UNIVERSITY OF CALIFORNIA, SAN DIEGO

Accelerating Conformational Sampling in Free Energy Calculations

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Chemistry

by

Mikolai Fajer

Committee in charge:

Professor J. Andrew McCammon, Chair
Professor John Crowell
Professor Katja Lindenberg
Professor José Onuchic
Professor Wei Wang

2011

The dissertation of Mikolai Fajer is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2011

DEDICATION

To my wife, Megan, with whom this would not be possible.

To my parents, Piotr and Liz, who helped me find this path.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My path up to this point has been affected by so many people, and most of all by my wife and parents. Without Megan's support, encouragement and gentle nudges who knows where I would be, so thank you. My parents gave me an early exposure to science, and I want to thank them for not grumbling too loudly when I decided to follow their footsteps. I must thank Adrian and Wei for introducing me to computational chemistry and for continually challenging the way I think about it. I also want to thank Andy for his invaluable guidance over the last five years, and I hope to live up to his high standards.

Additionally, many thanks to Andy for putting together such a vibrant laboratory group. First and foremost I must thank Donald Hamelberg, Ilja Khavrutskii, Robert Swift, Riccardo Baron, Jeff Wereszczynski, Kate Rogers, Juan Manuel Ortiz Sánchez and César Augusto de Oliveira for the collaborations that we have shared. Working with so many intelligent and dedicated people has made my graduate school career fun as well as challenging. The other members of the group have also been very helpful, and I want to specifically thank Morgan Lawrenz, Paul Gasper, Sarah Nichols, Yi Wang, Levi Pierce and Phineus Markwick for a variety of stimulating discussions.

Chapter 2 is a minimally modified reprint of the material as it appears in M. Fajer, D. Hamelberg, J. A. McCammon, *Replica-Exchange Accelerated Molecular Dynamics (REXAMD) Applied to Thermodynamic Integration.* J. Chem. Theory Comput. 4, 1565-1569 (2008). I was the principal investigator and first author of this paper.

Chapter 3 is a minimally modified reprint of the material as it appears in M. Fajer, R. Swift, J. A. McCammon, *Using Multistate Free Energy Techniques to Improve the Efficiency of Replica Exchange Accelerated Molecular Dynamics.* J. Comp. Chem. 30, 1719-1725 (2009). I was the principal investigator and first author of this paper.

Chapter 4 is a minimally modified reprint of the material as it appears in I. Khavrutskii, M. Fajer, J. A. McCammon, *Intrinsic Free Energy of the Conformational Transition of the KcsA Signature Peptide from Conducting to Nonconducting State.*

J. Chem. Theory Comput. 4, 1541-1554 (2008). I was the second author of this paper.

Chapter 5 is a preprint of *Calculation of Absolute Binding Free Energies by Thermodynamic Integration in AMBER12.* I was the principal investigator and the first author of this paper.

VITA

2006	B. S. in Chemistry , University of Florida
2006	B. S. in Physics , University of Florida
2011	Ph. D. in Chemistry, University of California, San Diego

PUBLICATIONS

M. Fajer, R. Baron, C. de Oliveira, K. Rogers, J. Ortiz Sánchez, J. Wereszczynski, T. Steinbrecher, R. Walker, J. A. McCammon, “Calculation of Absolute Binding Free Energies by Thermodynamic Integration in AMBER12.” *In preparation.*

M. Fajer, R. V. Swift, J. A. McCammon, “Using multistate free energy techniques to improve the efficiency of replica exchange accelerated molecular dynamics.” *J Comput Chem* 30 (11), 1719-25 (**2009**).

M. Fajer, D. H. Hamelberg, J. A. McCammon, “Replica-Exchange Accelerated Molecular Dynamics (REXAMD) Applied to Thermodynamic Integration.” *J Chem Theory Comput* 4 (10), 1565-9 (**2008**).

I. V. Khavrutskii, **M. Fajer**, J. A. McCammon, “Intrinsic Free Energy of the Conformational Transition of the KcsA Signature Peptide from Conducting to Nonconducting State.” *J Chem Theory Comput* 4 (9), 1541-54 (**2008**).

M. Fajer, H. W. Li, W. Yang, P. G. Fajer, “Mapping Electron Paramagnetic Resonance Spin Label Conformations by the Simulated Scaling Method.” *J Am Chem Soc* 45 (129), 13840-6 (**2007**).

H. W. Li, **M. Fajer**, W. Yang, “Simulated scaling method for localized enhanced sampling and simultaneous "alchemical" free energy simulations: A general method for molecular mechanical, quantum mechanical, and quantum mechanical/molecular mechanical simulations. ” *J Chem Phys* 126 (2), 024106 (**2007**).

M. Fajer, P. Fajer, K. Sale, “Molecular Modeling of Spin Labels." M. A. Hemmina and L. J. Berliner, editors. ESR Spectroscopy in Membrane Biophysics. Springer Series in Biological Magnetic Resonance (**2007**).

ABSTRACT OF THE DISSERTATION

Accelerating Conformational Sampling in Free Energy Calculations

by

Mikolai Fajer

Doctor of Philosophy in Chemistry

University of California, San Diego, 2011

Professor J. Andrew McCammon, Chair

Molecular dynamics are increasingly used to construct conformational ensembles of biochemical systems. The accuracy of these ensembles is determined by the accuracy of the underlying model and the extent of conformational sampling during the simulation. Biochemical systems can have motion on time scales that vary by several orders of magnitude, and these must all be described before a specific model can be validated. For this reason research into enhanced sampling methods that accelerate conformational sampling are vital to the progress of molecular dynamics.

This dissertation describes the validation and application of the replica exchange accelerated molecular dynamics (REXAMD) method in the context of free energy calculations. In chapters 2 and 3 the REXAMD method is validated using simple model systems. The convergence of REXAMD is shown to be an improvement over classical molecular dynamics. Additionally, various methods to improve the statistical behavior of REXAMD are investigated. In chapter 4 gradient-augmented Harmonic Fourier Beads, a minimum free energy pathway method, is used to study the conformational change of the ion selectivity peptide from the KcsA potassium channel. The robustness of various models, ranging from classical to quantum mechanical, is investigated and the importance of conformational sampling is observed. Finally in chapter 5, I propose a modification to the AMBER molecular dynamics package which allows the calculation of absolute binding free energies to be computed.

Chapter 1

Introduction

1.1 Molecular Models

Art is the lie that helps tell the truth.

Pablo Picasso

A scientific model is a simplified, and frequently mathematical, representation of a physical object that is used to assist in the analysis and prediction of experimental results. The equivalent experiments may be expensive or practically impossible, but a good model always offers new predictions that will help validate itself.

One of the most pervasive uses of modeling in biochemistry today is in structure prediction of biomolecules. In 1931 William Astbury reported the X-ray diffraction of several fibrous materials, including three different types of wool. [5] Of particular interest was the reversible change of diffraction patterns upon stretching the materials. Astbury referred to these as the α- and β-forms and showed that the amino acids repeated every 5.15 and 3.32 , respectively. Twenty years later Linus Pauling took his knowledge of amino acid structure and the hydrogen bond strength to propose the α-helix and β-strand polypeptide structures that maximized intermolecular hydrogen bonds and contained structural repeats close to those identified by Astbury. [6, 7] The application of molecular modeling to protein structure prediction from X-ray crystallography has continued to grow since Pauling's contribution. Currently diffraction patterns are converted to a three-dimensional model of electron density and then an atomistic model is proposed that reproduces the diffraction patterns.

