dynamics method[74] with the collision frequency 10.0 ps^{-1}. No bonds are constrained in the simulation. Mixing REMD is implemented by Fortran90 language. Its REMD code and ABMD code have been written separately to work with the Tinker software[75] in our previous papers.[76,77] Now they are integrated together as a module and inserted into the AMBER software.[71] For more discussions on the present method, please see the support information (Appendix S1)

Results and Discussion

We first show the simulation results of ALA dipeptide by the presented mixing REMD method. The total simulation time is 1000 ns. It has five replicas, including one biased replica, one equilibrium replica, and three REMD replicas. The biased replica is simulated at 700 K by the well-tempered ABMD method[59] on two collective variables: backbone angle Φ and Ψ. In the simulation, the biasing potential updates with a flooding time of 50 ps and a pseudotemperature of 5000 K. Every 1 ns, the biased replica directly copies its coordinates and velocities to the equilibrium replica at the same constant temperature. Then, the hottest REMD replica at the highest level (also 700 K) determines to accept or reject the state of the equilibrium replica every 20 ps (equilibrium time) by a standard Metropolis criterion[65] [eq. (2)]. All REMD replicas exchange with each other with a time period of 20 ps.

Figure 4 gives the distribution functions of three REMD replicas in the potential energy space in the mixing REMD simulation (black solid lines). They are simulated at 400, 529.150, and 700 K, respectively. To check the correctness of the data, we also show the distribution functions of three replicas in a standard 1000 ns REMD simulation[25,26] at the same temperatures (400, 529.150, and 700 K) with the same exchange time step (20 ps) (red dashed lines). The two data sets are in well agreement with each other.

Moreover, we also show the free energy surfaces on the collective variables of the molecule at the lowest temperature 400 K in Figure 5. Figures 5(a) and 5(b) are the surface of the mixing REMD simulation and the surface of the standard REMD simulation, respectively. Both surfaces are calculated by the PTWHAM code in Ref. [78], which is an implementation of the parallel-tempering WHAM method.[78] The results show that mixing REMD is efficient enough to sample all the minima in the whole collective variable space as well as the standard REMD. To make a future comparison, we project the two-dimensional (2D) free energy surfaces onto each collective variable (backbone angle Φ or Ψ) and plot the corresponding one-dimensional (1D) free energy profiles in Figures 5(c) and 5(d), respective. Here, the projection only involves the grids that are not 10.0 kcal/mol higher than the global minimum on the 2D surface. The 1D free energy profiles by the two methods are also in good agreement. These results indicate that the mixing REMD method is capable of producing the correct sampling data in the canonical ensemble.

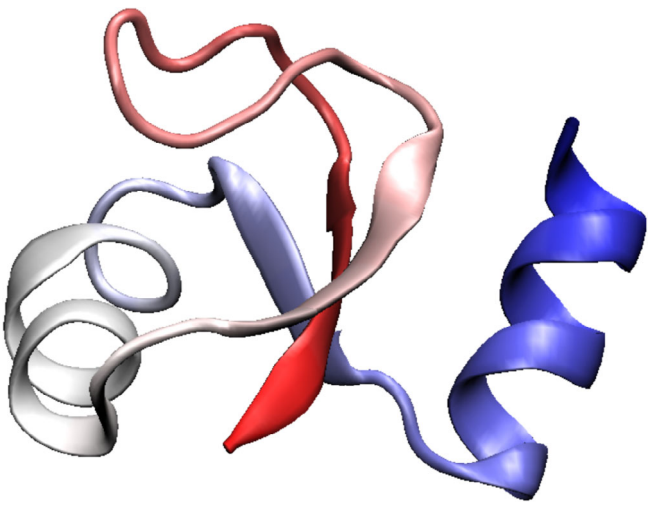

Figure 3. Native conformation of protein NTL9 in protein data bank[83] (pdb id: 2HBB[70]). The N- and C-terminal of the protein are rendered in red and blue, respectively. The figure is produced by Visual Molecular Dynamics (VMD).[84] [Color figure can be viewed at wileyonlinelibrary.com]

