THE UNIVERSITY OF CHICAGO

ENHANCED SAMPLING METHODOLOGIES FOR FREE ENERGY CALCULATIONS IN BIOMOLECULAR SYSTEMS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY
DONGHYUK SUH

CHICAGO, ILLINOIS
JUNE 2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my Ph.D. advisor Professor Benoît Roux for everything he has done for me. He is not only the great scientist himself, but also the best mentor among anyone can imagine. Whenever you find an obstacle in research, he fills you with great ideas and guides you to the light - right answer. Whenever you make a mistake, he does not blame you but fills you with courage and motivates you. He encourages you to think on your own, move the research in your own direction, and be creative; and thankfully, whenever you get stuck which happens most likely, he will be there to help. I am and will always be grateful for his support and guidance.

I also would like to thank my two other thesis committee, Professor Aaron Dinner and Professor Suri Vaikuntanathan. Both of them kindly gave me insightful discussions and counsels into my thesis. Again, I really appreciate your time and effort, Professor Aaron and Professor Suri!

The great Roux group experience would have never been the same without past and current lab members: Ahmed Rohaim, Avisek Das, Balasundaresan Dhakshnamoorth, Brian Radak, Chetan Rupakheti, Chris Boughter, Eliot Boulanger, Fabian Paul, Hadi Ramezani-Dakhel, Huan Rui, Hui Li, Jared Ostmeyer, Jing Li, Kevin Changhun Song, Lidong Gong, Lydia Blachowicz, Mikolai Fajer, Matthew Pond, Mukta Sharma, Nabil Faruk, Prithviraj Nandigrami, Rachael Youngworth, Rong Shen, Shahidul Islam, Sunhwan Jo, Jonathan Thirman, Wei Jiang, William Isley, Yilin Meng, Ying Li, Younghoon Koh, Yunjie Chen, and Ziwei He. And Chris Chipot, my dearest collaborator. All these people were extremely helpful at all times.

I am grateful for my friends from the university and church, who made my time here more pleasant and delightful. All or nothing; I cannot list all of you here, so rather will not name any. But I love you all.

I really appreciate unconditional love and support from my family, especially my parents, Jaewon Suh and Bokhee Lee. Again, I love you all.

Lastly, I would like to thank God.

ABSTRACT

We developed two enhanced sampling methods, one for configurational sampling of small molecules/peptides, and the other for calculating protein-protein binding free energies to solve interaction/recognition problems.

To enhance the configurational sampling which can be very computationally demanding with conventional Molecular Dynamics (MD) simulations, a Hybrid non-equilibrium MD/Monte Carlo propagator is developed in which we apply non-equilibrium work to the system in order to boost the Hamiltonian while a Metropolis-Hastings Monte Carlo (MC) step ensures sampling from the correct Boltzmann distribution. When the sampling of a biomolecular system gets stuck in a kinetic trap (metastable state), the Hybrid neMD/MC propagator helps it to escape. Specific biomolecular peptide systems were used to test validity and performance of the method.

We put forth a novel theoretical framework for binding free energy calculations between two proteins, leaning on the optimal curvilinear minimum free-energy path (MFEP) determined from the string method. The curvilinear path connects the fully bound state to the unbound state and is generated from quick simulations using an implicit solvent model, followed by application of the dynamic histogram analysis method (DHAM). In each of the simulations DHAM finds the free energy minimum; minima from all simulations jointly describe the MFEP. The curvilinear path avoids the free energy barrier that the rectilinear path connecting the same bound and unbound states has to cross, leading to a faster convergence of binding-free energy estimates. A host-guest system and a protein complex were used to test validity and performance of the method.

With these enhanced sampling methods developed, we tackle two biomolecular problems, namely dissociation of the HIV-1 Nef /SH3 and Colicin/Im9 complexes. The former shows an immense change in binding affinity if a single residue (R96I) is mutated. The binding affinity of both wildtype and R96I mutant complex are investigated using the string method. Additionally the bound complex is sampled with Hybrid neMD/MC propagator without any

restraints to find additional key residue interactions. Knowing these residues allowed us to explain the origin of cognate/non-cognate Colicin binding; Im9 (cognate interaction) shows 6 orders of magnitude stronger binding affinity than Im2 (non-cognate interaction). The binding affinity of both complexes are investigated using the string method.

CHAPTER 1

INTRODUCTION

1.1 Enhanced sampling method development

Molecular Dynamics (MD), which consists in integrating Newton's classical equations of motion to generate detailed trajectories, is a powerful tool to study complex atomic models of biomolecular systems.[1] MD simulations can help understand biological macromolecular systems by providing a wealth of information across a broad range of time scales, thus complementing experimental techniques that often have more limited spatial or temporal resolution. However, the detailed dynamics of all-atom models takes place on a complex and rugged energy surface, and as a consequence, sampling the relevant conformations is burdened by a host of slow process. In this thesis, we formulate and test novel methods to tackle two important challenges in simulations of biomolecular systems: enhanced configurational sampling, and the determination of binding free energies.

In chapter 2, we introduce a novel hybrid non-equilibrium molecular dynamics – Monte Carlo (neMD-MC) propagator.[2–5] This method was designed to help achieve a more efficient configurational sampling during the simulations of a complex system. Of great importance, we demonstrate that the statistical weight of the configurations generated by the hybrid neMD-MC propagator is consistent with the equilibrium Boltzmann distribution. Discovering all possible conformations and determining their associated statistical weight is of great help in studies of biomolecular systems. One example is the folding/unfolding processes of a small peptide or protein in solution, which could be highly beneficial in many applications such as drug design.

In chapter 3, we introduce a novel formulation of the absolute binding free energy based on the physical separation of the two molecules along the minimum free energy path (MFEP).[6–8] The goal of this method is to serve in computational studies aimed at understanding the nature of protein-protein association and recognition. While unbiased (brute-force) MD sim-

ulations provide the most detailed and most general conformational exploration, they often requires excessively long trajectories. This manifests itself as a slow convergence of statistical estimates, especially free energies. However, the generality of unbiased trajectories can be traded for faster convergence by restricting the sampling around to those conformations that dominate the observables that we are interested in. Technically, this is achieved by introducing restraining potential affecting selected degrees of freedom, or collective variables (CVs).[9] By biasing the simulations toward selected conformations, convergence of free energy calculations can be speeded up by several orders of magnitude. In our new string-based method to calculate the binding free energy of two proteins, we enhance the convergence by restraining along the MFEP.

1.2 Hybrid non-equilibrium MD-MC propagator

Chapter 2 introduces the hybrid neMD-MC propagator, which is adapted from the featured article [Suh, Radak, Chipot, and Roux, J. Chem. Phys. 148, 014101(2018)]. Classical molecular dynamics (MD) and Metropolis-Hasting Monte Carlo (MC) simulations based on detailed atomic models are powerful tools to study the properties of complex biomolecular systems.[1–3] While simulations based on realistic all-atom (AA) models arguably offer the most detailed information, such models evolve on a complex and rugged energy surface and their dynamics are often burdened by a host of slow processes. For this reason, achieving an adequate sampling of all the relevant configurations of a system from straight MD or MC simulations is often challenging.

A number of schemes have been proposed to enhance sampling and accelerate the exploration of configurational space by smoothing the underlying potential energy surface of a system.[10–25] Among those schemes, two simple and attractive ideas are the hyperdynamics or accelerated MD (aMD),[15,16] and the replica-exchange with solute tempering (REST2) algorithms.[22] However, to recover the Boltzmann equilibrium distribution of the proper Hamiltonian, one must either carry out a post-hoc re-weighting analysis, or generate the

simulation within the context of a Hamiltonian tempering replica-exchange scheme involving multiple copies of the system.[26] While the former suffers from most configurations being statistically meaningless after re-weighting, the latter substantially increases the computational cost of the simulation.

One avenue to address these issues is to carry out the propagation in such a way that the resulting configurational sampling reflects the proper Hamiltonian. For example, it is possible to reformulate the self-guided Langevin dynamics (SGLD) in such a way as to restore microscopic detailed balance[25]. In the present effort, we wish to explore hybrid dynamical propagation schemes that combine the strength of non-equilibrium molecular dynamics (neMD) and Metropolis Monte Carlo (MC) to achieve enhanced sampling.[27–32] In hybrid neMD-MC, the value of some chosen variable or coupling parameter is altered gradually in a time-dependent and controlled fashion, while the remaining degrees of freedom are allowed to evolve freely according to the normal equations of motion. The configuration generated by such a non-equilibrium "switching" trajectory is then treated as a candidate that must be either accepted or rejected via a Metropolis criterion to generate the equilibrium Boltzmann distribution.

The hybrid neMD-MC propagator designed here comprises an "equilibrium phase", a "boosting phase", and a "Metropolis MC step":

(*i*) Equilibrium phase: An atomic system is dynamically propagated for some period of time using standard equilibrium MD on the correct potential energy surface.

(*ii*) Boosting phase: The system is then propagated for a brief period of time τ via a time-dependent Hamiltonian that is evolved toward the perturbed potential energy surface and then back to the correct potential energy surface;

(*iii*) Metropolis MC step: The resulting configuration at the end of the neMD trajectory is then accepted or rejected according to a Metropolis criterion before returning to step 1. Using a symmetric switching schedule for ramping the Hamiltonian up and down,

as well as keeping detailed balance with symmetric two-end momentum reversal ensure that the algorithm strictly produces a Boltzmann equilibrium distribution.[33]

In contrast to the hybrid neMD-MC simulations guided by a coarse-grained model introduced previously,[31] the algorithm described above relies on a perturbed Hamiltonian, but does not require the construction of a coarse-grained (CG) model to generate the target configuration. The hybrid neMD-MC sampling propagator implemented here rests on two schemes during the boosting phase, the so-called hyperdynamics accelerated MD (aMD)[15,16] and the replica-exchange with solute tempering (REST2).[22] Nevertheless, the strategy allows virtually any number of variations. Furthermore, the hybrid propagator may be naturally combined with a number of enhanced sampling strategies and free energy techniques.[23] For example, preliminary results are also shown using a time-dependent bias along a collective variable determined via the adaptive biasing force (ABF) approach.[34]

The performance of the method is then illustrated with specific biomolecular systems. Our results indicate that the method can yield a significant speedup for biomolecular systems.

1.3 String method for protein-protein binding free energy calculation

Chapter 3 introduces the novel formulation of the binding free energy of two molecules based on the string method, which is adapted from the work [Suh, Jo, Jiang, Chipot, and Roux, J. Chem. Theory Comput. submitted, 2019].

A rigorous free-energy framework has been developed for computing absolute binding free energies, using a geometric route, resting on the calculation of the potential of mean force (PMF) for the separation of the two binding partners.[9,35,36] To reduce the space of accessible configurations and accelerate sampling, the binding process is decomposed in stepwise stages, wherein a number of geometric restraints are introduced and removed, and their contribution to the standard binding free energy is formally accounted for. These

restraints are aimed at restricting the conformational freedom and relative orientation and position of the two binding partners. This approach, which has proven a method of choice to describe protein association, relies traditionally on a uniaxial rectilinear separation path of the configurationally restrained bodies.

In an attempt to extend the work of Gumbart et al.,[36] we found that the PMF calculation for the separation of two proteins along a predefined uniaxial path, coinciding with the vector connecting their centers of mass, may converge very slowly. Although free energy is a state function and its calculation does not depend on the choice of the path, in practice, this choice may have a significant impact on the rate of convergence. In general, free-energy calculations are expected to converge faster when the model reaction coordinate coincides with minimum free-energy path.[6–8]

In this work, we explore the local free-energy surface that underlies protein–protein association, and identify the optimal curvilinear pathway, using the string method.[7,8] The calculation of the standard binding free energy is recast in an original and rigorous theoretical formalism that rests on the string pathway, and evaluated by means of enhanced-sampling simulations. Performance of the methodology is probed with the model host–guest complex formed by benzene associated to cucurbit[7]uril (CB[7]),[37] and the well characterized protein complex formed by extracellular ribonuclease barnase binding inhibitor barstar.[38–41] The barnase/barstar system in particular is one of the best model system for assessing the computational methodologies based on all-atom simulations. The complex has femtomolar affinity, which translates as –19.0 kcal/mol of binding free energy.[42] There are high resolution structures of the complex[38] and the effect of multiple site-directed mutations on binding affinity has been measured,[40,40] which makes them well suited for assessing the accuracy of the free energy methods.

1.4 Protein-Protein association/recognition problems

Chapter 4 discusses real biomolecular problems and our attempts to solve them with proposed enhanced sampling methods introduced in previous chapters.

The first problem we tackle is the dissociation of the HIV-1 Nef/SH3 complex that shows a substantial binding affinity change with a single residue substitution. The mutation on the SH3 domain (R96I) increases binding affinity tremendously and coincides with slight change in relative orientation of the binding partners.[43] Both wild-type and R96I mutant complex are simulated with the string method formulated above.[44] While simulations restrained to the string reveal the binding free energy of each complex, simulating the wild-type and mutant bound complexes with the hybrid neMD-MC propagator and without any restraints allows to find additional key residue interactions.

The second problem we tackle is dissociation of the Colisin/Im9 and Colisin/Im2 complexes that are a representative cognate vs. non-cognate binding partners. Although cognate Im9 and non-cognate Im2 immunity proteins have no significant structural difference, the latter has 6 orders of magnitude weaker binding affinity to Colicin E9.[45]The two proteins are separated along the optimal pathway revealed with our string method. Furthermore, we try to predict if given complexes are cognate or non-cognate from the steepness of the free energy difference along the separation pathway.

CHAPTER 2

HYBRID NON-EQUILIBRIUM MOLECULAR DYNAMICS - MONTE CARLO PROPAGATOR[1]

2.1 Introduction

While all-atom simulation can give detailed information of the model system, one can observe that the system with compound free energy surfaces gets stuck into the metastable state or kinetic trap. Several schemes have been proposed to address this problem by effectively smoothing the potential energy surface. However, in order to recover the proper Boltzmann equilibrium probability distribution, these approaches must then rely on statistical reweighting techniques, or generate the simulations within a Hamiltonian tempering replica-exchange scheme. The present work puts forth a novel hybrid sampling propagator combining Metropolis-Hastings Monte Carlo with proposed moves generated by non-equilibrium MD. This hybrid neMD-MC propagator comprises three elementary elements: *i*. An atomic system is dynamically propagated for some period of time using standard equilibrium MD on the correct potential energy surface; *ii*. The system is then propagated for a brief period of time during what is referred to as a "boosting phase", via a time-dependent Hamiltonian that is evolved toward the perturbed potential energy surface and then back to the correct potential energy surface; *iii*. The resulting configuration at the end of the neMD trajectory is then accepted or rejected according to a Metropolis criterion before returning to step 1. A symmetric two-end momentum reversal prescription is used at the end of the neMD trajectories to guarantee that the hybrid neMD-MC sampling propagator obeys microscopic detailed balance and rigorously yields the equilibrium Boltzmann distribution. The hybrid neMD-MC sampling propagator is designed and implemented to enhance the sampling by relying on the accelerated MD (aMD) and solute tempering schemes. It is also combined

1. Adapted from Suh, Radak, Chipot, and Roux, *J. Chem. Phys.* **148**, 014101 (2018).

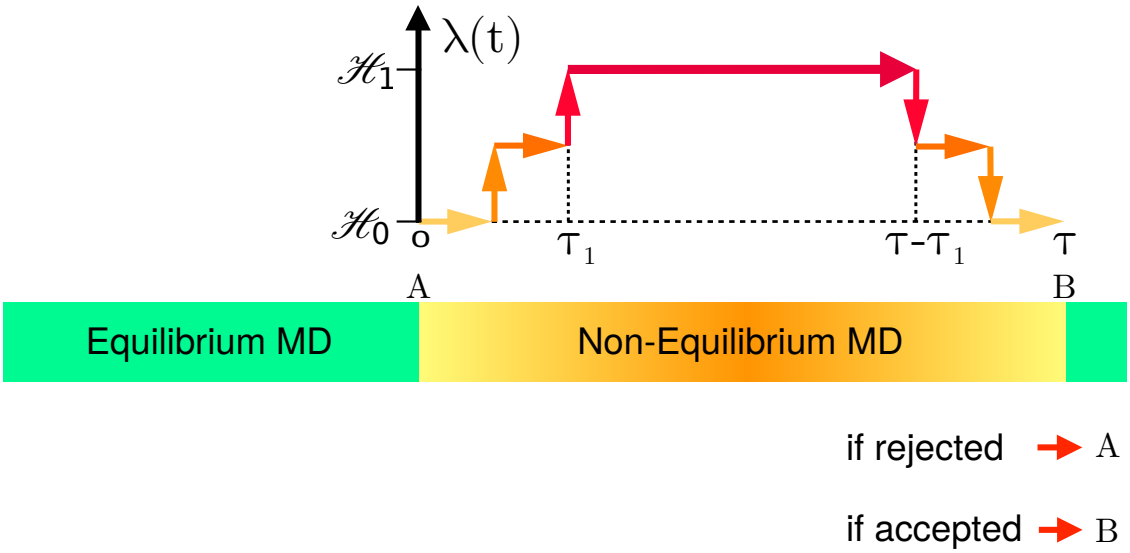

Figure 2.1: Hybrid neMD-MC propagator scheme comprising an equilibrium MD phase, a non-equilibrium MD boosting phase, and a acceptance or rejection via a Metropolis MC step.

with the adaptive biased force (ABF) sampling algorithm to examine. Illustrative tests with specific biomolecular systems indicate that the method can yield a significant speedup.

2.2 Theory and Computational Methods

2.2.1 Theory

In this section, we formulate the theoretical basis of the hybrid neMD-MC sampling propagator.

Let us consider a classical system with time-dependent Hamiltonian $H[\boldsymbol{x};\lambda(t)]$, where $\boldsymbol{x}$ represents all coordinates and momenta, and $\lambda(t)$ is a time-dependent coupling parameter.

Here, λ=0 denotes the unperturbed Hamiltonian, H_0, while λ=1 corresponds to the perturbed Hamiltonian, H_1, offering the maximum boost.

In the conventional hybrid neMD-MC scheme for alchemical changes to the Hamiltonian, one varies λ from 0 to 1 in a neMD trajectory in order to generate a proposed move from $\boldsymbol{x}$ to $\boldsymbol{x'}$. The final configuration at the end of the non-equilibrium switching period corresponds to a different Hamiltonian. For example, in constant-pH simulations based on a hybrid

neMD-MC algorithm, this could be the unprotonated or protonated state of a residue.[27,28]

In contrast, the boosting phase of the hybrid neMD-MC propagator starts and ends with the same unperturbed Hamiltonian, H_0. In other words, the system is first brought to a perturbed Hamiltonian, H_1, associated with a potential energy surface that is smoother and easier to sample ($\lambda = 0 \rightarrow \lambda = 1$), and then brought back to the unperturbed Hamiltonian H_0 ($\lambda = 1 \rightarrow \lambda = 0$).

A random walk in phase space for a system at thermodynamic equilibrium must obey microscopic detailed balance,

$$\begin{aligned} \pi(\boldsymbol{x}) T_{\mathrm{p}}(\boldsymbol{x} \rightarrow \boldsymbol{x'}) T_{\mathrm{a}}(\boldsymbol{x} \rightarrow \boldsymbol{x'}) \quad &= \\ & \pi(\boldsymbol{x'}) T_{\mathrm{p}}(\boldsymbol{x'} \rightarrow \boldsymbol{x}) T_{\mathrm{a}}(\boldsymbol{x'} \rightarrow \boldsymbol{x}) \end{aligned} \tag{2.1}$$

where $\pi(\boldsymbol{x})$ is the equilibrium probability of being in state $\boldsymbol{x}$, $T_{\mathrm{p}}(\boldsymbol{x} \rightarrow \boldsymbol{x'})$ is the transition probability for generating a proposed move from $\boldsymbol{x}$ to $\boldsymbol{x'}$, and $T_{\mathrm{a}}(\boldsymbol{x} \rightarrow \boldsymbol{x'})$ is the transition probability for accepting a proposed move from $\boldsymbol{x}$ to $\boldsymbol{x'}$. Here, $\pi(\boldsymbol{x}) = Q^{-1} \exp\left[-H(\boldsymbol{x})/k_{\mathrm{B}}T\right]$, where Q is the canonical partition function. After the boosting phase is completed, the energy difference $\Delta H = H[\boldsymbol{x}_{\mathrm{B}}, \lambda = 0] - H[\boldsymbol{x}_{\mathrm{A}}, \lambda = 0]$ is calculated, and a standard Metropolis acceptance criterion

$$T_{\mathrm{a}} = \min\left[1, e^{-\beta \Delta H}\right], \tag{2.2}$$

is used to enforce microscopic detailed balance ($\beta \equiv 1/k_{\mathrm{B}}T$ where k_{B} is Boltzmann's constant and T is the absolute temperature). The hybrid propagator with the different stages is depicted schematically in Figure 2.1.

A "symmetric two-ends momentum reversal" prescription is used in association with the non-equilibrium switches,[28,31,33] by which one randomly chooses with equal probability to carry out these trajectories as a "forward" or "backward" MD propagation. Forward propagation means that the non-equilibrium switch trajectory is simply continued using current positions and momenta, whereas backward propagation means that the momenta

of the system are flipped at the beginning and at the end of the non-equilibrium switch trajectory (i.e., it is equivalent to a propagation going back in time). It was shown previously that the symmetric two-ends momentum reversal prescription increases the sampling rate by reducing the likelihood that different regions of configurational space remain isolated from one another.[31,33] The energy difference ΔH in Eq. (2.2) may be further decomposed as a sum of two terms, q and w, where q is the heat exchange between the system and an external heat bath during propagation, and w is the non-equilibrium work done during the perturbation. Since it is not trivial to keep track of heat-exchange during stochastic integration (such as Langevin dynamics), we instead opt to use a deterministic integrator such that the heat exchange is formally zero. The energy difference between states A and B is equal to the non-equilibrium work applied during the trajectory augmented by the shadow work done. The shadow work corresponds to the non-equilibrium work caused by the integrator error associated with the finite time-step.[46] During the boosting phase, we used a schedule symmetric in time for perturbing the Hamiltonian by the coupling parameter, $\lambda(t)$,

$$\lambda(t;\tau,\tau_1) = \begin{cases} t/\tau_1, & (0 \le t < \tau_1) \\ 1, & (\tau_1 \le t < \tau - \tau_1) \\ (\tau - t)/\tau_1, & (\tau - \tau_1 \le t < \tau) \end{cases} \tag{2.3}$$

where τ is the total switching time and τ_1 is the ramping time (see Figure 2.1). The middle region, during which λ is a constant equal to 1, is the boosting phase that lasts for $\tau - 2\tau_1$.

