Computing Conformational Free Energy by Deactivated Morphing

Sanghyun Park,[1,*] Albert Y. Lau,[2] and Benoit Roux[2,3]

[1]*Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA*

[2]*Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL 60637, USA*

[3]*Biosciences Division, Argonne National Laboratory, Argonne, IL 60439, USA*

(Dated: April 17, 2008)

Abstract

Despite the significant advances in free energy computations for biomolecules, there exists no general method to evaluate the free energy difference between two conformations of a macromolecule that differ significantly from each other. A crucial ingredient of such a method is the ability to find a path between different conformations that allows an efficient computation of free energy. The free energy difference is the same no matter what path is taken, but not all paths are equally practical. In this paper we introduce a method named 'deactivated morphing' and apply it to two test systems: alanine dipeptide and deca-alanine, both in explicit water. An important feature of this method is the (shameless) use of nonphysical paths, which makes the method robustly applicable to conformational changes of arbitrary complexity.

*Corresponding author: sanghyun@mcs.anl.gov

I. INTRODUCTION

Many important phenomena in molecular biology involve large conformational changes of macromolecules. A quantitative understanding of these phenomena would not be complete without the knowledge of the free energy differences associated with the conformational changes. Free energy is a central concept in understanding properties of physical and chemical systems including biomolecules. Many ingenious algorithms have been invented and tremendous amount of computer time has been spent to compute free energies involved in biomolecular processes. However, there exists no general method to evaluate the free energy difference between two conformations of a macromolecule that differ significantly from each other.

To prevent potential confusion, we state our definition of 'conformation' right away: a conformation of a molecule is a specific assignment of the atomic coordinates of the molecule. In other words, a conformation is a point in a $3N$-dimensional space, with N being the number of atoms in the molecule. Given a number of different conformations of a molecule, we ask the question "what is the free energy difference between different conformations?" One thing to consider here is that biological molecules undergo incessant thermal fluctuations. We need to account for such fluctuations in the computation of conformational free energy. Therefore, a more sensible question is "what is the free energy difference between different conformation ensembles?"

It is, however, not always straightforward to decide how much fluctuation should be included in a conformation ensemble. For example, we may choose an ensemble of conformations whose root-mean-square deviation (RMSD) from a reference conformation is less than a certain threshold. But, the choice of the threshold must be system-dependent; larger thresholds are needed for larger systems. For certain systems, RMSD may not be the best criterion for defining conformation ensembles. Nevertheless, it seems reasonable to expect that if the conformations of interest correspond to centers of metastable free energy wells, the conformational free energy differences should not be overly sensitive to the precise definition of conformation ensembles. In this paper, we assume that a definition of conformation ensembles is already given and focus on the actual computation of free energy differences.

A computation of the free energy difference between two conformation ensembles requires a transformation of one ensemble to the other, unless there is a way to calculate the absolute free energies separately. One key property of free energy is that it is a state function: the free energy difference is the same no matter which transformation path is taken. This, however, does not mean that all paths are equally useful. Finding a path that allows an efficient computation of free

energy is often a crucial step. In general, if there is a big free energy barrier along a transformation path, that path is not so useful because a big barrier usually causes high uncertainties in computed free energies.

A physical path, if one can find one, can be a good choice. The fact that it is a physical path, along which the conformational change in question may actually happen with reasonable probability, indicates that all the barriers along the path are small enough to be surmounted by thermal fluctuations. Finding a physical path, however, is a very challenging task especially for complex conformational changes (see Ref [1] for review). An alternative is to use a nonphysical path. Nonphysical paths are easy to come up with; for example, linear interpolation between two conformations will produce a nonphysical path in most cases. Nonphysical paths, however, will most likely feature very high barriers due to clashes of atoms, stretching of chemical bonds, and so on. To recap, physical paths allow efficient computations of free energy, but they are hard to find; nonphysical paths are easy to find, but it is hard to compute free energy along them.

While physical paths can provide useful information regarding mechanism, kinetics, and so on, it is quite plausible that nonphysical paths may turn out to be more efficient for the purpose of free energy computations. In this paper, we present a method we call 'deactivated morphing'. In this method, the internal interactions of a macromolecule are completely turned off before a transformation is carried out along a nonphysical path. The internal interactions are the ones mostly responsible for high barriers along nonphysical paths. The absence of the internal interactions, therefore, makes nonphysical paths practical for free energy computations. Turning off the internal interactions, of course, is a serious business and must be done with great care. We achieve this by means of position restraints applied to each atom of the macromolecule.

In the next section, we describe the method of deactivated morphing (DM). We then present applications of DM to two test systems: alanine dipeptide (AlaD) and deca-alanine (Ala10), both in explicit water. In Section III, we compute the free energy differences among four representative conformations of AlaD and compare them to the results obtained with umbrella sampling. In Section IV, we compute the free energy difference between helix and hairpin conformations of Ala10. Possible improvements and future applications of DM are discussed in Sections V and VI.

II. DEACTIVATED MORPHING

In this section, we describe the rationale and the details of DM. The schematic diagram of Fig. 1 will be a guide throughout.