The same type of physical intuition that Pauling used to construct low energy conformations has been formalized under the name *molecular mechanics.* In molecular mechanics the atoms are treated as classical particles for which the Hamiltonian, a function that describes the energy and forces of any structure, can be written. These Hamiltonians are generally constructed to be pairwise-additive as seen in Equation 1.1. The collection of parameters (k_b, b_0, etc.) are referred to as a *forcefield.* The forcefields are constructed from fitting the parameters to experimental data of small compounds, for example the IR vibration frequencies to set the bond force constants. After these force-fields are constructed they are assumed to be *transferable*, or retain accuracy for larger collections of atoms than the compounds they were originally parametrized for.

$$\begin{aligned} U(\mathbf{r}) = \sum_{bonds} k_b (b - b_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} k_\chi (1 + \cos(n\chi - \delta)) \\ + \sum_{impropers} k_\psi (\psi - \psi_0)^2 + \sum \sum_{non\text{-}\ bonded} \epsilon_{ij} \left[\frac{R_{min,ij}}{r_{ij}} \right] + \left(\frac{-A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} \right) \end{aligned} \tag{1.1}$$

A more detailed history of molecular modeling can be found in Tamar Schlick's excellent book. [8]

1.2 Conformational Sampling

If the blind lead the blind, both will fall into a pit.

Matthew 15:14 (NIV)

A typical use of molecular modeling is to determine the conformations and dynamics of a molecule in solution with atomic resolution. The collection of conformations is called the *configuration ensemble*, and we know that the ensemble will obey the Boltzmann distribution (equation 1.2), where U is the potential energy and $\mathbf{q}$ is a specific conformation. The integral in equation 1.3 is over all conformational space $\mathcal{V}$ and as the number of atoms increases this space grows exponentially. The Boltzmann probability distribution $p(\mathbf{q})$ clearly shows that the low energy conformations will be much more probable than high energy conformations, so instead of looking at every possible conformation it is a good approximation to focus on the low energy conformations. The normalization constant Z is referred to as the configuration integral, and is a foundational quantity in statistical mechanics.

$$p\left(\mathbf{q}\right) = \frac{\exp\left[-\beta U\left(\mathbf{q}\right)\right]}{Z} \tag{1.2}$$

$$Z\left(N, V, T\right) = \int_{\mathcal{V}} \exp\left[-\beta U\left(\mathbf{q}\right)\right] d\mathbf{q} \tag{1.3}$$

The potential energy can be visualized as a surface, or a series of mountains and valleys. We want to learn about all of the valleys so we hire an intrepid explorer and ask him to visit each valley in turn. If you give him a map the length of the task will depend on the length of his stride, his endurance and the difficulty of the path. Now take away his map and he must wander about, not knowing in which direction the next valley lies or if the path he is currently on will come to a dead end. Finally blindfold him, cover his ears and hand him only a walking stick with which to feel about himself with. This is an accurate illustration of the difficulties in sampling molecular models. The length of our explorer's stride is the time step limit imposed by numerical integration schemes, and we must be careful to take small steps in order to retain accuracy. The explorer's endurance and path are reflections of statistical mechanics which tells us we are more likely to fall down slopes than climb them. The highly local nature of the Hamiltonian, equation 1.1, is the reduction of our explorer's senses so that we only know our immediate energy and forces but not what the energy and forces will be several feet in any direction without travelling there.

One assumption needs to be made explicit: the explorer can reach any point starting from any other point. This property is called *ergodicity* and is incredibly difficult to prove but is absolutely essential. If a system is indeed ergodic then we can leverage statistical mechanics and assume the *ergodic hypothesis*, equation 1.4, which states that the probability of a specific conformation $\mathbf{q}$ predicted from a simulation will converge to the value predicted by the Boltzmann distribution as the simulation length increases.

$$p\left(\mathbf{q}\right) = \lim_{t\to\infty} \frac{1}{T} \int_0^T \delta\left(\mathbf{q}, \mathbf{q}\left(t\right)\right) dt \tag{1.4}$$

We have now exchanged the need to test every possible conformation for the need to simulate an infinitely long time. This may not seem like an advantage, but let us study the limit in equation 1.4 more closely. The limit removes any bias that we have due to initial conditions because an infinite number of conformational transitions occur. The convergence behavior is thus determined by the kinetics of conformational transitions. If we can increase the rate of conformational transitions while still knowing the unperturbed statistics we can achieve convergence of equation 1.4 in a practical timescale. This is the motivation for *enhanced sampling* methods.

For example, in the metadynamics method a history dependent bias is applied to the system. [9] A series of Gaussians are added to Hamiltonian, which can be visualized as carrying around a large bag of cement and periodically laying cement under our explorer's feet. Over time the valley is filled in and the height between the cement and the mountain pass decreases. Transition state theory shows that the rate of a process depends on the free energy difference of the initial and transitions states, and that decreasing that difference will increase the rate. If you carefully record where you place the cement you can remove that bias from the simulation results and estimate what the original energy of each conformation would be. The extrapolation back to the unperturbed landscape will vary from method to method, but typically the less bias you apply the more accurate the extrapolation will be.

Many enhanced sampling methods additionally try to identify useful directions, called *collective variables*, to promote uphill movement in those directions. This is done to improve the efficiency of the methods since we only want to visit the transition paths and the basins, and not waste time climbing Mount Everest. These additional instructions can be very specific such as "enhance sampling only due north", or general such as "enhance sampling if we have been traveling in this direction for a while". The advantage of specific instructions is that they apply less bias and thus are more efficient, however they require prior knowledge of the system in order to know which direction is important. The choice of collective variable is thus very important. Laio *et al.* suggests finding as few collective variables as possible that clearly distinguish between the initial, intermediate and final states while still describing all of the relevant slow degrees of freedom. [1] More generic collective variables, for example the temperature or various energy terms, apply bias and ac-

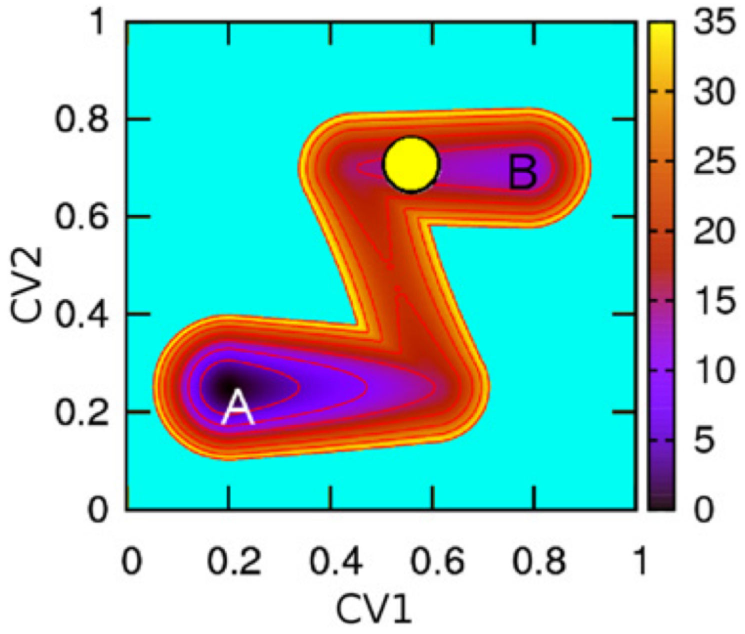

Figure 1.1: Hamiltonian lagging example taken from [1]. There is strong coupling between CV1 and CV2, and sampling in one will be dependent on sampling in the other.

celeration more broadly. This lowers the efficiency of the method but simplifies the selection of collective variables. Examples of these methods are simulated tempering [10], hyperdynamics [11], accelerated molecular dynamics [12], simulated scaling [13], and energy-space metadynamics [14].

