Chapter 16

Molecular Dynamics–Based Thermodynamic and Kinetic Characterization of Membrane Protein Conformational Transitions

Dylan Ogden and Mahmoud Moradi

Abstract

Molecular dynamics (MD) simulations are routinely used to study structural dynamics of membrane proteins. However, conventional MD is often unable to sample functionally important conformational transitions of membrane proteins such as those involved in active membrane transport or channel activation process. Here we describe a combination of multiple MD based techniques that allows for a rigorous characterization of energetics and kinetics of large-scale conformational changes in membrane proteins. The methodology is based on biased, nonequilibrium, collective-variable based simulations including nonequilibrium pulling, string method with swarms of trajectories, bias-exchange umbrella sampling, and rate estimation techniques.

Key words String Method, Umbrella Sampling, Nonequilibrium Pulling, Orientation Quaternion, Membrane Protein, Conformational Landscape, Transition Rate Estimation

1 Introduction

With advances in supercomputing technology, continued improvement of all-atom force fields, and increasing number of available structures of membrane proteins, molecular dynamics (MD) simulation technique [1–3] has emerged as a prominent computational method for determining the structural dynamics of membrane proteins in their membrane environment. MD is a technique that is routinely used to study the local fluctuations of membrane proteins around given functional states, often determined by X-ray crystallography, cryogenic electron microscopy (cryoEM), or homology modeling. The timescale gap between local fluctuations and large-scale conformational changes, however, has hindered the use of MD to study the functionally important conformational changes such as those involved in the state transition of transporters or activation of channels and receptors.

Ingeborg Schmidt-Krey and James C. Gumbart (eds.), *Structure and Function of Membrane Proteins*,
Methods in Molecular Biology, vol. 2302, https://doi.org/10.1007/978-1-0716-1394-8_16,

Conformational dynamics of proteins can be modeled as a diffusion in the protein conformational landscape [4–6]. The conformational free energy landscape has various basins and saddle points that represent the stable and transition states, respectively. A membrane protein, whether it is a channel, transporter, receptor, etc., is associated with many free energy minima, most of which are clustered around a few major free energy basins, representing the functional states of protein. For instance, a channel may have various closely related free energy minima around a large free energy basin that represents its active state, and one or a few free energy basins that represent its inactive state(s). Similarly, a membrane transporter has at least three free energy basins: one for the outward-facing (OF) state, one for the inward-facing (IF) state, and one for the occluded (Occ) state. Although the conformational landscape of a protein is generally very vast, most of this landscape is associated with large free energies and can be ignored. This is due to the presence of intra- and intermolecular forces that restrict the movement of atoms and molecular domains. These forces allow fluctuations around free energy basins and "rarely" allow the system to jump between these free energy basins by crossing the free energy barriers.

MD practitioners often use the available structures to study the local fluctuations of proteins around given free energy basins employing MD simulations of tens to hundreds of nanoseconds, and more recently up to several microseconds. Jumping between major free energy basins, however, rarely happens. In order to induce such jumps, one often needs to employ biased and/or nonequilibrium MD simulations. Many enhanced sampling techniques have been developed that can be employed to facilitate the sampling of rare events [7–18]. A handful of these methods are used routinely for the study of functionally relevant transitions; however, applying many of these methods to complex biological systems such as membrane proteins is challenging. Here we outline a number of methods to be used in finding and characterizing the membrane protein conformational transition pathways through the use of path-finding algorithms and enhanced sampling techniques. The specific techniques include: nonequilibrium pulling simulations (such as targeted MD, steered MD, and similar methods), path optimization algorithms (such as string method with swarms of trajectories [19, 20]), along-the-path free-energy calculations (such as bias-exchange umbrella sampling [20]), and transition rate estimation methods. The following section outlines the theory behind some of the techniques employed in this protocol.

2 Theory

2.1 Introduction

Dimensionality reduction is a necessary part of computational studies of protein dynamics due to the large number of degrees of freedom in the atomic coordinate space of proteins. Collective variable (colvar) based enhanced sampling techniques such as umbrella sampling (US) [7, 8] and its nonequilibrium counterparts [9–11, 13, 21, 22] effectively work within a reduced space. Collective variables can be defined intuitively to describe the slow degrees of freedom associated with functionally important protein conformational changes [23], for example, an interdomain molecular distance as in steered MD (SMD) [11] or the root-mean-square deviation (RMSD) from a reference structure as in targeted MD [9]. Several collective variable suites/modules [24–27] have recently been developed to allow for the system-specific design of collective variables such as path colvars [28, 29] and orientation colvars [26, 30].

An ideal collective variable could represent the "reaction coordinate" [31] as often described in transition state theory. However, even if a one-dimensional reaction coordinate exists, it is not known a priori. This has led to the development of several path-finding algorithms, which implicitly or explicitly approximate the reaction coordinate by the arc-length of a curve in the multidimensional space of atomic coordinates [32, 33] or colvars [4, 19]. Many of the colvar based enhanced sampling techniques implicitly or explicitly use a diffusion model to describe the effective dynamics in the colvar space [4, 5]. We have recently developed a Riemannian diffusion model for protein conformational dynamics that provides a robust framework for conformational free energy calculation methods and path-finding algorithms [34]. Unlike their Euclidean counterparts, the Riemannian potential of mean force (PMF) and minimum free energy path (MFEP) are *invariant under coordinate transformations* [34]. However, the protocol discussed here can be used with or without the Riemannian treatment of the colvar space.

Suppose that the dynamics of a high-dimensional atomic system ($\boldsymbol{x}$) can be simplified as effective dynamics in a reduced but generally multidimensional colvar space, $\boldsymbol{\zeta}$. The effective dynamics can be described by a Brownian motion in the $\boldsymbol{\zeta}$ space with an effective potential energy $G(\boldsymbol{\zeta})$, that is the PMF of the atomic system in the $\boldsymbol{\zeta}$ space:

$$G(\boldsymbol{\zeta}_0) = -k_B T \log\langle\delta(\boldsymbol{\zeta}(\boldsymbol{x}) - \boldsymbol{\zeta}_0)\rangle \quad (1)$$

is an ensemble average, k_B and T are the Boltzmann constant and temperature, respectively, and δ is the Dirac delta function.

One may sample the regions around a given point $\boldsymbol{\zeta}_0$ in the colvar space by adding a biasing term to the potential of the atomic system such as $U_0(\boldsymbol{\zeta}) = \frac{k}{2}\,(\boldsymbol{\zeta} - \boldsymbol{\zeta}_0)^2$, in which k is the force constant.

The free energy of the biased system (or the perturbed free energy) is:

$$F(\boldsymbol{\zeta}_0) = -k_B T \log \int d\boldsymbol{\zeta} \exp\left(-\beta(G(\boldsymbol{\zeta}) + U_0(\boldsymbol{\zeta}))\right) \tag{2}$$

where $\beta = (k_B T)^{-1}$. For large force constants, the PMF can be approximated using the perturbed free energy $F(\boldsymbol{\zeta})$. Otherwise, other methods can be used to estimate the PMF as briefly described below. If the biasing center is different in different simulations (or windows/images), we have $U_i(\boldsymbol{\zeta}) = \frac{k}{2}\,(\boldsymbol{\zeta} - \boldsymbol{\zeta}_i)^2$, where i is the window/image index.