In practice, the time-dependent coupling parameter $\lambda(t)$, was implemented via stepwise

variations with a sequence of $n_{\rm s}$ discrete steps,

$$\lambda(t;\tau,\tau_1) \;=\; \tag{2.4}$$

$$\begin{cases} 0, & 0 \le t < \tau_1/n_{\rm s} \\ 1/n_{\rm s}, & \tau_1/n_{\rm s} \le t < 2\tau_1/n_{\rm s} \\ \ldots & \\ (n_{\rm s}-1)/n_{\rm s}, & \tau_1(n_{\rm s}-1)/n_{\rm s} \le t < \tau_1 \\ 1, & \tau_1 \le t < \tau-\tau_1 \\ 1-1/n_{\rm s}, & \tau-\tau_1+\tau_1/n_{\rm s} \le t < \tau-\tau_1+2\tau_1/n_{\rm s} \\ \ldots & \\ 1/n_{\rm s}, & \tau-2\tau_1/n_{\rm s} \le t < \tau-\tau_1/n_{\rm s} \\ 0, & \tau-\tau_1/n_{\rm s} \le t < \tau \end{cases} \tag{2.5}$$

as depicted in Figure 2.1 (see also the Tcl pseudocode in Appendix).

The hybrid neMD-MC propagator algorithm was implemented according to the following steps:

(i) The system is dynamically propagated for some period of time using unbiased equilibrium MD with the unperturbed Hamiltonian.

(ii)
- The forward/backward direction for the switch is randomly chosen with equal probability for the neMD switching trajectory; the momenta are flipped if a backward propagation is chosen;
- The system is propagated via the time-dependent Hamiltonian $H[x,\lambda(t)]$ toward the perturbed potential energy surface ($\lambda = 0 \rightarrow \lambda = 1$) in a time τ_1;
- The system is propagated with the perturbed Hamiltonian for some time to enhance barrier crossings ($\lambda = 1$) for a time $\tau - 2\tau_1$;

– The system is propagated via a time-dependent Hamiltonian $H(t)$ evolved back toward the unperturbed Hamiltonian ($\lambda = 1 \rightarrow \lambda = 0$) in a time τ_1;

– The momenta are flipped again if the switch involved a backward propagation.

(*iii*) The resulting configuration is accepted or rejected according to the Metropolis-Hasting criterion: If the move is accepted, we repeat the cycle; if the move is rejected we return to the conformation at the end of step i.

The performance of the hybrid neMD-MC propagator depends on our ability to generate a proposition likely to help overcome the barriers in a rugged potential energy surface.

For this purpose, the parameters controlling the schedule, τ, τ_1, and the choice of the perturbed Hamiltonian, H_1, are critical. Two popular perturbation schemes designed to enhance sampling by deforming the potential energy surface were considered here for the boosting phase. The first one is the accelerated MD (aMD),[15,16]

$$U_{\mathrm{aMD}}[\boldsymbol{x}, \alpha, E] = \begin{cases} U[\boldsymbol{x}], & U[\boldsymbol{x}] \geq E \\ \\ U[\boldsymbol{x}] + \frac{(E - U[\boldsymbol{x}])^2}{\alpha + (E - U[\boldsymbol{x}])}, & U[\boldsymbol{x}] < E \end{cases} \tag{2.6}$$

where α is a tuning parameter that determines the depth of the modified potential energy basins lying below the minimum threshold energy E (the aMD potential is flat when $U[\boldsymbol{x}]$ is equal or smaller than E if α is zero, and increasingly unperturbed when α becomes very large).[15,16] The aMD prescription may be applied to the total potential energy function, or to various contributions such as the torsional potentials. In the present study,

we have used Eq. (2.6) to reduce the energy barriers in the torsional potentials. To construct a time-dependent Hamiltonian for the boosting phase of the hybrid-aMD propagator, the threshold energy E was kept fixed while the tuning parameter α in Eq. (2.6) was effectively replaced by $\alpha/\lambda(t)$, with $\lambda(t)$ following the symmetric schedule given in Eq. (2.3).

The singularity at the first step is avoided by setting E equal to zero until $\lambda(t)$ reaches its first non-zero value at time $t \geq \tau_1/n_\mathrm{s}$ following Eq. (2.4).

The second perturbation scheme considered here is the replica-exchange with solute tempering (REST2),[22]

$$U_{\mathrm{REST2}}[\boldsymbol{x}, \gamma] = U_\mathrm{v} + \gamma U_\mathrm{u} + \gamma^{0.5} U_\mathrm{uv}, \tag{2.7}$$

where U_v is the solvent potential energy, U_u is the solute potential energy, and U_uv is the solute-solvent interaction potential energy.

It should be noted that we use the acronym REST2 here to indicate solute interaction tempering, even in the absence of multiple replicas.

Let the coupling parameter γ_max correspond to the maximum solute tempering allowed, traditionally expressed as T/T_u in terms of an effective temperature T_u ascribed to the solute.[22]

To construct a time-dependent Hamiltonian for the boosting phase of the hybrid-REST2 propagator, the coupling parameter γ in Eq. (2.7) was effectively replaced by the time-dependent form $\gamma(t) = \lambda(t)(\gamma_{max} - 1) + 1$, with $\lambda(t)$ following the symmetric schedule given in Eq. (2.3).

2.2.2 Computational Methods

The hybrid neMD-MC propagator was tested on several biologically relevant molecular systems. The first objective of these tests is to show that the propagator does produce the correct Boltzmann equilibrium distribution. The second objective of these tests is to show that the propagator achieves convergence of the equilibrium properties more rapidly than simple unbiased MD. The hybrid neMD-MC propagator was implemented as a general Tcl script for the simulation program NAMD.[47] The Tcl script is given in the Appendix.

The first test case is deca-alanine in vacuum with a dielectric constant ϵ=20.

Eight individual 200-ns trajectories were generated using (i) equilibrium MD, (ii) two

hybrid propagators using either REST2 or aMD as a boost, (iii) two hybrid propagators randomly accepting proposed moves by means of mean-acceptance ratio from the corresponding propagators, and (iv) two accelerating schemes, namely REST2 and aMD.

The equations of motion were integrated with a time-step of 1 fs. The temperature was maintained at 300 K using Langevin dynamics with a damping coefficient of 1.0 ps^{-1}. Nonbonded short range interactions were truncated at 9 Å with a switching function effective from 8 Å. The parameters of the switching schedule were set to $t_{\mathrm{eqMD}} = 5$ ps, $\tau = 5$ ps, $\tau_1 = 2$ ps, number of switches $n_{\mathrm{s}} = 39$, maximum REST2 boost $\gamma = 0.8$, and maximum aMD boost $(E, \alpha) = (50, 10)$.

The CHARMM 22 protein force field[48] was employed to model deca-alanine, allowing a direct comparison with a previous theoretical study.[34]

A second test case considers again deca-alanine in vacuum, but this time with a dielectric constant, ϵ, of 1.

This system is normally difficult to sample using simple unbiased equilibrium MD trajectories because of the strong internal backbone-backbone hydrogen bonds, making the usage of some form of importance sampling strategy a necessity. Here, the adaptive biasing force (ABF) method was considered.[49,50] Results from ABF with standard MD and from ABF with hybrid neMD-MC boosted by aMD and REST2 were examined. Eight individual 40-ns trajectories were generated for each case, using α-helix and C_5-extended conformation as initial coordinates. After optimization, the switching schedule was set to $t_{\mathrm{eqMD}} = 7$ ps, $\tau = 2$ ps, $\tau_1 = 1$ ps, number of switches $n_{\mathrm{s}} = 18$, maximum REST2 boost $\gamma = 0.5$, and maximum aMD boost $(E, \alpha) = (40, 1)$.

Extensive computations based on the multiple-walker ABF method[49,50] were performed to provide a reference potential of mean force (PMF) that can be used to assess the converge of the different methods.

The computation was carried out with eight independent walkers, corresponding to an aggregated simulation time of 8 μs.

A third test case considers the folding/unfolding of the acetyl- and -NH_2 terminally capped $(AAQAA)_3$ peptide solvated in explicit water.

The 15 residue peptide was simulated using the hybrid-REST2 neMD-MC propagator. As a basis of comparison, the system was also simulated with simple unbiased MD, as well as multiple-copy REST2 and temperature replica-exchange MD simulations (TREMD).

For the optimization, different switching schedules were tested while varying τ, τ_1, the maximum boost (REST2 or aMD), and number of switches.

After optimization, the parameters of the switching schedule were set to $t_{\mathrm{eqMD}} = 2$ ps, $\tau = 24$ ps, $\tau_1 = 8$ ps, number of switches $n_{\mathrm{s}} = 790$, and maximum REST2 boost $\gamma = 0.7$. The system was equilibrated at 300 K with initial structure of complete α-helical conformation solvated with 3964 water molecules in a cubic cell of initial dimensions equal to $55 \times 55 \times 55$ $\mathrm{Å}^3$. The equations of motion were integrated with a time-step of 2 fs. The temperature was maintained at 300 K using Langevin dynamics with a damping coefficient of 1 ps^{-1}. Nonbonded short-range interactions were truncated at 12 Å with a switching function effective from 10 Å. The CHARMM 36 force field was used to describe the peptide, with the TIP3P water model.

The REST2 calculation was done with 10 replicas ranging up to the same maximum boost of $\gamma = 0.7$ as in the hybrid neMD-MC simulations and exchange attempts every 0.2 ps. The TREMD simulation was done with 32 replicas ranging from 278 K to 416 K and exchange attempts every 10 ps, matching exactly the protocol of Best et al.[51]. However, the total length of the present TREMD simulation is much shorter (1.875 ns per replica and 60 ns in total) than that of Best et al.[51] (150 ns per replica and 4.8 μs in total). To provide an objective comparison of the actual computational cost of these single- and multiple-copy approaches, the results are reported in terms of MD steps per replica times the number of replica involved.

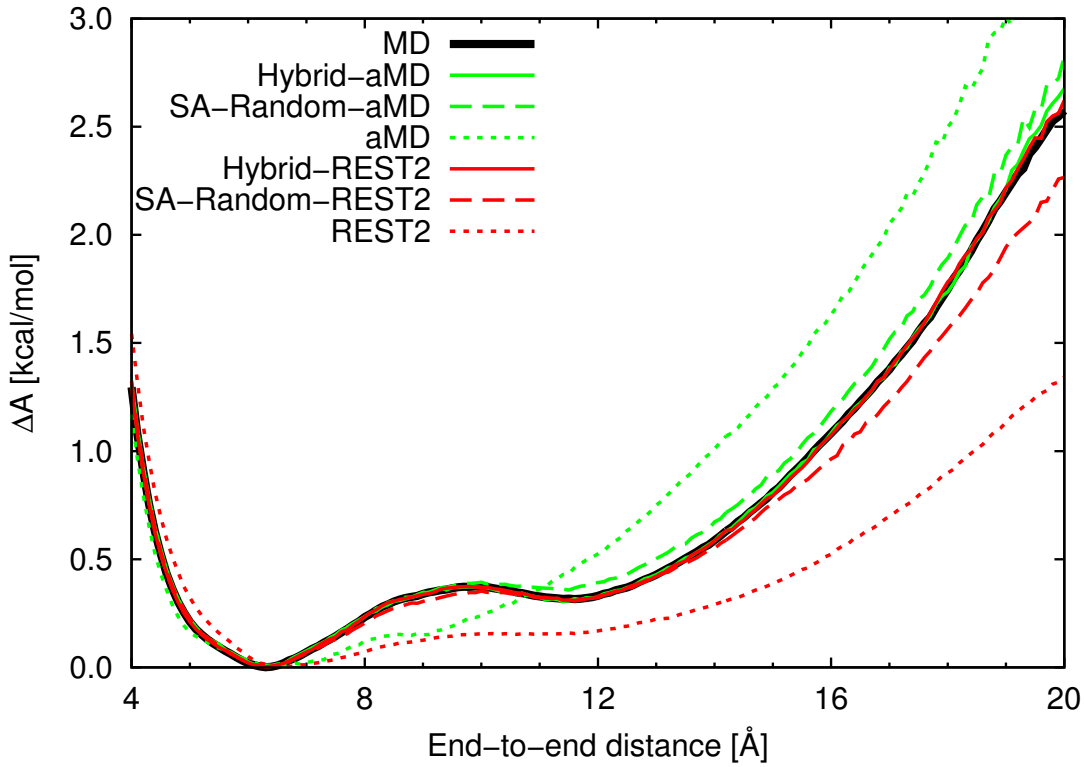

Figure 2.2: Validation test using deca-alanine with dielectric constant of 20. SA-Random stands for hybrid scheme that randomly accepts proposed moves using same mean acceptance ratio achieved from hybrid simulations.

2.3 Results and Discussion

2.3.1 Validation

We first examine the results for deca-alanine in vacuum with a dielectric constant ϵ=20. The main purpose of this test is to validate the hybrid neMD-MC propagator. Because of the high dielectric constant, this is a convenient toy model, easy to sample directly by brute-force MD. The PMF as a function of the end-to-end distance extracted from unbiased MD is shown in Figure 2.2. It is a double-well potential, with basins at 6 Å and 12 Å, separated by a small barrier of about 0.5 kcal/mol. The PMF calculated using the hybrid propagator is also shown in Figure 2.2. The comparison shows that the hybrid method correctly reproduces the PMF extracted from the unbiased MD trajectory. The small error bars corresponding to a 99% confidence interval are indicative of the high convergence. As expected, the PMFs generated

by pure REST2 and aMD without any reweighting do not reproduce the correct result. They biased PMFs nonetheless serve to display the overall free energy surface when the system is propagated at the highest boosting level using these two perturbation schemes designed to enhance sampling by deforming the potential energy surface. Lastly, the random acceptance simulations with the same mean acceptance ratio from the hybrid-aMD (0.1522±0.002) and hybrid-REST2 simulations (0.716±0.003) deviate from the correct PMF. This shows that the acceptance criterion with the symmetric two-ends momentum reversal is critical to obey detailed balance and obtain correct results.

2.3.2 Optimization

Generally, the efficiency of neMD-MC depends on maximizing the production of uncorrelated configurations while trying to minimize the effort spent to generate these.[32] The main drawback of neMD-MC simulations is that each new attempted MC move requires a non-equilibrium MD simulation; a low acceptance rate necessarily implies that a large fraction of computer time is discarded by the algorithm. Whereas long switches might yield a high acceptance probability, but are computationally prohibitive, short switches are computationally inexpensive, but expected to yield vanishingly low acceptance probabilities. The most efficient algorithm is obviously a compromise balancing between these two opposing factors. It is generally unclear as to how to systematically achieve this balance.

In conventional neMD-MC schemes the switch is meant to allow transitions between two different discrete states, thus, longer switch times lead to higher acceptance rates as the transformation approaches the adiabatic limit.[32] In a previous study of a constant-pH algorithm, an optimal switching time for changing the protonation state of propionic acid in explicit solvent was determined to be about 15 ps.[32] However, the hybrid propagator described here is not designed to produce transitions between discrete states. One implication is that a high acceptance probability, i.e., the number of accepted moves over the total number of attempted moves, does not guarantee that one is using the most effective

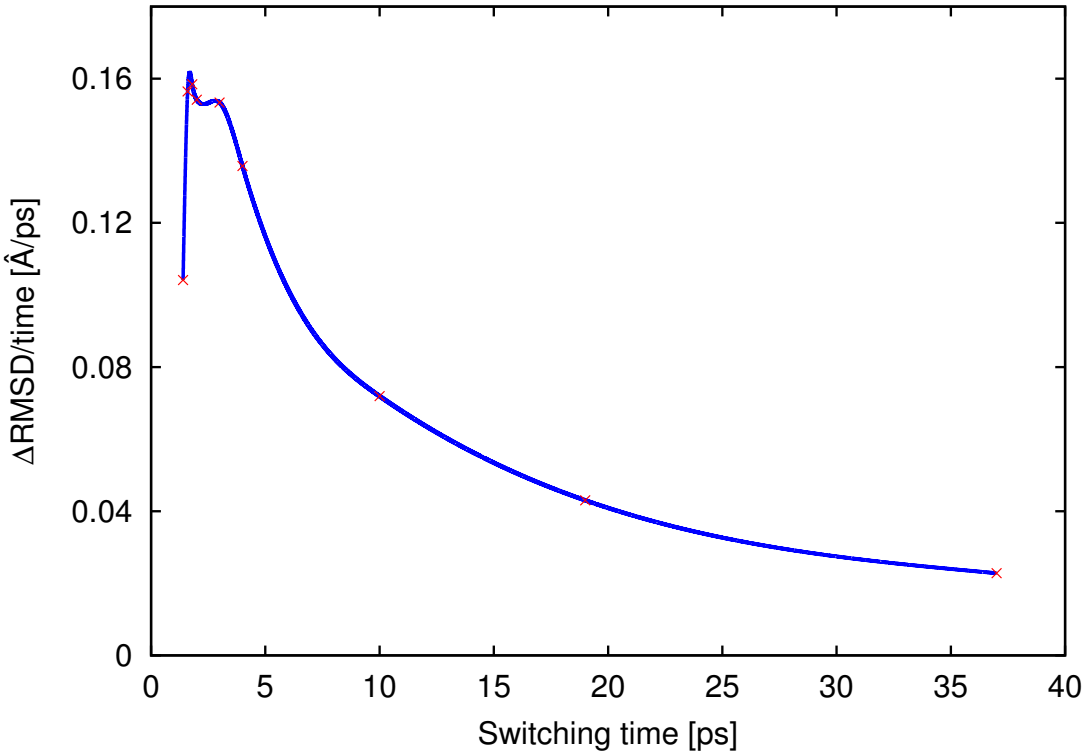

Figure 2.3: Conformational evolution speed for deca-alanine in vacuum with a dielectric constant of 5 as a function of the boosting schedule. The conformational evolution speed is root-mean-square deviation (RMSD) of the peptide (calculated from all heavy atoms) at a time t relative to its configuration at a previous time $t - \Delta T$. For each case, the time interval ΔT was set to its switching time. The blue line is a fitted function for visual guide.

perturbation scheme and boosting schedule. For example, the accepted candidate configuration may be too close to the starting configuration if the perturbation is too mild. Within the constraint of fixed budget of computer time, increasing indefinitely the switching time also implies a concomitant decrease in the time available to perform equilibrium MD. Under these premises, simply maximizing the acceptance probability is not a useful criterion for optimizing efficiency.

Several factors could affect the efficiency of the hybrid neMD-MC scheme described here. As a simple measure of efficiency of the hybrid neMD-MC propagation, we define a "conformational evolution speed" given by the net root-mean-square deviation (RMSD) of the peptide at a time t relative to the peptide configuration at time $t - \Delta T$. Monitoring the conformational evolution speed, while including the total cost to generate the non-equilibrium

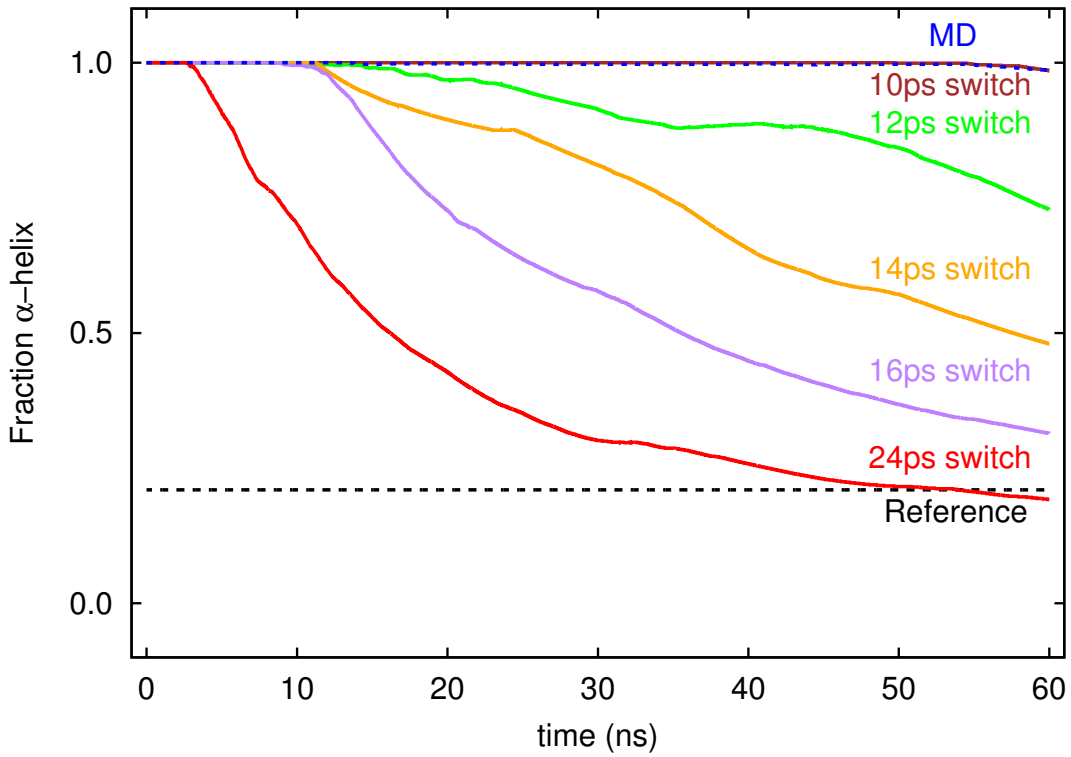

Figure 2.4: Optimization test for the $(AAQAA)_3$ system in explicit solvent by monitoring time cumulative average for the α-helicity. 4 independent trajectories were simulated for all cases, and 4 additional independent simulations were operated for MD and 24 ps switch. The peptide is defined to be in the α-helical conformation when the (ϕ, ψ) backbone dihedral angles of three consecutive residues fulfill the condition: $|-65^\circ - \phi| < 35^\circ$ and $|-37^\circ - \psi| < 30^\circ$.

switches, allows us to assess the net efficiency of the hybrid neMD-MC propagator in terms of the perturbation scheme and boosting schedule. As illustrated in Figure 2.3, it is indeed possible to identify a maximum efficiency in the case of the deca-alanine system. Based on this metric, a switching time of about 1.8 ps yields the most efficient hybrid neMD-MC propagator for the deca-alanine system.

For the $(AAQAA)_3$ system in explicit solvent, we monitored the fraction of secondary-structure elements in the peptide chain as a function of switching time. As illustrated in Figure 2.4, the efficiency is fairly poor when the switching time is too short. The simulation started with the peptide in an α-helical conformation, which persisted for an extended period of time. In contrast, the efficiency is considerably improved by employing a switching time of 24 ps. While it is possible that this could lead to further improvement, the behavior of the

system was not examined for switching times longer than 24 ps. In practice, the efficiency of the $(AAQAA)_3$ system in explicit solvent is already very good with 24 ps according to Figure 2.4. As shown in Table 2.1, the acceptance probability P_a is about 0.92 with a switching time of 24 ps. This is approaching the limit afforded by the shadow work, which would be present with normal equilibrium MD propagation.