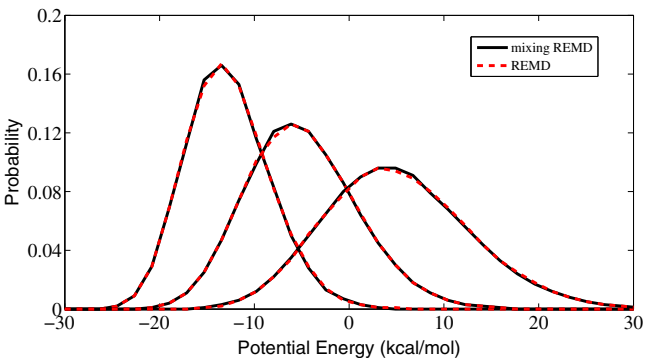

Figure 4. Potential energy distribution of three REMD replicas of ALA dipeptide in the simulation by the standard REMD method (red dashed lines) and the mixing REMD method (black solid lines). [Color figure can be viewed at wileyonlinelibrary.com]

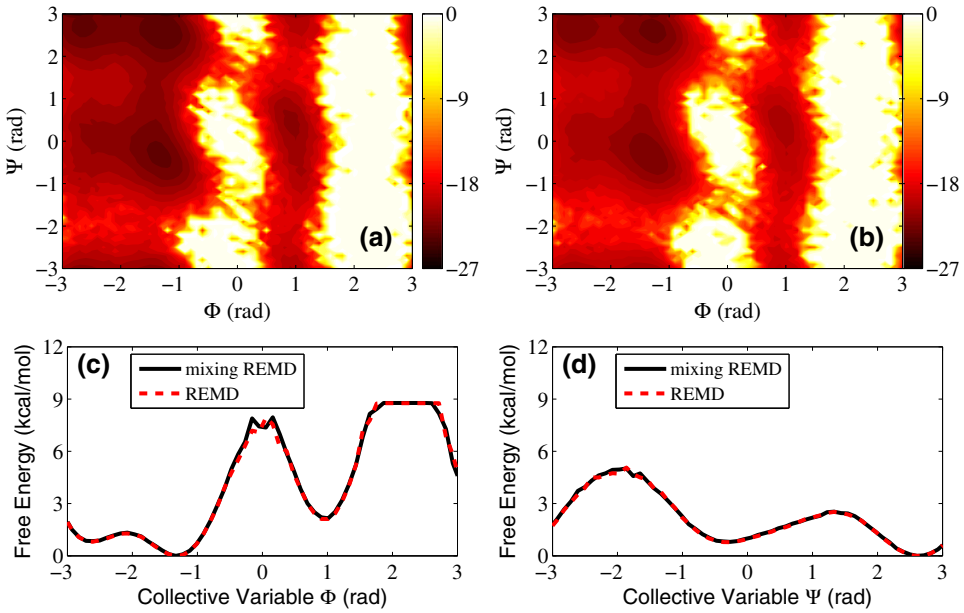

Figure 5. Free energy surfaces of ALA dipeptide on the two collective variables, backbone angle Φ and Ψ at 400 K. (a) Mixing REMD simulation. (b) Standard REMD simulation. The projections of the surfaces on each collective variable are shown in (c) and (d), respectively. Black solid lines: mixing REMD simulation. Red dashed lines: standard REMD simulation. [Color figure can be viewed at wiley onlinelibrary.com]

ALA dipeptide is a simple model to show the validity of the mixing REMD method. To further illustrate its fast sampling speed, we use a much more complicated model: protein NTL9 (Fig. 3). Because of the large degrees of freedom in the molecule, searching in its conformational space and constructing the free energy surface are very time consuming. Here, the free energy surface is represented by two collective variables. The first is C_α RMSD to the native state and the second is the number of contacts in the native state (contact criterion: residue–residue interval on the sequence larger than 10 residues and atom–atom distance smaller than 4.0 Å). Both collective variables are calculated by the nonequilibrium free energy module in AMBER16[71] (cv_type = "MULTI_RMSD" and "N_OF_BONDS").