Methods such as US rely on calculating the relative free energy of different points by biasing the system around those points in different simulations. Alternatively, the biasing center could change by time, for example, replacing $\boldsymbol{\zeta}_i$ by $\boldsymbol{\eta}(t)$. Examples of such simulations are SMD and targeted MD. Here, we refer to such methods as nonequilibrium pulling simulations, which may use any colvar(s) with any schedule of change (i.e., $\boldsymbol{\eta}(t)$ may or may not be linear in time). The biasing potential is described by: $U(\boldsymbol{\zeta}, t) = \frac{k}{2}\,(\boldsymbol{\zeta} - \boldsymbol{\eta}(t))^2$. The accumulated nonequilibrium work at any given time is: $W(t) = \int_0^t dt' \frac{\partial}{\partial t'} U(\boldsymbol{\zeta}, t')$. Nonequilibrium work can be used to estimate the perturbed free energy using the Jarzynski relation [35] or other nonequilibrium work relations [36, 37]. In this protocol, we use pulling simulations only to generate initial pathways for other simulation protocols. The nonequilibrium pulling simulations may use multidimensional colvars; however, a specific 1D pathway needs to be selected in order to perform the simulations.

Ideally, one may use a 1D collective variable for defining the effective dynamics as well as the biasing protocol. In practice, however, this may only be possible for extremely simple systems. A practical solution to this problem is to keep the collective variable space multidimensional, while sampling only around a particular pathway, represented by a 1D curve $\boldsymbol{\zeta}(\xi)$, parametrized by ξ. The choice of the pathway is obviously crucial here and determines the relevance of the free energy results to the transition of interest. Now $\xi(\boldsymbol{x})$ can be treated as a 1D colvar defined as a function of atomic coordinates $\boldsymbol{x}$, and $G(\xi_0)$ is the PMF associated with ξ_0,

$$\exp(-\beta G(\xi_0)) = \langle \delta(\xi(\boldsymbol{x}) - \xi_0) \rangle. \tag{3}$$

Assuming ξ dynamics can be effectively described by a diffusive model, we have,

$$\mathrm{d}\xi = \left(-\beta D(\xi) \frac{\mathrm{d}}{\mathrm{d}\xi} G(\xi) + \frac{\mathrm{d}}{\mathrm{d}\xi} D(\xi)\right) \mathrm{d}t + \sqrt{2D(\xi)}\,\mathrm{d}B, \tag{4}$$

in which $D(\xi)$ is a position-dependent diffusion constant, and $B(t)$ is a Wiener process such that $\langle B(t)\rangle = 0$ and $\langle B^2(t)\rangle = t$. Fokker–Planck (or Smoluchowski) equation associated with this process is

$$\frac{\partial}{\partial t}p(\xi, t|\xi_0, 0) = \frac{\partial}{\partial \xi}\Big(D\exp(-\beta G(\xi))\frac{\partial}{\partial \xi}\left(\exp(\beta G(\xi))p(\xi, t|\xi_0, 0)\right)\Big), \quad (5)$$

in which $p(\xi, t|\, \xi_0, 0)$ is the likelihood of finding the system at ξ after time t, given it was at ξ_0 at time 0. If two major free energy minima exist at points A and B, with no other basins outside the region spanning from A to B, the mean-first-passage time (MFPT) from A to B ($\overline{\tau}_{FP}$) can be estimated using the following relation [38]:

$$\overline{\tau}_{FP} = \int_{\xi_A}^{\xi_B} d\xi \frac{\int_{\xi_A}^{\xi} d\xi' \exp\left(-\beta G(\xi')\right)}{D(\xi)\exp\left(-\beta G(\xi)\right)}. \quad (6)$$

The aim of the protocols described here is to find the MFEP in a multidimensional colvar space, representing the most probable transition path between two free energy basins associated with two functional states of a protein. We start by generating an approximate path using nonequilibrium pulling simulations [23] (path generation), followed by path optimization in the multidimensional colvar space using string method [19, 20], followed by along-the-path free energy calculations using bias-exchange US [20], and finally followed by estimating the transition rate between the two states using the estimated free energies and diffusion constants.

2.2 String Method with Swarms of Trajectories (SMwST)

The SMwST algorithm [19, 20] starts from an initial string, defined by N points/images $\{\boldsymbol{x}_i\}$, where i is any integer from 1 to N. Colvar $\boldsymbol{\zeta}$ primarily defines the biasing potential, which is $U_i(\boldsymbol{\zeta}) = \frac{k}{2}\left(\boldsymbol{\zeta} - \boldsymbol{\zeta}_0\right)^2$ for M copies of image i. The initial values for the image centers are determined from the initial string: $\boldsymbol{\zeta}_i = \boldsymbol{\zeta}(\boldsymbol{x}_i)$. The SMwST algorithm consists of three iterative steps as follows. (Step I) Restraining: Each system is restrained for τ_R (restraining time) using the harmonic potential described above centered at the current image $\boldsymbol{\zeta}_i$. (Step II) Drifting: The simulations are continued after being released from restraints for τ_D (drifting time). (Step III) Reparameterization: The new center for each image i is determined by averaging over all observed $\boldsymbol{\zeta}(\boldsymbol{x})$ values of M systems associated with image i at time $\tau_R + \tau_D$ and using a linear interpolation algorithm to keep the image centers equidistant. By iterating over these steps, the string will converge to the zero-drift path, around which the string centers oscillate (upon convergence). The zero-drift path is an approximation of the MFEP [6, 34].

2.3 Bias Exchange Umbrella Sampling (BEUS)

Once the MFEP (parametrized by ξ) is known, $F(\xi)$ can be estimated using a generalization of US [39], termed BEUS [20]. Similar to the SMwST method, ξ is discretized and N umbrella windows/images are defined with biasing potentials $U_i(\boldsymbol{\zeta}) = \frac{k}{2}\left(\boldsymbol{\zeta} - \boldsymbol{\zeta}_0\right)^2$ for $i = 1, \ldots, N$. This scheme can be thought of as a 1D US along the reaction coordinate ξ with an additional restraint on the (shortest) distance from the $\boldsymbol{\zeta}(\xi)$ curve. Perturbed free energies $F_i = F(\boldsymbol{\zeta}_i)$ can be estimated (up to an additive constant) by self-consistently solving the equations [40–42]:

$$e^{-\beta F_i} = \sum\nolimits_t \frac{e^{-\beta U_i(\boldsymbol{\zeta}^t)}}{\sum_j T_j e^{-\beta\left(U_j(\boldsymbol{\zeta}^t) - F_j\right)}} \tag{7}$$

in which Σ_t sums over all collected samples (irrespective of which replica or image they belong to) and T_j is the number of samples collected for image j. With appropriate reweighting, PMF can be reconstructed in any arbitrary collective variable space, given sufficient sampling in that space. w^t, the unnormalized weight of configuration $\boldsymbol{x}^t$ can be estimated via [41]:

$$w^t = \left(\sum_i T_i e^{-\beta(U_i(\boldsymbol{\zeta}^t) - F_i)}\right)^{-1} \tag{8}$$

in which $\{F_i\}$ are estimated via Eq. (7). Alternatively [41], one may estimate $\{w^t\}$ and