2.3.3 Performance

Deca-alanine is an extremely well-characterized benchmark system for testing the ability of new sampling schemes to capture subtle conformational equilibria.[34,52] Its global free-energy minimum corresponds to a fully α-helical conformation resulting from the formation of robust $i \rightarrow i+4$ intra-molecular hydrogen bonds. When generating a PMF along its end-to-end distance, the metastable states that exist along the orthogonal degrees of freedom in the "rugged" region between 4 and 12 Å can dramatically slow the convergence of conformational sampling. Furthermore, there is a very large free energy difference of about 30 kcal/mol between the completely folded form (14 Å) and the family of unfolded structures (32 Å). As a result, biased simulations must be used to overcome the considerable imbalance in the equilibrium probability along the end-to-end distance. Here, we use the adaptive biasing force (ABF) method.[49,50] It should be emphasized that the necessity to adopt some form of

Table 2.1: Acceptance probability for the $(AAQAA)_3$ peptide[1]

τ (ps)	τ_1 (ps)	P_a
10	1	0.42
12	2	0.60
14	3	0.72
16	4	0.79
24	8	0.92

[1] τ is the total switching time, τ_1 is the ramping time, and the boosting phase lasts $\tau - 2\tau_1$; 4 neMD-MC simulations were carried out for a switching time of 10, 12, 14 and 16 ps, and 8 neMD-MC simulations were carried out for a switching time of 24 ps.

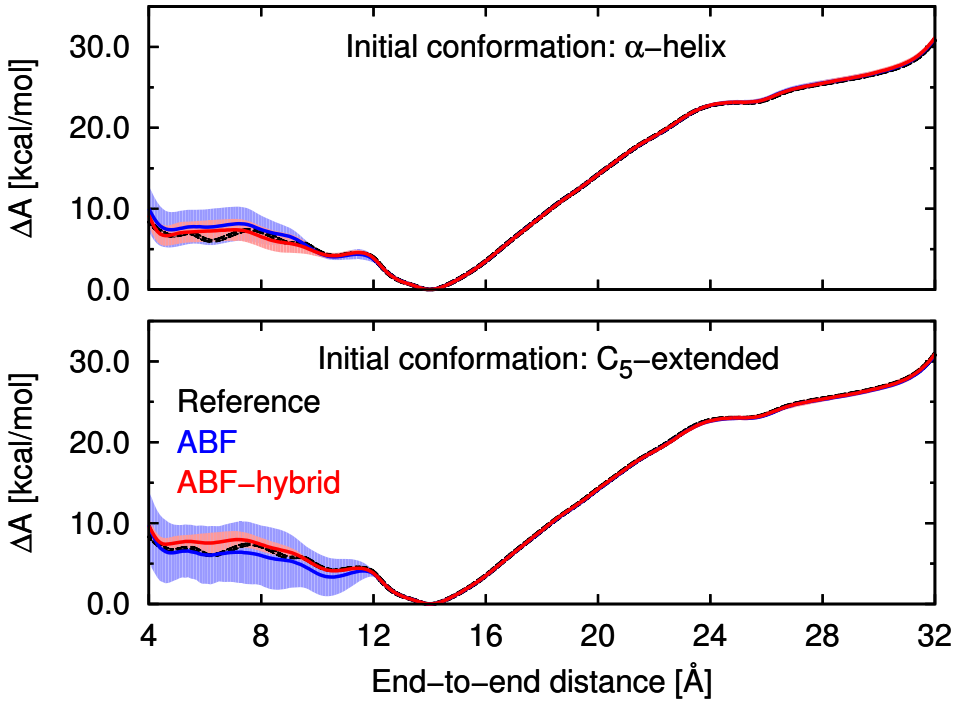

Figure 2.5: PMF of the end-to-end distance of tested on deca-alanine in vacuum computed using ABF and ABF with hybrid-(REST2, aMD), Results using α-helix (left) and C_5-extended (right) conformation as starting structure are shown. 8 independent trajectories of 40 ns were simulated using ABF and ABF-hybrid.

importance sampling method to explore the end-to-end distance of the deca-alanine system remains true, whether a hybrid neMD-MC propagator is used or not.

The main results are displayed in Figure 2.5. Comparison of the various methods reveals that the hybrid propagator with ABF clearly converged faster toward the reference PMF. While the standard deviations of the PMF based on eight independent simulations, either MD or the hybrid, are comparable in regions of the reaction pathway that are easily sampled (>12 Å), those for the hybrid method are certainly lower in the rugged region (<12 Å). This is indicative that the hybrid method facilitates the sampling of the numerous metastable states lying along orthogonal degrees of freedom in this region. Interestingly, when simulating multiple independent trajectories with pure ABF, we observed that some trajectories occasionally remained "trapped" in metastable regions for an extended fraction of the simulation time. This occurred for ABF with both starting conformations, leading

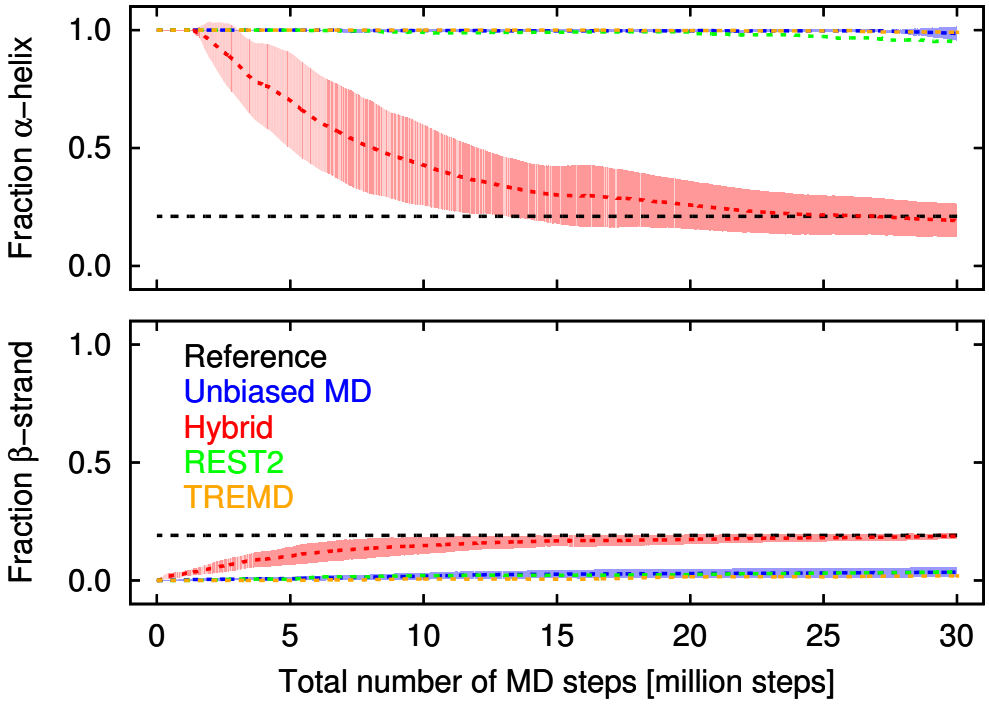

Figure 2.6: Cumulative time-average for the secondary structure elements of a $(AAQAA)_3$ peptide in explicit solvent using unbiased MD (single-copy), hybrid-REST2 neMD-MC (single-copy), REST2 with 10 replicas, and TREMD with 32 replicas. 8 independent trajectories of 60 ns were simulated using unbiased MD and the hybrid neMD-MC propagator, 6 ns each of 10 replicas were simulated using REST2 and 1.875 ns each of 32 replicas were simulated using TREMD. The total number of MD steps on the x-axis represents the actual computational costs of each approach as the number of steps per replica times the number of replica. The peptide is defined to be in the α-helical conformation when the (ϕ, ψ) backbone dihedral angles of three consecutive residues fulfill the condition: $|-65^\circ - \phi| < 35^\circ$ and $|-37^\circ - \psi| < 30^\circ$. The peptide is in the β-strand conformation when (ϕ, ψ) fulfill the condition: $|-140^\circ - \phi| < 40^\circ$ and $|150^\circ - \psi| < 30^\circ$.

to an over-stabilization of the metastable states and resulting in slow convergence. If one consider these trapped simulations as outliers and discard them, then ABF-MD performs overall better than ABF-hybrid. Importantly, this highly undesirable trapping phenomenon was not observed for the ABF-hybrid simulations. In assessing the overall efficiency of the ABF-hybrid simulations, it is important to note that no data is collected about the gradient during the neMD switching trajectories. Thus, the total amount of data accumulated to estimate the mean gradient is necessarily smaller than with ABF-MD. The hybrid scheme with ABF having only 70% of equivalent MD data points, yet performing better, underscores its power.

As a final illustrative test, we consider the folding/unfolding of the 15 residue peptide $(AAQAA)_3$ in explicit solvent. This system has been the object of extensive simulation studies and is very well characterized.[31,51] To highlight the conformational sampling challenge, the simulations were all started in the long-lived α-helix metastable state. Let us first compare the results from simple brute-force equilibrium MD and a hybrid neMD-MC propagator with REST2 boosting. Figure 2.6 shows the efficiency of the hybrid-REST2 propagator through time-cumulative averages of secondary structures (α-helix and β-strand). It is observed that the simulation converges towards the reference equilibrium average much faster with the hybrid propagator. This trend is systematic, as indicated by the 99% confidence interval estimated from the 8 independent trajectories. Rapid conformational fluctuations of the $(AAQAA)_3$ peptide are visibly occurring with the hybrid-REST2 propagator, as shown by the time-series of the individual simulations (Supplementary Figures S1-S4). Conversely, the systems simulated via simple brute-force equilibrium MD remain essentially trapped in the starting α-helical conformation.

While the brute-force equilibrium MD generated 13 times more data points than the hybrid propagator, the conformational space of the $(AAQAA)_3$ peptide was explored much more efficiently with the hybrid propagator. Remarkably, REST2 (10 replicas) and TREMD (32 replicas) result in a weaker performance than the single-copy hybrid-REST2 neMD-MC propagator. The conformational sampling of $(AAQAA)_3$ in solution remains inefficient with these multiple-copy approaches, despite the fact that there were frequent exchanges between the replicas (Supplementary Figures S5 and S6). The difference in performance is particularly glaring when comparing in terms of total computational cost of all approaches (Figure 2.6).

It is of interest to compare these results with the ABF calculations for the deca-alanine toy model discussed above.

In the case of the ABF-hybrid simulations, nearly 70% of the data produced correspond to equilibrium sampling and were accumulated to estimate the mean gradient.

Increasing the amount of non-equilibrium switches is not advantageous because it reduces the amount of data used to estimate the mean gradient needed in ABF. This points to the limitation associated with the inherent loss of data during the non-equilibrium switches in the context of ABF. When the data collected from simple brute-force equilibrium MD are sufficiently uncorrelated, as in the case of deca-alanine, then increasing the proportion of the sampling carried out via the hybrid propagator becomes statistically less efficient. Even if the hybrid propagator actually moves the system around conformational space rapidly, enough data points must still be accumulated to accurately determine the average gradient that is needed for the ABF method. This stands in contrast with the $(AAQAA)_3$ simulation, for which the fraction of helicity could be determined accurately even though only 7.7% of the data produced corresponds to equilibrium sampling. The hybrid propagator can efficiently explore the conformational space accessible to the peptide in solution and the result without relying on accumulated equilibrium data as in the case of ABF. Even though the brute-force equilibrium MD simulations include about 13 times more data points than the hybrid simulation, it is extremely inefficient to explore the conformational space accessible to the $(AAQAA)_3$ peptide in solution.

2.4 Conclusion

Powerful hybrid neMD-MC algorithms can be designed to boost and accelerate the conformational sampling of complex molecular systems. At the heart of the hybrid neMD-MC algorithms is a transient non-equilibrium perturbation of the system, which is aimed at overcoming the barriers in the rugged energy landscapes. Importantly, the neMD-MC algorithms robustly generates the correct Boltzmann equilibrium distribution. A vast range of time-dependent perturbations are possible to construct a family of hybrid neMD-MC propagators adapted to different situations. In the present test, we boosted the conformational sampling by relying on Hamiltonian perturbations based on aMD and REST2, but the strategy allows virtually any accelerated methods to be introduced during the neMD steps to efficiently

lower the free-energy barriers. The benchmark systems tested demonstrate the correctness and effectiveness of the hybrid propagator, emphasizing its faster convergence compared to equilibrium MD. Although the switching schedule was optimized heuristically in the present study, it is possible to infer it analytically when the Hamiltonian is perturbed linearly.[32] Depending on the problem at hand, hybrid neMD-MC propagators may advantageously be combined with a number of established strategies and free energy methods,[23] including umbrella sampling (US),[53] adaptive biasing force (ABF),[34] Hamiltonian tempering replica-exchange,[54] and alchemical free energy perturbation.[18] Future work will explore how the present hybrid neMD-MC propagators behave in the case of very complex biological objects and can be possibly tailored to address the rugged free-energy landscape underlying intricate processes therein.

CHAPTER 3

STRING METHOD FOR PROTEIN-PROTEIN BINDING FREE ENERGY CALCULATION[1]

3.1 Introduction

An approach based on the potential of mean force (PMF) for the reversible spatial separation of two binding macromolecules is the method of choice to quantitatively characterize the affinity of protein complexes.[9,35,36] Nonetheless, multiple challenges remain to render the current methodology reliable and computationally efficient in practice. In particular, the PMF calculation for the separation of two proteins along a predefined rectilinear path may be suboptimal and slowly convergent. Here, we put forth a novel theoretical framework for binding free energy calculations, leaning on the optimal curvilinear minimum free-energy path determined from the string method.[6–8,44] The proposed formalism is validated by comparing the results obtained using both rectilinear and curvilinear pathways for a prototypical host–guest complex formed by cucurbit[7]uril (CB[7]) binding benzene, and for the barnase–barstar protein complex. We find that the calculations following the traditional rectilinear pathway and the string-based curvilinear pathway agree quantitatively, but convergence is faster with the latter.

1. Adapted from Suh, Jo, Jiang, Chipot, and Roux, *J. Chem. Theory. Comput.* submitted, 2019.

3.2 Theory

3.2.1 String method formulation

The equilibrium binding constant K_{eq} between two proteins can be expressed as,[9,35,36,55,56]

$$K_{\mathrm{eq}} = \frac{\displaystyle\int_{\mathrm{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta U}}{\displaystyle\int_{\mathrm{unbound}} \mathrm{d}\mathbf{X}\ \delta(\mathbf{r}_{\mathrm{com}} - \mathbf{r}^*)\ \mathrm{e}^{-\beta U}} \tag{3.1}$$

where U is the potential energy function of the full system, the vector $\mathbf{r}_{\mathrm{com}} \equiv (r, \theta, \phi)$ represents the relative position of the center of mass of the two proteins, and the vector $\mathbf{r}^* \equiv (r^*, \theta^*, \phi^*)$ sets a reference corresponding to separated, noninteracting proteins. $\beta \equiv k_{\mathrm{B}}T$, where k_{B} is the Boltzmann constant, and T is the temperature. The denominator and the numerator of eq 3.1 represent, respectively, the initial and final states of the overall dissociation process, namely the protein–protein complex, and the two noninteracting proteins separated in the bulk solution. Although this expression for the equilibrium binding constant is rigorous, it cannot be easily evaluated using MD simulations with explicit solvent as it stands.[9] Our strategy consists in introducing a series of intermediate states between the initial (bound) and final (unbound) states, such that each contribution of the total free-energy difference can be calculated more easily in stages. To improve the convergence of the calculation, we also typically introduce a conformational restraint, u_{c}, based on the distance root-mean-square deviation (RMSD) acting on various parts of the protein structures — typically, backbone and side chains,[36] as explained below. Accordingly, the expression for

the binding constant can be rewritten as,

$$
\begin{aligned}
K_{\mathrm{eq}} &= \frac{\displaystyle\int_{\mathrm{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta U}}{\displaystyle\int_{\mathrm{unbound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}]}} \\
&\quad \times \frac{\displaystyle\int_{\mathrm{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}]}}{\displaystyle\int_{\mathrm{unbound}} \mathrm{d}\mathbf{X}\ \delta(\mathbf{r}_{\mathrm{com}}-\mathbf{r}^{*})\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}]}} \\
&\quad \times \frac{\displaystyle\int_{\mathrm{bound}} \mathrm{d}\mathbf{X}\ \delta(\mathbf{r}_{\mathrm{com}}-\mathbf{r}^{*})\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}]}}{\displaystyle\int_{\mathrm{unbound}} \mathrm{d}\mathbf{X}\ \delta(\mathbf{r}_{\mathrm{com}}-\mathbf{r}^{*})\ \mathrm{e}^{-\beta U}} \\
&= \mathrm{e}^{\beta\Delta G_{\mathrm{c}}^{\mathrm{bound}}} \times \left(\frac{\displaystyle\int_{\mathrm{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}]}}{\displaystyle\int_{\mathrm{unbound}} \mathrm{d}\mathbf{X}\ \delta(\mathbf{r}_{\mathrm{com}}-\mathbf{r}^{*})\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}]}} \right) \times \mathrm{e}^{-\beta\Delta G_{\mathrm{c}}^{\mathrm{unbound}}} \qquad (3.2)
\end{aligned}
$$

Formally, the middle term of eq 3.2 is the equilibrium binding constant for a conformationally restrained system with total potential energy function $U' = [U + u_{\mathrm{c}}]$. For the sake of simplicity, we will refer to this quantity as K'_{eq}. The terms on the left and on the right are the free-energy contributions to introduce the conformational restraint in the bound and in the unbound states, namely,

$$
e^{-\beta\Delta G_{\mathrm{c}}^{\mathrm{bound}}} = \frac{\displaystyle\int_{\mathrm{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}]}}{\displaystyle\int_{\mathrm{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta U}} \qquad (3.3)
$$

and,

$$
e^{-\beta\Delta G_{\mathrm{c}}^{\mathrm{unbound}}} = \frac{\displaystyle\int_{\mathrm{unbound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}]}}{\displaystyle\int_{\mathrm{unbound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta U}} \qquad (3.4)
$$

We now turn to the protein-protein separation process. Its pathway is handled in a subspace of six collective variables (CVs), $\mathbf{Z} = [\mathbf{r}, \mathbf{\Omega}]$, where $\mathbf{r} = (r, \theta, \phi)$ is the relative protein–protein center-of-mass position, and $\mathbf{\Omega} = (\Theta, \Phi, \Psi)$ is the relative protein–protein

orientation defined in terms of the three Euler angles. The associated restraining potential writes,

$$u(\mathbf{Z}) = u_{\mathrm{t}}(\mathbf{r}) + u_{\mathrm{o}}(\mathbf{\Omega}) \tag{3.5}$$

The internal coordinate system of the protein-protein complex is established using six reference points, three in the first protein (P_1, P_2, and P_3), and three in the second protein (L_1, L_2, and L_3). Using these reference points, the six relative translational and orientational degrees of freedom can be defined using spherical coordinates (r, θ, ϕ) and Euler angles (Θ, Φ, Ψ), where r is the $\overline{P_1L_1}$ distance, θ is the $\widehat{P_1L_1L_2}$ angle, ϕ is the $\angle(P_1L_1L_2L_3)$ dihedral angle, Θ is the $\widehat{P_2P_1L_1}$ angle, Φ is the $\angle(P_2P_1L_1L_2)$ dihedral angle, and Ψ is the $\angle(P_3P_2P_1L_1)$ dihedral angle. In the formulation based on the string method,[8,59] the separation pathway is represented by a chain of M states, or "images", in a subspace of CV, i.e., the path is $\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots \mathbf{Z}^{(M)}\}$. The last, M^{th} image corresponds to the unbound state, wherein $\mathbf{Z}^{(M)} \equiv [\mathbf{r}^*, \mathbf{\Omega}^*]$, with the given orientation $\mathbf{\Omega}^* = (\Theta^*, \Phi^*, \Psi^*)$, which will be further discussed below. The i^{th} image is simulated with the harmonic restraining potential,

$$u^{(i)}(\mathbf{Z}) = \frac{1}{2}\, \mathbf{k} \cdot \left(\mathbf{Z} - \mathbf{Z}^{(i)}\right)^2 \tag{3.6}$$

where $\mathbf{k} = (k_r, k_\theta, k_\phi, k_\theta, k_\Phi, k_\Psi)$, is the force constant, and $\mathbf{Z}^{(i)}$ denotes the set of reference values for the harmonic restraining potentials corresponding to the i^{th} image. The reference value for this image, $\mathbf{Z}^{(i)}$, is determined from the string method algorithm by moving in the direction of the mean force in the collective-variable space, and then reparametrizing the string assuming a Euclidian metric.[8,59] Using this definition, the equilibrium binding

constant $K'_{\rm eq}$ is written as,

$$\begin{aligned} K'_{\rm eq} &= \frac{\displaystyle\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta U'}}{\displaystyle\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U'+u^{(1)}]}} \\ &\quad \times \left(\frac{\displaystyle\int {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U'+u^{(1)}]}}{\displaystyle\int {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U'+u^{(2)}]}} \cdots \frac{\displaystyle\int {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U'+u^{(M-1)}]}}{\displaystyle\int {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U'+u^{(M)}]}} \right) \\ &\quad \times \frac{\displaystyle\int_{\rm unbound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U'+u^{(M)}]}}{\displaystyle\int_{\rm unbound} {\rm d}\mathbf{X}\ \delta(\mathbf{r}_{\rm com}-\mathbf{r}^*)\ {\rm e}^{-\beta U'}} \\ &= {\rm e}^{\beta\Delta G^{(1)}_{\rm bound}} \times \left(\prod_{i=1}^{M-1} {\rm e}^{-\beta\Delta G^{(i,i+1)}} \right) \times {\rm e}^{-\beta\Delta G^{(M)}_{\rm unbound}} \end{aligned} \tag{3.7}$$

The middle terms between the parentheses, which correspond to the progressive spatial separation of the complex from the bound to the unbound states, comprise simple relative free energies of the type,

$$e^{-\beta\Delta G^{(i,i+1)}} = \frac{\displaystyle\int {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U'+u^{(i)}]}}{\displaystyle\int {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U'+u^{(i+1)}]}} \tag{3.8}$$

These terms represent the free energy along the optimal string pathway for substituting the restraining potential of the $i^{\rm th}$ image on the CVs from the restraining potential of the $(i+1)^{\rm th}$ image. Such free-energy differences can be evaluated with confidence using the histogram-less weighted histogram analysis method (WHAM),[60] or Bennett acceptance ratio (BAR),[61] assuming there is sufficient overlap between two contiguous images, thereby obviating the need for an explicit separation PMF.