The assumption we have made thus far is that we know all of the important directions, but what happens if we miss one? This issue is referred to as Hamiltonian lagging, and is illustrated in Figure 1.1. [15, 1] Here we have failed to recognize the importance of the y-axis and are only biasing along the x-axis. During a transition from A to B we are unlikely to move along the y-axis since it is still uphill, and thus incorrectly conclude that B is unfavorable relative to A.

1.3 Free Energy

There is no free lunch.

Milton Friedman

Free energy is the amount of work that a system can perform. Accurate prediction of free energy changes can determine ligand-protein affinity, protein conformational transitions, and a whole host of other important biophysical phenomena. For example, the association constant K_a is a function of the Gibbs free energy change upon binding ΔG_b (equation 1.5).

$$K_a = C^\circ \exp\left[-\beta \Delta G_b^\circ\right] \tag{1.5}$$

The Helmholtz free energy of a state can be determined solely from the partition function as shown in equation 1.6. The configuration space $\mathcal{V}$ can now refer to a well-defined subspace instead of the entire configuration space as in section 1.2. For example, in protein-ligand binding the bound state ($\mathcal{V}_1$) would refer to all of the conformations with the ligand in the binding cavity and the unbound state ($\mathcal{V}_0$) to the conformations at large inter-molecular separation. Once these two volumes are defined the free energy difference can be expressed as a ratio of the corresponding configuration integrals (equation 1.7). This formalism applies to chemical reactions, molecular association and intra-molecular rearrangement.

$$A = -\beta^{-1} \ln \frac{1}{N!\Lambda^{3N}} Z\left(N, V, T\right) \tag{1.6}$$

$$\Delta A = -\beta^{-1} \ln\left(Z_1/Z_0\right) \tag{1.7}$$

Free energy is a state function and is thus independent of the path taken between two states. The most physically analogous path is to construct a potential of mean force (PMF) along one or more collective variables that connects the two states. In the case of protein-ligand binding this would be the intermolecular separation. The bound and unbound states can then be defined as the attractive portion of the PMF and the volume at large separation distances, respectively. [16] The sampling along the collective variable usually needs to be enhanced in order to achieve convergence of the entire PMF. The most commonly used method is umbrella sampling [17], although biased sampling methods like metadynamics [18] and non-equilibrium methods like steered molecular dynamics [19] are becoming more popular. Identifying the minimum free energy path between two states is also very important in determining rate limiting steps. Early work on the nudged elastic band method [20] has continued to be refined. [21, 22]

It is also possible to take non-physical, or *alchemical*, paths between two states. The typical alchemical path has two Hamiltonians that describe the two different states. For example, in protein-ligand binding it is common practice to use a decoupled or gas state ($\lambda = 1$) to complete the thermodynamic cycle (figure 1.2). [23] Another possibility is to have two different force-fields. [24] The Hamiltonians are then mixed as a function of the alchemical parameter λ. There are many paths to move from one alchemical endpoint to the other, and the linear mixing in equation 1.8 is just the simplest case. The most important property that distinguishes alchemical paths is their statistical uncertainty, and the decoupling of the van der Waals interactions in particular has been a source of numerical and statistical challenges. [25, 26, 27, 28]

$$H_{\text{alchemical}}\left(\mathbf{q},\lambda\right) = \left(1-\lambda\right) H_0\left(\mathbf{q}\right) + \lambda H_1\left(\mathbf{q}\right) \tag{1.8}$$

Computation of the alchemical free energy difference is done with either thermodynamic integration (TI) [29], thermodynamic perturbation (TP) [30], or one of their derivatives. These competing methods differ in philosophy and practical accuracy. Thermodynamic perturbation determines the work to instantly perturb from one state to another. This method requires a moderate amount of overlap between the two alchemical states in order to be accurate, and additional simulations utilizing intermediate states will need to be performed if there is poor overlap. In thermodynamic integration a series of simulations are constrained at different λ values and the gradient with respect to λ is computed. This gradient, $\langle \partial H_{\text{alchemical}}/\partial\lambda\rangle_{\lambda}$, can then be integrated to determine the free energy difference between the two states. The statistical uncertainty of TI arises first from convergence of the gradient and second from the numerical integration of an *a priori* undetermined function. Thermodynamic integration may require more simulations in order to achieve the same level of accuracy as thermodynamic perturbation. [31]

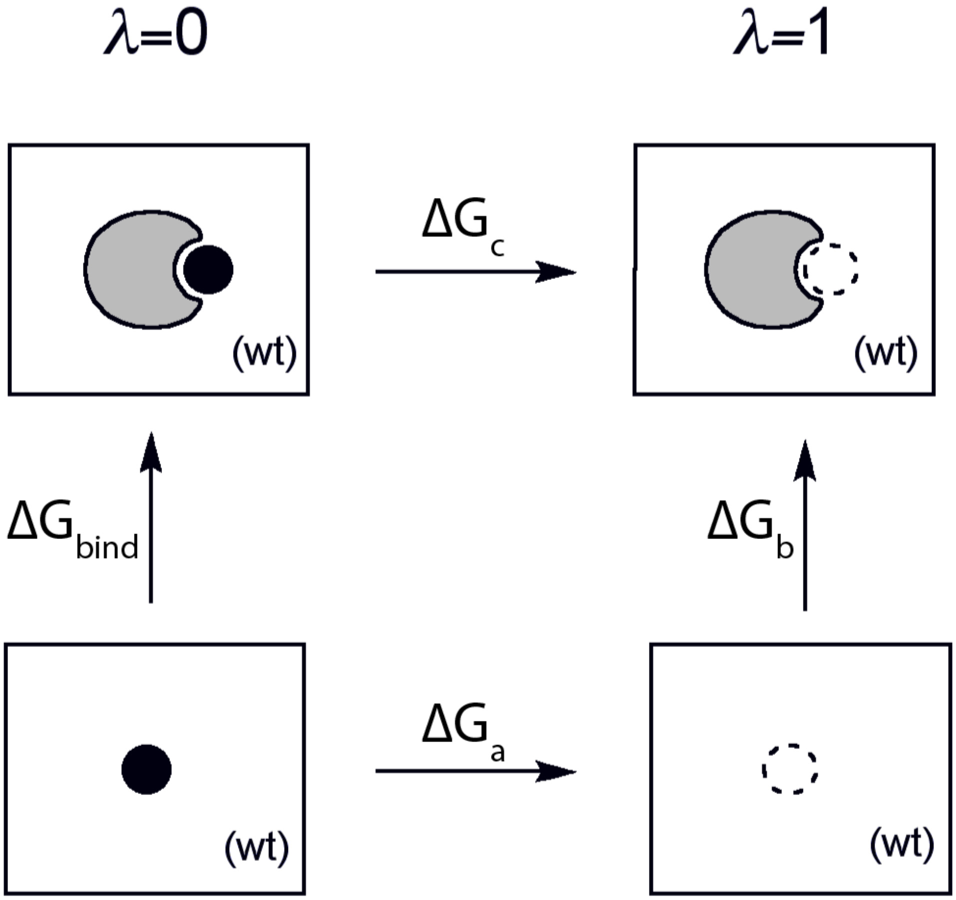

Figure 1.2: Schematic of the double-decoupling alchemical pathway. Two simulations are performed to determine ΔG_a and ΔG_c. The ΔG_b is zero. Thus $\Delta G_{\mathrm{bind}} = \Delta G_a - \Delta G_c$

Both thermodynamic integration and thermodynamic perturbation require accurate sampling of the mixed Hamiltonian, and the challenges described in section 1.2 are only complicated by the addition of the λ parameter. For this reason enhanced sampling methods are frequently used in free energy calculations. There is also quite a bit of work on how to make the most efficient use of data collected from the simulations. The evolution of better estimators for free energy perturbation has been particularly successful, moving from the original uni-directional estimator to the Bennett acceptance ratio [32] and multi-state Bennett acceptance ratio methods [33, 4].