$\{F_i\}$ by iteratively solving Eq. (8) and:

$$e^{-\beta F_i} = \sum_t w^t e^{-\beta U_i(\boldsymbol{\zeta}^t)} \tag{9}$$

The PMF in terms of $\boldsymbol{\eta}(\boldsymbol{x})$, an arbitrary collective variable, is estimated (up to an additive constant) as:

$$G(\boldsymbol{\eta}_0) = -\beta^{-1} \log\left(\sum_t w^t K(\boldsymbol{\eta}(\boldsymbol{x}^t) - \boldsymbol{\eta}_0)\right) \tag{10}$$

in which K is a kernel function. The above estimator is not accurate if the sampling in $\boldsymbol{\eta}(\boldsymbol{x})$ is not converged which is the case if $\boldsymbol{\eta}(\boldsymbol{x})$ is associated with slow dynamics and is not strongly correlated with $\boldsymbol{\zeta}$. For the special case of $\boldsymbol{\eta} = \boldsymbol{\zeta}$, the perturbed free energies $\{F_i\}$ can be used directly to estimate the PMF only within the stiff-spring approximation.

Finally, for averaging an arbitrary quantity $A(\boldsymbol{x})$ along the pathway $\boldsymbol{\zeta}(\xi)$, one may use the weighted average $\overline{A}(t) = \sum_t w^t A(\boldsymbol{x}^t)\delta(\boldsymbol{\zeta}^t - \boldsymbol{\zeta}(\xi))$. However the unweighted estimator $\overline{A}_i = \langle A(\boldsymbol{x}^t)\rangle_i$ is more efficient. $\overline{\sigma}_i{}^2 = \frac{\left\langle A^2(\boldsymbol{x}^t) - \overline{A}^2\right\rangle_i}{g}$ provides an estimate for the variance, given $g = 1 + 2\frac{\tau_{ac}^A}{\tau_{lag}}$ is the statistical inefficiency in which τ_{ac}^A is the autocorrelation time associated with quantity A and τ_{lag} is the lag time between the data points used in the analysis [23].

2.4 Transition Rate Estimation

Discretizing Relation (5) results in [43]:

$$\boldsymbol{P}(\delta t) = (\mathbf{1} + \boldsymbol{R}\delta t)\boldsymbol{P}(0), \tag{11}$$

where $\boldsymbol{P}(t)$ is a vector with elements $P_i = p(\xi, t| \xi_0, 0)$, δt is a small-time step, and $\boldsymbol{R}$ is a tridiagonal matrix with elements $R_{i\ i} = -R_{i\ i+1} - R_{i\ i-1}$, and:

$$R_{i\ i\pm 1} = \delta\xi^2 D(\xi_{i\pm\frac{1}{2}})\exp(-\beta(G(\xi_i) - G(\xi_{i\pm\frac{1}{2}}))), \tag{12}$$

where $\delta\xi = \xi_{i+1} - \xi_i$ for any i. More generally, for any lag time Δt and any time t, we have

$$\boldsymbol{P}(t + \Delta t) = \exp(\boldsymbol{R}\Delta t)\boldsymbol{P}(t), \tag{13}$$

which implies that the likelihood of finding a system at bin j at time $t + \Delta t$, given it was at bin i at time t, is proportional to $\exp(\boldsymbol{R}\Delta t)_{i\ j}$. Therefore, assuming neither $G(\xi)$ nor $D(\xi)$ is known, one may find both, as in Ref. [43], by maximizing the likelihood $L = \Pi_\alpha(\exp(\boldsymbol{R}\Delta t))_{i_\alpha\ j_\alpha}$ (Π_α runs over all observations of trajectories starting at the bin i_α at a given time t and being found at the bin j_α at time $t + \delta t$). Assuming $G(\xi)$ is known, one may find $D(\xi)$ using a similar maximum likelihood approach [44]. For any given $D(\xi)$, $\boldsymbol{R}$ can be evaluated, resulting in the log-likelihood,

$$l = \sum_\alpha \log((\exp(\boldsymbol{R}\Delta t))_{i_\alpha\ j_\alpha}), \tag{14}$$

which can be maximized using a Metropolis Monte Carlo algorithm. We first estimate the factors $\exp(-\beta(G(\xi_i) - G(\xi_{i\pm 1/2})))$ in $R_{i\ i\pm 1}$, where $G(\xi_i)$ is determined for $i = 1, 2, \ldots, N$ from the BEUS simulations and $G(\xi_{i+1/2})$ is estimated by interpolation. An arbitrary series $D_{i+1/2}, 1, \ldots, N - 1$ can be used as an initial guess for $D(\xi_{i+1/2})$. $R_{i\ i\pm 1}$ and $R_{i\ i}$ values are then calculated to estimate the log-likelihood l. For a faster convergence, one may start with the estimates of $\boldsymbol{R}$ associated with the $\Delta t \to 0$ limit of Relation (12) (i.e., Relation (11)) to maximize the log-likelihood in (14). It is easy to show that the following values for $R_{i\ i\pm 1}$ maximize the log-likelihood in (14) at the $\Delta t \to 0$ limit:

$$R_{i\ i\pm 1} = \frac{1}{\Delta t}\frac{N_{i\ i\pm 1} + N_{i\pm 1\ i}}{N_{i\ i}\exp(-\beta\Big(G(\xi_{i\pm i}) - G(\xi_i)\Big) + N_{i\pm 1\ i\pm 1}}, \tag{15}$$

in which $N_{i\ j}$ is the number of observed jumps from bin i to j with lag time Δt. Diagonal values of $\boldsymbol{R}$ can be also estimated using $R_{i\ i} = -R_{i\ i+1} - R_{i\ i-1}$, while the other elements are zero. For an arbitrary lag time Δt, the log-likelihood in Relation (14) can be evaluated using the values of $\boldsymbol{N}$ matrix as

$$l = \sum_\alpha N_{i\ j}\log((\exp(\boldsymbol{R}\Delta t))_{i\ j}). \tag{16}$$

Starting from $\Delta t \rightarrow 0$ limit of $\boldsymbol{R}$, one can use a Metropolis Monte Carlo algorithm to maximize the log-likelihood l in Relation (16). $D_{i+1/2}$ can be estimated using

$$D_{i+1/2} = \delta\xi^2 R_{i\ i\pm 1} \exp(\beta(G(\xi_i) - G(\xi_{i\pm 1/2}))). \tag{17}$$