To build the free energy surface, three typical simulations are carried out. The first is an RECT-like simulation. It has three replicas that are simulated by the well-tempered ABMD method[59] with a biasing potential on the two collective variables. Although with a different biasing method to the original RECT[46] (well-tempered ABMD[59] vs. well-tempered metadynamics[47,48]), it uses the same collective variable tempering idea. So, from now on, this simulation will be referred to as modified RECT simulation for convenience. The level parameter that represents the sampling ability of the replica is pseudotemperature,[59] which is similar to the boosting temperature in the original RECT.[46] The flooding time for all the replicas is 20 ps and the pseudotemperatures are 0, 2500, and 5000 K. Obviously, the first replica is equivalent to an unbiased one in a normal MD simulation. The environment temperatures of all the replicas are the same of 300 K. Exchange between the neighboring replicas is tried every 40 ps according to the following probability function[46]

$$P_{ij} = \min\left\{1, \frac{\exp\left(-V^{(i)}(\mathrm{r}_j)/k_BT\right)\cdot\exp\left(-V^{(j)}(\mathrm{r}_i)/k_BT\right)}{\exp\left(-V^{(i)}(\mathrm{r}_i)/k_BT\right)\cdot\exp\left(-V^{(j)}(\mathrm{r}_j)/k_BT\right)}\right\} \quad (4)$$

Here, T is the temperature and $V^{(j)}(\mathbf{r}_i)$ is the potential energy of the ith replica at jth biasing potential level. Each level is corresponding to an adaptive biasing potential that is constructed by the well-tempered ABMD method with a particular pseudotemperature.[59] The second simulation for protein NTL9 is a standard REMD simulation.[25,26] It also has three replicas. However, they are placed at 300, 314.643, and 330 K, respectively. Replica exchange is tried by eq. (3) every 40 ps. No biasing potential is used for the replicas in the simulation.

RECT improves the sampling efficiency by an adaptive biasing potential[46] and REMD gets this by a high temperature.[25,26] They are chosen here because of their easy usability. They do not need an initial guess of a reference parameter or potential before the formal simulation. However, both of them have a global level parameter to represent the sampling abilities of the replicas, like the boosting temperature in RECT and the environment temperature in REMD. A large level parameter improves the sampling efficiency on the rare events. In a space of the level parameter, the replicas must be close enough to each other to satisfy the fast exchange requirement. This demands a large number of replicas, especially for a large molecule with high free energy barriers.

The last simulation is performed by the mixing REMD method in this work. In fact, it is an integrated method of the well-known methods like ABMD,[1–3] REMD,[25,26] R-REMD,[45] RECT,[46] and Monte Carlo method.[58] There are five replicas in this simulation. One for the ABMD simulation at 330 K, one for the equilibrium simulation at 330 K and three for REMD simulation at 300, 314.643, and 330 K, respectively (same to previous REMD simulation). The biased ABMD replica is also simulated by the well-tempered ABMD method.[59] Its flooding time is 20 ps, pseudotemperature is 5000 K, and collective variables are C_α RMSD and number of native contacts. All these biasing parameters are the same as the replica at the highest level of the modified RECT simulation. The state variables of the biased replica are directly sent to the equilibrium replica every 1 ns and further sent to the hottest REMD replica every 40 ps (equilibrium time)

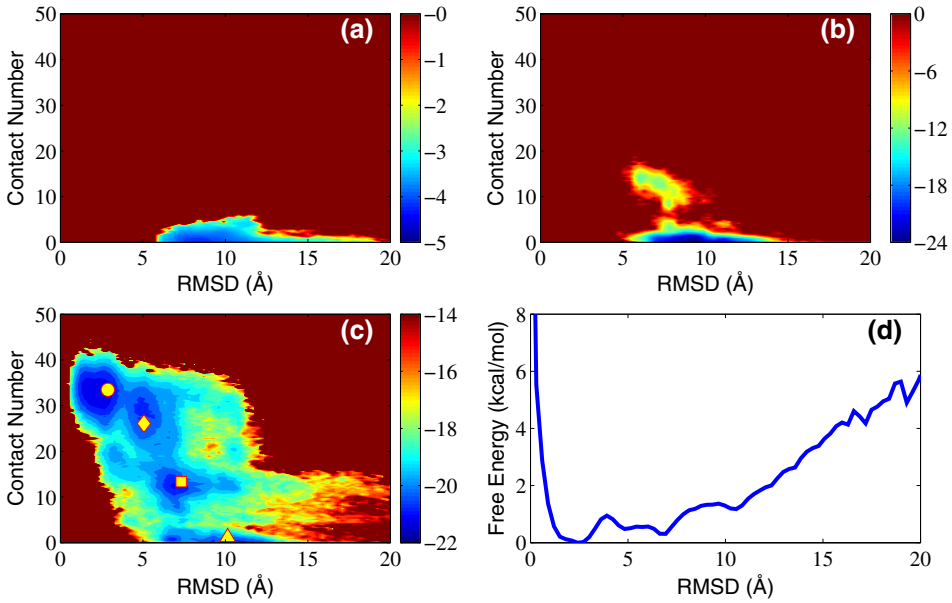

Figure 6. Free energy surfaces of the protein NTL9 in the (a) REMD simulation, (b) modified RECT simulation, and (c) mixing REMD simulation. All the free energy units are kcal/mol. Panel (d) is a projection of the free energy surface of mixing REMD to its first collective variable C_α RMSD. [Color figure can be viewed at wileyonlinelibrary.com]

according to the Metropolis acceptance criterion[65] in eq. (2). Exchange time period for the REMD replicas is 40 ps.