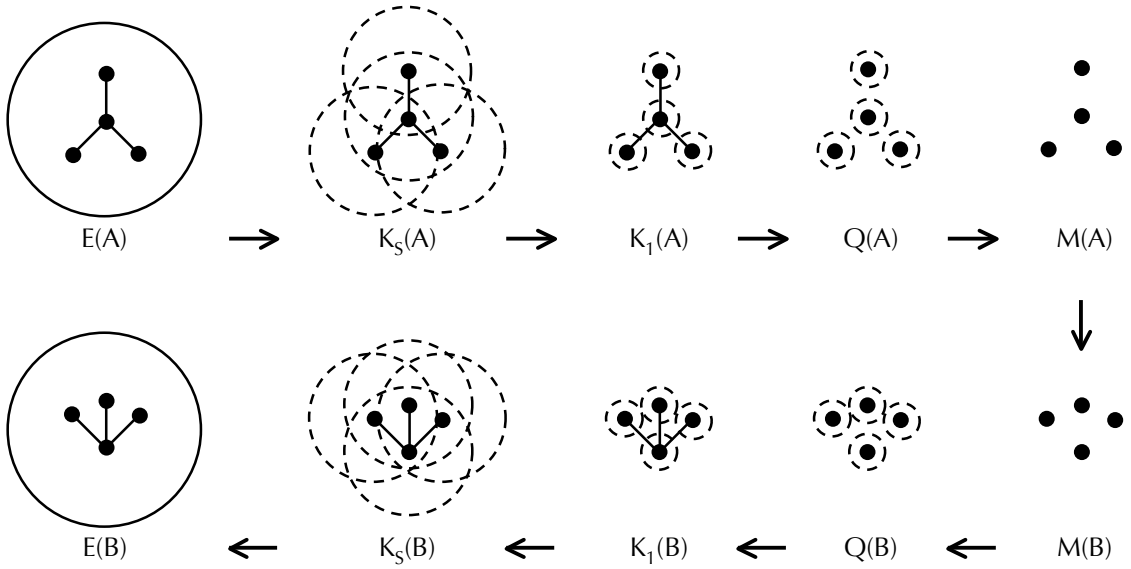

FIG. 1: Deactivated morphing. An ensemble around conformation A is gradually transformed into another ensemble around conformation B through restraining, deactivating, and morphing. Dashed circles represent position restraints.

A. Conformational free energy

Suppose we are asked to compute the free energy difference between two conformation ensembles of a protein in water,[1] one restricted within a basin around conformation A and the other around conformation B. Let E(A) and E(B) denote these ensembles around conformation A and B, respectively. We will also refer to them as 'states' at times in the sense that they may be considered thermodynamic states. A basin, for example, can be defined as the set of conformations that are within some cutoff RMSD ξ from a reference conformation. The solid circles in Fig. 1 signify such basins.

The free energy F_{E} of each ensemble is then given as[2]

$$e^{-\beta F_{\mathrm{E}}} = \int d\mathbf{X}\, d\mathbf{Y}\, \Theta(\mathbf{X})\, e^{-\beta U(\mathbf{X},\mathbf{Y})} \,. \tag{1}$$

The variable $\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ denotes the protein coordinates (i.e., the protein conformation) and $\mathbf{Y} := (\mathbf{y}_1, \ldots, \mathbf{y}_{N'})$ denotes the water coordinates, where N and N' are the total number of atoms of protein and water, respectively. The potential energy U can be decomposed into the

[1] It can be any solute in any solvent, but we find it comforting to use the terminology of protein and water.

[2] Even though it is straightforward to generalize the framework to other types of thermodynamic ensembles, for notational simplicity we stay with the canonical ensemble framework. For the simulations featured in Sections III and IV, the isothermal-isobaric ensemble framework is the relevant one.

internal interactions of protein, the internal interactions of water, and the interactions between protein and water:

$$U(\mathbf{X}, \mathbf{Y}) = U_{\mathrm{p}}(\mathbf{X}) + U_{\mathrm{w}}(\mathbf{Y}) + U_{\mathrm{pw}}(\mathbf{X}, \mathbf{Y}) \; . \tag{2}$$

The function $\Theta(\mathbf{X})$ indicates whether a conformation $\mathbf{X}$ belongs to the basin. For example, if the basin is defined within a cutoff RMSD ξ from a reference conformation $\hat{\mathbf{X}}$, $\Theta(\mathbf{X})$ equals 1 if $\mathrm{RMSD}(\mathbf{X}, \hat{\mathbf{X}}) \leq \xi$ and 0 otherwise, where the RMSD is defined as

$$\begin{aligned} \mathrm{RMSD}(\mathbf{X}, \hat{\mathbf{X}}) &:= \left[\frac{1}{N} (\mathbf{X} - \hat{\mathbf{X}})^2 \right]^{1/2} \\ &= \left[\frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \hat{\mathbf{x}}_n)^2 \right]^{1/2} \; . \end{aligned} \tag{3}$$

The reference conformation $\hat{\mathbf{X}}$ is taken to be A or B, depending on which ensemble we are dealing with.

Note that here we define the RMSD *without* alignment, which makes subsequent simulations more convenient. As a consequence, conformation ensembles defined in terms of the RMSD do not cover translation and rotation. We expect, however, that the removal of translation and rotation will only have a marginal effect on the total free energy difference as long as the two conformations are of similar overall sizes.

Given the two ensembles E(A) and E(B), what we need to compute is the free energy difference $F_{\mathrm{E(B)}} - F_{\mathrm{E(A)}}$. In general, it is practically impossible to compute the free energy difference unless the two ensembles overlap significantly, which is what makes free energy a very expensive quantity to compute. In the present case, there will be virtually no overlap unless the conformational change in question is nothing but trivial. We need to have intermediate states along a path that connects A and B so that significant overlaps exist between all the neighboring states. Such overlaps allow us to compute free energy differences using bidirectional free energy perturbation, i.e. the Bennett acceptance ratio (BAR) method [2], as described in Section II E.[3]

As we declared in Introduction, here we use nonphysical paths such as the ones given by linear interpolation between two conformations. Transformation along a nonphysical path generally causes many energetically unfavorable events, such as clashes of atoms, stretching of bonds, and so on, which lead to large free energy barriers along the path and high uncertainties in computed free energies. In other words, it is very difficult to secure significant overlaps along a nonphysical

[3] An alternative would be to perform thermodynamic integration along the chosen path.

path. We get around this problem by completely deactivating the internal interactions of protein before carrying out a transformation. A new problem then arises: we need to secure an overlap between the states before and after the deactivation. We get around that problem by restraining the protein atoms.

The entire procedure of DM consists of the following steps, as illustrated in Fig. 1. Apply weak restraints on the atoms of protein in conformation A [state $\mathrm{K}_S(\mathrm{A})$] such that there is good overlap with state E(A). Gradually strengthen the restraints [states $\mathrm{K}_S(\mathrm{A}), \mathrm{K}_{S-1}(\mathrm{A}), \ldots, \mathrm{K}_1(\mathrm{A})$] until the protein atoms are more or less fixed. Switch off the internal interactions of protein [state Q(A)], and fix the protein atoms at conformation A [state M(A)]. Morph conformation A [state M(A)] into conformation B [state M(B)]. Unfix the protein atoms and apply strong restraints on them [state Q(B)]. Switch on the internal interactions of protein [state $\mathrm{K}_1(\mathrm{B})$]. Gradually weaken the restraints [states $\mathrm{K}_1(\mathrm{B}), \mathrm{K}_2(\mathrm{B}), \ldots, \mathrm{K}_S(\mathrm{B})$] until there is good overlap with state E(B). We now explain the details of restraining, deactivating, and morphing.