$D(\xi_i)$ can be estimated by interpolation $(D_{i+1/2} + D_{i-1/2})/2$. Finally, the MPFT ($\overline{\tau}_{FP}$) can be estimated numerically using

$$\overline{\tau}_{FP} = \sum_{i=1}^{N} \frac{\sum_{j=1}^{i} \exp\left(-\beta G(\xi_j)\right)}{D(\xi_i) \exp\left(-\beta G(\xi_i)\right)}. \tag{18}$$

3 Methods

3.1 Initial Preparation

1. Begin by preparing a membrane-embedded, water-solvated model of protein using one of its available structures. The suggested protocol here may use information from multiple structures in the next steps, but only one initial model needs to be prepared for all MD simulations (*see* **Notes 1** and **2**).
2. Before employing any biased or nonequilibrium simulations, it is important to run an equilibrium, unbiased simulation of the protein as in a conventional MD simulation. The next steps of the protocol will suffer particularly in terms of convergence if they are initiated from unequilibrated structures.
3. The length of the initial equilibration simulation can vary on how quickly a stable conformation can be reached. This is typically examined by monitoring the RMSD of the protein (*see* **Note 3**).
4. The last snapshot of the equilibrium MD simulation can be used as the initial conformation for the pulling simulations. If longer simulations have been performed, multiple snapshots may be used to examine the reproducibility as long as the selected structures resides in the equilibrated region.

3.2 Path Generation: Nonequilibrium Pulling Simulations

1. The choice of colvars should be specific to the protein of interest. A set of colvars used successfully to induce the transition of interest in one protein may not be applicable to another protein. Some common examples include the RMSD with respect to a target structure (as in targeted MD) and the distance between the mass centers of two specific molecules or molecular domains (as in SMD). The orientation based colvars have particularly proven to be very effective in describing the orientations of transmembrane helices or helical bundles of transmembrane proteins and are highly recommended as an alternative to RMSD and distance (*see* **Note 4**).

2. When determining what colvar to use, it is important to be familiar with the proposed conformational transition of the protein, in particular the target state. It is always useful to have a model of the target state even if the model is not complete or not accurate. The target model will not be used to run an actual MD simulation, so it can typically only contain the C_α atoms of the protein or the domains that will be steered (e.g., transmembrane helices).

3. Once a target model is available, one may run a targeted MD simulation with the target model to generate a transition path. The targeted MD based transition pathways are not often reliable since they typically generate pathways that are not close to the MFEP. However, they could provide a reference to compare other protocols that are based on other colvars. Multistep targeted MD simulations also provide an alternative method for generating initial pathways, if intermediate target models are available (*see* **Note 5**).

4. The number of colvars used is completely dependent on how simple or complex the transition pathway may be. Multiple colvars may be used in a single nonequilibrium pulling protocol; however, a specific schedule needs to be provided for changing the center of bias for each colvar. In other words, the system is steered along a specific path in the colvar space, if multiple colvars are used.

5. The choice of force constant is also dependent on both the colvar type—the unit of force constant is that of energy divided by the square of the colvar unit—and the particular transition of interest—the barriers that need to be crossed as the system is driven along the predetermined pathway is particularly important. The force constant should be large enough to induce the conformational transition of interest. If the force constant is too large, however, the simulation could become unstable, and the molecular system could undergo deformation or distortion. One may need to start with an educated guess and perform a short simulation to determine whether or not the colvars change according to their schedule. If not, the force constant can be increased. If there is very little deviation from the schedule, the force constant can be lowered to allow some deviation without introducing a delay in the schedule that increases significantly over time (*see* **Note 6**).

6. The simulation time is also dependent on the choice of the colvars, the desired transition, and even the force constant chosen. It is reasonable to start with relatively short simulations (a few nanoseconds) to roughly determine the quality of the protocol and fine tune the parameters; however, the final simulation that will seed the next step of our protocol (i.e., SMwST)

should be long enough (at least 100 ns) to allow for relaxation of orthogonal degrees of freedom not involved in the colvars used at least to some extent.

7. Since pulling simulations are relatively inexpensive, it is advantageous to repeat and try many different protocols to identify one or more that may lead to a reasonable transition pathway for the given protein of interest (*see* Fig. 1).
8. A reasonable protocol must satisfy the following four criteria:
   (a) The desired transition must occur (*see* **Note 7**). To monitor this, one may use various measures depending on the particular transition of interest. For instance, for the activation of a channel, one may monitor the number of water molecules within the transmembrane region of protein (*see* Fig. 1) or measure the pore radius using programs such as HOLE [45].

   (b) The protocol must not introduce undesired structural distortions such as major secondary structural changes (unless it is part of the transition mechanism).

   (c) The protocol should not require large amounts of work (e.g., over 500 kcal/mol). Large work indicates that the generated transition pathway is too far from equilibrium and it may not easily relax to a converged pathway in the next step (i.e., SMwST).

   (d) Finally, it is important to follow up the pulling simulations with an equilibrium simulation to release the system of all restraints and to ensure that the system will maintain a relatively stable conformation following the transition. This is to be done by slowly releasing all restraints and then allowing the system to converge toward a new equilibrium state. A protocol may satisfy all three criteria above but the final conformation may not maintain the desired functional state (e.g., the channel may not stay open). In this case, the protocol is not considered a successful protocol (*see* **Note 8**).

3.3 Path Optimization: String Method with Swarms of Trajectories

1. After performing the pulling simulations and the post equilibrium simulations, an initial set of conformations along the generated transition pathway can be extracted from the nonequilibrium trajectories to initiate the SMwST algorithm (*see* **Note 9**).
2. The number of conformations extracted as snapshots from the trajectory files will be the number of images used in the algorithm. The number of images to be used in the algorithm may depend on the complexity of the pathway that is to be refined (*see* **Note 10**).