Mixing REMD method has the same advantage as RECT and REMD. No initial guess of a reference parameter or potential is required before the simulation. And more importantly, it does not have a global level parameter to differentiate the sampling abilities of the unbiased replicas. It is true that REMD replicas in the mixing REMD simulation are placed at different temperatures as those in a standard REMD simulation. However, this is mainly for the generation of the unbiased sampling data, not for the fast sampling purpose. The mission for passing the explicit free energy barriers in the collective variable space is up to the biased replicas. So, unlike a standard REMD,[25,26] mixing REMD does not have to insert many replicas in the temperature space to reach the upper limit of the temperature. It is more flexible to users.

All the simulations last for 2000 ns long and start from an extremely extended state, which is greatly different from the native state of the protein. The free energy surfaces of the protein in the three simulations, REMD, modified RECT, and mixing REMD, are shown in Figures 6(a)–6(c), respectively. All the surfaces are constructed from the unbiased sampling data by PTWHAM code in Ref. [78]. It must be noted that REMD and mixing REMD simulations have three unbiased replicas, but modified RECT simulation only has one. This is because modified RECT performs the replica exchange in a different pseudotemperature space. Among the replicas, only the one stays at the ground level is simulated without a biasing potential. However, at the same time, modified RECT has two biased replicas, which are more than that in the REMD and mixing REMD simulation.

From the figure, we find that REMD is the most inefficient one of the three simulations [Fig. 6(a)]. All its replicas are trapped into a free energy minimum at the bottom from the beginning to the end. The upper limit of the temperature 330 K is not high enough to let the replicas escape from the minimum in a limited simulation time. This is a common sampling problem of REMD. Simply increasing the temperature range is not always a good solution. Because a wide temperature range separates the replicas far away from each other. Successful exchanges between the replicas become difficult. "Hot" replicas at high levels cannot effectively pull the "cool" replicas at low levels.

As to this point, applying a biasing potential is more practical. Figure 6(b) shows that modified RECT simulation has a broader free energy surface than REMD. However, in the whole 2000 ns simulation time, it only finds one more free energy minimum at the top of the initial one. As we discussed in the Introduction section, RECT also requires fast exchanges between the replicas. A small exchange rate might slow down the sampling speed. Moreover, the biasing potentials of the replicas grow continuously in the simulation, except the unbiased one at the ground level. This potential gap further prevents the unbiased replica jumping to the high levels in the simulation.

Compared to REMD and modified RECT, mixing REMD performs much better. It shows a more comprehensive free energy surface [Fig. 6(c)]. Some important information is revealed here. For example, there are four distinguishable free energy minima from the unfolded state at the bottom right corner to the native state at the up left corner. These minima indicate a potential folding pathway of the protein. To get a better view on this pathway, we divide all the sampling data at 300 K (200,000 points) into four clusters.

In the past, there were two well-known clustering algorithms, named k-medoids[79,80] and k-means.[81] For k-medoids algorithm,[79,80] at every iteration step, it first selects a random data point (structure) in each cluster as a temporary center. Then, it calculates a global average distance function D for all the data points according to these temporary center points.[80]

$$D = \frac{1}{N_{\text{tot}}} \sum_i \sum_k (c_k(i) - c_{k,i})^2 \qquad (5)$$

Here, N_{tot} is the total number of data points (structures) in the trajectory. i and k are the indices of the data points and the collective variables, respectively. $c_k(i)$ is the value of the kth