Chapter 2

Replica-Exchange Accelerated Molecular Dynamics (REXAMD) Applied to Thermodynamic Integration

Abstract

Accelerated molecular dynamics (AMD) is an efficient strategy for accelerating the sampling of molecular dynamics simulations, and observable quantities such as free energies derived on the biased AMD potential can be reweighted to yield results consistent with the original, unmodified potential. In conventional AMD the reweighting procedure has an inherent statistical problem in systems with large acceleration, where the points with the largest biases will dominate the reweighted result and reduce the effective number of data points. We propose a replica exchange of various degrees of acceleration (REXAMD) to retain good statistics while achieving enhanced sampling. The REXAMD method is validated and benchmarked on two simple gas-phase model systems, and two different strategies for computing reweighted averages over a simulation are compared.

2.1 Introduction

Free energy is one of the most important quantities in biophysics. The calculation of free energy using molecular dynamics simulations is complicated by the dependence on the amount of the relevant phase space sampled. The complication is more pronounced when two alchemical free energy end points differ by more than a few trivial moieties. The use of restraints to restrict the phase space has proven useful in the convergence of thermodynamic integration, umbrella sampling, and the Bennett acceptance ratio techniques. [34, 35, 36] Another approach is to enhance phase space sampling instead of restricting the phase space, and often relies on the modification of the original Hamiltonian during molecular dynamics simulations. [13, 37] Accelerated molecular dynamics (AMD), which conventionally modifies the energy landscape by adding a bias to states below an energy threshold, E_{cut} (Equation 2.1 and 2.2), is an example of the Hamiltonian modification approach and has proven capable of efficiently generating canonical ensembles consistent with experiments on the millisecond timescale. [12, 38]

$$V^{*}\left(r, E_{cut}, \alpha\right) = V\left(r\right) + \begin{cases} 0 & V\left(r\right) \geq E_{cut} \\ \Delta V\left(r, E_{cut}, \alpha\right) & V\left(r\right) < E_{cut} \end{cases} \tag{2.1}$$

$$\Delta V\left(r, E_{cut}, \alpha\right) = \frac{\left(E_{cut} - V\left(r\right)\right)^{2}}{\alpha + \left(E_{cut} - V\left(r\right)\right)} \tag{2.2}$$

A potential problem with modifying the Hamiltonian occurs when reweighting an observable O^{*} from the accelerated simulation to find O on the original potential (Equation 2.3 for AMD). If the simulation is highly accelerated and involves a large range of boost factors ΔV, the reweighted average will be dominated by the relatively few points/structures with large values of ΔV in the limit of finite sampling. This statistical problem has recently been quantified as a reduction in the effective number of data points in the simulation. [39] Thus there is a tradeoff between the degree of acceleration and the statistical precision in AMD simulations. The calculation of free energies using thermodynamic integration computes $\langle dV/d\lambda \rangle_{\lambda_i}$ over the course of a simulation, and the calculation of free energy is very sensitive to the statistical accuracy of the computed averages.

$$\langle O \rangle = \frac{\langle O^{*} exp \left[\beta \Delta V \left(r \right) \right] \rangle}{\langle exp \left[\beta \Delta V \left(r \right) \right] \rangle} \tag{2.3}$$

In order to take advantage of the sampling efficiency of the AMD method as well as maintain the statistical relevance of every data point, we propose using a replica-exchange framework to couple varying degrees of acceleration. The low degrees of acceleration will not be prone to the reweighting problem, and can still take advantage of the high acceleration through replica exchanges. This replica-exchange accelerated molecular dynamics (REXAMD) is a member of the Hamiltonian replica-exchange (HREM) class of simulations, varying from other HREM techniques in the specific Hamiltonian modification scheme. A similar REXAMD approach has recently been applied to studying the effects of neighboring side chains on peptide backbone conformations in short peptides. [40] We demonstrate the REXAMD approach by increasing the convergence rate of thermodynamic integration (TI) for two simple gas-phase model systems, although the method could utilize other free energy calculation methods instead of TI.

2.2 Computational Detail

First some terms should be defined. *State* is used to denote a specific level in the replica-exchange scheme. For example, in temperature replica-exchange each state corresponds to a specific temperature, and in REXAMD each state is a modified Hamiltonian described by a set of boost parameters. The term *replica* is used to denote the individual structures that are exchanged between the various REXAMD states. The term *simulation* refers to a specific setup of REXAMD, and the term *run* refers to an instance of a simulation. Simulation is also used to identify the average and standard error computed from multiple runs.

The current replica-exchange framework is a Python program that launches a modified AMBER8 accelerated molecular dynamics simulation [12] for each replica in between Metropolis Monte Carlo exchanges (Equation 2.4). The Monte Carlo (MC) exchanges occur every 1000 molecular dynamics (MD) steps and the pairs that attempt exchanges alternate every other MC period. For example, in a simulation with four states (labeled s0-s3) the simulation would execute 1000 MD steps, attempt MC exchanges between the s0-s1 states and the s2-s3 states, execute 1000 MD steps, attempt a MC exchange between the s1-s2 states, and repeat. The molecular dynamics simulations used a 1 fs^{-1} time step and were coupled to a 300K Langevin thermostat with a collision frequency of 10.0 ps^{-1}. The Python program reset the seed number for the AMBER random number generator after every MC exchange.

$$p_{ex} = \begin{cases} 1 & \Delta\left(i,j\right) \leq 0 \\ exp\left[-\beta\Delta\left(i,j\right)\right] & \left(i,j\right) > 0 \end{cases} \tag{2.4}$$

$$\Delta\left(i,j\right) = \Delta V\left(r_j,\alpha_i\right) + \Delta V\left(r_i,\alpha_j\right) - \Delta V\left(r_i,\alpha_i\right) - \Delta V\left(r_j,\alpha_j\right) \tag{2.5}$$

The boosting scheme is identified as a suffix added to the REXAMD acronym as follows: REXAMDt denotes a boost only to the torsional potential, and REXAMDtT denotes a dual boost scheme applied to the torsional and total potentials. [41] The *-rw* suffix indicates the reported results are from the reweighting of the most accelerated state in a specific simulation. When the *-rw* suffix is not present the result is coming from the least accelerated state, which in this paper is always no acceleration.

In order to separate the effect of acceleration from the effect of using M replicas, the REXREG control simulations are a replica-exchange between identical regular dynamics potentials. Note that this makes the acceptance probability of MC exchange in Equation 2.4 identically equal to one. The REXREG simulations are analogous to M independent runs from the same starting point with different initial velocities and taking an average result from the M runs.

The replica-exchange efficiency will be monitored based on two criteria. The first criterion is the average acceptance ratio of the replica-exchanges over the course of a run and gives a rough idea of how capable the given replica-exchange scheme is at mixing replicas. The second criterion is the observed relative frequency RMSD metric. [3] This metric compares the observed population frequency of the replicas against the idealized case where each of M replicas spends $1/M$ of the total time in any given state of the system. The RMSD metric varies from zero for the ideal mixing to $\sqrt{M-1}/M$ for no mixing. The observed relative frequency metric is more detailed than the average acceptance ratio in monitoring the mixing efficiency of the replica-exchange simulation.