B. Restraining

In the restraining step (states $\mathrm{K}_1, \ldots, \mathrm{K}_S$), we apply harmonic restraints on the atoms of protein:

$$U_{\mathrm{K}_i}(\mathbf{X}, \mathbf{Y}) = U(\mathbf{X}, \mathbf{Y}) + \frac{\kappa_i}{2}(\mathbf{X} - \hat{\mathbf{X}})^2 \,, \tag{4}$$

where $i = 1, \ldots, S$. The corresponding free energy is given by

$$e^{-\beta F_{\mathrm{K}_i}} = \int d\mathbf{X}\, d\mathbf{Y}\, e^{-\beta U_{\mathrm{K}_i}(\mathbf{X},\mathbf{Y})} \,. \tag{5}$$

Again, the reference conformation $\hat{\mathbf{X}}$ is either conformation A or B. Each state K_i is associated with a different spring constant κ_i. We choose κ_1 big enough that the deactivation can be performed with good overlap, and choose κ_S small enough that states E and K_S overlap significantly. The overlap between neighboring states determines the intermediate spring constants, $\kappa_2, \ldots, \kappa_{S-1}$, and hence the total number of spring constants to be used. Notice that the protein atoms are restrained in the absolute space, which removes translation and rotation.

C. Deactivating

In the deactivation step, two things happen: the internal interactions of protein are turned off and the protein atoms are fixed at the reference positions. Fixing of the protein atoms is useful

for the subsequent morphing step. M denotes the state after these two operations have been completed:

$$U_{\mathrm{M}}(\mathbf{Y}) = U_{\mathrm{w}}(\mathbf{Y}) + U_{\mathrm{pw}}(\hat{\mathbf{X}}, \mathbf{Y}) \tag{6}$$

$$e^{-\beta F_{\mathrm{M}}} = \int d\mathbf{Y}\, e^{-\beta U_{\mathrm{M}}(\mathbf{Y})} \,. \tag{7}$$

Notice that the water still interacts with the protein, but only through the fixed reference positions $\hat{\mathbf{X}}$, which are parameters, not dynamical variables.

The free energy difference between K_1 and M cannot be computed directly from simulations of K_1 and M because these two states are defined in terms of two different sets of dynamical variables, $(\mathbf{X}, \mathbf{Y})$ for state K_1 and $\mathbf{Y}$ for state M. Therefore, we introduce an intermediate state Q:

$$U_{\mathrm{Q}}(\mathbf{X}, \mathbf{Y}) = U_{\mathrm{w}}(\mathbf{Y}) + U_{\mathrm{pw}}(\hat{\mathbf{X}}, \mathbf{Y}) + \frac{\kappa_1}{2}(\mathbf{X} - \hat{\mathbf{X}})^2 \tag{8}$$

$$e^{-\beta F_{\mathrm{Q}}} = \int d\mathbf{X}\, e^{-\beta \frac{\kappa_1}{2}(\mathbf{X}-\hat{\mathbf{X}})^2} \int d\mathbf{Y}\, e^{-\beta[U_{\mathrm{w}}(\mathbf{Y}) + U_{\mathrm{pw}}(\hat{\mathbf{X}}, \mathbf{Y})]} \tag{9}$$

In this state, the protein atoms only see the restraining potentials (with the same spring constant κ_1 as in state K_1) and the water interacts with the fixed positions $\hat{\mathbf{X}}$ instead of $\mathbf{X}$. Consequently, $\mathbf{X}$ and $\mathbf{Y}$ are totally decoupled.

Because K_1 and Q share the same set of dynamical variables, the free energy difference between them can be computed with BAR. The sampling of $\mathbf{X}$ and $\mathbf{Y}$ can be done independently as Eq. 9 indicates: we sample $\mathbf{Y}$ from a simulation of water with the protein coordinates fixed at $\hat{\mathbf{X}}$, and sample $\mathbf{X}$ from the Gaussian distributions. From Eqs. 7 and 9, the free energy difference between Q and M is given by

$$e^{-\beta(F_{\mathrm{Q}} - F_{\mathrm{M}})} = \int d\mathbf{X}\, e^{-\beta \frac{\kappa_1}{2}(\mathbf{X}-\hat{\mathbf{X}})^2} \,. \tag{10}$$

Notice that this free energy difference is in fact independent of $\hat{\mathbf{X}}$, as the parameter $\hat{\mathbf{X}}$ can be removed by a simple change of variables. When we consider the entire DM procedure (Fig. 1), $F_{\mathrm{Q(A)}} - F_{\mathrm{M(A)}}$ therefore *exactly* cancels out with $F_{\mathrm{Q(B)}} - F_{\mathrm{M(B)}}$.

D. Morphing

The morphing step consists of transforming state M(A) into M(B), or vice versa. Because the protein atoms no longer interact with each other, any transformation path can be used as long

as it does not cause abrupt disturbances in protein–water interactions. Here we simply choose a linear interpolation path:

$$\hat{\mathbf{X}} = (1-\lambda)\,\hat{\mathbf{X}}_{\mathrm{A}} + \lambda\,\hat{\mathbf{X}}_{\mathrm{B}} \quad \text{with} \quad 0 \le \lambda \le 1\,, \tag{11}$$

where $\hat{\mathbf{X}}_{\mathrm{A}}$ and $\hat{\mathbf{X}}_{\mathrm{B}}$ denote the conformations A and B, respectively.

Along a chosen transformation path, we arrange a number of intermediate states associated with different values of λ. We then use BAR to compute the free energy differences between neighboring states. The overlap between neighboring states dictates how many intermediate states are needed. Because the protein coordinates are no longer dynamical variables, they are not a factor in securing overlaps, which allows us to take fewer intermediate states than we would need if we performed morphing with restraints on the protein coordinates instead of fixing them.