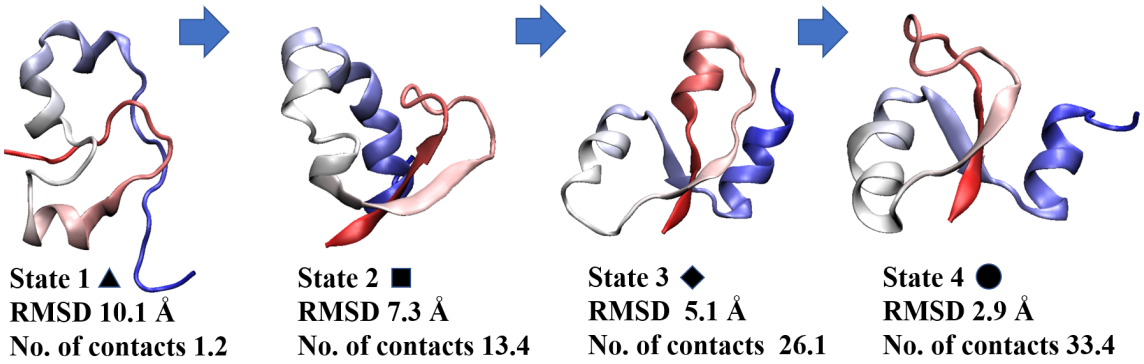

Figure 7. Representative structures of the four minima on the free energy surface of the protein NTL9 in the mixing REMD simulation. The positions of the minima are marked in Figure 6(c) by the triangle (state 1), square (state 2), diamond (state 3), and circle (state 4), respectively. The figure is produced by VMD.[84] [Color figure can be viewed at wileyonlinelibrary.com]

collective variable of the data point *i*. $c_{k,i}$ is the *k*th collective variable of the temporary center point that is nearest to point *i*. If the average distance function *D* is smaller than previous iteration step, these temporary center points will be accepted as the formal centers of the clusters. And all the other data points are reassigned to them subsequently. Otherwise, a new selection of the temporary center points will be used for the calculation of the average distance function again. k-means algorithm is slightly different to k-medoids. It does not attempt to choose the center points of the clusters in a random way. It simply calculates the average positions of the clusters at every iteration step and then assigns the rest data points (structures) to them.[81]

Based on k-medoids[79,80] and k-means,[81] some improved clustering methods have been proposed, like a simple and fast k-medoids method in Ref. [82]. In the method, a center of a cluster is only determined by its own data points (same as k-means) and it locates at a real data point, not an average position (same as k-medoids). We use the clustering method in our work. At every iteration step, a local average distance function *d*(*i*) is calculated for each data point *i* to its neighbors in a particular cluster.[82]

$$d(i) = \frac{1}{n-1}\sum_{j\neq i}\sum_{k}(c_k(i)-c_k(j))^2 \qquad (6)$$

Here, *n* is the number of data points (structures) included in the current cluster. *i* and *j* are the indices of the data points with a maximum of *n*. *k* is the index of the collective variable. $c_k(i)$ and $c_k(j)$ are the values of the *k*th collective variable of the data points *i* and *j*, respectively. When the calculation is over, the data point with a smallest distance function *d* is set as the center of the cluster. This is the only difference to k-means algorithm.[81] The ways of assigning the data points to the cluster centers are the same.

As a result of clustering, four center points of the clusters are found at (10.1, 1.2), (7.3, 13.4), (5.1, 26.1), and (2.9, 33.4). They are marked by the triangle (state 1), square (state 2), diamond (state 3), and circle (state 4) from bottom to top, respectively [Fig. 6(c)]. The structures corresponding to the four states are shown in Figure 7. It presents that the protein collapses into a coiled structure first at the beginning of the folding (state 1). The state only contains some α-helix fragments. Then two β-sheets emerge at the N-terminal of the protein (state 2). In the third stage (state 3), the long helix at C-terminal (in blue color) breaks into one β-sheet and one α-helix. Now all the secondary structures in the protein have formed, including two α-helices (at the left and the right side) and three β-sheets (at middle). These fragments are packed in the same way as that in the native state (Fig. 3). Meanwhile, an incorrect helix shows up at the top. However, it turns into a correct loop structure finally in state 4. In this stage, all the secondary structures extend their lengths on both ends to stabilize the protein structure.

Conclusions

Free energy calculation for biomolecules is not easy in practice. Poor sampling on the free energy landscape might prevent the molecule passing the high barriers. Till now, three kinds of strategies are proposed. The first is the temperature acceleration strategy. Coupling to a high temperature allows a molecule to increase the kinetic energy and climb up the barriers. The second is the biasing potential acceleration strategy. The artificial force from the potential drives the molecule over the barriers even if in a normal temperature environment. Such a biasing potential can be prepared before the simulation or updated on the fly during the simulation. The last one is a combination of the two strategies above. These methods have been proved effective in many applications.