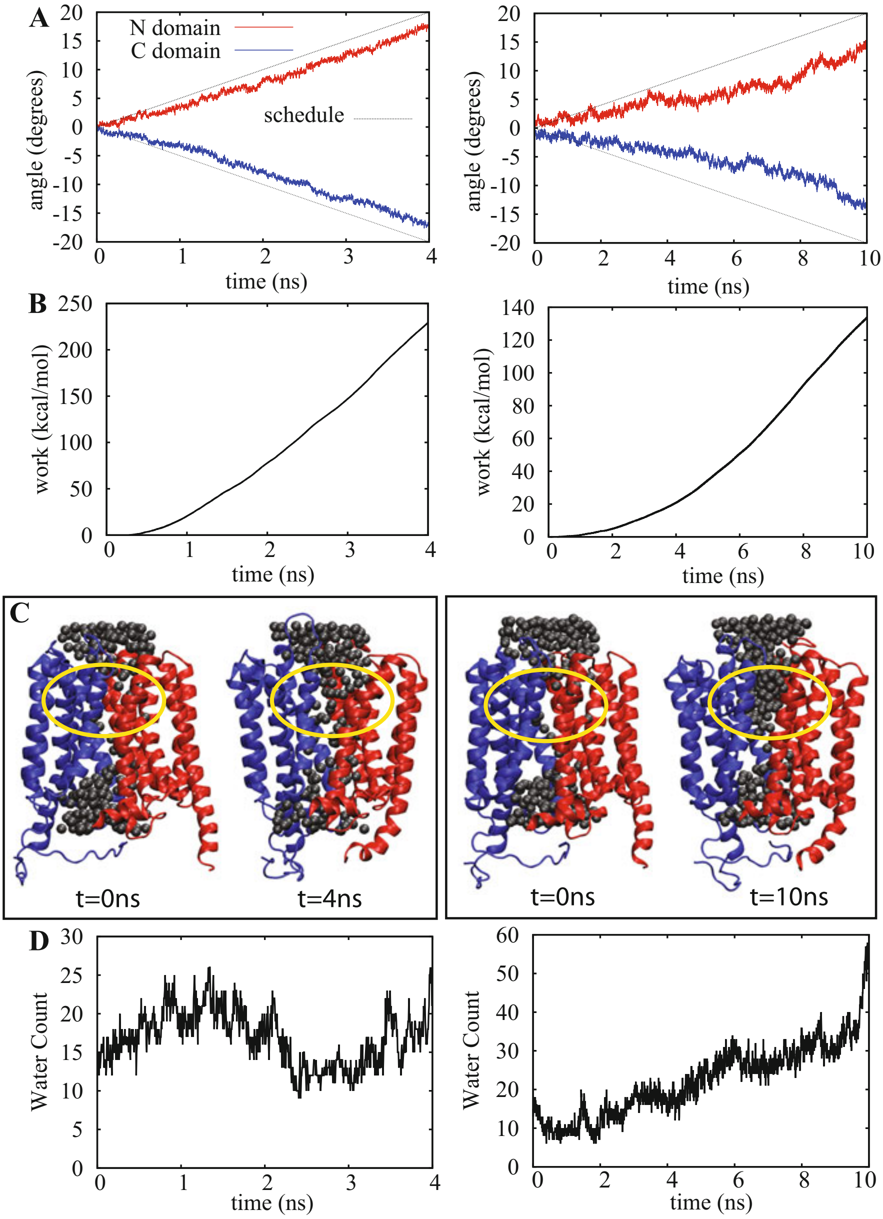

Fig. 1 Comparing two nonequilibrium pulling protocols to induce an IF→OF transition in membrane transporter GlpT [20]. Both protocols induce rotational changes on the N- and C-bundle domains of GlpT using spin angles (left) or orientations (right) of the two domains. Both protocols use a force constant of 3 kcal/mol deg^2. (**a**) Time series of spin (left) and orientation (right) angles. The black line shows the schedule of the time-dependent center of harmonic potentials. (**b**) Time series of nonequilibrium work. (**c**) Snapshots of protein at the beginning and end of simulations. The water molecules within the transmembrane region are shown. (**d**) The number of water molecules in the periplasmic side of the transmembrane region as a function of time

3. SMwST may be performed as a series of simulations, as originally implemented or it may be run in parallel as a single job as currently implemented in NAMD (using a TCL script) and Amber (using the NFE suite). In the parallel version, each image is represented by a number of independent copies that are first restrained to stay around the current image center (the restraining step) and then released (drifting stage). The total number of replicas is determined by the number of copies of each image multiplied by the number of images (*see* **Note 11**).
4. The collective variables to be used in the string method simulations may or may not be similar to those used in the pulling simulations. Obviously, there needs to be more than one colvar to have a meaningful path optimization. There is no particular limitation on the number of colvars used as long as the colvar space represents a smooth space (*see* **Note 12**).
5. The force constants to be used for the restraining step may be at least as large as those used in previous pulling simulations if the same colvars are used. However, it is recommended to use larger force constants at this stage to ensure that the restrained conformations reach and stay around the desired image center during short restraining simulations. Distortion is unlikely in these short simulations and thus larger force constants can be employed (*see* **Note 13**).
6. The number of iterations to be employed in the string method simulations will be dependent on the initial pathway generated. The closer to the MFEP the initial pathway is, the faster the SMwST simulations converge. It is thus important to ensure a reasonable pulling protocol is used to generate the initial pathway before employing computationally costly SMwST simulations.
7. The convergence of the SMwST can be monitored using measures such as string RMSD from a reference structure such as initial string or final string. The string RMSD between two strings (of N images) is defined as the root mean square distance of the individual colvars. For instance, if n colvars are used, the string RMSD between strings i and j would be $\sqrt{\frac{1}{nN}\sum_{k=1}^{N}\sum_{l=1}^{n} d\left(\xi_{k,l,i},\xi_{k,l,j}\right)^2}$, where $d(a, b)$ is the distance between colvars a and b, and $\xi_{k,\ l,\ i}$ is the l'th colvar of the k'th image of string i (*see* Fig. 2).

3.4 Free Energy Calculations: Bias-Exchange Umbrella Sampling

1. The converged SMwST string provides an approximation for the MFEP. To estimate the free energy along this pathway, the US simulations can be carried out using the converged SMwST image centers as the center of US windows with the same colvars used for the SMwST simulations and the last snapshots of SMwST simulations as the initial conformations of the US simulations.

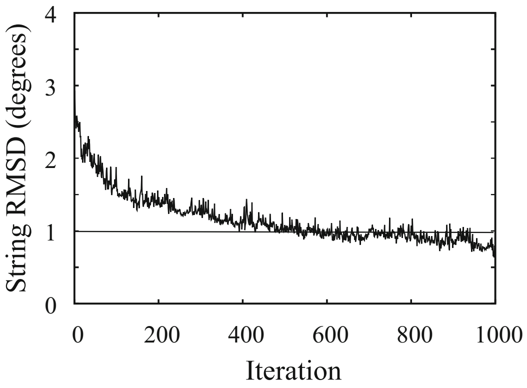

Fig. 2 Monitoring the convergence of the SMwST algorithm. String RMSD at each iteration with respect to the last string from SMwST simulations of IF→OF GlpT transition using 12 transmembrane orientations. Comparing the horizontal line with measured RMSD string shows that after about 400 iterations, the SMwST does not significantly change

2. The BEUS scheme is recommended over the conventional US scheme since it allows for the diffusion of the individual replicas along the pathway through the exchange procedure of the BEUS scheme. The faster and more reliable convergence is expected as a result.
3. Rather than using one copy per image/window, one may use multiple copies of each image in the BEUS scheme. This is particularly convenient since the SMwST algorithm already provides us with multiple copies of each image if the fully parallel version is used (*see* **Note 14**).
4. BEUS uses a similar restraining bias as that used by the SMwST simulations, but unlike the restraining used in SMwST, the force constant cannot be too large. The restraining is used to keep the protein from drifting away from the current image centers as in the SMwST method; however, an additional feature of BEUS is that the biasing potentials are also used to determine the exchange criteria. Therefore, the force constant cannot be too high, otherwise, there will not be adequate exchange between the neighboring windows. The force constant can be selected such that all replicas have 10 to 50% of successful exchange at every attempt (*see* **Note 15**).
5. By employing a nonparametric reweighting scheme as discussed above in Subheading 2, the perturbed free energies can be estimated to reconstruct the free energy profile (*see* Fig. 3) and various other ensemble averages along the MFEP. The perturbed free energies are only a good estimate of the PMF along the path, if the force constant is large enough (*see* **Note 16**).