The free-energy difference $\Delta G^{(1)}_{\text{bound}}$ associated with restraining potential $u^{(1)}$, namely,

$$e^{\beta G^{(1)}_{\text{bound}}} = \frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta U'}}{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U'+u^{(1)}]}} \tag{3.9}$$

is calculated by staging the orientational and translational restraints,

$$\begin{aligned} e^{\beta \Delta G^{(1)}_{\text{bound}}} &= \frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U']}}{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U'+u^{(1)}_{\text{t}}]}} \times \frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U'+u^{(1)}_{\text{t}}]}}{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U'+u^{(1)}_{\text{t}}+u^{(1)}_{\text{o}}]}} \\ &= \mathrm{e}^{\beta \Delta G^{(1)}_{\text{bound,t}}} \times \mathrm{e}^{\beta \Delta G^{(1)}_{\text{bound,o}}} \end{aligned} \tag{3.10}$$

where,

$$e^{-\beta \Delta G^{(1)}_{\text{bound,t}}} = \frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U'+u^{(1)}_{\text{t}}]}}{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta U'}} \tag{3.11}$$

and

$$e^{-\beta \Delta G^{(1)}_{\text{bound,o}}} = \frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U'+u^{(1)}_{\text{t}}+u^{(1)}_{\text{o}}]}}{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U'+u^{(1)}_{\text{t}}]}} \tag{3.12}$$

The free-energy contribution corresponding to the release of the restraint for the last, M^{th} image, wherein the two proteins are separated in solution, is,

$$\begin{aligned} e^{-\beta \Delta G^{(M)}_{\text{unbound}}} &= \frac{\displaystyle\int_{\text{unbound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U'+u^{(M)}]}}{\displaystyle\int_{\text{unbound}} \mathrm{d}\mathbf{X}\ \delta(\mathbf{r}_{\text{com}} - \mathbf{r}^*)\ \mathrm{e}^{-\beta U'}} \\ &= F_{\text{t}} \times \mathrm{e}^{-\beta \Delta G^{(M)}_{\text{unbound,o}}} \end{aligned} \tag{3.13}$$

where,

$$
\begin{aligned}
e^{-\beta\Delta G^{(M)}_{\text{unbound,o}}} &= \frac{\displaystyle\int_{\text{unbound}} \mathrm{d}\mathbf{X}\ \delta(\mathbf{r}_{\text{com}} - \mathbf{r}^*)\ \mathrm{e}^{-\beta[U' + u_{\text{t}}^{(M)} + u_{\text{o}}^{(M)}]}}{\displaystyle\int_{\text{unbound}} \mathrm{d}\mathbf{X}\ \delta(\mathbf{r}_{\text{com}} - \mathbf{r}^*)\ \mathrm{e}^{-\beta U' + u_{\text{t}}^{(M)}}} \\
&= \frac{\displaystyle\int \mathrm{d}\mathbf{\Omega}\ \mathrm{e}^{-\beta u_{\text{o}}^{(M)}(\mathbf{\Omega})}}{\displaystyle\int \mathrm{d}\mathbf{\Omega}}
\end{aligned}
$$

$$
= \frac{1}{4\pi^2} \int_0^{\pi} \mathrm{d}\Theta \sin\Theta \int_0^{2\pi} \mathrm{d}\Phi \int_0^{2\pi} \mathrm{d}\Psi\ \mathrm{e}^{-\beta[k_\Theta(\Theta-\Theta^*)^2 + k_\Phi(\Phi-\Phi^*)^2 + k_\Psi(\Psi-\Psi^*)^2]} \quad (3.14)
$$

and,

$$
\begin{aligned}
F_{\text{t}} &= \frac{\displaystyle\int_{\text{unbound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U' + u_{\text{t}}^{(M)}]}}{\displaystyle\int_{\text{unbound}} \mathrm{d}\mathbf{X}\ \delta(\mathbf{r}_{\text{com}} - \mathbf{r}^*)\ \mathrm{e}^{-\beta U'}} \\
&= \int d\mathbf{r}\ \mathrm{e}^{-\beta u_{\text{t}}^{(M)}(\mathbf{r})} \\
&= \int_0^{\infty} \mathrm{d}r\ r^2 \int_0^{\pi} \mathrm{d}\theta \sin\theta \int_0^{2\pi} \mathrm{d}\phi\ \mathrm{e}^{-\beta[k_r(r-r^*)^2 + k_\phi(\phi-\phi^*)^2 + k_\theta(\theta-\theta^*)^2]} \qquad (3.15)
\end{aligned}
$$

with respect to the reference values (r^*, ϕ^*, θ^*) corresponding to the M^{th} image. The translational factor, F_{t}, has dimensions of Å^3, which must be converted to a concentration in mole per liter in order to define the standard binding free energy, i.e., $e^{-\beta G^\circ_{\text{t}}} = F_{\text{t}}\ C^\circ = F_{\text{t}}/1661$. The remaining free-energy contributions from the various distance RMSD-based conformational restraints are staged for the bound and unbound states, and the final result can be expressed as,

$$
\begin{aligned}
K_{\text{eq}}^{\text{string}} &= \mathrm{e}^{-\beta[\Delta G^{(M)}_{\text{unbound,t}} - \Delta G^{(1)}_{\text{bound,t}}]} \times \mathrm{e}^{-\beta[\Delta G^{(M)}_{\text{unbound,o}} - \Delta G^{(1)}_{\text{bound,o}}]} \\
&\quad \times \prod_{i=1}^{M-1} \mathrm{e}^{-\beta \Delta G^{(i,i+1)}} \times \mathrm{e}^{-\beta[\Delta G^{(M)}_{\text{unbound,c}} - \Delta G^{(1)}_{\text{bound,c}}]} \qquad (3.16)
\end{aligned}
$$

3.2.2 PMF-based formulation

In the conventional PMF-based method,[9] the physical separation of the binding partners is treated by calculating the PMF, $w(r)$, along the one-dimensional Euclidian distance r between their centers of mass, restrained along a rectilinear axis "a" with potential $u_{\mathrm{a}}(\theta,\phi) = k_\phi(\phi-\phi^*)^2+k_\theta(\theta-\theta^*)^2$. This rectilinear axial restraint is part of the complete translational restraining potential, u_{t}, which is used in the string pathway determination, i.e., $u_{\mathrm{t}}(\mathbf{r}) = u_r(r) + u_{\mathrm{a}}(\theta,\phi)$ with $u_r(r) = \frac{1}{2}\ k_r(r-r^*)^2$.

It is of interest to relate the expression of K_{eq} from the PMF-based approach with eq 3.16 arising from the string method in the case of a straight path (i.e., when the separation of the two proteins is realized along a pre-defined rectilinear axis). Under these premises, it ought to be noted that the restraining potentials have the following properties, $u_{\mathrm{a}}^{(1)} = \cdots = u_{\mathrm{a}}^{(M)} \equiv u_{\mathrm{a}}$, and $u_{\mathrm{o}}^{(1)} = \cdots = u_{\mathrm{o}}^{(M)} \equiv u_{\mathrm{o}}$, whereas $u_{\mathrm{r}}^{(1)} \neq \cdots \neq u_{\mathrm{r}}^{(M)}$. To better appreciate the similarity and differences between the two formulations, a side-by-side comparison of the formal expressions reads,

$$
\begin{aligned}
K_{\rm eq}^{\rm string} = & \frac{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta U}}{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}]}} \times \frac{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}]}}{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm t}^{(1)}]}} \\
& \times \frac{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm t}^{(1)}]}}{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm t}^{(1)}+u_{\rm o}]}} \\
& \times \frac{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm t}^{(1)}+u_{\rm o}]}}{\int_{\rm unbound} d\mathbf{1}\ \delta(\mathbf{r}_{\rm com}-\mathbf{r}^*) \int {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm o}]}} \\
& \times \frac{\int_{\rm unbound} {\rm d}\mathbf{X}\ \delta(\mathbf{r}_{\rm com}-\mathbf{r}^*)\ {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm o}]}}{\int_{\rm unbound} {\rm d}\mathbf{X}\ \delta(\mathbf{r}_{\rm com}-\mathbf{r}^*)\ {\rm e}^{-\beta[U+u_{\rm c}]}} \\
& \times \frac{\int_{\rm unbound} {\rm d}\mathbf{X}\ \delta(\mathbf{r}_{\rm com}-\mathbf{r}^*)\ {\rm e}^{-\beta[U+u_{\rm c}]}}{\int_{\rm unbound} {\rm d}\mathbf{X}\ \delta(\mathbf{r}_{\rm com}-\mathbf{r}^*)\ {\rm e}^{-\beta U}}
\end{aligned}
$$

$$
\begin{aligned}
K_{\rm eq}^{\rm PMF} = & \frac{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta U}}{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}]}} \times \frac{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}]}}{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm a}]}} \\
& \times \frac{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm a}]}}{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm a}+u_{\rm o}]}} \\
& \times \boxed{\frac{\int_{\rm bound} {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm a}+u_{\rm o}]}}{\int_{\rm unbound} d\mathbf{1}\ \delta(\mathbf{r}_{\rm com}-\mathbf{r}^*) \int {\rm d}\mathbf{X}\ {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm o}]}}} \\
& \times \frac{\int_{\rm unbound} {\rm d}\mathbf{X}\ \delta(\mathbf{r}_{\rm com}-\mathbf{r}^*)\ {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm o}]}}{\int_{\rm unbound} {\rm d}\mathbf{X}\ \delta(\mathbf{r}_{\rm com}-\mathbf{r}^*)\ {\rm e}^{-\beta[U+u_{\rm c}]}} \\
& \times \frac{\int_{\rm unbound} {\rm d}\mathbf{X}\ \delta(\mathbf{r}_{\rm com}-\mathbf{r}^*)\ {\rm e}^{-\beta[U+u_{\rm c}]}}{\int_{\rm unbound} {\rm d}\mathbf{X}\ \delta(\mathbf{r}_{\rm com}-\mathbf{r}^*)\ {\rm e}^{-\beta U}}
\end{aligned}
\tag{3.17}
$$

To evaluate the boxed expression (right-hand-side), we define the function,[9]

$$
\rho(\mathbf{r}'_{\rm com}) \equiv \int {\rm d}\mathbf{X}\ \delta(\mathbf{r}_{\rm com}-\mathbf{r}'_{\rm com})\ {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm o}]} \tag{3.18}
$$

where $\rho(\mathbf{r}^*) \equiv \rho(r^*,\theta^*,\phi^*)$. At a sufficiently large distance r^*, this function becomes independent of θ^* and ϕ^*, and $\rho(\mathbf{r}^*) \equiv \rho(r^*,0,0)$. The boxed expression in eq 3.17 can be rewritten as,

$$\boxed{\frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{a}}+u_{\mathrm{o}}]}}{\displaystyle\int_{\text{unbound}} d\mathbf{X}\ \delta(\mathbf{r}_{\mathrm{com}}-\mathbf{r}^*)\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}]}}} = \frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{r}_{\mathrm{com}}\ \rho(\mathbf{r}_{\mathrm{com}})\ \mathrm{e}^{-\beta u_{\mathrm{a}}}}{\rho(\mathbf{r}^*)}$$

$$= \frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{r}_{\mathrm{com}}\ \rho(\mathbf{r}_{\mathrm{com}})\ \mathrm{e}^{-\beta u_{\mathrm{a}}}}{\rho(r^*,0,0)}$$

$$= \frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{r}_{\mathrm{com}}\ \rho(\mathbf{r}_{\mathrm{com}})\ \mathrm{e}^{-\beta u_{\mathrm{a}}}}{\rho(r^*,0,0)\dfrac{\displaystyle\int \mathrm{d}\mathbf{r}_{\mathrm{com}}\ \delta(r-r^*)\ \mathrm{e}^{-\beta u_{\mathrm{a}}}}{\displaystyle\int \mathrm{d}\mathbf{r}_{\mathrm{com}}\ \delta(r-r^*)\ \mathrm{e}^{-\beta u_{\mathrm{a}}}}}$$

$$= S^* \frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{r}_{\mathrm{com}}\ \rho(\mathbf{r}_{\mathrm{com}})\ \mathrm{e}^{-\beta u_{\mathrm{a}}}}{\rho(r^*,0,0)\displaystyle\int \mathrm{d}\mathbf{r}_{\mathrm{com}}\ \delta(r-r^*)\ \mathrm{e}^{-\beta u_{\mathrm{a}}}}$$

$$= S^* \frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{r}_{\mathrm{com}}\ \rho(\mathbf{r}_{\mathrm{com}})\ \mathrm{e}^{-\beta u_{\mathrm{a}}}}{\displaystyle\int \mathrm{d}\mathbf{r}_{\mathrm{com}}\ \delta(r-r^*)\ \rho(\mathbf{r}_{\mathrm{com}})\ \mathrm{e}^{-\beta u_{\mathrm{a}}}}$$

$$= S^* \int_{\text{bound}} dr' \frac{\displaystyle\int \mathrm{d}\mathbf{r}_{\mathrm{com}}\ \delta(r-r')\rho(\mathbf{r}_{\mathrm{com}})\ \mathrm{e}^{-\beta u_{\mathrm{a}}}}{\displaystyle\int \mathrm{d}\mathbf{r}_{\mathrm{com}}\ \delta(r-r^*)\rho(\mathbf{r}_{\mathrm{com}})\ \mathrm{e}^{-\beta u_{\mathrm{a}}}}$$

$$= S^* \int_{\text{bound}} dr'\ \mathrm{e}^{[w(r')-w(r^*)]}$$

$$= S^* I^* \tag{3.19}$$

where

$$I^* = \int_{\text{bound}} dr'\ \mathrm{e}^{-\beta[w(r')-w(r^*)]} \tag{3.20}$$

with $w(r)$, the one-dimensional PMF defined as a function of the distance, r, separating the centers of mass, and calculated in the presence of the restraining potential $[u_{\mathrm{c}} + u_{\mathrm{a}} + u_{\mathrm{o}}]$ (the upper limit of the integral in r' should correspond to some maximum distance in order to capture the bound state). S^* is the integral of the angular potential u_{a} when r is equal

to r^*, which is equivalent to a surface term,

$$\begin{aligned} S^* &= \int \mathrm{d}\mathbf{r}_{\mathrm{com}}\ \delta(r - r^*)\ \mathrm{e}^{-\beta u_{\mathrm{a}}(\theta,\phi)} \\ &= r^{*2} \int_0^{\pi} \mathrm{d}\theta \sin\theta \int_0^{2\pi} \mathrm{d}\phi\ \mathrm{e}^{-\beta u_{\mathrm{a}}(\theta,\phi)} \end{aligned} \tag{3.21}$$

It follows that the binding constant can be expressed as,[9]

$$K_{\mathrm{eq}}^{\mathrm{PMF}} = S^* I^*\, \mathrm{e}^{-\beta[G_{\mathrm{unbound,c}} - G_{\mathrm{bound,c}}]} \times e^{-\beta[G_{\mathrm{bound,o}} - G_{\mathrm{unbound,o}}]} \times e^{\beta G_{\mathrm{bound,a}}} \tag{3.22}$$

The string counterpart of the boxed expression in eq (3.17) can be also recast with $u_{\mathrm{t}}^{(1)} = u_r^{(1)} + u_{\mathrm{a}}$ and r^* is the distance separating the two centers of mass corresponding to the last, M^{th} image of the pathway,

$$
\begin{aligned}
\frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\text{c}}+u_{\text{t}}^{(1)}+u_{\text{o}}]}}{\displaystyle\int_{\text{unbound}} d\mathbf{1}\ \delta(\mathbf{r}_{\text{com}}-\mathbf{r}^*)\int \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\text{c}}+u_{\text{o}}]}} &= \frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{r}_{\text{com}}\ \rho(\mathbf{r}_{\text{com}})\ \mathrm{e}^{-\beta[u_r^{(1)}+u_{\text{a}}]}}{\rho(\mathbf{r}^*)} \\
&= \frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{r}_{\text{com}}\ \rho(\mathbf{r}_{\text{com}})\ \mathrm{e}^{-\beta[u_r^{(1)}+u_{\text{a}}]}}{\rho(r^*,0,0)} \\
&= \frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{r}_{\text{com}}\ \rho(\mathbf{r}_{\text{com}})\ \mathrm{e}^{-\beta[u_r^{(1)}+u_{\text{a}}]}}{\rho(r^*,0,0)\dfrac{\displaystyle\int \mathrm{d}\mathbf{r}_{\text{com}}\ \delta(r-r^*)\ \mathrm{e}^{-\beta u_{\text{a}}}}{\displaystyle\int \mathrm{d}\mathbf{r}_{\text{com}}\ \delta(r-r^*)\ \mathrm{e}^{-\beta u_{\text{a}}}}} \\
&= S^*\frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{r}_{\text{com}}\ \rho(\mathbf{r}_{\text{com}})\ \mathrm{e}^{-\beta[u_r^{(1)}+u_{\text{a}}]}}{\rho(r^*,0,0)\displaystyle\int \mathrm{d}\mathbf{r}_{\text{com}}\ \delta(r-r^*)\ \mathrm{e}^{-\beta u_{\text{a}}}} \\
&= S^*\frac{\displaystyle\int_{\text{bound}} \mathrm{d}\mathbf{r}_{\text{com}}\ \rho(\mathbf{r}_{\text{com}})\ \mathrm{e}^{-\beta[u_r^{(1)}+u_{\text{a}}]}}{\displaystyle\int \mathrm{d}\mathbf{r}_{\text{com}}\ \delta(r-r^*)\ \rho(\mathbf{r}_{\text{com}})\ \mathrm{e}^{-\beta u_{\text{a}}}} \\
&= S^*\int_{\text{bound}} dr'\frac{\displaystyle\int \mathrm{d}\mathbf{r}_{\text{com}}\ \delta(r-r')\rho(\mathbf{r}_{\text{com}})\ \mathrm{e}^{-\beta[u_r^{(1)}+u_{\text{a}}]}}{\displaystyle\int \mathrm{d}\mathbf{r}_{\text{com}}\ \delta(r-r^*)\rho(\mathbf{r}_{\text{com}})\ \mathrm{e}^{-\beta u_{\text{a}}}} \\
&= S^*\int dr'\ \mathrm{e}^{-\beta u_r^{(1)}}\ \mathrm{e}^{-\beta[w(r')-w(r^*)]} \\
&= S^*I^*\ \frac{\displaystyle\int dr'\ \mathrm{e}^{-\beta u_r^{(1)}}\ \mathrm{e}^{-\beta w(r')}}{\displaystyle\int dr'\ \mathrm{e}^{-\beta w(r')}} \\
&= S^*I^*\ \mathrm{e}^{-\beta\Delta G^{(1)}_{\text{bound},r}} \qquad (3.23)
\end{aligned}
$$

where $\Delta G^{(1)}_{\text{bound},r}$ is the free energy incurred to introduce the distance restraining potential $u_r^{(1)}$ at the first image of the pathway, that is in the bound state. The S^* term is evaluated at the last, M^{th} image of the pathway. Leaning on the above developments, the expression of the equilibrium constant obtained with the traditional PMF-based approach can be reconciled

with that of the string method for a rectilinear path according to,

$$\begin{aligned} K_{\mathrm{eq}}^{\mathrm{string}} &= \mathrm{e}^{-\beta[\Delta G_{\mathrm{unbound,c}} - \Delta G_{\mathrm{bound,c}}]} \times e^{-\beta[\Delta G_{\mathrm{unbound,o}} - \Delta G_{\mathrm{bound,o}}]} \\ &\quad \times e^{\beta \Delta G^{(1)}_{\mathrm{bound,t}}} \times F_{\mathrm{t}} \times \prod_{i=1}^{M-1} \mathrm{e}^{-\beta \Delta G^{(i,i+1)}} \\ &= \mathrm{e}^{-\beta[\Delta G_{\mathrm{unbound,c}} - \Delta G_{\mathrm{bound,c}}]} \times e^{-\beta[\Delta G_{\mathrm{unbound,o}} - \Delta G_{\mathrm{bound,o}}]} \\ &\quad \times e^{\beta \Delta G^{(1)}_{\mathrm{bound,t}}} \times S^* I^* \times \mathrm{e}^{-\beta \Delta G^{(1)}_{\mathrm{bound},r}} \\ &= \mathrm{e}^{-\beta[\Delta G_{\mathrm{unbound,c}} - \Delta G_{\mathrm{bound,c}}]} \times e^{-\beta[\Delta G_{\mathrm{unbound,o}} - \Delta G_{\mathrm{bound,o}}]} \\ &\quad \times e^{\beta \Delta G_{\mathrm{bound,a}}} \times S^* I^* \\ &= K_{\mathrm{eq}}^{\mathrm{PMF}} \end{aligned} \tag{3.24}$$

noting that $\Delta G^{(1)}_{\mathrm{bound,t}} = \Delta G^{(1)}_{\mathrm{bound},r} + \Delta G_{\mathrm{bound,a}}$.

3.2.3 Removal of restraints

The conformational restraints are defined as distance RMSDs, ξ, of the two binding proteins relative to their conformation in the bound state. The RMSD restraining potential, u_{c}, comprise four contributions,

$$u_{\mathrm{c}} = u_{\mathrm{c,bb1}}(\xi_{\mathrm{bb1}}) + u_{\mathrm{c,bb2}}(\xi_{\mathrm{bb2}}) + u_{\mathrm{c,sc1}}(\xi_{\mathrm{sc1}}) + u_{\mathrm{c,sc2}}(\xi_{\mathrm{sc2}}) \tag{3.25}$$

$$= K_{\mathrm{c}}(\xi_{\mathrm{bb1}})^2 + K_{\mathrm{c}}(\xi_{\mathrm{bb2}})^2 + K_{\mathrm{c}}(\xi_{\mathrm{sc1}})^2 + K_{\mathrm{c}}(\xi_{\mathrm{sc2}})^2 \tag{3.26}$$

where $u_{\mathrm{c,bb1}}$, $u_{\mathrm{c,bb2}}$, $u_{\mathrm{c,sc1}}$ and $u_{\mathrm{c,sc2}}$ are the RMSD restraining potentials acting on the backbone of protein 1, the backbone of protein 2, the interfacial side chains of protein 1, the interfacial side chains of protein 2, respectively. In practice, the free energy associated with

these restraints can be evaluated sequentially,

$$e^{-\beta\Delta G_{\mathrm{c,sc2}}} = \frac{\int \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c,bb1}}+u_{\mathrm{c,bb2}}+u_{\mathrm{c,sc1}}+u_{\mathrm{c,sc2}}]}}{\int \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c,bb1}}+u_{\mathrm{c,bb2}}+u_{\mathrm{c,sc1}}]}} \tag{3.27}$$

$$e^{-\beta\Delta G_{\mathrm{c,sc1}}} = \frac{\int \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c,bb1}}+u_{\mathrm{c,bb2}}+u_{\mathrm{c,sc1}}]}}{\int \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c,bb1}}+u_{\mathrm{c,bb2}}]}} \tag{3.28}$$

$$e^{-\beta\Delta G_{\mathrm{c,bb2}}} = \frac{\int \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c,bb1}}+u_{\mathrm{c,bb2}}]}}{\int \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c,bb1}}]}} \tag{3.29}$$

$$e^{-\beta\Delta G_{\mathrm{c,bb1}}} = \frac{\int \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c,bb1}}]}}{\int \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta U}} \tag{3.30}$$

The expressions are identical for the bound and unbound states. It is worth noting that perturbation due to the side-chain restraints is treated in the presence of the backbone restraints, although there is a certain arbitrariness in how one chooses to order the various restraints. The free-energy contributions arising from the conformational restraints possess all the same form, for the bound and the unbound proteins, and can be evaluated by means of PMF calculations,

$$\begin{aligned} e^{-\beta\Delta G_{\mathrm{c}}} &= \frac{\int \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta[U^*+u_{\mathrm{c}}(\xi)]}}{\int \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta U^*}} \\ &= \frac{\int d\xi\ \mathrm{e}^{-\beta[w(\xi)+u_{\mathrm{c}}(\xi)]}}{\int d\xi\ \mathrm{e}^{-\beta w(\xi)}} \end{aligned} \tag{3.31}$$

where $w(\xi)$ is the PMF along the distance RMSD, calculated under potential U^*, which includes the remaining conformational restraints corresponding to the sequential process embodied in eq 3.30. To summarize, the treatment of the conformational free-energy contributions requires eight individual PMF calculations, namely four for the bound state, and four for the unbound state.