In this work, we propose an integrated mixing REMD method that belongs to the third strategy. The entire simulation system is divided into three subsystems. Some replicas of a molecule sample quickly in the collective variable space by an adaptive biasing potential (biased replicas[1–3]). Others stay at various temperatures with the original all-atom force field (unbiased REMD replicas[25,26]). And they are communicated by the replicas in an equilibrium simulation (equilibrium replicas). Calculations show that this integrated method works fine in the free energy calculation of different molecules.

We want to note that the REMD part of the method cannot further enhance the sampling of its bias part in a predefined collective variable space. Adaptive biasing potential flattens the original free energy surface more efficiently than a higher

temperature. In this work, the purpose of the REMD part is to accelerate the sampling in the whole conformational space, including those degrees of freedom that are not defined in the biasing potential. Of course, the REMD part cooperates with the bias part. It still keeps the fast sampling ability on the selected collective variables.

Another purpose of the REMD part is for the generation of the unbiased sampling data in the simulation. Compared to the biased sampling data, unbiased data are easier to use in the clustering analysis or the calculation of the thermodynamics quantities. The data do not need to be reweighted by the biasing potential, especially when the adaptive biasing potential is not yet converged in the simulation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under grant nos. 31770773 and 31370848.

Keywords: collective variable · free energy calculation · replica exchange molecular dynamics · adaptively biased molecular dynamics

How to cite this article: H. Zhang, Q. Gong, H. Zhang, C. Chen. *J. Comput. Chem.* **2019**, *40*, 1806–1815. DOI: 10.1002/jcc.25834

Additional Supporting Information may be found in the online version of this article.