E. BAR and the overlap of ensembles

We use BAR to compute the free energy differences between neighboring states. A free energy computation using BAR between two states proceeds as follows [2]. A set of microstates $\{\mathbf{R}_1, \ldots, \mathbf{R}_{L_1}\}$ are sampled from state 1 with potential energy function $U_1(\mathbf{R})$, and another set of microstates $\{\mathbf{R}_{L_1+1}, \ldots, \mathbf{R}_{L_1+L_2}\}$ are sampled from state 2 with $U_2(\mathbf{R})$. In the present case, a microstate is a collection of protein and water coordinates, $\mathbf{R} = (\mathbf{X}, \mathbf{Y})$, except for the morphing procedure where $\mathbf{R} = \mathbf{Y}$. The free energy difference $\Delta F := F_2 - F_1$ is then obtained by solving

$$e^{\beta\Delta F} = \sum_{l=1}^{L_1+L_2} \left[L_1\, e^{-\beta\Delta F} + L_2\, e^{-\beta\Delta U(\mathbf{R}_l)} \right]^{-1}\,, \tag{12}$$

where $\Delta U := U_2 - U_1$.

As mentioned above, the free energy difference ΔF thus obtained is reliable only if there is significant overlap between the two states. One way to inspect this overlap is to look at the overlap between $\rho_1(\Delta U)$ and $\rho_2(\Delta U)$, the distributions of ΔU sampled from state 1 and 2 respectively [3, 4]. This is a natural way because ΔU is the quantity through which the sampled microstates are incorporated into BAR. Another interesting property is that the two curves, $\rho_1(\Delta U)$ and $\rho_2(\Delta U)$, always intersect at $\Delta U = \Delta F$.

We use BAR for all the steps of DM except for the computation of $F_{\mathrm{E}} - F_{\mathrm{K}_S}$. It is certainly viable to use BAR here as well, but there is a slight inconvenience in simulating state E. Without any restraints, the protein may stray too far away from the reference conformation. For an efficient sampling, we need to push it back whenever the protein moves out of the RMSD boundary,

which requires computing the RMSD during the simulation. Instead, in the spirit of the weighted histogram analysis method (WHAM) [5, 6] or the multistate version of BAR [7, 8], we collect all the microstates sampled from $\mathrm{K}_1, \ldots, \mathrm{K}_S$ to estimate $F_{\mathrm{E}} - F_{\mathrm{K}_S}$:

$$e^{-\beta(F_{\mathrm{E}} - F_{\mathrm{K}_S})} = \sum_{l=1}^{L_1+\cdots+L_S} \Theta(\mathbf{X}_l) \left[\sum_{i=1}^{S} L_i \frac{e^{-\beta \frac{\kappa_i}{2}(\mathbf{X}_l - \hat{\mathbf{X}})^2}}{e^{-\beta(F_{\mathrm{K}_i} - F_{\mathrm{K}_S})}} \right]^{-1} \tag{13}$$

where $\{\mathbf{X}_1, \ldots, \mathbf{X}_{L_1}\}$ is a set of conformations sampled from K_1, $\{\mathbf{X}_{L_1+1}, \ldots, \mathbf{X}_{L_1+L_2}\}$ is from K_2, and so on. The function $\Theta(\mathbf{X})$ is the same as the one that appears in Eq. 1; its presence indicates that only the conformations that are within the basin are included in the sum.

F. Breakdown of conformational free energy

The three parts of DM — restraining, deactivating, and morphing — provide a breakdown of conformational free energy. The restraining part ($F_{\mathrm{K}_1} - F_{\mathrm{E}}$) mostly measures the conformational entropy of protein, the deactivating part ($F_{\mathrm{Q}} - F_{\mathrm{K}_1}$) the internal energy of protein, and the morphing part ($F_{\mathrm{M(B)}} - F_{\mathrm{M(A)}}$, etc.) the contribution of water–protein interactions. We note that this breakdown is not exact; the restraining and deactivating procedures certainly affect both protein and water. Nevertheless, we expect that it should be a reasonable classification of the dominant contributions of each part.

III. ALANINE DIPEPTIDE

AlaD is chosen as the first test system. From an umbrella sampling simulation, we constructed a free energy map on the ϕ-ψ space shown in Fig. 2. The ϕ-ψ map revealed four minima, around $(-90, -60)$, $(-90, 150)$, $(60, 60)$, and $(60, -120)$, from which we select four representative conformations, labeled A to D. Conformation ensembles are defined within a cutoff RMSD of 0.5 Å from these conformations. Using DM, we compute the free energy differences among these conformation ensembles. Although the conformational changes involved here are rather trivial, we choose this problem as a test because of the possibility to compare the results of DM and umbrella sampling.

A simulation box (Fig. 3) was prepared with an AlaD molecule solvated in 275 TIP3P [9] water molecules. States are set up as described in the preceding section and illustrated in Fig. 1. For the restraining part, we use 15 states, $\mathrm{K}_1, \ldots, \mathrm{K}_{15}$, with exponentially distributed spring constants that are listed in Fig. 4. Morphing is done by linear interpolation (Eq. 11) with 11 different values