The thermodynamic integration of the model systems was computed using a linear scaling of an all-atom potential (Equation 2.6). Gaussian quadrature integration was used to evaluate the thermodynamic integral from a finite number of $\langle dV/d\lambda \rangle_{\lambda_i}$ calculated at specific λ_i values (Equation 3.8). The Gaussian quadrature points and weights were taken from the AMBER8 manual. [42] Two strategies were used to calculate $\langle dV/d\lambda \rangle_{\lambda_i}$ at each λ_i. The first strategy, reweighted periods, calculated the reweighted average of each block of 1000 MD steps in between MC exchanges. These reweighted averages were then averaged together over a complete run to yield $dV/d\lambda$ for a specific $\langle dV/d\lambda \rangle_{\lambda_i}$. The assumption behind this approach is that the $dV/d\lambda$ values sampled during 1 ps give rise to $\langle dV/d\lambda \rangle_{\lambda_i}$ for a local region of the conformational space. The replica exchange will then balance the occurrence of the local regions. The second strategy, reweighted run, takes an instantaneous $dV/d\lambda$ and its corresponding ΔV from the MD step immediately prior to a MC exchange. These values are then used to compute a reweighted $\langle dV/d\lambda \rangle_{\lambda_i}$ for the entire simulation. This approach virtually guarantees uncorrelated $dV/d\lambda$ values at the expense of the number of points being considered in the average. In both strategies each λ_i was simulated ten times with different random seeds and velocities. An average and standard error for each $\langle dV/d\lambda \rangle_{\lambda_i}$ is then determined and combined into the overall ΔG. The average ΔG is only reported to the first significant digit of the standard error.

$$V(\lambda) = (1 - \lambda) V_0 + \lambda V_1 \tag{2.6}$$

$$\Delta G = \int_0^1 \langle dV/d\lambda \rangle_{\lambda_i} d\lambda \approx \sum_i w_i \langle dV/d\lambda \rangle_i \tag{2.7}$$

Two model systems were studied to validate and benchmark the REXAMD method. Both model systems are symmetric alchemical mutations where the product has an identical structure to the reactant and thus the ΔG is zero and independent of the force field. Model system A (MSA) is a gas-phase alchemical mutation from ethane-to-ethane (Figure 2.1A). This system will serve as a positive control to show that REXAMD can reproduce the results of an ergodic regular molecular dynamics

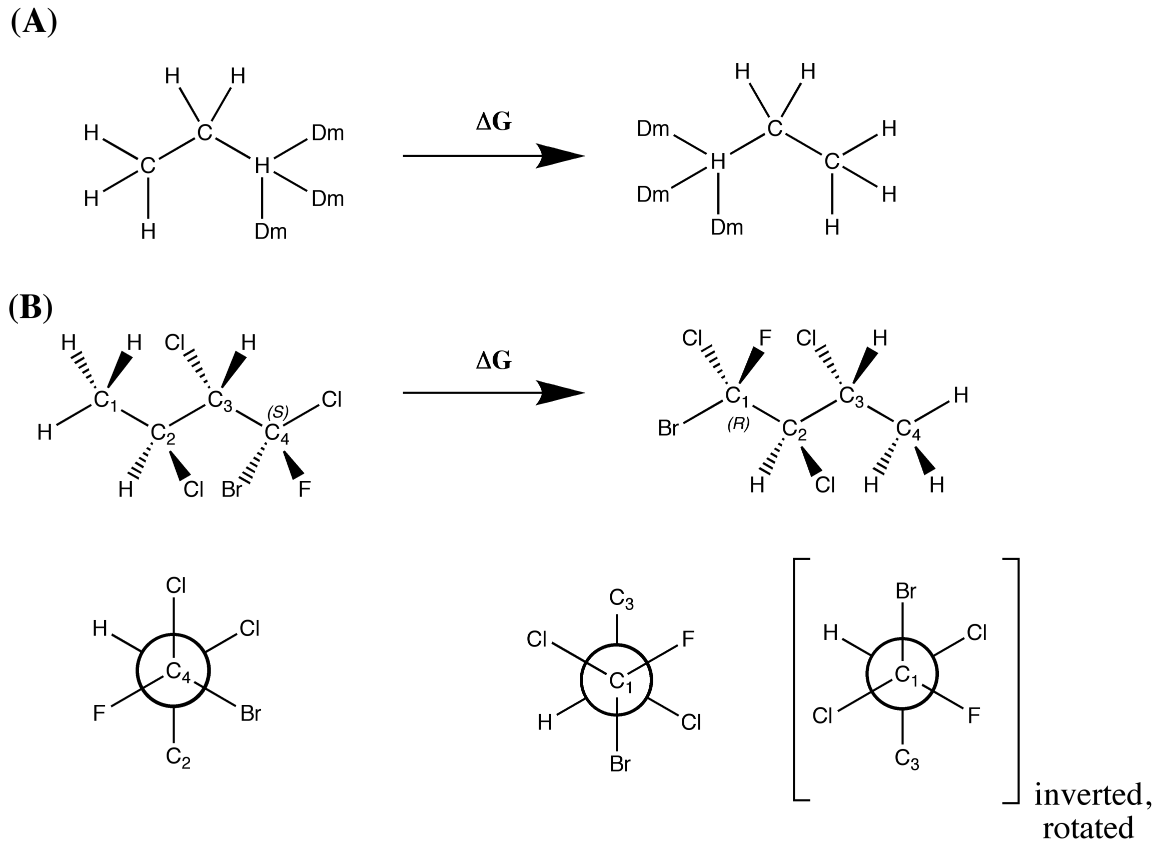

Figure 2.1: Structure of the model systems (A) and (B). The Dm atoms indicate a dummy atom with no nonbonded interactions.

simulation. The relative simplicity of the system and the low transition barriers guarantees that the regular molecular dynamics (REXREG) is able to sample the entire conformational space in a short timescale. The thermodynamic integration for MSA uses a 9-point Gaussian quadrature. The MSA REXAMDt simulations used only two replicas, an unmodified potential and an accelerated potential with a torsional boost (E_{cut} of 5.0 kcal/mol, of 2.0 kcal/mol). Each run was simulated for 8 million MD steps, or the equivalent of 8 ns for an unmodified potential.

Model System B (MSB) is a highly halogenated butane (Figure 2.1B). The initial conformation of the system is in a different rotameric state for the two λ_i endpoints, as seen in the Newman projections in Figure 2.1B, and thus requires proper conformational sampling to yield the correct . The chlorine atoms attached to C2 and C3 were added to make the rotameric sampling more difficult, requiring acceler-

Table 2.1: Summary of the Replica-Exchange Efficiency. The average and standard deviation of the acceptance ratios are from the ten runs and the M states. The average and standard deviation of the rmsd of the relative occupancy of the M replicas over the M states, as defined by [3] are reported.

simulation	acceptance ratio	observed relative frequency rmsd
MSA REXAMDt (8 ns, 2 states)	39.3 ± 2.2%	0.00565 ± 0.00420
MSB REXAMDtT (20 ns, 8 states)	40.2 ± 16.0%	0.00827 ± 0.00231

ation to achieve the correct answer within the current timescale of 20 ns. The dual boosting scheme was used for this model system in order to accelerate the large van der Waals interactions experienced in this system. In order to increase the difficulty of converging to the correct result, we are only using a 3-point Gaussian quadrature. The boost parameters for the eight replicas in the MSB REXAMDtT simulations are shown in Table S2.1, and are labeled from s0 to s7 in terms of increasing boost.

2.3 Results and Discussion

2.3.1 Model System A

In MSA both the REXREG and REXAMDt simulations were able to efficiently and exhaustively explore the conformational space (data not shown), and the replica mixing was quite efficient (Table 2.1) within the 8 ns runs. The exhaustive sampling resulted in converged ΔG values within the first ns of the REXREG and REXAMDt simulations (Figure 2.1). The ΔG results from the entire 8 ns are summarized in Table 2.2. Recall that *MSA REXAMDt* refers to result taken from the non-accelerated state and *MSA REXAMDt-rw* refers to the reweighted results of the accelerated state.