3.2.4 Minimum free-energy path

The string method can be used to determine the optimal minimum free energy pathway (MFEP), $\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots \mathbf{Z}^{(M)}\}$, between two states defined in the subspace of CVs $\mathbf{Z}$. To clarify the significance of the MFEP, let us first define the free energy surface associated with the set of CVs,

$$\mathrm{e}^{-\beta \mathcal{W}(\mathbf{Z})} = C \int d\mathbf{X}\, \delta(\tilde{\mathbf{Z}}(\mathbf{X}) - \mathbf{Z})\, \mathrm{e}^{-\beta U(\mathbf{X})} \tag{3.32}$$

where C is a constant, and $\tilde{\mathbf{Z}}(\mathbf{X})$ indicates a vector-valued function that maps every configuration $\mathbf{X}$ of the system onto the CVs. The free energy surface $\mathcal{W}(\mathbf{Z})$ that we use to describe the separation of a protein-protein complex is a function of six CVs. By definition, the MFEP is a curve in the space of the six CVs that connects the two minima A and B of $\mathcal{W}(\mathbf{Z})$ and to which the vector $\mathbf{M}(\mathbf{Z}) \cdot \langle \mathcal{F}(\mathbf{Z}) \rangle$ is everywhere tangent,[8,59,62] i.e. it is a curve satisfying

$$[\mathbf{M}(\mathbf{Z}) \cdot \nabla \mathcal{W}(\mathbf{Z})]^{\perp} = 0 \tag{3.33}$$

where $\langle \mathcal{F}(\mathbf{Z}) \rangle$ is the mean force equal to minus the gradient of the free energy surface, $-\nabla \mathcal{W}(\mathbf{Z})$, and the superscript $\perp$ indicates projection in the directions perpendicular to the curve. The quantity $\mathbf{M}(\mathbf{Z})$ is a metric tensor with the dimension of a diffusion coefficient accounting for the curvilinear nature of the CVs.[7,8,63] As the curvilinear string pathway here is utilized as a support to characterize equilibrium binding rather than kinetic factors, $\mathbf{M}$ will be assumed to be constant and isotropic in the following for the sake of simplicity.

In general, the two ends of the string, $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(M)}$, may be held fixed or allowed to adapt freely within the region A and B associated with the end-points. For instance, the first image (region A) represented by the collective variable $\mathbf{Z}^{(1)}$ is held fixed since it corresponds to the known structure of the bound complex. In the context of a free energy calculation carried out in bulk solution, the mean force, $\langle \mathcal{F}(\mathbf{Z}) \rangle$, should tend to zero as the two proteins are increasingly separated in solution. This means that the free energy surface does not depend on the relative orientation of the two proteins once they are separated in bulk solution. However, the relative orientation of the two unbound proteins (region B) is not known a priori, which means that this must be determined by building up the separation pathway progressively. This observation can be exploited here to adapt the string method for a binding process, such that it will yield a more effective separation pathway for the binding free energy calculation.

Normally, the standard string method algorithm converges toward a minimum free energy pathway (MFEP) according to a steepest descent principle iteratively from the mean force, calculated from simulations that are harmonically restrained at specific values of the CVs, $\mathbf{Z}^{(i)}$. By virtue of the linearity of the harmonic restraints applied to the different images, the mean force can be simply evaluated from the mean deviation away from the reference value of the CVs, i.e., $\langle \mathcal{F}(\mathbf{Z}) \rangle = -K(\langle \mathbf{Z}^{(i)} \rangle - \mathbf{Z}^{(i)}_{\text{ref}})$. In practice, a straightforward method to optimize the string consists in simply displacing the reference value of the CVs for image i, $\mathbf{Z}^{(i)}_{\text{ref}}$, toward the average $\langle \mathbf{Z}^{(i)}_{\text{ref}} \rangle$ determined from the simulation. This step is then followed by standard reparametrization,[8,59] and the iteration cycle is repeated until convergence to yield the MFEP. We refer to this as the "average CV method" for optimizing the string.

Here, to obtain a more robust and effective solution for the separation process and reduce the amount of unnecessary rotational tumbling motions of the macromolecules in solution, the free energy surface function of the six CVs, $\mathcal{W}(\mathbf{Z})$, is estimated locally using the information from the restrained simulations. To compute a MFEP of protein binding, a search is started from the bound complex as image number 1 and additional images are progressively

added along the MFEP until the proteins are sufficiently separated in bulk. New images are added along the MFEP based on the estimated free energy landscape, $\mathcal{W}(\mathbf{Z})$, calculated from the restrained simulations of the existing images. Once the two proteins are sufficiently separated, e.g., about 10-12 Å apart, the free energy landscape ix expected to become very flat. At such large separation, variations in the separation axis (θ, ϕ) and the Euler angles (Θ, Φ, Ψ) should therefore be avoided. In practice, the CVs $(\theta^{(i)}, \phi^{(i)}, \Theta^{(i)}, \Phi^{(i)}, \Psi^{(i)})$ for images $i \geq n$, should be restricted to the value $(\theta^{(n)}, \phi^{(n)}, \Theta^{(n)}, \Phi^{(n)}, \Psi^{(n)})$, where n is chosen such that the proteins are considered sufficiently separated.

To build the string of images, the free energy surface is constructed iteratively, by using the restrained image simulations, starting from the first image $\mathbf{Z}^{(1)}$. An exhaustive search in the space of the six CVs would be computationally prohibitive. Instead, the search can be simplified because the goal is to find the MFEP along the separation of the protein complex. Therefore, the procedure consists in adding images along the direction of increasing r. The CVs of the new images is determined by assuming an increase in the separation distance, r, with $\mathbf{Z}_{\text{new}} = \{\mathbf{Z}_{\text{new}} | r_{\text{new}} = r_{\text{old}} + \Delta r$, where $(\theta_{\text{new}}, \phi_{\text{new}}, \Theta_{\text{new}}, \Phi_{\text{new}}, \Psi_{\text{new}})$ are chosen to minimize $\mathcal{W}(\mathbf{Z}_{\text{new}})$ at a fixed r_{new}. The parameter Δr is adjusted to control the magnitude of the free energy increment between adjacent images, $||\mathcal{W}(\mathbf{Z}_{\text{new}}) - \mathcal{W}(\mathbf{Z}_{\text{old}})||$. In practice, five new images were built per cycle to accelerate the construction of the separation pathway. Moreover, the calculation of a free energy surface in such a high dimensional space – $\mathcal{W}(\mathbf{Z})$ is a function of six CVs – can be a difficult task using traditional umbrella sampling algorithms requiring accumulating histograms bins for the entire space. To overcome these difficulties, we adopted the dynamic histogram analysis method (DHAM), which relies on a transition count matrix, to estimate $\mathcal{W}(\mathbf{Z})$.[64] Because the transition count matrix converges faster than accumulated histograms, DHAM results in a more robust estimation of the free energy surface. The transition matrix can be sparsely constructed when a short lag time was used, which also helps the computations. In practice, a free energy surface with six-dimensional parameter can be computed rapidly with DHAM using a reasonable grid spacing to define

the set of discrete conformational states. A grid spacing of 5° for various angular parameters and 0.2 Å were used. The initial string of the separation pathway was calculated using an implicit solvent model, and then smoothed using an interpolation algorithm before each image was solvated with explicit water molecules.

To accelerate convergence, the initial guess for the optimized curvilinear path was generated with implicit solvent starting from the equilibrated bound-state complex. The bound-state complex was first prepared and equilibrated for 1 ns with implicit solvation in the presence of RMSD and distance restraints. The temperature was maintained at 300 K using Langevin dynamics with a damping coefficient of 1.0 ps^{-1} with dielectric constant of 80. The path finding algorithm was implemented by incrementally adding images from the bound state to the fully separated state according to the following protocol:

(*i*) Using the partial string with n images, $\{\mathbf{Z}^{(1)}[\mathbf{r}^{(1)}, \mathbf{\Omega}^{(1)}], \ldots, \mathbf{Z}^{(n)}[\mathbf{r}^{(n)}, \mathbf{\Omega}^{(n)}]\}$, we append m additional windows (typically $m = 5$) by incrementing the center-of-mass distance incremented from the n^{th} window (the orientational restraints are kept unchanged), yielding the augmented string, $\{\mathbf{Z}^{(1)}[\mathbf{r}^{(1)}, \mathbf{\Omega}^{(1)}], \ldots, \mathbf{Z}^{(n)}[\mathbf{r}^{(n)}, \mathbf{\Omega}^{(n)}], \mathbf{Z}^{(n+1)}[\mathbf{r}^{(n)} + \Delta r, \mathbf{\Omega}^{(n)}], \ldots,$
$\mathbf{Z}^{(n+m)}[\mathbf{r}^{(n)} + m\Delta r, \mathbf{\Omega}^{(n)}]\}$.

(*ii*) The $(n + m)$ images for the string are equilibrated and sampled with quadratic CVs restraints centered on the reference positions, $\mathbf{Z}^{(i)}$.

(*iii*) A local six-dimensional free energy surface is generated using DHAM, and the $(n+m)$ images are reset to the local minimum on this free energy surface. The string of $(n+m)$ images is then reparametrized.

Once the two molecules are fully separated, the M images along the string are solvated with explicit water molecules and the string pathway is refined further following steps (*ii*) and (*iii*). An alternative method that can generate an initial string pathway faster is possible by replacing the DHAM in step (*iii*) by an averages method as discussed above, This method

iteratively shifts the reference center of each image toward its mean calculated from the trajectory with CV restraints. The reference center at the i^{th} iteration is,

$$\{\mathbf{Z}_{\text{ref}}^{(1)}, \ldots, \mathbf{Z}_{\text{ref}}^{(M)}\}^{(i)} = \epsilon\{\langle\mathbf{Z}_{\text{ref}}^{(1)}\rangle, \ldots, \langle\mathbf{Z}_{\text{ref}}^{(M)}\rangle\}^{(i)} + (1-\epsilon)\{\mathbf{Z}_{\text{ref}}^{(1)}, \ldots, \mathbf{Z}_{\text{ref}}^{(M)}\}^{(i-1)} \tag{3.34}$$

where ϵ is a mixing factor that controls the rate of convergence of the iteration (if ϵ is close to zero then the reference center from the previous iteration is weighted more heavily).

3.3 Computational methods

3.3.1 Host-guest system

The AMBER force field parameters for the cucurbit[7]uril molecule were taken from the previous study of Gilson and co-workers and converted to the CHARMM format.[65,66] CHARMM-GUI was used to solvate the cucurbit[7]uril with benzene system. The dimensions of the simulation system is 33 Å × 33 Å × 47 Å. The CHARMM PARAM36 force field was used.[67–69] The system was solvated by 1664 water molecules that were represented by the TIP3P model.[70] Both the rectilinear and curvilinear separation pathways comprise 28 images,and each image was simulated for 2 ns, for a total of 56 ns. The sampling of the restrained simulations were enhanced by using replica-exchange MD simulation (US/H-REMD).[71]

The string pathway was initially refined using the average CV method with ϵ of 0.4 as in eq 3.34, implicit solvent model of dielectric constant $\epsilon = 80$ was used, 4 starting images and 6 iterations of 4 images addition were done as shown in protocol to generate a 28-image string, and 20 iterations of string refinement were carried out. The CB[7] system at the last image along the curvilinear string pathway is shown in Figure 3.1. For the end-state bound complex, the free energy associated with the orientational and translational restraints was calculated with US/H-REMD simulations (11 windows decreasing the magnitude of the restraining force constants) using the multiple-copy algorithm (MCA) implemented in NAMD.[72] For

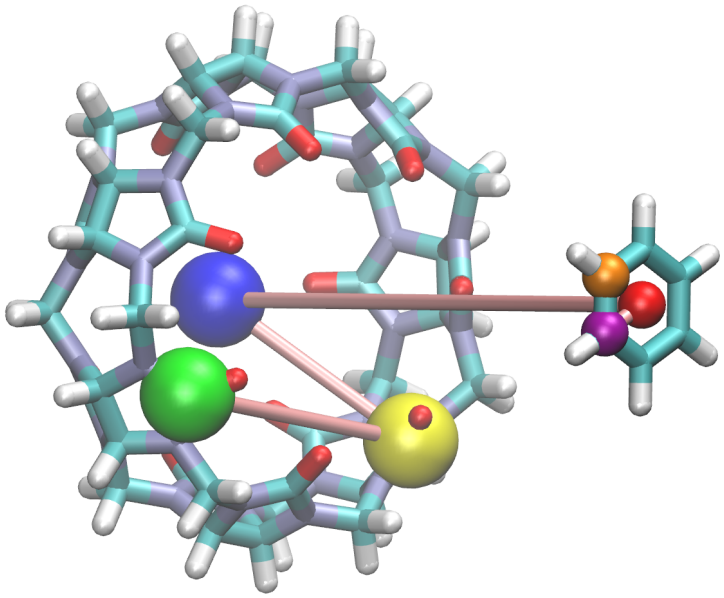

Figure 3.1: The CB[7] system for the last image along the straight rectilinear separation string pathway. The reference points P_1 (blue), P_2 (yellow), and P_3 (green), L_1 (red), L_2 (purple), and L_3 (orange) serving to set the CVs for the relative position and orientation of the two molecules are shown. The point P_1 (blue) is set at the center of mass of selected atoms of the back ring of the cucurbit[7]uril molecule, and the point L_1 (red) is set at the center of mass of the benzene molecule.

the end-state unbound complex, the free energies associated with releasing the orientational restraints were calculated analytically.

All MD simulations were carried out with the NAMD 2.11b simulation program.[72–74] The temperature and the pressure were maintained at 300 K and 1 atom, using Langevin dynamics with a damping coefficient of 1.0 ps^{-1} and Langevin piston. A time-step of 2 fs was used and the bonded distance of hydrogen atoms was rigidly constrained. A particle mesh Ewald (PME) algorithm was used to treat the long-range electrostatic interactions, with a grid spacing of 1 point per Å.[75] Real-space nonbonded interactions were smoothly switched from 10 Å to 12 Å.

3.3.2 Barnase–barstar complex

The bound-state complex for the barnase–barstar complex was first prepared and equilibrated for 1 ns with implicit solvation in the presence of RMSD and distance restraints to

build all the simulation systems. Three separate simulation systems were then constructed, including the barnase–barstar complex in solution, and the isolated barnase and barstar in solution. The crystallographic barnase–barstar complex structure with high-resolution was to taken from RCSB database (PDB:1BRS) initial configuration.[38] Chains C and F were taken from the structure, which of the three unit-cell complexes are expected to be most representative of the native state. The N-terminal alanine and glutamine residues were added to barnase, and residues 40 and 82 in barstar were mutated back to cysteine in order to produce a wild-type complex. Standard protonation states were assigned based on a pKa calculation using propKa with an assumed pH of 8.0.[76] CHARMM-GUI was used to solvate protein in a water box along with 2–6 neutralizing sodium or chloride ions.[67] The final system size were approximately 25,000 atoms for each monomeric barnase and barstar protein, and 50,000 atoms for the barnase–barstar complex. The dimensions of the simulation system is 70 Å $\times$ 70 Å $\times$ 108, which ensures that the two proteins do not interact across the periodic boundaries at full separation. The CHARMM PARAM36 force field was used[68,69] with the rigid TIP3P water model.[70]

To restrict fluctuations of the side chains while the proteins are separated, RMSD restraints were introduced on those residues known to be critical for binding or ones near them. Specifically, one restraint on the RMSD of residues 27, 59, 60, 83, 85, 87, and 102 of barnase and one on the RMSD of residues 29, 31, 33, 35, 38, 39, 42, and 76 of barstar were added. In total, four conformational restraints were applied, two on the backbones (denoted by subscripts Bn,c and Bs,c) and two on the specific set of residues at the interface (Bn,res and Bs,res). Additionally, five angular restraints were applied, three on the orientation of barstar relative to barnase (denoted by subscript o, encompassing the Θ, Φ, and Ψ Euler angles) and two on its relative position (denoted by a, which includes the polar angles θ and ϕ).

The string pathway for the barnase–barstar separation was generated using the DHAM protocol. The six points used to construct the CVs mapping the relative position and

orientation of barnase and barstar along the separation string pathway are defined by the center-of-mass of: all Cα of barnase (P_1), Cα of residue 70-73, 89-92 of barnase (P_2), Cα of residue 85-88, 97-100 of barnase (P_3), all Cα of barstar (L_1), Cα of residue 13-24 of barstar (L_2), and Cα of residue 66-79 of barstar (L_3). The initial generation of the string accounted for solvation implicitly with a simple uniform dielectric constant of $\epsilon = 80$. Starting with five images, 49 iterations were carried out each appending five additional images to the string to generate a 250-image string. At this point, all images were solvated with 17295 water molecules and four sodium ions were added to neutralize the simulation system. The entire string with explicit solvent was then further refined with three additional iterations. The final string was equilibrated with the solute scaling REST2,[77] with four value of the scaling parameter. The barnase and barstar at the last image along the curvilinear string pathway is shown in Figure 3.2. Five additional images were inserted to foster the swapping

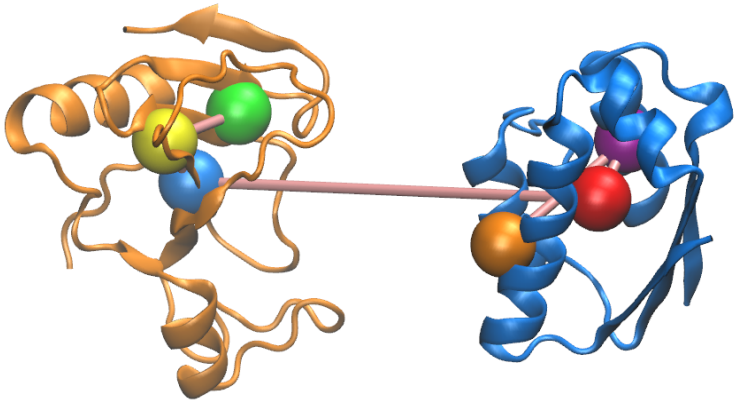

Figure 3.2: Barnase and barstar for the last image of the curvilinear separation string pathway. Barnase is on the left with reference points P_1 (blue), P_2 (yellow), and P_3 (green), and and barstar is on the right with reference points L_1 (red), L_2 (purple), and L_3 (orange). These points serve to set the CVs for the relative position and orientation of the two proteins. The reference points are defined by the center-of-mass of: all Cα of barnase (P_1), Cα of residue 70-73, 89-92 of barnase (P_2), Cα of residue 85-88, 97-100 of barnase (P_3), all Cα of barstar (L_1), Cα of residue 13-24 of barstar (L_2), and Cα of residue 66-79 of barstar (L_3).

probability between neighboring images, resulting in an optimized 255-image string pathway. The system was sampled using H-REMD simulations, including 255 images with six evenly-spaced solute scaling (from 1.0 to 0.75) for a total of 1530 replicas. The residues 27, 31, 35, 37, 38, 56, 58-60, 82-85, 101-104, and 106 of barnase and the residues 27, 29-36, 38-40, 42-47,

72, 73, and 76 of barstar were affected by the REST2 solute scaling. H-REMD simulations were carried out for 18 ns, yielding a total of 27.54 μs of MD aggregate data. For comparison, the binding free energy calculation of the wild-type barnase and barstar was carried out by following the spatial separation of the associated complex along a rectilinear (straight) and along a curvilinear string pathway. The binding free energy calculation for the single and double mutants of barnase and barnase was carried out using only the curvilinear string pathway. The system comprising 255 images and eight evenly-spaced solute scaling for a total of 1530 replicas was sampled using H-REMD simulations. H-REMD simulations were carried out for 10 ns each, yielding a total of 20.40 μs.

The free energies for releasing the various RMSD conformational restraints were calculated using umbrella sampling (US) with H-REMD simulations (US/H-REMD) comprising 16 US windows with different offset varying from 0.0 to 3.0 Å RMSD. The US/H-REMD simulations were carried out for 1.5 ns per replica. For each case, there is a total of eight RMSD conformational restraint free energies to calculate (two proteins: barnase and barstar, two atom selections: backbone and side chain, and two end-states: bound and unbound), corresponding to a total computational cost of 192 ns per system. The free energies for releasing the orientational and positional restraints were calculated using free energy perturbation (FEP) with alchemical λ replica-exchange MD simulations (FEP/λ-REMD) by switching off the restraining force constants. The FEP/λ-REMD comprising 11 windows were carried out for 1.5 ns, yielding a total of 16.5 ns per restraint. In the case of the orientational and positional restraints, simulations are required only for the bound complex because the restraint for the unbound complex are handled analytically. This yields a total of 49.5 ns to calculate the free energy $\Delta G_{\mathrm{bound,o}}, \Delta G_{\mathrm{bound,a}}, \Delta G_{\mathrm{bound,t}}$.

All MD simulations were carried out with the NAMD 2.11b simulation program.[72–74] For the explicit solvent systems, the temperature and the pressure were maintained at 300 K and 1 atom, using Langevin dynamics with a damping coefficient of 1.0 ps^{-1} and Langevin piston. A time-step of 2 fs was used and chemical bonds involving hydrogen atoms were

rigidly constrained. A particle mesh Ewald (PME) algorithm was used to treat the long-range electrostatic interactions, with a grid spacing of 1 point per Å.[75] Real-space nonbonded interactions were smoothly switched from 10 Å to 12 Å. To compare with previous study that used neutral termini, zwiterionic N- and C- terminal were alchemically transformed into neutral N- and C- terminal. A dual-topology protocol was used where both termini are present at all times but do not interact with each other.[78]The alchemical FEP/MD calculation was carried out with 21 evenly-spaced coupling constant λ (0, 0.05, ..., 0.95, 1.00). Each λ-window was simulated for 1.5 ns for a total of 31.5 ns.

To estimate the magnitude of the barnase–barstar interactions remaining in the maximum separation, a finite-distance correction was calculated using the Poisson-Boltzmann (PB) equation. The calculations were carried out using the PBEQ module[79] of the CHARMM program[80] version 42a1 with the optimized atomic Born radii.[81] For for each string pathway, the PB finite-distance correction was calculated by averaging over MD configurations taken every 200 ps. The PB calculation was performed using focusing technique, with a first grid size of 1 Å followed by a finer grid size of 0.45 Å. The protein dielectric constant was set to 12 and the solvent dielectric constant was set to 80 (the finite-distance correction does not vary significantly when changing the dielectric constant of the protein).