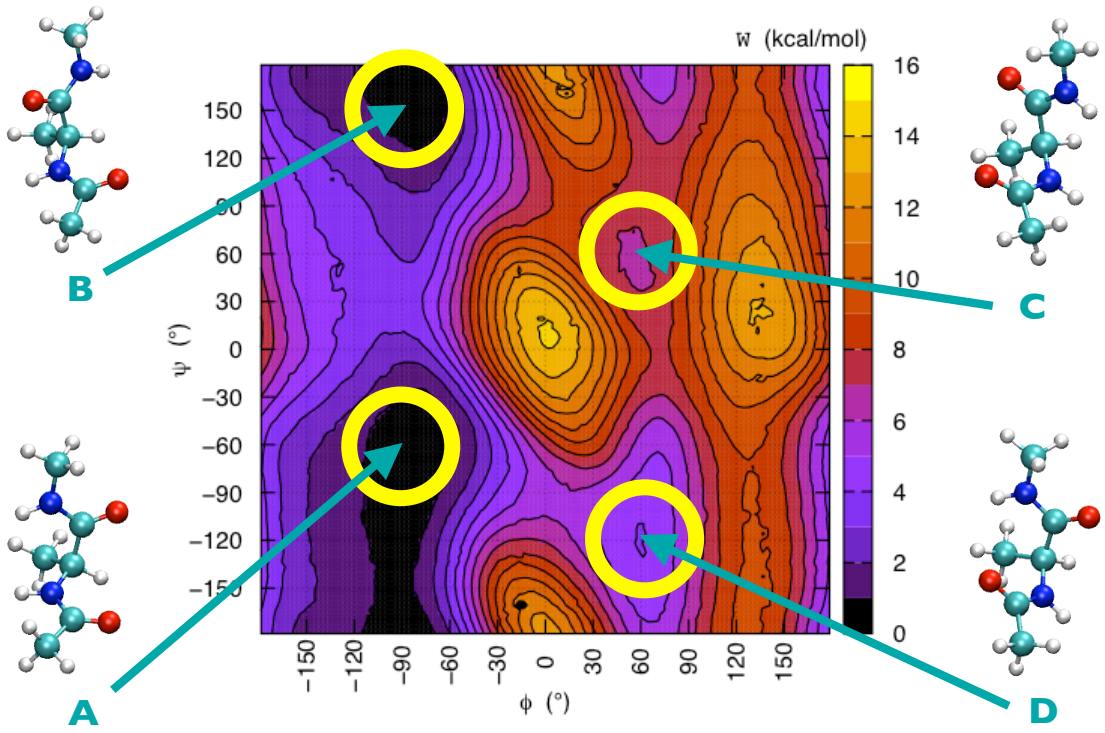

FIG. 2: Four conformations of AlaD selected from the ϕ-ψ map. The four circles roughly correspond to the boundaries of the basins within 0.5 Å RMSD from the four reference conformations. The free energy map was constructed from an umbrella sampling simulation using 144 windows spaced every 30° in ϕ and ψ. For each window, we collected data from a 750 ps run after a 250 ps equilibration period.

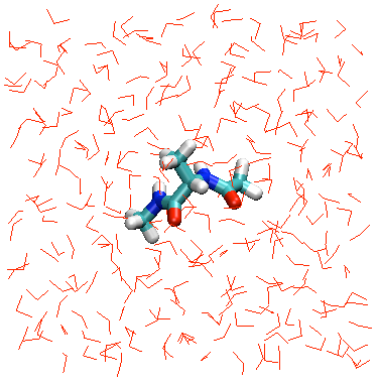

FIG. 3: The simulation box containing an AlaD molecule and 275 water molecules.

of λ. Between conformation A and B, for example, $\lambda = 0$ corresponds to conformation A, $\lambda = 1$ conformation B, and $\lambda = 0.1, 0.2, \ldots, 0.9$ nine intermediate conformations. For each state, we ran a 1 ns molecular dynamics (MD) simulation. Atomic coordinates were sampled every 100 fs, and were grouped into ten blocks in temporal order. The first block is considered an equilibration

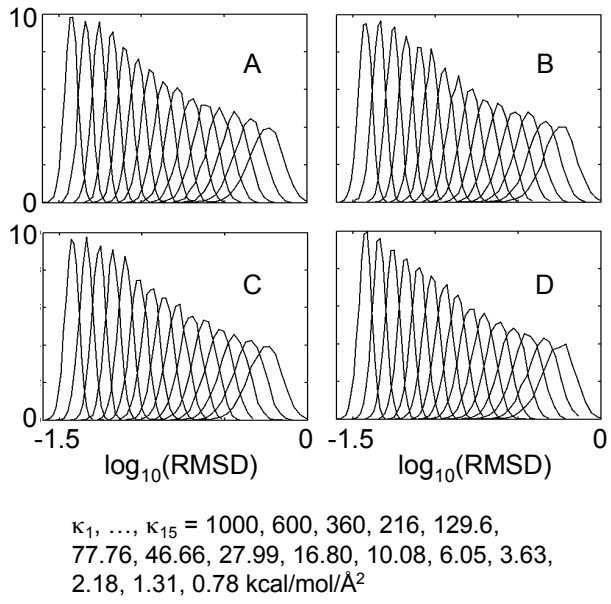

FIG. 4: Overlap of states in the restraining procedure for AlaD. Each panel contains 15 distributions of the RMSD (in Å) from the reference conformation, A, B, C, and D, for 15 different restraining states, $\mathrm{K}_1, \ldots, \mathrm{K}_{15}$. The spring constants, $\kappa_1, \ldots, \kappa_{15}$, used in these states are listed at the bottom.

period and is discarded. From each of the remaining nine blocks, we calculate free energy differences by Eqs. 12 and 13, and report our estimates as $\mu \pm 2\sigma/\sqrt{9}$, where μ and σ are the average and the standard deviation of the nine blocks.

For MD simulations, we used the software NAMD [10] with the CHARMM22 force field [11]. All simulations were run at constant temperature (300 K) and pressure (1 atm) using the Langevin thermostat and the Langevin-piston barostat [12]. Long-range interactions were treated with a cutoff distance of 12 Å, and a time step of 1 fs was used. All the molecular figures in this paper were generated with the software VMD [13].

The choices for the arrangement of states, such as the 15 restraining states and 9 intermediate states for morphing, were made such as to ensure overlaps between states. We check overlaps by plotting the distributions of ΔU (see Sec. II E). In Fig. 4, as an illustration, we show the overlaps of the 15 restraining states. Between restraining states, ΔU depends only on the RMSD,

$$U_{\mathrm{K}_i}(\mathbf{X}, \mathbf{Y}) - U_{\mathrm{K}_j}(\mathbf{X}, \mathbf{Y}) = \frac{\kappa_i - \kappa_j}{2} N[\mathrm{RMSD}(\mathbf{X}, \hat{\mathbf{X}})]^2 \,, \tag{14}$$

which allows us to inspect the overlaps of the ΔU distributions by plotting the RMSD distributions.