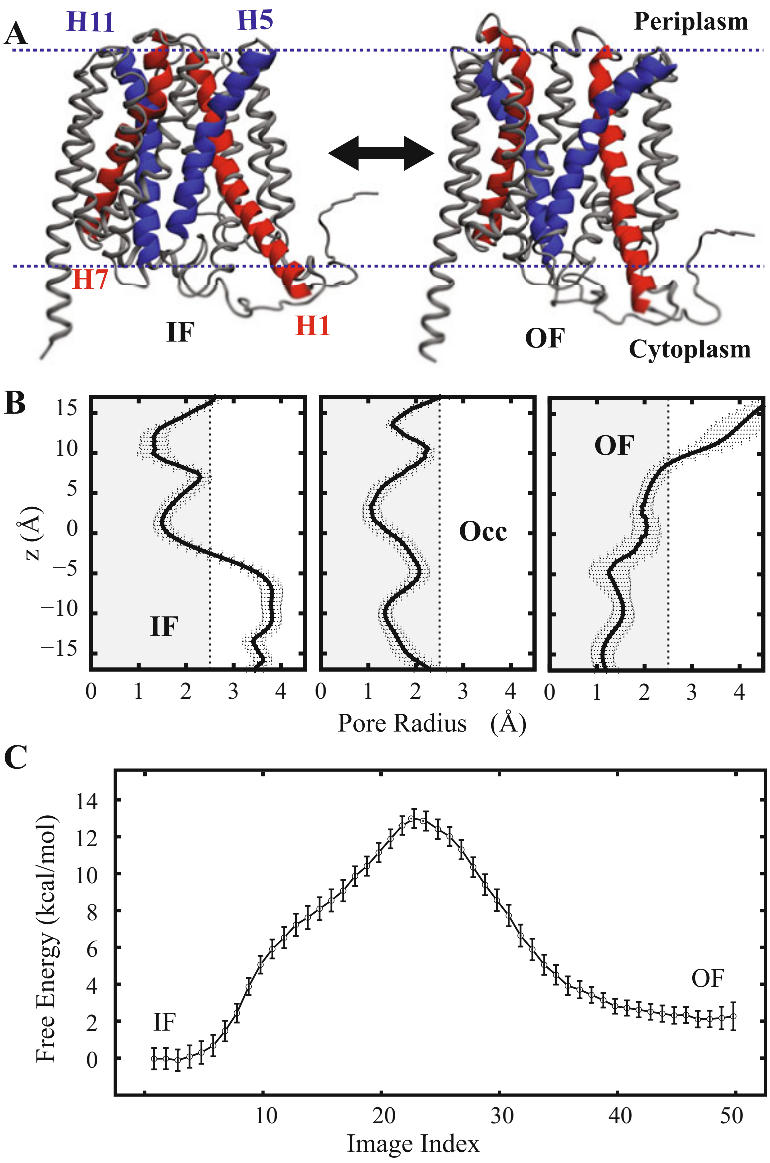

Fig. 3 BEUS simulation results of IF→OF GlpT transition. (**a**) Snapshots of the first ($i = 1$) and last ($i = 50$) images, representing the IF and OF state, respectively. (**b**) The pore radius along the pore based on the snapshots of the first (IF) and last (OF) as well as an intermediate (Occ) image. The latter represents an occluded state. (**c**) The perturbed free energies estimated from the BEUS simulations [20]

6. An alternative method to estimate the PMF is to use the weighting factors (w^{t}'s) to construct a PMF along a given 1D collective variable such as the first principal component obtained from the principal component analysis of C_α atoms of proteins. The latter method has an advantage over the former method in that it does not require the large force constant condition.

3.5 Transition Rate Estimation

1. Assuming that we have identified the MFEP relatively accurately in a relevant colvar space such that the effect dynamics of the system can be assumed to be diffusive along the identified MFEP, we can use a 1D diffusion model to describe the effective kinetics of the system as described in Subheading 2 above.
2. To determine the transition rate, the free energy profile (or the PMF) along the MFEP is needed, which is already obtained from BEUS simulations. In addition, the position-dependent diffusion constant along the MFEP is also needed to accurately estimate the transition rate using Relation (6). The diffusion constant estimation can be carried out by estimating inter-image transition rates measured using unbiased simulations along the MFEP.
3. Multiple copies of conformations per image/window will be extracted from BEUS trajectories to initiate these unbiased simulations. If multiple copies of images were used in BEUS simulations as recommended, one may simply use the last snapshots of all BEUS trajectories as the starting point for these unbiased simulations.
4. Although these are unbiased simulations, it is recommended to collect the colvar values during the simulations to monitor jumps between the images. These are the same colvars used in the BEUS simulations and the colvar values can be used to first assign an image to each sampled conformation at any given time and then count the number of transitions between different images. One may then build an empirical transition matrix based on these counts. The empirical transition matrix will be dependent on the lag time used to collect the data (or to count the jumps). It is recommended to use multiple lag times to determine the behavior of the estimated transition rates as a function of the lag time (*see* **Note 17**).
5. Once an empirical transition rate is constructed for a given lag time, a Metropolis Monte Carlo algorithm will be used as described in Subheading 2, to estimate diffusion constants $D(\xi_i)$ from the empirical transition matrix and the BEUS estimate free energies $G(\xi_i)$ and eventually estimate the MFPT and the overall transition rate (*see* Fig. 4 and **Note 18**).

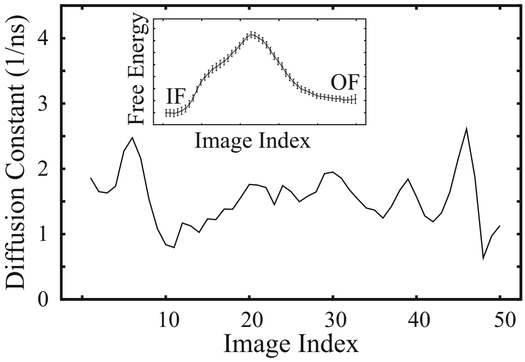

Fig. 4 The diffusion constant as a function of image index estimated based on GlpT BEUS simulations (for free energies) and follow-up equilibrium simulations using a lag time of 0.5 ns, which was determined to be the optimum lag time. The MFPT estimated based on these calculations for the IF→OF GlpT transition is approximately 6 s

4 Notes

Initial Preparation:

1. In the case of membrane transporters, a typical crystal structure, cryo-EM structure, or homology model will be in an inward-facing (IF), outward-facing (OF), or occluded (Occ) state. The choice of the starting point is often based on the quality and reliability of the structure. For example, a homology model is less reliable than a crystal structure; a mutant or engineered crystal structure is less reliable than a wild-type one.
2. The choice of lipid composition, salt type and concentration, protonation states of titratable amino acids, force field parameters, temperature, box size, and other MD simulation parameters is determined at this stage. Care must to be taken in making these choices as any changes of these parameters in the next steps may complicate the interpretability of the results.
3. Although it is common to monitor the RMSD of the protein with respect to initial frame (or preferably initial model that usually represents the known (e.g., crystal) structure), it is recommended to also monitor the RMSD with respect to the last frame. If the RMSD with respect to the last frame stays small for a long enough period, it is a much stronger evidence for the stability of the final conformation than if the RMSD with respect to the initial frame stays constant.