3.4 Results and Discussion

A novel formulation based on the string method was developed to utilize a physical separation pathway of two molecules to calculated their binding free energy. This string-based formulation differs from the PMF-based method, which assumes a rectilinear separation pathway.[9] The final expression of the PMF-based binding free energy is eq 3.22. The most distinctive part of this method is the integral I^* displayed in eq 3.20, which involves the PMF along the one-dimensional Euclidian distance r restrained along a rectilinear (straight) axis.[9] In contrast, the string-based calculation does not require a similar integral involving a one-dimensional PMF. The most critical part of the string-based calculation leading to

the final expression eq 3.16 is the series of relative free energies between neighboring images along the string pathway that is given in eq 4.2.

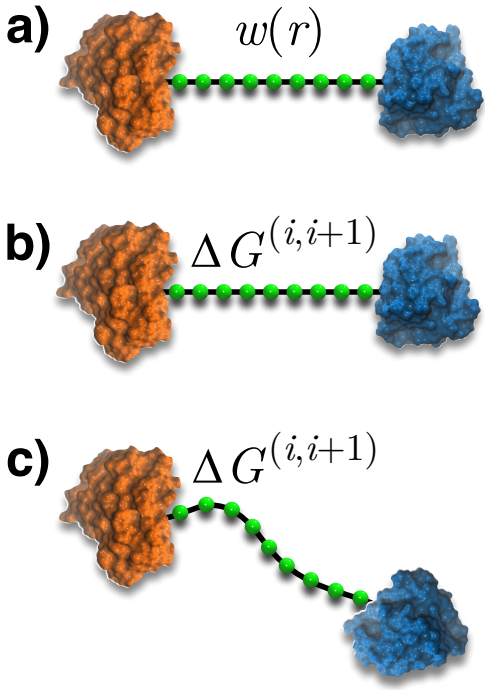

Figure 3.3: Schematic representation of the three different methods to calculate the binding free energy of two molecules. (a) and (b) assume the same straight rectilinear separation pathway. In (a), $w(r)$ is calculated from the biased umbrella sampling windows following the PMF-based formulation of eq 3.22,[9] while in (b) the relative free energy $\Delta G^{(i,i+1)}$ is calculated via eq 4.2 from the biased windows corresponding to the images along the straight pathway following the string-based formulation of eq 3.16. In (c), the relative free energy $\Delta G^{(i,i+1)}$ is calculated via eq 4.2 from the biased windows corresponding to the images along the optimized curvilinear string pathway following the string-based formulation of eq 3.16.

In the special case where all the images along a string are artificially kept along a rectilinear pathway, it is possible to directly relate the PMF-based and string-based formulations. This is illustrated in Figure 3.3 where the PMF $w(r)$ is calculated from the biased umbrella sampling windows following the PMF-based formulation in (a), and the relative free energy $\Delta G^{(i,i+1)}$ is calculated from the biased windows corresponding to the images along the straight pathway following the string-based formulation in (b). In this case the relative free energies between different images from eq 4.2 is related to the the one-dimensional PMF,

$w(r)$, in the following manner,

$$e^{-\beta\Delta G^{(i,i+1)}} = \frac{\int dr\ \mathrm{e}^{-\beta[w(r)+u_r^{(i+1)}(r)]}}{\int dr\ \mathrm{e}^{-\beta[w(r)+u_r^{(i)}(r)]}} \tag{3.35}$$

If the force constant of the restraining potential $u_r^{(i)}(r)$ is very large, then the relative free energy between neighboring images will be closely related to the difference in the PMF, $\Delta G^{(i,i+1)} \approx w(r^{(i+1)}) - w(r^{(i)})$. In this case, the cumulative $\sum_{i=1} \Delta G^{(i,i+1)}$ is expected to essentially follow $w(r)$ as a function of $r^{(i)}$. However, if the restraining force constant is weak compared to $w(r)$, then the cumulative free energy will differ from the PMF. In the following we will examine and test the string-based formulation to determine the binding free energy for two host-guest complexes of different nature.

3.4.1 Host-guest system

The string-based formulation to determine the binding free energy of two molecules was first probed on a smaller benchmark system to examine its performance and accuracy. For this purpose, we chose the association of benzene with cucurbit[7]uril (CB[7]), a popular host-guest system for examining computational methods to calculate binding affinity.[82,83] We first examine the results of CB[7] system in explicit solvent. The string pathway was generated through both the DHAM and average CV methods. The variation of the six CVs along the pathway is shown in Figure 3.4. Both methods yield qualitatively similar paths with some differences, although the pathway from the DHAM method seems to be slightly smoother. This may be related to the fact that the images were constructed directly at the minima of a free energy surface. Nevertheless, as our primary objective with this system is to validate the formalism, the pathway determined from the averaging method was used for the subsequent calculations.

Three approaches used to calculate the binding free energy are compared, namely (1)

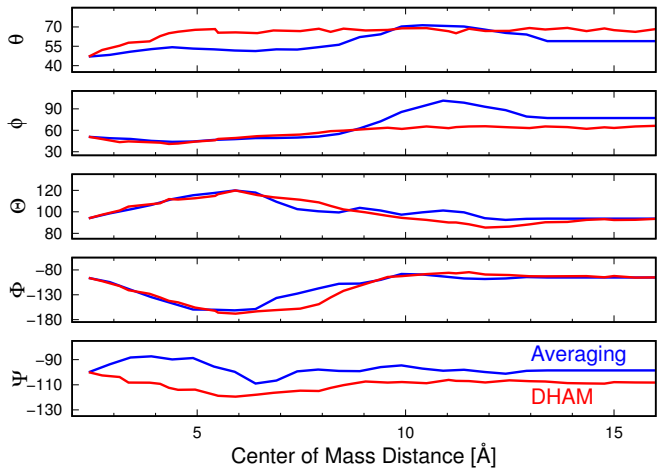

Figure 3.4: Variations of the collective variables along the minimum free energy path obtained for the CB[7] system.

the PMF-based formulation with a rectilinear separation path (eq 3.22)[9], (2) the present string formulation but with a rectilinear separation path (eq 3.16), and (3) the present string formulation with a curvilinear separation path (eq 3.16). The results of the calculations as well as the various free energy components for each of the three approaches are given in Table 3.1. It may be noted that, because the binding free-energy calculations are executed with restraints that single out one given orientation of the benzene molecule, a symmetry contribution of $k_{\mathrm{B}}T\ln$ (12) term must added to the final binding free energy. Based on this comparison, it is manifest that the three different approaches yield essentially identical result. Similarly, the corresponding free energy components are essentially within statistical error (not all components from the different methods may be compared).

Figure 3.5 shows that both the rectilinear and curvilinear paths converge to the same separation free energy. Expectedly, the free energy barrier along the curvilinear string pathway is smaller because it is optimized to follow the MFEP. There is a large barrier around $r = 5.5$ Å in the one-dimensional PMF along the rectilinear separation pathway shown in Figure 3.5, which is completely avoided when following the curvilinear pathway. In the case of the rectilinear (straight) pathway, the PMF differs significantly from the cumulative free

Table 3.1: Results for the CB[7] system

	PMF method (eq 3.22)	String method (eq 3.16)	
Separation path:	rectilinear	rectilinear	curvilinear
Contribution			
$\Delta\Delta G_{\mathrm{o}}$	+0.5	+0.5	+0.5
$\Delta\Delta G_{\mathrm{t}}$		+3.7	+3.7
$k_{\mathrm{B}}T\ln(S \times I \times C^{\circ})$	−11.7		
ΔG_{sepR}		−15.5	−15.4
$k_{\mathrm{B}}T\ln(12)$	−1.5	−1.5	−1.5
$\Delta G^{\circ}_{\mathrm{bind}}$	−12.7	−12.8	−12.7

Errors estimated from bootstrapping are on the order of 0.05 kcal/mol for each component.

energy between successive images, $\sum_{i=1} \Delta G^{(i,i+1)}$. This may be confusing since these results are calculated from exactly the same set of simulations. However, as shown by eq 3.35, the relative free energy between neighboring images will be closely related to the difference in the PMF only if the force constant of the restraining potential is very large. The differences observed in Figure 3.5 are due to the fact that the force constant of the restraining potential for the distance is not very large in these simulations.

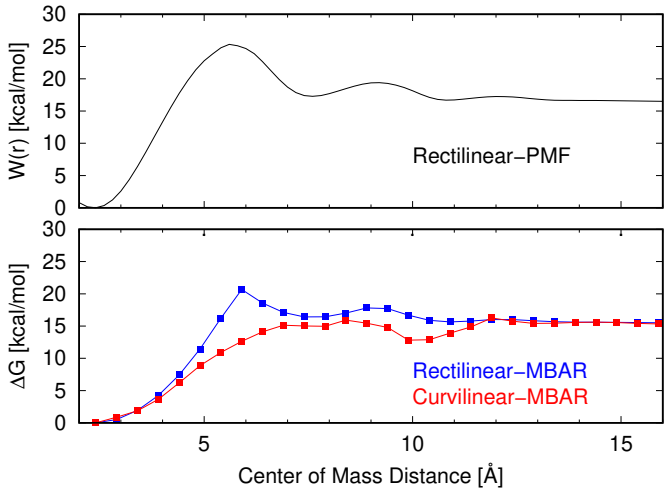

Figure 3.5: Separation free energy for the CB[7] system. (Top) The one-dimensional PMF as a function of the distance r along the straight axis . (Bottom) Cumulative free energy changes along the rectilinear (blue) and curvilinear (red) string pathways calculated from eq 4.2.

This example strengthen our confidence in the validity of the three approaches. Furthermore, all three methods yield quantitatively consistent results. While the original PMF-based

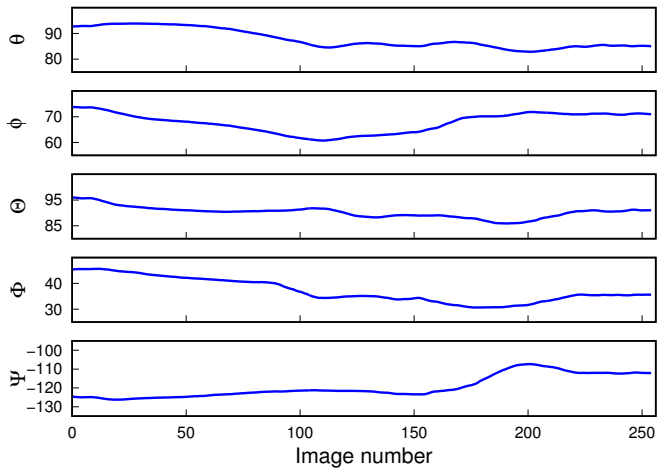

Figure 3.6: Variations of the collective variables along the minimum free energy path obtained for the wild-type barnase–barstar system.

formulation with a rectilinear separation path is formally correct,[9] and can yield correct results (Table 3.1), it relies on the arbitrary choice of a given axis to carry out the separation of the two molecules. Even when the choice appears intuitively obvious, the system can encounter large barriers during the separation of the two binding partners. The example of the CB[7] system shows that the issue can occur even with exceedingly simple host-guest complexes. Generally, it may be expected that the presence of high free energy barriers will affect the rate of convergence of the results, even if formally the converged answer is not supposed to be affected by any arbitrary choice. Those issues can be avoided by using an optimized MFEP curvilinear pathway determined by the string method that follows the local minima on the free free energy surface.

3.4.2 Barnase–barstar complex

The pathway for the separation of the barnase–barstar complex was generated by optimizing the images of the string on the local free energy surface using the DHAM method. The variations of the CVs along the string is shown in Figure 3.6. As the separation between the two proteins increase, the variation in the CVs becomes smaller.

Table 3.2: Barnase–barstar wildtype system free energies components

	PMF method (eq 3.22)	String method (eq 3.16)	
Separation pathway:	rectilinear	rectilinear	curvilinear
Contributions			
$\Delta\Delta G_{\mathrm{Bs,bb}}$	+0.9	+0.9	+0.9
$\Delta\Delta G_{\mathrm{Bn,bb}}$	−0.7	−0.7	−0.7
$\Delta\Delta G_{\mathrm{Bs,sc}}$	+6.7	+6.7	+6.7
$\Delta\Delta G_{\mathrm{Bn,sc}}$	+3.2	+3.2	+3.2
$\Delta\Delta G_{\mathrm{o}}$	+5.8	+5.8	+5.8
ΔG_{a}	-0.4	-0.4	-0.4
ΔG_{r}		-0.3	-0.3
F_{t}		+2.6	+2.6
$k_{\mathrm{B}}T\ln(S\times I\times C^{\circ})$	−24.0		
ΔG_{sepR}		−26.4	−24.2
$\Delta G_{\mathrm{PB-correction}}$	−2.5	−2.5	−4.4
$\Delta G^{\circ}_{\mathrm{bind}}$	**−11.0**	**−11.1**	**−10.8**

Errors estimated from bootstrapping are on the order of 0.05 kcal/mol for each free energy component.

The main results are provided in Table 3.2. The final binding free energy, $\Delta G^{\circ}_{\mathrm{bind}}$, is -11.0, -11.1, -10.8 for the three methods tested here. Otherwise noted, the error estimates based on bootstrapping are confined within 0.05 kcal/mol for each component. RMSD conformational restraints ($\Delta\Delta G_{\mathrm{Bs,bb}}, \Delta\Delta G_{\mathrm{Bn,bb}}, \Delta\Delta G_{\mathrm{Bs,sc}}, \Delta\Delta G_{\mathrm{Bn,sc}}$) and orientational restraints $\Delta\Delta G_{\mathrm{o}}$ converged fast. Because both paths share the same first image, the free energy components associated with the restraints $\mathbf{Z}^{(1)}[\mathbf{r}^{(1)}, \mathbf{\Omega}^{(1)}]$ are identical.

For a quantitative comparison of the different methods, it is important to account for the fact that the two proteins may retain some significant interactions at the last image of the string, taken to represent the unbound state. As the two proteins are separated by a fairly large distance at the last image of the string, any remaining interactions should be dominated by long-range electrostatics effects. In this context, a reasonable approximation to the magnitude of the interactions is Poisson-Boltzmann (PB) theory. A similar strategy has been previously used by Hunenberger and McCammon to analyze the artifacts of periodicity in explicit solvent simulations.[84] The long-range correction is provided in Table 3.2. It is clear that accounting for these long-range interactions is necessary to obtain consistent results from the three computational approaches. Interestingly, the correction is larger for

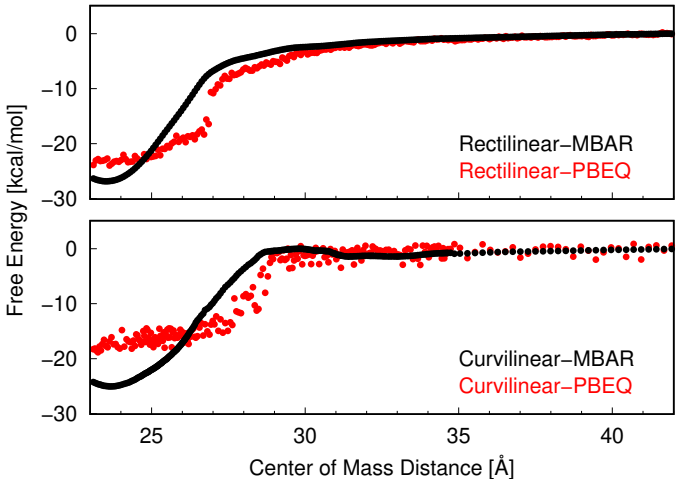

Figure 3.7: Cumulative free energy changes along the reaction pathway for the wild-type barnase–barstar system calculated with MBAR (black) and PB (red). The last window of MBAR results is set to 0 and the average of last 10 windows is set to 0 for PB results.

the curvilinear string pathway (-4.4 kcal/mol) compared to the rectilinear pathway (-2.5 kcal/mol). Even though both systems have identical center-of-mass protein-protein distance at maximum separation (last image of the string), the remaining electrostatic interaction differ considerably. The more favorable interaction at maximum separation between the two protein for the curvilinear pathway is consistent with the observation that the relative position and orientation of the two proteins is actually optimized on the local free energy surface along the MFEP. While one might be able to reduce the magnitude of the remaining long-range interactions by separating the two proteins further, this would impact the computational cost by necessitating additional images along the string and an increase of the size of the simulation system. Alternatively, one may be able to shield the long-range interactions by increasing the salt concentration, though this may not be desirable as it may not match the conditions used in experiments. Ultimately, estimating the long-range correction via a PB calculation offers a practical and valid approach. The higher variance of the continuum PB interactions along the curvilinear path shown in Figure 3.7 makes it clear that the relative orientation of two proteins is a key factor that affects the long-range interaction.[84] Remarkably, Figure 3.7 shows that a simple continuum electrostatic approximation begins

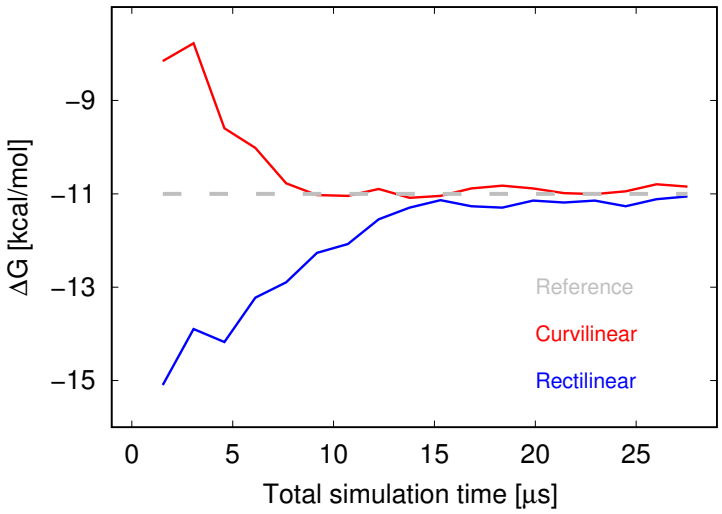

Figure 3.8: Total binding free energy change along cumulative simulation time. 20% or 1 ns/window, whichever smaller of earlier trajectory is discarded as equilibration.

to accurately match the free energy profile determined from explicit-solvent simulations at a center-of-mass distance of about 29 Å, namely when the two proteins are separated by about 5 Å away from the bound complex (from 24 Å to 29 Å). This observation suggests that a PB approximation may help engineer practical protocols seeking to reduce the computational cost of protein-protein binding free energy calculations.

A central issue that this work aims to address is to compare the rate of convergence of the full calculation for the different methods. The final estimated binding free energy calculated from the rectilinear and optimized curvilinear pathways is shown in Figure 3.8. In both cases the configurational sampling was enhanced using replica-exchange with solute tempering (REST2). Figure 3.8 shows that the convergence is indeed faster using the curvilinear path. While both paths yields -10.8 and -11.1 kcal/mol for total binding free energy, the result from the curvilinear pathway appears to converge around 8 μs of total simulation time, whereas the result from the rectilinear pathway starts to converge only around 15 μs. This suggests that the curvilinear pathway converges in about 50% of the total simulation time needed for a rectilinear (straight) pathway.

It is of great interest to compare the present results with experiment and with other MD

calculations. Experimentally, the barnase–barstar complex has a binding free energy on the order of −19.0 kcal/mol, while the present results are consistently around -11 kcal/mol. One possible factor that may affect the result is the pH and the protonation state of His102 in barnase.[38–40] It has been suggested that His102 must be protonated in the bound complex and that the proteins do not bind when His102 is neutral.[38–40]. ΔpKa was calculated with a continuum electrostatic PB equation for both crystal structure and snapshots from simulated trajectory. Surprisingly, estimates calculated with PB using both the crystal structure and snapshots from simulated trajectory indicate that protonation of His102 is unfavorable to binding. The calculation yield a negative ΔpKa of -8.3 and positive $\Delta\Delta G^{\circ}_{\text{bind}}$ of 11.4 kcal/mol regardless of which neutral forms of histidine is considered. This result is consistent with alchemical explicit solvent FEP calculations to examine the protonation of His102. The free energy difference for the three systems (bound, unbound, and a 3-residue reference system Ala-His-Ala) is -14.72, -26.22, -19.05 kcal/mol, respectively. Therefore, both PB and alchemical FEP are indicative that the binding affinity of the barnase–barstar complex decreases when His102 is protonated.