Summarized in Fig. 5 are the results of DM. The free energies of E(A) and E(B) are about

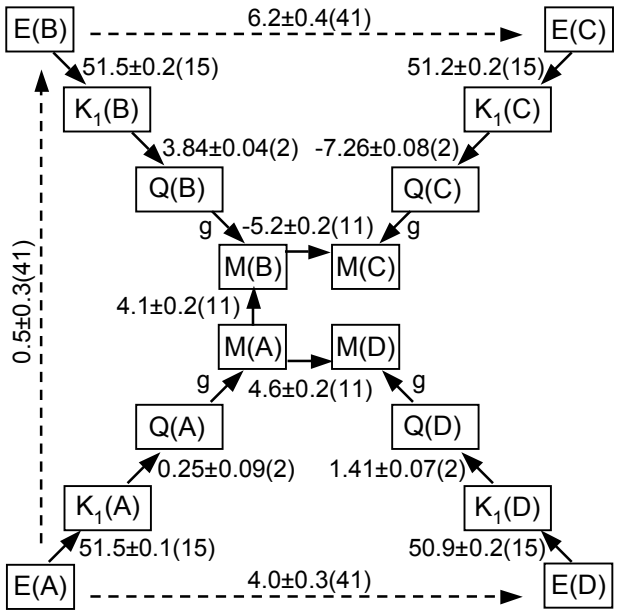

FIG. 5: The outcome of DM applied to the four conformations of AlaD. The numbers next to arrows denote the free energy differences (in kcal/mol) associated with the transitions. The numbers next to dashed arrows are the total free energy differences between the conformation ensembles. The numbers in parentheses are the simulation time (in ns) used to compute the free energy differences. The free energy difference between Q and X, denoted by g, cancels out exactly.

the same; $F_{\mathrm{E(C)}}$ and $F_{\mathrm{E(D)}}$ are $\sim$6 kcal/mol and $\sim$4 kcal/mol higher, respectively. These results agree well with the umbrella sampling results (Fig. 2). Also shown in Fig. 5 is a breakdown of the free energy differences as described in Sec. II F. The restraining free energy ($F_{\mathrm{K_1}} - F_{\mathrm{E}}$) is almost the same for the four conformations, which indicates that the four conformations are equally flexible. According to the deactivating free energy ($F_{\mathrm{Q}} - F_{\mathrm{K_1}}$), conformation B has the lowest internal energy followed by D, A, and C. And, the morphing free energy ($F_{\mathrm{M(B)}} - F_{\mathrm{M(A)}}$, etc.) suggests that water prefers conformation C followed by A, B, and D. Upon inspecting the four conformations (Fig. 2), we offer the following speculation. It appears that when the two oxygens point to the opposite directions (B and D), they participate in forming internal hydrogen bonds, thereby lowering the internal energy. On the other hand, water seems to favor the conformations in which the two oxygens point to the same direction (A and C) because the oxygens in these conformations are more accessible to form hydrogen bonds with water.

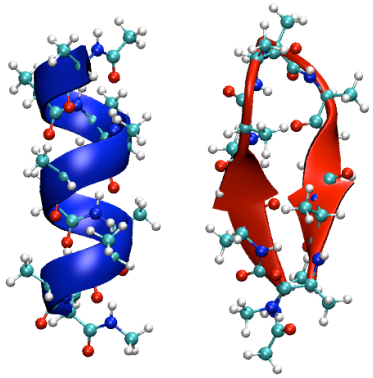

FIG. 6: Two conformations of Ala10, helix and hairpin.

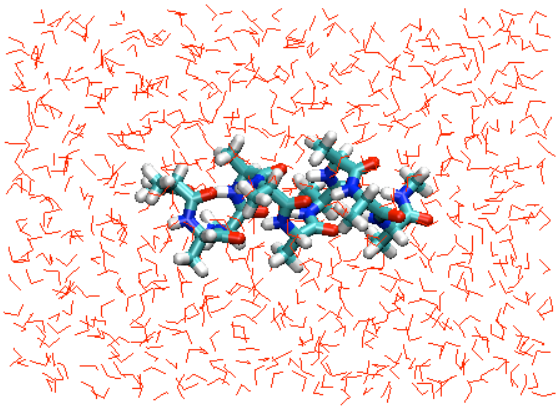

FIG. 7: The simulation box containing an Ala10 molecule and 676 water molecules.

IV. ALANINE DECAMER — HELIX AND HAIRPIN

As another test system, we choose Ala10. Two conformation ensembles are defined within a cutoff RMSD of 2 Å from the helix and hairpin conformations shown in Fig. 6, labeled A and B respectively. The objective is to compute the free energy difference between these two conformation ensembles. Although Ala10 is a simple system by the standards of biology, the conformational change between helix and hairpin is by no means trivial. We expect that physical-path-based methods will have quite a difficulty attacking this problem. Below we demonstrate that DM can solve it without too much trouble.

Shown in Fig. 7 is the simulation box containing an Ala10 molecule, capped with acetyl and N-methyl groups, and 676 water molecules. The computational procedure is the same as the preceding section except that, because of the bigger size of Ala10, here we use more states to ensure overlaps between neighboring states and run longer simulations for some of the states. We use

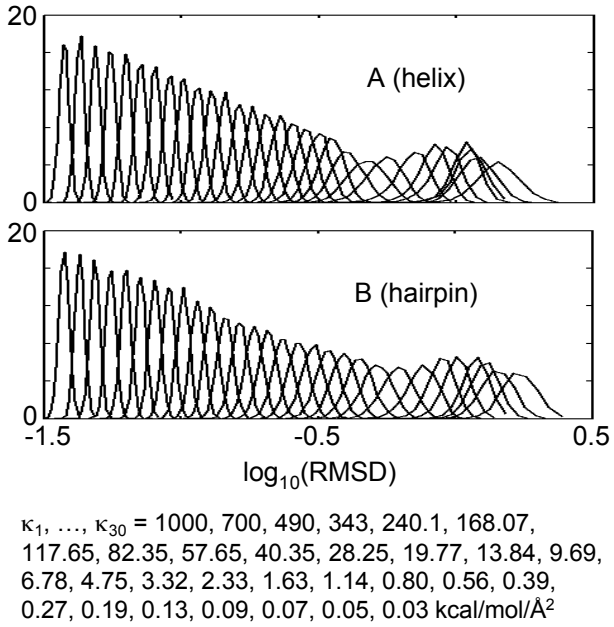

FIG. 8: Overlap of states in the restraining procedure for Ala10. Each panel contains 30 distributions of the RMSD (in Å) from the reference conformation, helix and hairpin, for 30 different restraining states, $\mathrm{K}_1, \ldots, \mathrm{K}_{30}$. The spring constants, $\kappa_1, \ldots, \kappa_{30}$, used in these states are listed at the bottom.