Nonequilibrium Pulling:

4. The orientation based colvars as implemented in NAMD, LAMPS, and Amber are based on the orientation quaternion formalism that measure the rotation of a semirigid-body

molecule or molecular domain with respect to the same molecule or molecular domain in a reference conformation. The colvar may fully describe all three rotational degrees of freedom (that is in the form of a unit orientation quaternion) or only describe the angle of rotation (orientation angle) or the angle of rotation with respect to a specific axis (tilt angle or spin angle). A 1D orientation based colvar such as an orientation, spin, or tilt angle is typically easier to use particularly at this stage of protocol. In SMwST/BEUS simulations, the orientation colvar that contains all three degrees of freedom is more appropriate to use.

5. Another method to employ the pulling simulations is the definition of individual initial, final, and hypothetical intermediate state along the transition pathway. This will be carried out in many individual targeted simulations until the final state has been reached. Targeting of the intermediates allows to target local minima which would otherwise not be sampled by employing a single target. The intermediate may be based on available crystal structures (e.g., the Occ state of a transporter) or they may be generated using other modeling techniques (e.g., coarse-graining or isotropic network models).

6. Although a short simulation could be quite informative with regards to the choice of the force constant, one needs to also examine whether the force constant is large enough for the other stages of the transition. Due to the presence of many metastable states and barriers, it is quite possible for the system to get trapped in one of the metastable states along the pathway and may never reach the final desired state. Increasing the force constant may help overcoming such issues; however, oftentimes, changing the choice of the colvars or their schedule is more effective.

7. The first step to monitor whether the desired transition has occurred is to monitor the colvars and how they follow their imposed schedule. However, it is important to note that the final targeted value may not be always reached, which admittedly, such is the case for TMD simulations. As long as the final conformation has the desired functional features (e.g., it is an open channel), the desired transition is considered to have been induced.

8. The postpulling equilibrium MD simulations must be long enough to allow for the relaxation of the system. If the system relaxes to a conformation that is at the desired functional state but significantly differs from the initial target used for pulling simulations, one may use the equilibrated conformation as a new target to generate an alternative transition path.

Path Optimization:

9. The extracted snapshots maybe extracted from the nonequilibrium pulling simulations in an equitemporal manner. In addition, one may include a few snapshots from the post-pulling equilibrium simulation trajectory as well. This often helps speeding up the convergence of the SMwST simulations.
10. If the number of images is small, the reparameterization step may introduce large changes to the image centers that cannot be easily achieved during the restraining step. Using too many images, on the other hand, introduces unnecessary curvatures. The ideal number of images is highly dependent on the nature and number of the colvars. A typical number of images would be between 50 and 200.
11. Depending on the computational resources available, one may use a hybrid version of serial and parallel SMwST. This is again implemented in both NAMD (smwst script) and Amber (NFE suite). In this version, fewer number of copies (or only one copy) per image is used, but each copy generates more than one sample before averaging the drift and updating the image centers. One may choose to use one copy per image and 20 samples per copy, or 20 copies per image and only one sample per copy, or anything in between, for example, 5 copies per image and 4 samples per copy. The recommended number of copies × number of samples is at least 20.
12. One may even use the atomic coordinate space (of select atoms, e.g., C_α's). The orientation based colvars, however, provide a smoother space and are expected to provide a faster and more reliable convergence.
13. It is important to monitor the progress of the SMwST simulations, particularly in the first few iterations, to ensure the appropriateness of the parameters chosen. For instance, one should check whether all copies of each image end up around the desired image center during the restraining stage before they are released. For instance, one may plot both the image centers and the actual colvar values of the sampled conformations in different 2D colvar spaces to ensure the sampled colvar values are closely distributed around each image center at the end of each restraining stage.

Free Energy Calculations:

14. Using multiple copies of an image/window in a BEUS simulation, given the availability of supercomputers that allow executing large jobs, not only allows for faster sampling but more importantly also allows for a better uncertainty estimation. The multiple copies of images are effectively independent simulations and provide uncorrelated data points for unbiased uncertainty estimation.

15. Having adequate sampling overlap between the neighboring windows is important in US simulations. Similarly, having adequate exchange rate is important in BEUS simulations. Care must be taken in choosing the number of images and the force constants to ensure each image is close enough to its neighboring image, allowing some exchange between the neighboring replicas. This can be remedied by either lowering the force constant between neighboring images that may be experiencing lower exchange rates or by adding more images allowing for more evenly spaced neighboring images and promoting a much greater exchange rate between neighboring images. The above parameters can be optimized iteratively using short runs with the goal of achieving similar rates of exchange between neighboring replicas.
16. Within the stiff-spring approximation, that is, the large force constant assumption, the perturbed free energy and the PMF are equal. However, the first-order correction of the approximation can be estimated by *some posteriori* from the estimated perturbed free energy $F(\xi)$ as: $\frac{1}{2\beta k}\left(\beta\left(\frac{d}{d\xi}F(\xi)\right)^{2} - \frac{d^{2}}{d\xi^{2}}F(\xi)\right)$.

Transition Rate Estimation:

17. If the lag time is too short, the data will be too correlated and the diffusion constants and transition rates will be overestimated. If the lag time is too long, given the limited simulation time, few transitions will be observed and the estimated diffusion constants and transition rates will be associated with large errors. Using multiple lag times allows for identifying the optimum lag time.
18. The simulation time is again system dependent; however, since the estimated free energies already provide relative interimage transition rates $\frac{R_{i\ i+1}}{R_{i+1\ i}} = \exp(-\beta(G(\xi_i) - G(\xi_{i+1})))$, the only information needed to fully construct the transition rate matrix is the downhill interimage transition rates. Without the BEUS free energy estimates, both uphill and downhill interimage transition rates need to be estimated, which requires considerably more time.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant numbers 1940188 and 1945465. This research is also supported by the Arkansas Biosciences Institute. This work used the Extreme Science and Engineering Discovery Environment (allocation MCB150129), which is supported by

National Science Foundation grant number ACI-1548562. This research is also supported by the Arkansas High Performance Computing Center which is funded through multiple National Science Foundation grants and the Arkansas Economic Development Commission.