[1] V. Babin, C. Roland, C. Sagui, *J. Chem. Phys.* **2008**, *128*, 134101.
[2] M. Moradi, V. Babin, C. Roland, T. A. Darden, C. Sagui, *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 20746.
[3] V. Babin, C. Sagui, *J. Chem. Phys.* **2010**, *132*, 104108.
[4] A. Laio, M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562.
[5] A. Laio, A. Rodriguez-Fortea, F. L. Gervasio, M. Ceccarelli, M. Parrinello, *J. Phys. Chem. B* **2005**, *109*, 6714.
[6] G. Bussi, A. Laio, M. Parrinello, *Phys. Rev. Lett.* **2006**, *96*, 090601.
[7] E. Darve, A. Pohorille, *J. Chem. Phys.* **2001**, *115*, 9169.
[8] E. Darve, D. Rodriguez-Gomez, A. Pohorille, *J. Chem. Phys.* **2008**, *128*, 144120.
[9] S. Park, D. L. Ensign, V. S. Pande, *Phys. Rev. E* **2006**, *74*, 066703.
[10] G. M. Torrie, J. P. Valleau, *J. Comput. Phys.* **1977**, *23*, 187.
[11] J. Wang, Y. Gu, H. Y. Liu, *J. Chem. Phys.* **2006**, *125*, 094907.
[12] D. T. Major, J. Gao, *J. Chem. Theory Comput.* **2007**, *3*, 949.
[13] E. Weinan, W. Q. Ren, E. Vanden-Eijnden, *Phys. Rev. B* **2002**, *66*, 052301.
[14] L. Maragliano, A. Fischer, E. Vanden-Eijnden, G. Ciccotti, *J. Chem. Phys.* **2006**, *125*, 024106.
[15] S. Marsili, A. Barducci, R. Chelli, P. Procacci, V. Schettino, *J. Phys. Chem. B* **2006**, *110*, 14011.
[16] L. Rosso, P. Minary, Z. Zhu, M. E. Tuckerman, *J. Chem. Phys.* **2002**, *116*, 4389.
[17] J. B. Abrams, M. E. Tuckerman, *J. Phys. Chem. B* **2008**, *112*, 15742.
[18] Z. Zhang, Y. Shi, H. Liu, *Biophys. J.* **2003**, *84*, 3583.
[19] L. Maragliano, E. Vanden-Eijnden, *Chem. Phys. Lett.* **2006**, *426*, 168.
[20] C. F. Abrams, E. Vanden-Eijnden, *Chem. Phys. Lett.* **2012**, *547*, 114.
[21] Y. Hu, W. Hong, Y. Shi, H. Liu, *J. Chem. Theory Comput.* **2012**, *8*, 3777.
[22] S. A. Paz, C. F. Abrams, *J. Chem. Theory Comput.* **2015**, *11*, 5024.
[23] Y. Q. Gao, *J. Chem. Phys.* **2008**, *128*, 064105.
[24] L. Xie, L. Shen, Z. Chen, M. Yang, *J. Chem. Phys.* **2017**, *146*, 024103.
[25] Y. Sugita, Y. Okamoto, *Chem. Phys. Lett.* **1999**, *314*, 141.
[26] Y. Sugita, A. Kitao, Y. Okamoto, *J. Chem. Phys.* **2000**, *113*, 6042.
[27] M. Andrec, A. K. Felts, E. Gallicchio, R. M. Levy, *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6801.
[28] W. Li, J. Zhang, J. Wang, W. Wang, *J. Am. Chem. Soc.* **2008**, *130*, 892.
[29] K. Ostermeir, M. Zacharias, *Biochim. Biophys. Acta* **2013**, *1834*, 847.
[30] E. Rosta, G. Hummer, *J. Chem. Phys.* **2009**, *131*, 165102.
[31] J. Zhang, M. Qin, W. Wang, *Proteins: Struct. Funct. Bioinf.* **2006**, *62*, 672.
[32] S. Kannan, M. Zacharias, *Proteins: Struc. Funct. Bioinf.* **2007**, *66*, 697.
[33] G. Xu, J. Wang, H. Liu, *J. Chem. Theory Comput.* **2008**, *4*, 1348.
[34] M. Fajer, D. Hamelberg, J. A. McCammon, *J. Chem. Theory Comput.* **2008**, *4*, 1565.
[35] W. Jiang, M. Hodoscek, B. Roux, *J. Chem. Theory Comput.* **2009**, *5*, 2583.
[36] W. Jiang, B. Roux, *J. Chem. Theory Comput.* **2010**, *6*, 2559.
[37] M. Moradi, V. Babin, C. Roland, C. Sagui, *J. Chem. Phys.* **2010**, *133*, 125104.
[38] R. Laghaei, N. Mousseau, G. Wei, *J. Phys. Chem. B* **2011**, *115*, 3146.
[39] E. Rosta, M. Nowotny, W. Yang, G. Hummer, *J. Am. Chem. Soc.* **2011**, *133*, 8934.
[40] M. Deighan, J. Pfaendtner, *Langmuir* **2013**, *29*, 7999.
[41] D. Hamelberg, J. Mongan, J. A. McCammon, *J. Chem. Phys.* **2004**, *120*, 11919.
[42] D. Bucher, L. C. T. Pierce, J. A. McCammon, P. R. L. Markwick, *J. Chem. Theory Comput.* **2011**, *7*, 890.
[43] Y. Miao, F. Feixas, C. Eun, J. A. McCammon, *J. Comput. Chem.* **2015**, *36*, 1536.
[44] B. Ensing, M. D. Vivo, Z. Liu, P. Moore, M. L. Klein, *Acc. Chem. Res.* **2006**, *39*, 73.
[45] A. Okur, D. R. Roe, G. Cui, V. Hornak, C. Simmerling, *J. Chem. Theory Comput.* **2007**, *3*, 557.
[46] A. Gil-Ley, G. Bussi, *J. Chem. Theory Comput.* **2015**, *11*, 1077.
[47] A. Barducci, G. Bussi, M. Parrinello, *Phys. Rev. Lett.* **2008**, *100*, 020603.
[48] J. F. Dama, M. Parrinello, G. A. Voth, *Phys. Rev. Lett.* **2014**, *112*, 240602.
[49] G. Bussi, F. L. Gervasio, A. Laio, M. Parrinello, *J. Am. Chem. Soc.* **2006**, *128*, 13435.
[50] M. Deighan, M. Bonomi, J. Pfaendtner, *J. Chem. Theory Comput.* **2012**, *8*, 2189.
[51] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, J. M. Rosenberg, *J. Comput. Chem.* **1992**, *8*, 1011.
[52] I. C. Yeh, M. S. Lee, M. A. Olson, *J. Phys. Chem. B* **2008**, *112*, 15064.
[53] M. R. Shirts, J. D. Chodera, *J. Chem. Phys.* **2008**, *129*, 124105.
[54] E. Rosta, G. Hummer, *J. Chem. Theory Comput.* **2015**, *11*, 276.
[55] L. S. Stelzl, A. Kells, E. Rosta, G. Hummer, *J. Chem. Theory Comput.* **2017**, *13*, 6328.
[56] P. Tiwary, M. Parrinello, *J. Phys. Chem. B* **2015**, *119*, 736.
[57] P. Tiwary, B. J. Berne, *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 2839.
[58] N. Metropolis, S. Ulam, *J. Am. Stat. Assoc.* **1949**, *44*, 335.
[59] M. Moradi, V. Babin, C. Roland, C. Sagui, *J. Phy.: Conf. Ser.* **2015**, *640*, 012020.
[60] L. Konermann, J. X. Pan, *Biochemistry* **2010**, *49*, 3477.
[61] J. F. Dama, G. M. Hocky, R. Sun, G. A. Voth, *J. Chem. Theory Comput.* **2015**, *11*, 5638.
[62] C. Jarzynski, *Phys. Rev. Lett.* **1997**, *78*, 2690.
[63] S. Park, F. Khalili-Araghi, E. Tajkhorshid, K. Schulten, *J. Chem. Phys.* **2003**, *119*, 3559.
[64] L. Y. Chen, D. A. Bastien, H. E. Espejel, *Phys. Chem. Chem. Phys.* **2010**, *12*, 6579.
[65] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, *J. Chem. Phys.* **1953**, *21*, 1087.
[66] J. Apostolakis, P. Ferrara, A. Caflish, *J. Chem. Phys.* **1999**, *110*, 2099.
[67] G. D. Leines, B. Ensing, *Phys. Rev. Lett.* **2012**, *109*, 020601.
[68] L. Maragliano, E. Vanden-Eijnden, *Chem. Phys. Lett.* **2007**, *446*, 182.
[69] V. Ovchinnikov, M. Karplus, E. Vanden-Eijnden, *J. Chem. Phys.* **2011**, *134*, 085103.
[70] H. Taskent-Sezgin, J. Chung, V. Patsalo, S. J. Miyake-Stoner, A. M. Miller, S. H. Brewer, R. A. Mehl, D. F. Green, D. P. Raleigh, I. Carrico, *Biochemistry* **2009**, *48*, 9040.
[71] D. A. Case, D. S. Cerutti, T. E. Cheatham, III., T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K. M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Saguli, C. L. Simmerling, W. M. Botello-Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, L. Xiao, D. M. York, P. A. Kollman, AMBER 2017; University of California: San Francisco, **2017**.
[72] N. Hai, D. R. Roe, C. Simmerling, *J. Chem. Theory Comput.* **2013**, *9*, 2020.
[73] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, *J. Chem. Theory Comput.* **2015**, *11*, 3696.
[74] R. J. Loncharich, B. R. Brooks, R. W. Pastor, *Biopolymers* **1992**, *32*, 523.
[75] P. Ren, J. W. Ponder, *J. Phys. Chem. B* **2003**, *107*, 5933.