Clearly, the calculated binding free energy differs considerably from the experimental value. Given the consistent results obtained from the rectilinear and curvilinear pathway, it is reasonable to conclude that issues of convergence are not the main cause of this large discrepancy. While the poor agreement with the experimental value is disappointing, it should not be entirely surprising that the details of the atomic force field will have a major impact on the binding free energy. For example, three arginine–aspartic acid pairs (Arg59-Asp76, Arg83-Asp39, Arg87-Asp39) provide extremely strong electrostatic interactions in the bound barnase–barstar complex. During simulations of the complex, these residues remain within an average minimum distance of 1.73, 1.77, 1.75 Å, respectively (calculated from the minimum distance of the closest atoms from each residue). The importance of Asp39 and Arg87 is also highlighted below in calculations based on the single and double mutants (D39A, R87A, and D39A/R87A). The sensitivity of the results to the force field

can be illustrated by considering the effect of a small modification of the Lennard-Jones parameters affecting the core repulsion between the key atoms of the positively charged arginine side chain (NC2) and the negatively charge carboxylate group (OC). The impact of such modification of the force field can simply be estimated with free energy perturbation (FEP) using the ensemble of configurations from the unperturbed complex as a reference system,

$$e^{-\Delta G_{\mathrm{pert}}/k_{\mathrm{B}}T} = \left\langle e^{-[U_{\mathrm{pert}}-U_{\mathrm{ref}}]/k_{\mathrm{B}}T} \right\rangle_{(U_{\mathrm{ref}})} \tag{3.36}$$

In the standard CHARMM force field, the R_{min} for these two atoms 3.55 Å. Reducing the R_{min} of this atom pairs by 0.1 Å to 3.45 Å shifts the binding affinity of those residues by -2.3, -1.7, -1.4 kcal/mol, yielding a total of -5.4 kcal/mol respectively. Such a change would shift the binding free energy to about -16.4 kcal/mol, which is closer to the experimental value. Interestingly, a calculation by Pan and co-workers based on the AMBER force field reported binding free energy of 19.2 kcal/mol, which is very close to the experimental result.[57]. One may note that the Lennard-Jones R_{min} for of the AMBER FF99SB force field for the corresponding pair of atoms 3.4852 Å, which might explain why the binding affinity for barnase–barstar complex is more favorable in this calculation. Another calculation based on the AMBER FF99SB force field by Noé and co-workers reported a binding free energy of 12-19 kcal/mol.[58] While the system simulated displays minor differences for a few residues on barstar (the cysteines at position 40 and 82 were substituted by alanine), their effect on the binding free energy should be minimal. This calculation was carried out in the context of a Markov State Model (MSM) analysis of the binding association of barnase and barstar, and for this reason there is a considerable uncertainty on the estimated binding free energy. In fact, the result depends on the MSM lag time, with estimates ranging over 8-22 kcal/mol when all lag times were considered.[58]

A previous calculations by Gumbart et al reported a binding free energy of -21.0 kcal/mol.[36] While the force field is the same as in the present study, a direct comparison is difficult because those calculations were carried out with neutral N- and C-terminal

whereas the present calculations were based on the zwiterionic form of the proteins, with charged N- and C-terminal. This issue can be addressed by using FEP to estimate the free energy change implied by the differences in the N- and C-terminal. Using a continuum electrostatic PB calculation, we estimate that converting the zwiterionic N- and C-termini into neutral N- and C-termini will increase the binding free energy by -5.8 kcal/mol, which would yield a total binding free energy of -16.8 kcal/mol. The impact implied by the differences in the N- and C-terminal can also calculated by using alchemical FEP simulations, yielding a stabilization of -6.7 kcal/mol (FEP gives 203.7 kcal/mol for the bound state and 210.4 kcal/mol for the unbound state). Adding this perturbation to present results yield a binding free energy on the order of -17.7 kcal/mol, which is in closer agreement with the previous result of Gumbart et al.[36] The calculated binding free energy remains smaller than the previous estimate of -21.0 kcal/mol, although it is likely that the total amount of sampling is responsible for the discrepancy; the present calculations is based on 27.54 μs of sampling, whereas Gumbart et al. used a total of 212 ns of sampling.[36]

3.4.3 Results for mutant cycles

To further validate the computational methodology and the accuracy of the force field, we investigate the influence of single and double mutations on the association of barnase–barstar.[40,41,85] Mutant cycle are often used to determine if two residues on each proteins are close each other in the bound complex; proximal pairs tend to contribute cooperatively to the binding affinity, whereas the distant pairs contribute independently.[40,41] The effective residue-residue interaction is commonly quantified by considering $\Delta\Delta G = G(a,b) - G(a^*,b) - G(a,b^*) + G(a^*,b^*)$, where a and b represent the wild-type residues on barnase and barnase, and a^* and b^* represent the mutated residues on barnase and barnase, respectively. This type of analysis is often used to derive geometric constraints used to determine the structure of protein-protein complexes.[86,87] However, many central issues have not been address by the mutant cycle experiments. For example, under what set of conditions can the quantity

$\Delta\Delta G$ be interpreted simply as an effective interaction energy between localized residues? This type of question that is not easily answered by experiments alone. Conclusions have been drawn from simple models,[88] but a deeper understanding will be gained from detailed atomic computations. As an example, it has been concluded that the interaction between residue Arg87 on barnase and Asp39 on barstar contributes cooperatively the binding by more than 6 kcal/mol.[41] Preliminary PMF-based binding calculations were carried out on the wild type (WT) complex, single mutants D39A (barstar) and R87A (barnase), and double mutant D39A,R87A. The results given in Table 3.3 are very encouraging. For instance, the change caused by the double and single mutants is correctly reproduced, indicating that the computations are likely to yield meaningful results. Furthermore, it is important to study a number of mutant cycles (single and double mutations) to assess the validity of the computational method. By recapitulating the double mutant experiments *in silico*, simulations can generate new knowledge that was previously not available. The barnase–barstar system provides a unique opportunity to examine the influence of mutations on the binding affinity due to the large amount of experimental data.

Table 3.3: Barnase–barstar mutant system free energies components.

Contribution	wild-type	D39A	R87A	D39A/R87A
$\Delta\Delta G_{\mathrm{Bs,c}}$	+0.9	−0.2	−0.4	+0.6
$\Delta\Delta G_{\mathrm{Bn,c}}$	−0.7	−0.4	+0.3	−0.1
$\Delta\Delta G_{\mathrm{Bs,res}}$	+6.7	+0.9	+3.6	+3.0
$\Delta\Delta G_{\mathrm{Bn,res}}$	+3.2	+2.6	+1.6	+2.0
$\Delta\Delta G_{\mathrm{o}}$	+5.4	+5.2	+5.7	+5.5
$\Delta\Delta G_{\mathrm{t}}$	+2.3	+2.2	+2.3	+2.4
ΔG_{sepR}	−24.2	−13.6	−15.8	−15.3
$\Delta G_{\mathrm{PB-correction}}$	−4.4	−3.6	−3.1	−3.2
$\Delta G^{\circ}_{\mathrm{bind}}$	−10.8	−6.9	−5.8	−5.1
$\Delta G^{\circ}_{\mathrm{exp}}$	−19.0	−11.3	−13.5	−11.9

The computational results for all four mutant barnase–barstar complexes are compared in Table 3.3. The experimental results are taken from the work of Schreiber and Fersht.[40,41] Binding is clearly less favorable for all three mutants, but the order of the binding affinity does not exactly match. Nevertheless, the effective residue-residue interaction extracted from

the mutant cycle, $\Delta\Delta G$, is reasonable. The $\Delta\Delta G$ extracted from the experimental data is -19+11.3+13.5-11.9 = -6.1 kcal/mol,[40,41] and the $\Delta\Delta G$ extracted from the calculation is -10.8+6.9+5.8-5.1 = -3.2 kcal/mol. The moderate agreement shows the validity and the sensitivity of the computational methodology to extract an effective residue-residue interaction in the context of four related systems.

3.5 Conclusion

This study formulated a new theoretical framework to calculate the binding free energy of using a physical separation of two binding partners along a curvilinear pathway determined by the string method. The theoretical relationship between the new method and the PMF-based method was clarified. The new formalism was validated by comparing the results obtained using both rectilinear and curvilinear pathways for a prototypical host–guest complex formed by CB[7] binding benzene, and the barnase–barstar protein complex. The minimum free energy path (MFEP) was defined and optimized in a 6-dimensional subspace with collective variables $\mathbf{Z}[\mathbf{r}^*, \mathbf{\Omega}^*]$. The string was optimized through either a local free energy surface with DHAM, or by successively averaging the CVs of the images along the string. Since the rectilinear path can be used to calculate the total binding free energy in both ways, as if the path were 'straight string', it was shown with host-guest system that two approaches result in the same binding free energy as well as the curvilinear path. The barnase–barstar system was used to demonstrate that the binding free energy calculated on the basis of a curvilinear pathway converges faster than its rectilinear counterpart. One complete mutant cycle was simulated to examine the ability of the method to capture correctly the effective residue-residue coupling compared to the experimental result. The effect of long-range electrostatic interactions was considered using PB calculations. Improved algorithms to find minimum free energy separation path could save significant amount of simulation time, especially when there is a high free energy barrier along the chosen rectilinear path.

CHAPTER 4

PROTEIN-PROTEIN ASSOCIATION/RECOGNITION PROBLEMS

4.1 Introduction

The protein-protein interaction governs many biological processes and plays a key role in a living cell, such as cell communications. Thanks to the crystallography, high resolution structural data became widely available. However, even with the knowledge of protein structures in the living cell was not sufficient to understand the function of highly organized cellular systems, and now we seek for deep understanding upon dynamical networks arise from specific protein-protein interactions.

Some proteins bind together, others do not, despite of the structural similarity. Knowing when and why proteins can specifically associate and bind together would be extremely helpful. This must begin with the statistical mechanics concept of binding free energy ΔG_{b}. A computational approach based on atomistic MD simulation and free energy methodology would be advantageous to tackle this problem. Unlike docking/scoring methods that shows high variance of performance from one system to another, all-atomistic MD simulation does not rely on any particular empirical assumptions about the binding, boasting steady performance.[89]

One might observe slow convergnece of the binding simulation with the complex free energy surface. To resolve the issue, we would like to disect the problem into a few steps where a set of geometric and conformational restraints is applied to reduce the degrees of freedom of two proteins and their relative orientation. This lessens sampling noise and accelerates the convergence of the resulting free energy profile.[9,35,36,44] The total binding free energy can be calculated from a sum of individual free energy components of adding/releasing restraints and the separation. The sampling rely on the data harvested from a collection of copies or "replicas" of the molecular system with Hamiltonian-tempering replica-exchange

MD ($\mathcal{H}$-REMD)simulations.[71,72,90,91] Such multiple copy algorithms (MCAs) offer a general and powerful strategy to enhance the sampling efficiency of conventional MD simulations.[71,72,90–92]

With these methods, two specific biomolecular questions we would like to tackle are HIV-1 Nef /SH3 and Colicin/Im9 complexes.

4.2 Methods

Again, how to obtain main target equilibrium binding constant K_{eq} between two proteins is explained in section 3.2.

In previous chapter 3, we introduced string method formulation of protein-protein binding free energy which is based on the physical separation of one protein from the other along a curvilinear pathway.[44] The binding free energy is determined in a step-by-step procedure in which the bound proteins are progressively separated along a curvilinear pathway in the presence of orientational and conformational restraining potentials. Initially, conventional PMF method that is separating two bound partners along a predefined rectilinear pathway was developed and used by Roux and co-workers[9,35,36]. However, free-energy calculations are expected to converge faster when the model reaction coordinate coincides with minimum free-energy path.[6–8] Utilizing the curvilinear path found with DHAM method, we expect the simulation to converge more rapidly.[44]

In our formulation based on the string method,[8,59] the curvilinear separation pathway is represented by the collective variables along the path, each window having $\mathbf{Z} = (\mathbf{r}, \mathbf{\Omega})$. The path of N images in 6 dimensional collective variables is $\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots \mathbf{Z}^{(N)}\}$, where the relative position of two proteins is defined with center-of-mass distance with two polar angles $\mathbf{r} = (r, \theta, \phi)$, and the relative orientation is defined with three Euler angles $\mathbf{\Omega} = (\Theta, \Phi, \Psi)$.

These relative positional and orientational degrees of freedom are defined with six internal reference points in two proteins, three in the first protein (P_1, P_2, and P_3), and three in the second protein (P_1', P_2', and P_3'). The associated restraining potential in collective variables

is $u(\mathbf{Z}) = u_{\mathrm{t}}(\mathbf{r}) + u_{\mathrm{a}}(\theta,\phi) + u_{\mathrm{o}}(\mathbf{\Omega})$. The last, N^{th} image corresponds to the unbound state, wherein $\mathbf{Z}^{(M)} \equiv (\mathbf{r}^*, \mathbf{\Omega}^*)$, with the given orientation $\mathbf{\Omega}^* = (\Theta^*, \Phi^*, \Psi^*)$. The i^{th} image is simulated with the harmonic restraining potential, $u^{(i)}(\mathbf{Z}) = \frac{1}{2}\ \mathbf{k} \cdot \left(\mathbf{Z} - \mathbf{Z}^{(i)}\right)^2$, where $\mathbf{k} = (k_r, k_\theta, k_\phi, k_\theta, k_\Phi, k_\Psi)$, is the force constant, and $\mathbf{Z}^{(i)}$ denotes the set of reference values for the harmonic restraining potentials corresponding to the i^{th} image. Using this definition, the equilibrium binding constant K'_{eq} of the system with RMSD restraining potential is written as,

$$K'_{\mathrm{eq}} \quad = \quad \mathrm{e}^{\Delta G^{(1)}_{\mathrm{bound}}/k_{\mathrm{B}}T} \times \left(\prod_{i=1}^{N-1} \mathrm{e}^{-\Delta G^{(i,i+1)}//k_{\mathrm{B}}T} \right) \times \mathrm{e}^{-\Delta G^{(N)}_{\mathrm{unbound}}/k_{\mathrm{B}}T} \tag{4.1}$$

The middle terms between the parentheses correspond to simple relative free energies of the type,

$$e^{-\Delta G^{(i,i+1)}/k_{\mathrm{B}}T} = \frac{\int d\mathbf{X}\ \mathrm{e}^{-[U' + u^{(i)}]}/k_{\mathrm{B}}T}{\int d\mathbf{X}\ \mathrm{e}^{-[U' + u^{(i+1)}]/k_{\mathrm{B}}T}} \tag{4.2}$$

These terms represent the free energy along the optimal string pathway for substituting the restraining potential of the i^{th} image on the collective variables from the restraining potential of the $(i+1)^{\mathrm{th}}$ image. These free energy differences between neighboring windows can be evaluated using the histogram-less weighted histogram analysis method (WHAM),[60] or Bennett acceptance ratio (BAR),[61] assuming there is sufficient overlap between two contiguous images, thereby obviating the need of an explicit separation PMF.

To enhance the sampling of the images along the curvilinear pathway, MD is combined with a replica-exchange (RE) algorithm.[44,71,90–92] In the REMD approach, multiple copies of the molecular system are simulated concurrently, but all with different conditions, e.g., different temperatures or Hamiltonians.[92] Attempts are periodically made to exchange configurations or parameters between different systems using a Metropolis Monte Carlo acceptance criterion, insuring Boltzmann-weighted statistics. Furthermore, the images along the string are sampled with the Hamiltonian-tempering replica-exchange MD solute scaling

(REST2).[77]

In summary, to calculate binding free energies we will use a novel computational method developed in our lab.[44] Briefly, the method seeks to calculate the reversible work needed to physically separate the two proteins in solution,[36] but rather than using a rectilinear pathway as used previously,[9,35] we will determine the optimal separation pathway of the two proteins using the string method. The free energy between the two end-states of the string will be calculated by combining the information of multiple biased "images" of the two proteins restrained at regularly interspaced distances between monomers along the pathway.[44] Those biased simulations are akin to umbrella sampling, though they will be used to extract free energy steps between adjacent images. Finally, additional simulations are needed to release the end states and obtain the proper standard binding free energy between the two proteins.[71] To speed up the convergence of these calculations and sample more efficiently, $\mathcal{H}$-REMD is employed.[72,93]

4.3 Future directions

HIV-1 Nef/SH3 Complex

HIV–1 Nef/SH3 complex is an example of single residue substitution leads to a large change in K_{d}. Protein Nef of HIV–1 forms a complex with both the SH3 domain of Fyn, a tyrosine kinase that binds Nef poorly, and a variant in which a point substitution has been introduced to mimic Hck, which binds Nef much more tightly than Fyn.[43] This makes the HIV–1 Nef/SH3 complex an interesting system because most mutation causes decrease in binding affinity whereas the mutation introduced to SH3 domain (R96I) increases binding affinity significantly. The relative orientations of binding are affected by the R96I mutation. The key interaction residues discovered by structural analysis and the simulation of both wildtype/R96I mutant will be compared to gain additional knowledge on the mechanism of increased binding affinity. To find out why R96I mutation imposes a critical impact to the

binding affinity, the two systems of two proteins (wildtype and R96I mutant) are physically separated along an optimal pathway determined by the string method.[44,59]

Colicin/Im9 Complex

The colicin/immunity protein system provides some remarkable examples of cognate vs. noncognate assemblies.[45] The strain of *Escherichia coli* that makes colicin E9, endowed with a DNase activity, also produces the Im9 immunity protein that inhibits it very efficiently; a different strain produces the Im2 immunity protein, which has a much lower affinity for E9.[94] Among the non-cognate interactions Im2 shows the strongest binding toward the E9 DNase domain with a K_{d} of 10^{-8} M, 6 orders of magnitude weaker than that of the cognate immunity protein Im9. Interestingly, the structure of Im2 and Im9 have no significant differences, and the bound orientation with respect to the DNase are very similar each other. For a computational methodology, this protein complex is challenging because the bound conformation and the unbound conformation is very about 1.5 Å measured by the interface RMSD (i-RMSD) whereas the other protein complexes have i-RMSD of values less than 1 Å. Additional sampling method might be necessary for efficient sampling of the conformational transition upon binding. To find out if given protein complex is cognate or non-cognate, the two proteins are physically separated along an optimal pathway determined by the string method.[44,59] Whether the complex is cognate or non-cognate is to be detected with the steepness of the free energy difference when two proteins are being separated.

The string protein separation pathway for both protein complexes has been generated and all the images along the pathway have been solvated and equilibrated. The extensive sampling to be carried out with $\mathcal{H}$-REMD simulation is ready to go.

APPENDIX

Appendix A: Effect of pH on the observed binding free energy

We consider the noncovalent binding of of a ligand L to a protein P (the treatment is valid also for two proteins). But there is a group in one of the two molecules that can be protonated/deprotonated when the complex is formed and the pH can affect the equilibrium binding process. The equilibrium when the protein is unprotonated may be expressed as $L + P^{\mathrm{u}} \rightleftharpoons LP^{\mathrm{u}}$, with the equilibrium constant

$$K_{\mathrm{b}}^{\mathrm{u}} = \frac{[LP^{\mathrm{u}}]}{[L][P^{\mathrm{u}}]}$$

Correspondingly, the equilibrium when the protein is protonated may be expressed as $L + P^{\mathrm{p}} \rightleftharpoons LP^{\mathrm{p}}$, with the equilibrium constant

$$K_{\mathrm{b}}^{\mathrm{u}} = \frac{[LP^{\mathrm{u}}]}{[L][P^{\mathrm{u}}]}$$

The protonation/deprotonation equilibrium can be expressed in the following way

$$\theta_P = \frac{[P^{\mathrm{p}}]}{[P^{\mathrm{u}}]} = 10^{(\mathrm{pKa(P)}-\mathrm{pH})}$$

and

$$\theta_{LP} = \frac{[LP^{\mathrm{p}}]}{[LP^{\mathrm{u}}]} = 10^{(\mathrm{pKa(LP)}-\mathrm{pH})}$$

from the Henderson-Hasselbalch equation. We want to know the fraction of bound protein (regardless of the ionization state),

$$\begin{aligned} f &= \frac{[LP^{\mathrm{u}}] + [LP^{\mathrm{p}}]}{[P_{\mathrm{tot}}]} \\ &= (1 + \theta_{LP}) \frac{[LP^{\mathrm{u}}]}{[P_{\mathrm{tot}}]} \end{aligned}$$

We have the condition

$$\begin{aligned} [P_{\mathrm{tot}}] &= [P^{\mathrm{u}}] + [P^{\mathrm{p}}] + [LP^{\mathrm{u}}] + [LP^{\mathrm{p}}] \\ &= (1 + \theta_P)[P^{\mathrm{u}}] + (1 + \theta_{LP})[LP^{\mathrm{u}}] \end{aligned}$$

From this condition we have

$$[P^{\mathrm{u}}] = \frac{1}{(1 + \theta_P)} \left([P_{\mathrm{tot}}] - (1 + \theta_{LP})[LP^{\mathrm{u}}] \right)$$

We can combine this with the equilibrium equation ($[LP^{\mathrm{u}}] = K_{\mathrm{b}}^{\mathrm{u}}[L][P^{\mathrm{u}}]$) to get

$$\begin{aligned} [LP^{\mathrm{u}}] &= K_{\mathrm{b}}^{\mathrm{u}}[L] \frac{1}{(1 + \theta_P)} \left([P_{\mathrm{tot}}] - (1 + \theta_{LP})[LP^{\mathrm{u}}] \right) \\ &= \frac{1}{(1 + \theta_P)} K_{\mathrm{b}}^{\mathrm{u}}[L][P_{\mathrm{tot}}] - \frac{(1 + \theta_{LP})}{(1 + \theta_P)} K_{\mathrm{b}}^{\mathrm{u}}[L][LP^{\mathrm{u}}] \end{aligned}$$

$$[LP^{\mathrm{u}}] + \frac{(1 + \theta_{LP})}{(1 + \theta_P)} K_{\mathrm{b}}^{\mathrm{u}}[L][LP^{\mathrm{u}}] = \frac{1}{(1 + \theta_P)} K_{\mathrm{b}}^{\mathrm{u}}[L][P_{\mathrm{tot}}]$$

$$[LP^{\mathrm{u}}] \left(1 + \frac{(1 + \theta_{LP})}{(1 + \theta_P)} K_{\mathrm{b}}^{\mathrm{u}}[L] \right) = \frac{1}{(1 + \theta_P)} K_{\mathrm{b}}^{\mathrm{u}}[L][P_{\mathrm{tot}}]$$

$$\frac{[LP^{\rm u}]}{[P_{\rm tot}]} = \frac{\frac{1}{(1+\theta_P)}K_{\rm b}^{\rm u}[L]}{\left(1+\frac{(1+\theta_{LP})}{(1+\theta_P)}K_{\rm b}^{\rm u}[L]\right)}$$

$$\begin{aligned} f &= \frac{\frac{(1+\theta_{LP})}{(1+\theta_P)}K_{\rm b}^{\rm u}[L]}{1+\frac{(1+\theta_{LP})}{(1+\theta_P)}K_{\rm b}^{\rm u}[L]} \\ &= \frac{K_{\rm b}^{\rm eff}[L]}{1+K_{\rm b}^{\rm eff}[L]} \end{aligned}$$

where the effective (apparent) equilibrium binding constant is

$$K_{\rm b}^{\rm eff} = \frac{(1+\theta_{LP})}{(1+\theta_P)}K_{\rm b}^{\rm u}$$

or equivalently,

$$K_{\rm b}^{\rm eff} = \frac{1+10^{({\rm pKa(LP)-pH})}}{1+10^{({\rm pKa(P)-pH})}}K_{\rm b}^{\rm u}$$

In the case of barstar-barnase, it is assumed that the pKa of His102 is 6.4 for the isolated protein. At pH 8, then $10^{({\rm pKa(P)-pH})}$ is 0.0039 and is essentially negligible. The sidechain His102 is neutral and unprotonated for the isolated protein. However, it is believed that the sidechain is ionized and protonated in the complex. This will increase the apparent equilibrium binding constant in the following way

$$K_{\rm b}^{\rm eff} \approx 10^{({\rm pKa(LP)-pH})}K_{\rm b}^{\rm u}$$

and the apparent (effective) binding free energy as

$$\Delta G_{\rm b}^{\rm eff} \approx -k_{\rm B}T\,\ln(10)\,({\rm pKa(LP)-pH}) + \Delta G_{\rm b}^{\rm u}$$

So you can see that the apparent binding free energy will be more negative is the pKa(LP)

has been shifted to be larger than the pH=8.

These considerations are important to interpret the results of the free energy calculations. Let $U_{\mathrm{p}}(\mathbf{r})$, $U_{\mathrm{u}}(\mathbf{r})$, represent the potential energy function for the protonated (p) and unprotonated (u) system, respectively. Let us define the free energy difference $\Delta G = G_{\mathrm{p}} - G_{\mathrm{u}}$,

$$e^{-\beta \Delta G} = \frac{\int d\mathbf{r}\ e^{-\beta U_{\mathrm{p}}(\mathbf{r})}}{\int d\mathbf{r}\ e^{-\beta U_{\mathrm{u}}(\mathbf{r})}} \tag{4.3}$$

It is assumed that, when the molecule is unprotonated, the proton is transferred to a reference ionizable group located far away in solution (i.e., the number of particles in the system is the same for the two states). In practice, the proton is treated as a dummy non-interacting particle for the unprotonated state. That is, the dummy H atom is never annihilated but its interaction with the environment are turned off. In this alchemical process, a number of internal MM covalent terms (bonds, angles, dihedrals) are kept to avoid the problem of a wandering non-interacting free proton in the simulation; their influence cancels out in the treatment of standard state.