References

1. Hansson T, Oostenbrink C, van Gunsteren WF (2002) Molecular dynamics simulations. Curr Opin Struct Biol 12:190–196
2. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. Nat Struct Biol 265:654–652
3. Karplus M, Kuriyan J (2005) Molecular dynamics and protein function. Proc Natl Acad Sci U S A 102:6679–6685
4. Maragliano L, Fischer A, Vanden-Eijnden E et al (2006) String method in collective variables: minimum free energy paths and isocommittor surfaces. J Chem Phys 125:24106
5. E W, Vanden-Eijnden E (2010) Transition-path theory and path-finding algorithms for the study of rare events. Annu Rev Phys Chem 61:391
6. Johnson ME, Hummer G (2012) Characterization of a dynamic string method for the construction of transition pathways in molecular reactions. J Phys Chem B 116:8573–8583
7. Torrie GM, Valleau JP (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. J Comp Phys 23:187–199
8. Northrup SH, Pear MR, Lee CY et al (1982) Dynamical theory of activated processes in globular proteins. Proc Natl Acad Sci U S A 79:4035–4039
9. Schlitter J, Engels M, Krüger P et al (1993) Targeted molecular dynamics simulation of conformational change—application to the T-R transition in insulin. Mol Simulation 10:291–308
10. Huber T, Torda AE, van Gunsteren WF (1994) Local elevation: a method for improving the searching properties of molecular dynamics simulation. J Comput Aided Mol Des 8:695
11. Izrailev S, Stepaniants S, Balsera M et al (1997) Molecular dynamics study of unbinding of the avidin-biotin complex. Biophys J 72:1568–1581
12. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314:141–151
13. Laio A, Parrinello M (2002) Escaping free energy minima. Proc Natl Acad Sci U S A 99:12562–12566
14. Darve E, Rodríguez-Gómez D, Pohorille A (2008) Adaptive biasing force method for scalar and vector free energy calculations. J Chem Phys 128:144120
15. Abrams CF, Vanden-Eijnden E (2010) Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics. Proc Natl Acad Sci U S A 107:4961–4966
16. Templeton C, Chen SH, Fathizadeh A et al (2017) Rock climbing: a local-global algorithm to compute minimum energy and minimum free energy pathways. J Chem Phys 147:152718
17. Chong LT, Saglam AS, Zuckerman DM (2017) Path-sampling strategies for simulating rare events in biomolecular systems. Curr Opin Struct Biol 43:88–94
18. Laio A, Panagiotopoulos AZ, Zuckerman DM (2018) Preface: special topic on enhanced sampling for molecular systems. J Chem Phys 149:072001
19. Pan AC, Sezer D, Roux B (2008) Finding transition pathways using the string method with swarms of trajectories. J Phys Chem B 112:3432–3440
20. Moradi M, Enkavi G, Tajkhorshid E (2015) Atomic-level characterization of transport cycle thermodynamics in the glycerol-3-phosphate:phosphate antiporter. Nat Commun 6:8393
21. Hummer G, Kevrekidis IG (2003) Coarse molecular dynamics of a peptide fragment: free energy, kinetics, and long-time dynamics computations. J Chem Phys 118:10762
22. Moradi M, Tajkhorshid E (2013) Driven metadynamics: reconstructing equilibrium free energies from driven adaptive-bias simulations. J Phys Chem Lett 4:1882–1887
23. Moradi M, Tajkhorshid E (2014) Computational recipe for efficient description of large-scale conformational changes in biomolecular systems. J Chem Theory Comp 10:2866–2880

24. Bonomi M, Branduardi D, Bussi G et al (2009) PLUMED: a portable plugin for free energy calculations with molecular dynamics. Comput Phys Commun 180:1961

25. Babin V, Karpusenka V, Moradi M et al (2009) Adaptively biased molecular dynamics: an umbrella sampling method with a time-dependent potential. Int J Quantum Chem 109:3666–3678

26. Fiorin G, Klein ML, Hénin J (2013) Using collective variables to drive molecular dynamics simulations. Mol Phys 111:3345

27. Sidky H, Colón YJ, Helfferich J et al (2018) Ssages: software suite for advanced general ensemble simulations. J Chem Phys 148:044104

28. Branduardi D, Gervasio FL, Parrinello M (2007) From a to b in free energy space. J Chem Phys 126:054103

29. Berteotti A, Cavalli A, Branduardi D et al (2009) Protein conformational transitions: the closure mechanism of a kinase explored by atomistic simulations. J Am Chem Soc 131:244–250

30. Moradi M, Tajkhorshid E (2013) Mechanistic picture for conformational transition of a membrane transporter at atomic resolution. Proc Natl Acad Sci U S A 110:18916–18921

31. Legoll F, Leliévre T (2010) Effective dynamics using conditional expectations. Nonlinearity 23:2131

32. Czerminski R, Elber R (1989) Reaction path study of conformational transitions and helix formation in a tetrapeptide. Proc Natl Acad Sci U S A 86:6963–6967

33. Mills G, Jónsson H (1994) Quantum and thermal effects in dissociative adsorption: evaluation of free energy barriers in multidimensional quantum systems. Phys Rev Lett 72:1124–1127X

34. Fakharzadeh A, Moradi M (2016) Effective Riemannian diffusion model for conformational dynamics of biomolecular systems. J Phys Chem Lett 7:4980–4987

35. Jarzynski C (1997) Nonequilibrium equality for free energy differences. Phys Rev Lett 78:2690–2693

36. Crooks GE (2000) Path-ensemble averages in systems driven far from equilibrium. Phys Rev E 61:2361–2366

37. Hummer G, Szabo A (2001) Free energy reconstruction from nonequilibrium single-molecule pulling experiments. Proc Natl Acad Sci U S A 98:3658–3661

38. Lifson S, Jackson JL (1962) On the self-diffusion of ions in a polyelectrolyte solution. J Chem Phys 36:2410–2414

39. Torrie GM, Valleau JP (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. J Comput Phys 23:187

40. Kumar S, Bouzida D, Swendsen RH et al (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. I. the method. J Comp Chem 13:1011–1021

41. Bartels C (2000) Analyzing biased Monte Carlo and molecular dynamics simulations. Chem Phys Lett 331:446

42. Shirts MR, Chodera JD (2008) Statistically optimal analysis of samples from multiple equilibrium states. J Chem Phys 129:124105

43. Hummer G (2005) Position-dependent diffusion coefficients and free energies from bayesian analysis of equilibrium and replica molecular dynamics simulations. New J Phys 7:34

44. Singharoy A, Chipot C, Moradi M et al (2017) Chemomechanical coupling in hexameric protein-protein interfaces harness energy within V-type ATPases. J Am Chem Soc 139:293–310

45. Smart OS, Neduvelil JG, Wang X et al (1996) HOLE: a program for the analysis of the pore dimensions of ion channel structural models. J Mol Graph 14:354–360