The statistical precision can be monitored in terms of the number of values that were used in computing $\langle dV/d\lambda \rangle_{\lambda_i}$. For example, applying the reweighted run strategy to the REXAMDt simulation yields a total of 80,000 data points for each $\langle dV/d\lambda \rangle_{\lambda_i}$ (ten 8 ns trajectories). This strategy resulted in a ΔG of 0.02 ± 0.02 kcal/mol. In order to produce the same number of points when using the reweighted

Table 2.2: ΔG Summary of MSA Thermodynamic Integration Results. The units are in kcal/mol. The average and standard error from ten simulations are reported for each simulation type.

reweighting strategy	REXREG	REXAMDt	REXAMDt-rw
periods	+0.002 ± 0.001	-0.001 ± 0.001	-0.001 ± 0.001
runs	-0.04 ± 0.01	0.02 ± 0.02	-0.01 ± 0.03

periods strategy we consider only the first 8 ps of the ten duplicate runs for each λ_i, which yields a ΔG of 0.02 ± 0.03 kcal/mol. Note the similarity in both the accuracy and precision of these two results, indicating that exhaustive sampling occurs below the picosecond timescale. The slower ΔG convergence of the reweighted run strategy versus the reweighted periods strategy is due to the slower rate of data collection for the reweighted run strategy.

The REXAMDt-rw simulations also exhibit high accuracy and precision (Figure 2.2 and Table 2.2). The average boost applied over the MSA REXAMDt simulations from all of the λ_i values was 2.0 ± 0.9 kcal/mol. The small range of boosts (standard deviation of 0.9 kcal/mo) is predicted to have a relatively small effect on the reweighted precision as predicted by Shen and Hamelberg. [39] The reweighted periods strategy reduces the effective number of instantaneous $dV/d\lambda$ values from 80 million to 16 million for each $\langle dV/d\lambda\rangle_{\lambda_i}$, and the REXAMDt-rw simulations exhibit marginally worse accuracy/precision than the REXAMDt (Table 2.2). A similar effect is observed in the reweighted runs strategy (a reduction from 80,000 to approximately 15,000).

2.3.2 Model System B

The 20ns MSB REXAMDtT simulations are well mixed (Table 2.1, Figure S2.1, Figure S2.2). The regular molecular dynamics (REXREG) was unable to efficiently sample the conformational space and still shows a substantially non-zero after the 20ns for both the reweighted periods and reweighted runs strategies (Table 2.3). The slow convergence of the REXREG result can also be seen in the block averaging of $\langle dV/d\lambda\rangle_{\lambda_i}$ in Figure 2.3. In contrast, the REXAMDtT simulations were able to

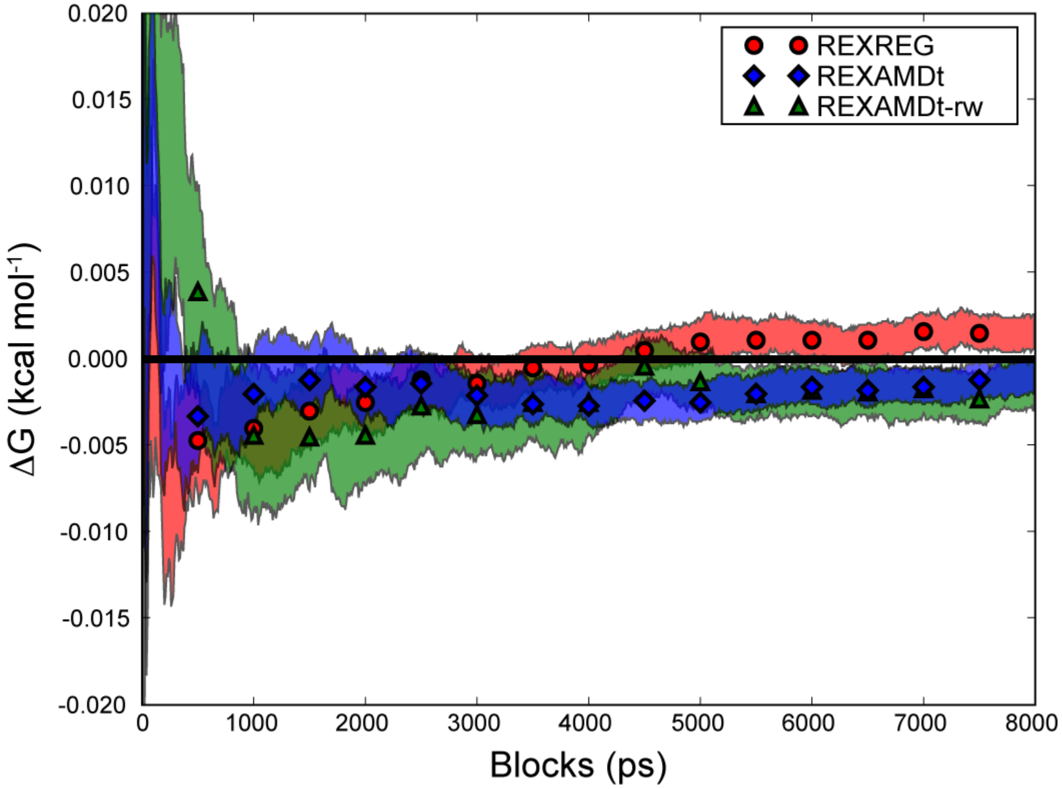

Figure 2.2: Block average of the MSA thermodynamic integration results when using the reweighted periods strategy. The symbols show the average value of each simulation type, and the shaded region shows the standard error from the ten duplicate runs.

Table 2.3: ΔG Summary of MSB Thermodynamic Integration Results. The units are in kcal/mol. The average and standard error from ten simulations are reported for each simulation type.

reweighting strategy	REXREG	REXAMDt	REXAMDt-rw
periods	+0.12 ± 0.08	+0.04 ± 0.01	+0.08 ± 0.06
runs	+0.16 ± 0.07	0.03 ± 0.04	-9 ± 7

efficiently sample the conformational space (data not shown). The was consistently within 0.1 kcal/mol of zero after 2.9 and 5.5 ns for the reweighted periods strategy and the reweighted runs strategy, respectively.

The reweighting procedure was applied to the state with the highest degree of acceleration, s7, because this state is the most independent of the other states in terms of convergence. The most accelerated state is also expected to have the highest range of ΔV boost factors and therefore exhibit the largest reweighting problem. [39] This prediction can be seen in the poor accuracy and precision of the of reweighted runs for REXAMDtT-rw (Table 2.3, Figure 2.4). The effective numbers of data points for the s7 states are shown in Table S-I and demonstrate the source of the poor statistics. For example, the λ_i of 0.5 simulations had a standard deviation of boost values of 13 kcal/mol, and only 30 of the 200,000 data points from the ten duplicate runs contributed to $\langle dV/d\lambda \rangle_{\lambda=0.5}$.

The reweighted periods strategy for REXAMDtT-rw has at least one effective point in each 1 ps period, and therefore at least 200,000 data points for each $\langle dV/d\lambda \rangle_{\lambda}$ when the ten duplicate runs are considered. Compared to the reweighted runs strategy, the increase of the effective number of points results in the increase of the accuracy and precision of the computed ΔG by two orders of magnitude (Table 2.3). The effective number of points is still less than that of REXAMDtT, which has 200 million data points, and the accuracy and precision of REXAMDtT are still better than those of REXAMDtT-rw when using the same averaging strategy (Table 2.3, Figure 2.3).