In principle, ΔG ought to be directly related to the experimental pKa of the site in the molecule. However, to match the experimental pKa value with a molecular mechanical (MM) force field, it is necessary to introduce an offset constant C,

$$10^{\mathrm{pKa}} = \frac{\int d\mathbf{r}\ e^{-\beta [U_{\mathrm{p}}(\mathbf{r}) - C]}}{\int d\mathbf{r}\ e^{-\beta [U_{\mathrm{u}}(\mathbf{r})]}} \tag{4.4}$$

The use of such an empirical offset constant C is necessary to calibrate the method in order to match the absolute pKa of an ionizable group because the MM potential energy function is not designed to account the true quantum mechanical proton affinity.

$$\mathrm{pKa} = \frac{1}{k_{\mathrm{B}} T \ln(10)} \left(-\Delta G + C\right) \tag{4.5}$$

Here, C is an offset constant C that depends only on the type of the titratable site, but

is otherwise assumed to not vary with the sequence and the conformation of the protein. We could determine the empirical offset constant C for a simple reference system (like a dipeptide in solution) and then assume that this value is transferable to the protein. Under this assumption, the pKa shift for the sidechain for the isolated protein P relative to the reference system is

$$\mathrm{pKa}(P) - \mathrm{pKa}(\mathrm{ref}) = -\frac{1}{k_{\mathrm{B}}T\ln(10)}\left(\Delta G(P) - \Delta G(\mathrm{ref})\right) \tag{4.6}$$

and the corresponding shift in the ligand-protein complex LP is

$$\mathrm{pKa}(LP) - \mathrm{pKa}(\mathrm{ref}) = -\frac{1}{k_{\mathrm{B}}T\ln(10)}\left(\Delta G(LP) - \Delta G(\mathrm{ref})\right) \tag{4.7}$$

where $\Delta G(\mathrm{ref})$, is the free energy to protonate the sidechain in the reference system, $\Delta G(P)$ is the free energy to protonate the sidechain in the isolated protein, and $\Delta G(LP)$ is the free energy to protonate the sidechain in the ligand-protein complex.

Appendix B: DHAM

The PMF in the six-dimensional space of collective variables was determined using the dynamic histogram analysis method (DHAM). The biased Markov transition probability matrices, $M^{(k)}$ is related to the underlying unbiased Markov transition probability matrix, M, as follows,

$$M_{ij}^{(k)} = f_i^{(k)} c_{ij}^{(k)} M_{ij} \tag{4.8}$$

where $c_{ij}^{(k)}$ and $f_i^{(k)}$ are the coefficient matrix and the normalization factor for simulation k, respectively. A maximum likelihood solution for unbiased Markov transition probability matrix is shown to be,

$$M_{ij} = \frac{\sum_{k=1}^{N_{\text{sim}}} T_{ij}^{(k)}}{\mu_i + \sum_{k=1}^{N_{\text{sim}}} n_i^{(k)} f_i^{(k)} c_{ij}^{(k)}} \tag{4.9}$$

where $T_{ij}^{(k)}$ is a transition count matrix in bin i and j and $n_i^{(k)}$ is a number of sample in bin i in simulation k. μ_i is a normalization constant that makes $\sum_j M_{ij} = 1$. f_i and μ_i can be solved explicitly, however, an approximate solution can be used instead. In US, each window has a different bias potential $u^{(k)}$ applied, which typically is in the form of harmonic potential, which uses $f_i^{(k)}$

$$u_i^{(k)} = \frac{1}{2} K^{(k)} (x_i - x^{(k)})^2 \tag{4.10}$$

where $x^{(k)}$ is the center of the harmonic potential and x_i is the i th bin along the reaction coordinate. $K^{(k)}$ is the force constant. At the limit of very short time step, the presence of a biasing potential changes the transition probability to

$$M_{ij}^{(k)} \propto M_{ij} \exp\left[-\frac{1}{2 k_{\text{B}} T} \left(u_j^{(k)} - u_i^{(k)} \right) \right] \tag{4.11}$$

which is similar to a spatial discretization of the Smoluchowski diffusion equation. The approximate solution assumes the following relationship for a short lag time, which is derived by imposing detailed balance condition, $M_{ij}p_i = M_{ji}p_j$ and $M_{ij}^{(k)}p_i^{(k)} = M_{ji}^{(k)}p_j^{(k)}$,

$$f_i^{(k)}c_{ij}^{(k)} = \exp\left[-\left(u_j^{(k)} - u_i^{(k)}\right)/2k_{\mathrm{B}}T\right] \tag{4.12}$$

Thus, at the limit of short lag time,

$$M_{ij} = \frac{M_{ij}^{\mathrm{unnorm}}}{\sum_{l=1}^{N_{\mathrm{bin}}} M_{il}^{\mathrm{unnorm}}} \tag{4.13}$$

and

$$M_{ij}^{\mathrm{unnorm}} = \frac{\sum_{k=1}^{N_{\mathrm{sim}}} T_{ij}^{(k)}}{\sum_{k=1}^{N_{\mathrm{sim}}} n_i^{(k)} \exp\left[-\left(u_j^{(k)} - u_i^{(k)}\right)/2k_{\mathrm{B}}T\right]} \tag{4.14}$$

REFERENCES

[1] M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nat. Struct. Mol. Biol.* **9**, 646–652 (2002).

[2] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).

[3] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57**, 97–109 (1970).

[4] J. P. Nilmeier, G. E. Crooks, D. D. L. Minh, and J. D. Chodera. Nonequilibrium candidate monte carlo is an efficient tool for equilibrium simulation. *Proc. Natl. Acad. Sci. U.S.A* **108**(45), E1009–E1018 (2011).

[5] D. Suh, B. K. Radak, C. Chipot, and B. Roux. Enhanced configurational sampling with hybrid non-equilibrium molecular dynamics-Monte Carlo propagator. *J Chem Phys* **148**(1), 014101 Jan 2018.

[6] L. Onsager and S. Machlup. Fluctuations and irreversible processes. *Phys. Rev.* **91**, 1505–1512 (1953).

[7] W. E, W. Ren, and E. Vanden-Eijnden. String method for the study of rare events. *Phys. Rev. B* **66**(5), 052301 (2002).

[8] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* **125**, 24106 Jul 2006.

[9] H. Woo and B. Roux. Calculation of absolute protein-ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 6825–6830 May 2005.

[10] T. Huber, A. E. Torda, and W. F. van Gunsteren. Local elevation: a method for improving the searching properties of molecular dynamics simulation. *J. Comput. Aided Mol. Des.* **8**(6), 695–708 December 1994.

[11] V. Leone, F. Marinelli, P. Carloni, and M. Parrinello. Targeting biomolecular flexibility with metadynamics. *Curr. Opin. Struct. Biol.* **20**(2), 148–154 (2010).

[12] X. Wu and B. R. Brooks. Self-guided langevin dynamics simulation method. *Chem. Phys. Lett.* **381**(3), 512–518 (2003).

[13] A. Damjanović, X. Wu, B. R. Brooks, et al. Backbone relaxation coupled to the ionization of internal groups in proteins: a self-guided langevin dynamics study. *Biophys. J.* **95**(9), 4091–4101 (2008).

[14] G. Koenig, X. Wu, and B. Brooks. Crossing energy barriers with self-guided langevin dynamics. In *Eur. Biophys. J. Biophys.*, volume 40, pages 108–109. SPRINGER 233 SPRING ST, NEW YORK, NY 10013 USA, (2011).

[15] A. F. Voter. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* **78**(20), 3908 (1997).

[16] D. Hamelberg, J. Mongan, and J. A. McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120**, 11919 (2004).

[17] M. Fajer, D. Hamelberg, and J. A. McCammon. Replica-exchange accelerated molecular dynamics (rexamd) applied to thermodynamic integration. *J. Chem. Theory Comput.* **4**(10), 1565–1569 (2008).

[18] W. Jiang and B. Roux. Free energy perturbation hamiltonian replica-exchange molecular dynamics (fep/h-remd) for absolute ligand binding free energy calculations. *J. Chem. Theory Comput.* **6**(9), 2559–2565 (2010).

[19] L. Maragliano and E. Vanden-Eijnden. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.* **426**(1), 168–175 (2006).

[20] H. Lei and Y. Duan. Improved sampling methods for molecular simulation. *Curr. Opin. Struct. Biol.* **17**(2), 187–191 April 2007.

[21] M. Christen and W. F. van Gunsteren. On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. *J. Comput. Chem.* **29**(2), 157–166 (2008).

[22] L. Wang, R. Friesner, and B. J. Berne. Replica exchange with solute scaling: A more efficient version of replica exchange with solute tempering (rest2). *J. Phys. Chem. B* **115**, 9431–9438 (2011).

[23] A. Mitsutake, Y. Mori, and Y. Okamoto. Enhanced sampling algorithms. *Methods Mol. Biol.* **924**, 153–195 (2013).

[24] R. C. Bernardi, M. C. Melo, and K. Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta* **1850**(5), 872–877 (2015).

[25] X. Wu, B. R. Brooks, and E. Vanden-Eijnden. Self-guided langevin dynamics via generalized langevin equation. *J. Comp. chem.* **37**(6, SI), 595–601 MAR 5 2016.

[26] M. Arrar, C. A. F. de Oliveira, M. Fajer, W. Sinko, and J. A. McCammon. w-REXAMD: A Hamiltonian replica exchange approach to improve free energy calculations for systems with kinetically trapped conformations. *J. Chem. Theory Comput.* **9**, 18–23 (2013).

[27] H. Stern. Molecular simulation with variable protonation states at constant ph. *J. Chem. Phys.* **126**, 164112 (2007).

[28] H. Stern. Erratum: "molecular simulation with variable protonation states at constant ph" [j. chem. phys.126, 164112 (2007)]. *J. Chem. Phys.* **127**, 079901 (2007).

[29] A. J. Ballard and C. Jarzynski. Replica exchange with nonequilibrium switches. *Proc. Natl. Acad. Sci. U.S.A* **106**(30), 12224–12229 JUL 28 2009.

[30] J. P. Nilmeier, G. E. Crooks, D. D. L. Minh, and J. D. Chodera. Nonequilibrium candidate Monte Carlo is an efficient tool for equilibrium simulation. *Proc. Natl. Acad. Sci. USA* **108**, E1009–E1018 (2011).

[31] Y. Chen and B. Roux. Enhanced sampling of an atomic model with hybrid nonequilibrium molecular dynamics–Monte Carlo simulations guided by a coarse-grained model. *J. Chem. Theory Comput.* **11**, 3572–3583 (2015).

[32] B. K. Radak and B. Roux. Efficiency in nonequilibrium molecular dynamics monte carlo simulations. *J. Chem. Phys.* **145**, 134109 (2016).

[33] Y. Chen and B. Roux. Efficient hybrid non-equilibrium molecular dynamics - Monte Carlo simulations with symmetric momentum reversal. *J. Chem. Phys.* **141**, 114107 (2014).

[34] J. Comer, J. C. Gumbart, J. Hénin, T. Lelièvre, A. Pohorille, and C. Chipot. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. *J. Phys. Chem. B* **119**, 1129–1151 (2015).

[35] J. C. Gumbart, B. Roux, and C. Chipot. Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *J. Chem. Theor. Comp.* **9**, 794–802 (2013).

[36] J. C. Gumbart, B. Roux, and C. Chipot. Efficient Determination of Protein-Protein Standard Binding Free Energies from First Principles. *J. Chem. Theory Comput.* **9**, 3789–3798 (2013).

[37] J. Lagona, P. Mukhopadhyay, S. Chakrabarti, and L. Isaacs. The cucurbit[n]uril family. *Ang. Chem.-Int. Ed.* **44**(31), 4844–4870 (2005).

[38] A. Buckle, G. Schreiber, and A. Fersht. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution. *Biochemistry* **33**, 8878–8889 (1994).

[39] G. Schreiber and A. R. Fersht. Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. *Biochemistry* **32**(19), 5145–5150 May 1993.

[40] G. Schreiber, A. M. Buckle, and A. R. Fersht. Stability versus function: Two competing forces in the evolution of barstar. *Structure* **2**, 945–951 (1994).

[41] G. Schreiber and A. R. Fersht. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J Mol Biol* **248**, 478–486 (1995).

[42] R. W. Hartley. Directed mutagenesis and barnase-barstar recognition. *Biochemistry* **32**(23), 5978–5984 Jun 1993.

[43] C.-H. Lee, K. Saksela, U. A. Mirza, B. T. Chait, and J. Kuriyan. Crystal Structure of the Conserved Core of HIV-1 Nef Complexed with a Src Family SH3 Domain. *Cell* **85**, 931–942 June 1996.

[44] D. Suh, S. Jo, W. Jiang, C. Chipot, and B. Roux. String method for Protein-Protein Binding Free Energy Calculation. *J Chem Theory Comput* ((submitted)).

[45] K. Okabayashi, A. Hasegawa, and T. Watanabe. Microreview: capsule-associated genes of Cryptococcus neoformans. *Mycopathologia* **163**(1), 1–8 January 2007.

[46] D. A. Sivak, J. D. Chodera, and G. E. Crooks. Using nonequilibrium fluctuation theorems to understand and correct errors in equilibrium and nonequilibrium simulations of discrete Langevin dynamics. *Phys. Rev. X* **3**, 011007 (2013).

[47] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).

[48] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).

[49] J. Comer, J. Phillips, K. Schulten, and C. Chipot. Multiple-replica strategies for free-energy calculations in namd: Multiple-walker adaptive biasing force and walker selection rules. *J. Chem. Theory Comput.* **10**, 5276–5285 (2014).

[50] K. Minoukadeh, T. Lelièvre, and C. Chipot. Potential of mean force calculations: A multiple-walker adaptive biasing force approach. *J. Chem. Theory Comput.* **6**, 1008–1017 (2010).

[51] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, Jr. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles. *J. Chem. Theory Comput.* **8**, 3257–3273 (2012).

[52] S. Park, F. Khalili-Araghi, E. Tajkhorshid, and K. Schulten. Free energy calculation from steered molecular dynamics simulations using jarzynski's equality. *J. Chem. Phys.* **119**, 3559–3566 (2003).

[53] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23**, 187–199 (1977).

[54] W. Jiang, Y. Luo, L. Maragliano, and B. Roux. Calculation of free energy landscape in multi-dimensions with Hamiltonian-exchange umbrella sampling on petascale supercomputer. *J. Chem. Theory Comput.* **8**, 4672–4680 (2012).

[55] Y. Deng and B. Roux. Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *J. Chem. Theor. Comp.* **2**, 1255–1273 (2006).

[56] Y. Q. Deng and B. Roux. Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B* **113**, 2234–2246 (2009).

[57] A. C. Pan, D. Jacobson, K. Yatsenko, D. Sritharan, T. M. Weinreich, and D. E. Shaw. Atomic-level characterization of protein-protein association. *Proc. Natl. Acad. Sci. U.S.A.* Feb 2019.

[58] N. Plattner, S. Doerr, G. De Fabritiis, and F. Noe. Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat Chem* **9**(10), 1005–1011 10 2017.

[59] A. Pan, D. Sezer, and B. Roux. Finding transition pathways using the string method with swarms of trajectories. *J. Phys. Chem. B* **112**, 3432–3440 Mar 2008.

[60] S. Kumar, D. Bouzida, R. Swendsen, P. Kollman, and J. Rosenberg. The Weighted Histogram Analysis Method for free-energy calculations on biomolecules. I. The method. *J. Comp. Chem.* **13**, 1011–1021 (1992).

[61] C. Bennett. Efficient estimation of free-energy differences from Monte-Carlo data. *J. Comp. Chem.* **22**, 245–268 (1976).

[62] M. E. Johnson and G. Hummer. Characterization of a dynamic string method for the construction of transition pathways in molecular reactions. *J Phys Chem B* **116**(29), 8573–8583 Jul 2012.

[63] L. Maragliano, B. Roux, and E. Vanden-Eijnden. Comparison between Mean Forces and Swarms-of-Trajectories String Methods. *J. Chem. Theor. Comp.* **10**(2), 524–533 FEB 2014.

[64] E. Rosta and G. Hummer. Free energies from dynamic weighted histogram analysis using unbiased Markov state model. *J Chem Theory Comput* **11**(1), 276–285 Jan 2015.

[65] A. T. Fenley, N. M. Henriksen, H. S. Muddana, and M. K. Gilson. Bridging Calorimetry and Simulation through Precise Calculations of Cucurbituril-Guest Binding Enthalpies. *J Chem Theory Comput* **10**(9), 4069–4078 Sep 2014.

[66] N. M. Henriksen, A. T. Fenley, and M. K. Gilson. Computational Calorimetry: High-Precision Calculation of Host-Guest Binding Thermodynamics. *J Chem Theory Comput* **11**(9), 4377–4394 Sep 2015.

[67] S. Jo, W. Jiang, H. Sun Lee, B. Roux, and W. Im. CHARMM-GUI Ligand Binder for Absolute Binding Free Energy Calculations and Its Application. *J. Chem. Inf. and Mod.* **53**, 267–277 (2013).

[68] A. J. MacKerell, D. Bashford, M. Bellot, R. Dunbrack, J. Evanseck, M. Field, S. Fischer, J. Gao, H. Guo, D. J.-M. S. Ha, L. Kuchnir, K. Kuczera, F. Lau, C. Mattos, S. Michnick, T. Ngo, D. Nguyen, B. Prodhom, W. Reiher III, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).

[69] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. Mackerell. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone Φ, Ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J Chem Theory Comput* **8**(9), 3257–3273 Sep 2012.

[70] 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**, 926–935 (1983).

[71] W. Jiang, Y. Luo, L. Maragliano, and B. Roux. Calculation of free energy landscape in multi-dimensions with hamiltonian-exchange umbrella sampling on petascale supercomputer. *J. Chem. Theor. Comp.* **8**, 4672–4680 (2012).

[72] W. Jiang, J. C. Phillips, L. Huang, M. Fajer, Y. Meng, J. C. Gumbart, Y. Luo, K. Schulten, and B. Roux. Generalized Scalable Multiple Copy Algorithms for Molecular Dynamics Simulations in NAMD. *Comput Phys Commun* **185**(3), 908–916 Mar 2014.

[73] J. Phillips, G. Zheng, S. Kumar, and L. Kal. Namd: Biomolecular simulation on thousands of processors. In *Proceedings of the IEEE/ACM SC2002 Conference*, page 277. IEEE Press, (2002).

[74] J. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comp. Chem.* **26**, 1781–1802 (2005).

[75] T. Darden, A. Toukmaji, and L. Pedersen. Long-range electrostatic effects in biomolecular simulations. *Journal de Chimie Physique et de Physico-Chimie Biologique* **94**, 1346–1364 (1997).

[76] M. Olsson, C. Sondergaard, M. Rostkowski, and J. Jensen. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK(a) Predictions. *J Chem Theory Comput* **7**, 525–537 (2011).

[77] S. Jo and W. Jiang. A generic implementation of replica exchange with solute tempering (REST2) algorithm in NAMD for complex biophysical simulations. *Comp. Phys. Comm.* **197**, 304–311 DEC 2015.

[78] S. Dixit and C. Chipot. Can absolute free energies of association be estimated from molecular mechanical simulations? The biotin-streptavidin system revisited. *J. Phys. Chem. A* **105**, 9795–9799 (2001).

[79] W. Im, D. Beglov, and B. Roux. Continuum solvation model: Electrostatic forces from numerical solutions to the Poisson-Bolztmann equation. *Comp. Phys. Comm.* **111**, 59–75 (1998).

[80] B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. Charmm: The biomolecular simulation program. *J. Comput. Chem.* **30**(10, Sp. Iss. SI), 1545–1614 JUL 30 2009.

[81] 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**, 5239–5248 (1997).

[82] H. S. Muddana, A. T. Fenley, D. L. Mobley, and M. K. Gilson. The SAMPL4 host-guest blind prediction challenge: an overview. *J. Comput. Aided Mol. Des.* **28**(4), 305–317 Apr 2014.

[83] K. I. Assaf, M. Florea, J. Antony, N. M. Henriksen, J. Yin, A. Hansen, Z. W. Qu, R. Sure, D. Klapstein, M. K. Gilson, S. Grimme, and W. M. Nau. HYDROPHOBE Challenge: A Joint Experimental and Computational Study on the Host-Guest Binding of Hydrocarbons to Cucurbiturils, Allowing Explicit Evaluation of Guest Hydration Free-Energy Contributions. *J Phys Chem B* **121**(49), 11144–11162 12 2017.

[84] P. Hunenberger and J. McCammon. Effect of artificial periodicity in simulations of biomolecules under Ewald boundary conditions: a continuum electrostatics study. *Biophysical Chemistry* **78**, 69–88 (1999).

[85] C. Frisch, G. Schreiber, C. Johnson, and A. Fersht. Thermodynamics of the interaction of barnase and barstar: changes in free energy versus changes in enthalpy on mutation. *J Mol Biol* **267**, 696–706 (1997).

[86] M. A. Eriksson and B. Roux. Modeling the structure of agitoxin in complex with the Shaker K+ channel: a computational approach based on experimental distance restraints extracted from thermodynamic mutant cycles. *Biophys. J.* **83**, 2595–2609 (2002).

[87] B. Gilquin, S. Braud, M. A. L. Eriksson, B. Roux, T. D. Bailey, B. T. Priest, M. L. Garcia, A. Menez, and S. Gasparini. A variable residue in the pore of kv1 channels is critical for the high affinity of blockers from sea anemones and scorpions. *Journal of Biological Chemistry* **280**(29), 27093–27102 (2005).

[88] R. Loewenthal, J. Sancho, T. Reinikainen, and A. Fersht. Long-range surface charge-charge interactions in proteins. Comparison of experimental results with calculations from a theoretical method. *J Mol Biol* **232**, 574–583 (1993).

[89] P. L. Kastritis and A. M. J. J. Bonvin. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J. Proteome Res.* **9**(5), 2216–2225 May 2010.

[90] W. Jiang, M. Hodoscek, and B. Roux. Computation of Absolute Hydration and Binding Free Energy with Free Energy Perturbation Distributed Replica-Exchange Molecular Dynamics . *J. Chem. Theory Comput.* **5**, 2583–2588 (2009).

[91] W. Jiang and B. Roux. Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (FEP/H-REMD) for Absolute Ligand Binding Free Energy Calculations. *J. Chem. Theor. Comp.* **6**, 2559–2565 (2010).

[92] Y. Sugita and Y. Okamoto. Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. *Chem. Phys. Lett.* **329**, 261–270 (2000).

[93] S. Jo, C. Chipot, and B. Roux. Efficient Determination of Relative Entropy Using Combined Temperature and Hamiltonian Replica-Exchange Molecular Dynamics. *J. Chem. Theor. Comput.* **11**(5), 2234–2244 MAY 2015.

[94] A. H. Keeble, N. Kirkpatrick, S. Shimizu, and C. Kleanthous. Calorimetric dissection of colicin DNase–immunity protein complex specificity. *Biochemistry* **45**(10), 3243–3254 March 2006.