30 restraining states with exponentially distributed spring constants listed in Fig. 8. We ran 3 ns simulations for states $\mathrm{K}_{26}, \ldots, \mathrm{K}_{30}$ and 2 ns for states $\mathrm{K}_{20}, \ldots, \mathrm{K}_{25}$. For all the other states, we ran 1 ns simulations. Longer simulations were necessary in order to reduce the uncertainties of free energy estimates for the restraining states with soft springs because those states span relatively large phase space. The overlaps between the restraining states are shown in Fig. 8.

In the first example, we were able to deactivate AlaD in one step; in other words, we did not need any intermediate states between K_1 and Q. In the case of Ala10, we find that the deactivation procedure must be done in four steps in order to secure overlaps of neighboring states. Residues 1 and 2 are deactivated first; 3, 4 and 5 are next; then 6, 7, and 8; followed by 9 and 10.[4] We use more states for morphing as well. Between helix ($\lambda = 0$) and hairpin ($\lambda = 1$), we use 49 intermediate conformations ($\lambda = 0.02, 0.04, \ldots, 0.98$). Figure 9 shows the free energy profile for the morphing procedure along with several conformations along the way. The molecule becomes significantly smaller around the middle of the morphing pathway, which seems to cause the large dip (~ 80 kcal/mol) in the free energy profile. Nevertheless, we are able to obtain the total free energy difference involved in morphing with a reasonable accuracy: 3.6 ± 0.7 kcal/mol.

[4] Notice that residues 1 and 10 are bigger than the rest because of the capping groups.

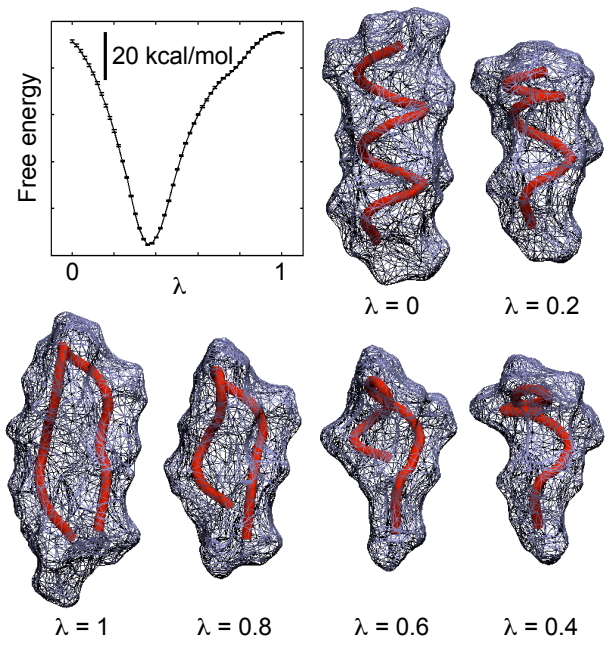

FIG. 9: Morphing between helix and hairpin. The free energy profile over the course of morphing is shown as a function of the parameter λ in Eq. 11. Out of 51 conformations used, six are shown; meshes represent molecular surfaces and red tubes are traces of backbone.

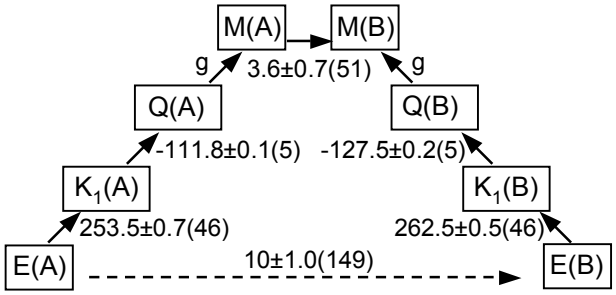

FIG. 10: The outcome of DM applied to the helix (A) and hairpin (B) conformations of Ala10. The numbers next to arrows denote the free energy differences (in kcal/mol) associated with the transitions. The number next to the dashed arrow is the total free energy difference between the two conformation ensembles. The numbers in parentheses are the simulation time (in ns) used to compute the free energy differences. The free energy difference between Q and M, denoted by g, cancels out exactly.

Figure 10 summarizes the results of DM along with a breakdown of free energy as described in Sec. II F. According to the restraining free energy ($F_{\mathrm{K}_1} - F_{\mathrm{E}}$), the hairpin has larger conformational entropy, by 9.0 kcal/mol, than the helix. The deactivation free energy ($F_{\mathrm{Q}} - F_{\mathrm{K}_1}$) indicates that the internal energy of the helix is 16.0 kcal/mol lower than that of the hairpin. And the morphing free energy ($F_{\mathrm{M(B)}} - F_{\mathrm{M(A)}}$) suggests that water prefers the helix by 3.6 kcal/mol. Putting these three

TABLE I: The protein energy $U_{\rm p}$ and the deactivation free energy $F_{\rm Q} - F_{\rm K_1}$ (in kcal/mol). The numbers in parentheses are the relative values, with conformation A as a reference.

	Conformation	$-U_{\rm p}$	$F_{\rm Q} - F_{\rm K_1}$
AlaD	A	11.8 (0.0)	0.3 (0.0)
	B	14.7 (2.9)	3.8 (3.5)
	C	4.3 (-7.5)	-7.3 (-7.6)
	D	12.1 (0.3)	1.4 (1.1)
Ala10	A (helix)	-68.8 (0.0)	-111.8 (0.0)
	B (hairpin)	-84.5 (-15.7)	-127.5 (-15.7)

components together, the helix wins by 10 kcal/mol. This is not a tiny difference in free energy, as it implies that the equilibrium population of the helix is 10^7 times higher than that of the hairpin.

V. POSSIBLE IMPROVEMENTS

The results of the two tests above are satisfactory. But, at the same time, they suggest that there is plenty of room for improvement in the methodology of DM. We list a few possibilities in the following.

We used exponentially distributed spring constants for the restraining states. With such choices, however, the overlap tends to be smaller between the states with stiffer springs and larger between those with softer springs (Figs. 4 and 8). An optimal choice would yield the same degree of overlap throughout the entire range of spring constants. Another issue with the restraining procedure is that restraining states with soft springs may require long simulations because they span relatively large phase space. It may be more efficient to use restraining potentials such as $(\kappa/2)N[\mathrm{RMSD}(\mathbf{X}, \hat{\mathbf{X}}) - \zeta]^2$ with a set of values for ζ, although that would require computation of the RMSD during simulations.