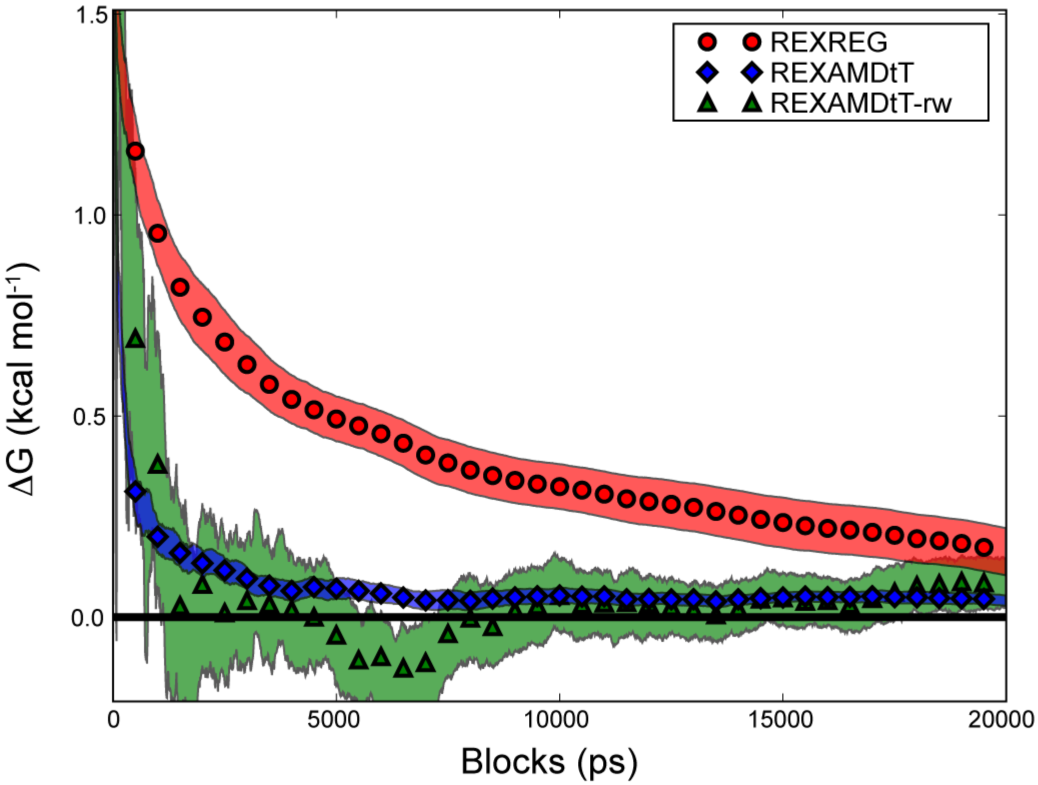

Figure 2.3: Block average of the MSB thermodynamic integration results when using the reweighted periods strategy. The symbols show the average value of each simulation type, and the shaded region shows the standard error from the ten duplicate runs.

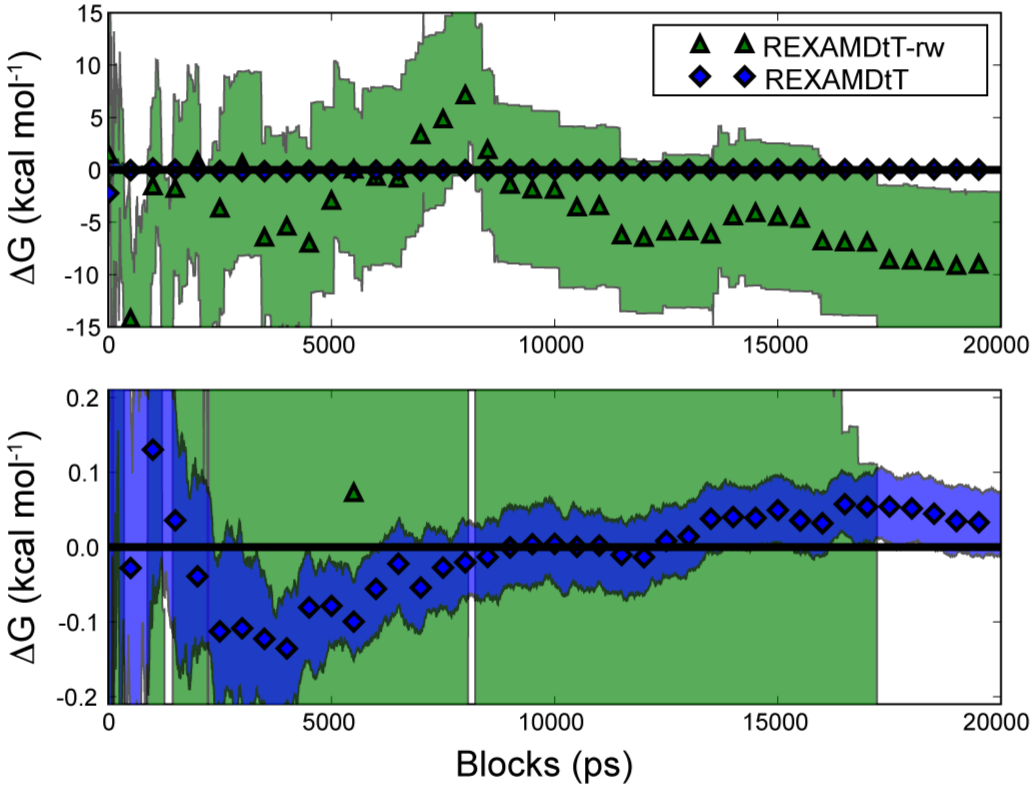

Figure 2.4: Block average of the MSB thermodynamic integration results from the reweighted runs strategy shown on two different scales. The symbols show the average value of each simulation type, and the shaded region shows the standard error for each simulation type. The top plot shows the REXAMDtT-rw results on scale and shows how poor the statistics are after reweighting. The bottom plot shows the REXAMDtT results on scale and shows how quickly the REXAMD technique converges to within statistical accuracy.

2.4 Conclusion

The REXAMD method has been shown to efficiently accelerate conformational sampling while avoiding the statistical reweighting problem inherent in AMD. The REXAMD method was validated on the simple model system A. In the more complex model system B the dual boost REXAMD scheme showed marked improvement over the regular molecular dynamics approach, as well as better statistical accuracy and precision in comparison to the reweighted results of the accelerated replicas. We are currently researching the application of this method to more complicated systems.

Acknowledgement

Chapter 2, in full, is a minimally modified reprint of the material as it appears in M. Fajer, D. Hamelberg, J. A. McCammon, âĂIJReplica-Exchange Accelerated Molecular Dynamics (REXAMD) Applied to Thermodynamic Integration.âĂİ J. Chem. Theory Comput. 4, 1565-1569 (2008). The dissertation author was the principal investigator and first author of this paper.