[76] C. Chen, Y. Huang, *J. Comput. Chem.* **2016**, *37*, 1565.
[77] C. Chen, *J. Comput. Chem.* **2017**, *38*, 2298.
[78] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, K. A. Dill, *J. Chem. Theory Comput.* **2007**, *3*, 26.
[79] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, W. C. Swopea, *J. Chem. Phys.* **2007**, *126*, 155101.
[80] G. R. Bowman, An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation, Springer Science+Business Media, Dordrecht, the Netherlands, **2014**, p. 7.
[81] L. T. Da, K. S. Fu, D. A. Silva, X. Huang, Protein Conformational Dynamics; Springer International Publishing: Switzerland, **2014**, p. 29.
[82] H. S. Park, C. H. Jun, *Expert Syst. Appl.* **2009**, *36*, 3336.
[83] P. W. Rose, B. Beran, C. X. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, J. D. Westbrook, J. Young, B. Yukich, C. Zardecki, H. M. Berman, P. E. Bourne, *Nucleic Acids Res.* **2011**, *39*, D392.
[84] W. Humphrey, A. Dalke, K. Schulten, *J. Mol. Graph.* **1996**, *14*, 33.

Received: 28 November 2018
Revised: 15 March 2019
Accepted: 17 March 2019
Published online on 3 April 2019