In order to ensure overlap of states, deactivation may have to be done in multiple steps. We were able to deactivate AlaD in one step, but the deactivation of Ala10 required four steps. We find that, with the choice of $\kappa_1 = 1000\,\mathrm{kcal/mol/\AA^2}$, about 30 atoms can be deactivated at a time. As we move on to larger systems, the deactivation procedure may become quite expensive. The results from our examples suggest that there may be a way to approximate the deactivation free energy. As shown in Table I, the values of $F_{\rm Q} - F_{\rm K_1}$ and $-U_{\rm p}$ are substantially different, an indication that the deactivation of $U_{\rm p}$ also affects the protein entropy and the protein–water interactions.

If we compare the relative values among the different conformations, however, we see that the deactivation free energy $F_{\mathrm{Q}} - F_{\mathrm{K}_1}$ can be replaced by $-U_{\mathrm{p}}$, the negative of the internal protein energy, without losing too much accuracy. It remains to be seen whether this approximation is indeed robust for diverse systems.

As we move on to larger systems and more complex conformational changes, the morphing procedure also gets more demanding. In the present examples, we chose linear interpolation paths for morphing. But, it is possible that some other paths may be more efficient. For the case of Ala10, in particular, the total free energy difference for morphing is only 3.6 kcal/mol, but the entire free energy profile during morphing spans ∼80 kcal/mol. If we can find a path along which the range of free energy is smaller, the computation of morphing free energy would likely require fewer intermediate states and therefore be less expensive. Based on that morphing is performed after U_{p} is turned off, what we need is a path that minimally perturbs the protein–water interactions. How to construct such a path is a topic for future research.

Considering that the protein–water interactions are mainly responsible for the morphing free energy, it is a reasonable attempt to estimate the morphing free energy using continuum solvent models such as Poisson-Boltzmann. In fact, a Poisson-Boltzmann calculation of the helix and hairpin conformations of Ala10 yields an energy difference of 2.8 kcal/mol, which is quite close to the morphing free energy of 3.6 kcal/mol we obtained with DM. [The Poisson-Boltzmann calculations were carried out using the PBEQ module of CHARMM [14]. The space-dependent dielectric constant $\epsilon(\mathbf{r})$ and the atomic charges were mapped onto a cubic grid of 120^3 points with a 0.5 Å spacing, and the finite-difference Poisson equation was solved numerically using an over-relaxation algorithm. The atomic charges were taken from the force field used in the MD simulations. The dielectric constant was set to 80 in the bulk solvent region and was set to 1 inside the protein. Ionic salt concentration was set to zero. A set of atomic Born radii optimized from free energy simulations with explicit solvent [15] was used to setup the solvent-protein dielectric boundary, incorporating the re-entrant surface based on a 1.4 Å probe radius. The calculated solvation free energy of the helix and hairpin conformer is -33.88 and -31.04 kcal/mol, respectively.] It remains to be seen whether Poisson-Boltzmann calculations will continue to be as accurate for systems with more diverse charge distributions.

VI. CONCLUSIONS

In this paper, we have introduced a method named deactivated morphing (DM) for computing conformational free energy of macromolecules. An important characteristic of DM is the use of nonphysical paths, which makes the method robustly applicable to conformational changes of arbitrary complexity. There are numerous problems for which DM can potentially be useful. For example, it can be used to study the energetics of allosteric changes of proteins triggered by signals such as ligand binding. Another area of application is protein structure prediction, in which DM can be useful for ranking putative structures based on atomistic force fields. We hope that this method will be much improved in the future and open the door for computational study of a broad range of problems in molecular biophysics.

Acknowledgments

We would like to thank Chris Jarzynski for a stimulating discussion. SP was supported by the Director's Fellowship of the Argonne National Laboratory under the U.S. Department of Energy Contract No. DE-AC02-06CH11357.

[1] R. Elber, *Long-timescale simulation methods*, Curr. Opin. Struct. Biol. **15**, 151 (2005).

[2] C. H. Bennett, *Efficient estimation of free energy differences from Monte Carlo data*, J. Comp. Phys. **22**, 245 (1976).

[3] G. E. Crooks, *Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems*, J. Stat. Phys. **90**, 1481 (1998).

[4] S. Park, *Comparison of the serial and parallel algorithms of generalized ensemble simulations: An analytical approach*, Phys. Rev. E **77**, 016709 (2008).

[5] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, *The weighted histogram analysis method for free-energy calculations on biomolecules*, J. Comput. Chem. **13**, 1011 (1992).

[6] M. Souaille and B. Roux, *Extension to the weighted histogram analysis method: Combining umbrella sampling with free energy calculations*, Comp. Phys. Commun. **135**, 40 (2001).

[7] A. Kong, P. McCullagh, X. L. Meng, D. Nicolae, and Z. Tan, *A theory of statistical models for Monte Carlo integration*, J. R. Statist. Soc. B **65**, 585 (2003).

[8] M. R. Shirts and J. D. Chodera, *Statistically optimal analysis of multiple equilibrium simulations*, Submitted (2008).

[9] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *Comparison of simple potential functions for simulating liquid water*, J. Chem. Phys. **79**, 926 (1983).

[10] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, *Scalable molecular dynamics with NAMD*, J. Comput. Chem. **26**, 1781 (2005).

[11] A. D. MacKerell Jr, D. Bashford, M. Bellott, R. L. Dunbrack Jr, J. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, et al., *All-hydrogen empirical potential for molecular modeling and dynamics studies of proteins using the CHARMM22 force field*, J. Phys. Chem. B **102**, 3586 (1998).

[12] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks, *Constant pressure molecular dynamics simulation: The Langevin piston method*, J. Chem. Phys. **103**, 4613 (1995).

[13] W. Humphrey, A. Dalke, and K. Schulten, *VMD – Visual Molecular Dynamics*, J. Mol. Graphics **14**, 33 (1996).

[14] W. Im, D. Beglov, and B. Roux, *Continuum solvation model: Electrostatic forces from numerical solutions to the Poisson-Bolztmann equation*, Comp. Phys. Commun. **111**, 59 (1998).

[15] M. Nina, D. Beglov, and B. Roux, *Atomic radii for continuum electrostatics calculations based on molecular dynamics free energy simulations*, J. Phys. Chem. B **101**, 5239 (1997).