Supplementary Material

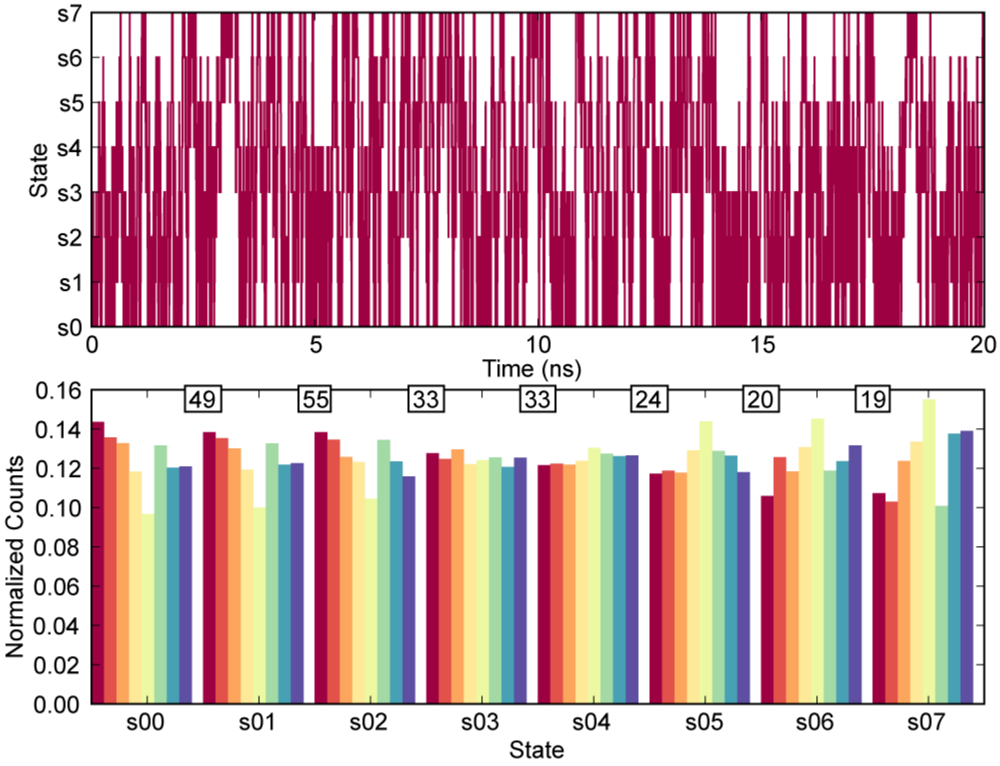

Figure S2.1: A representative example of the replica exchange details for the MSB REXAMDtT simulations. The top plot shows a time series of a single replica traveling through the various states. The boxed numbers between the states are the overall acceptance probability for the pairs of states. For example, the acceptance probability over the 20 ns simulation for the s0-s1 pair was 49.0%. The bottom plot shows the observed relative frequency of each replica (different colors) in each state. The overall acceptance probability is also shown between the state pairs.

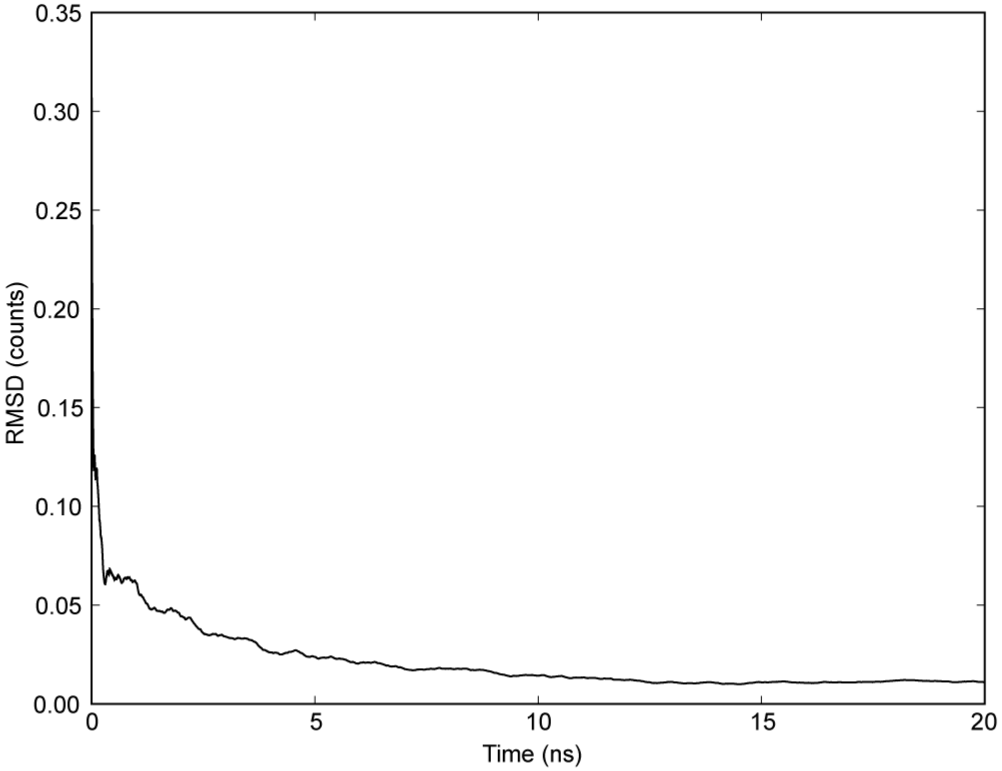

Figure S2.2: A representative time series of the RMSD of the relative occupancy of the M replicas over the M states for the MSB REXAMDtT simulations, as defined by [3].

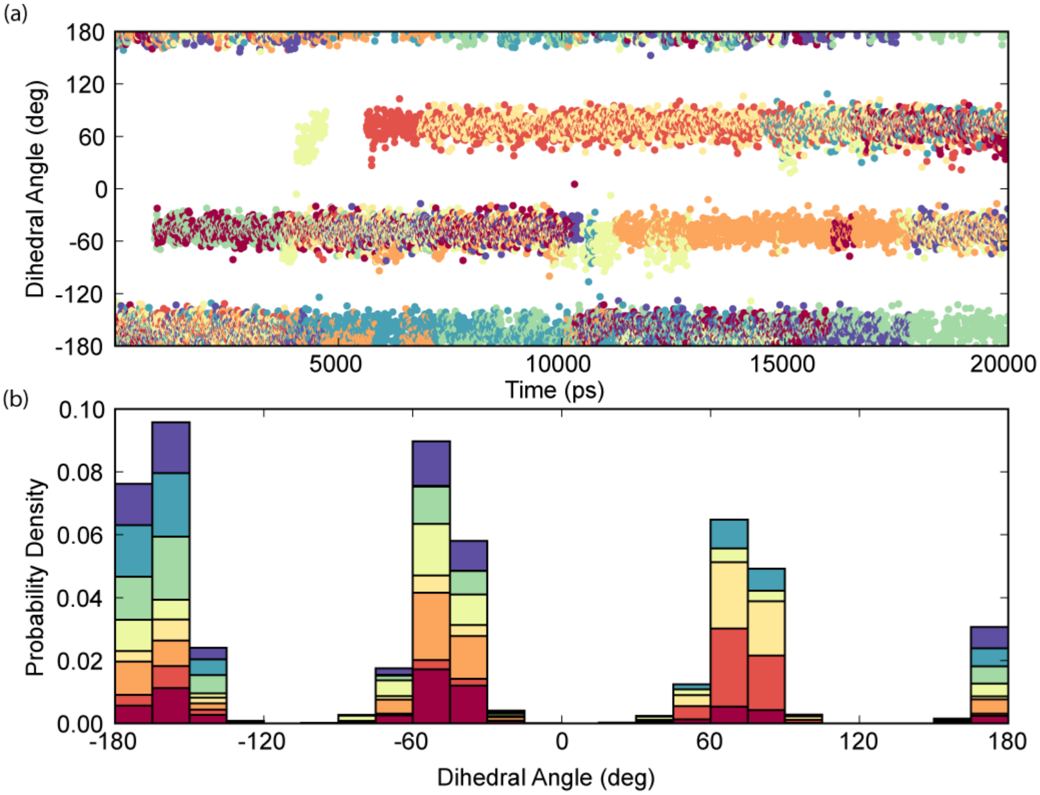

Figure S2.3: A representative example of the conformational sampling of REXREG for the halogenated dihedral angle of MSB (C2-C3-C4-Cl) at $\lambda = 0.11270$. (a) Time series for the 8 replicas shown in different colors. Conformational transitions do occur, but are rare. (b) Histogram of the 8 replicas, with the contribution from each replica shown in different colors. Note the unequal contribution from each replica, indicating non-ergodic sampling.

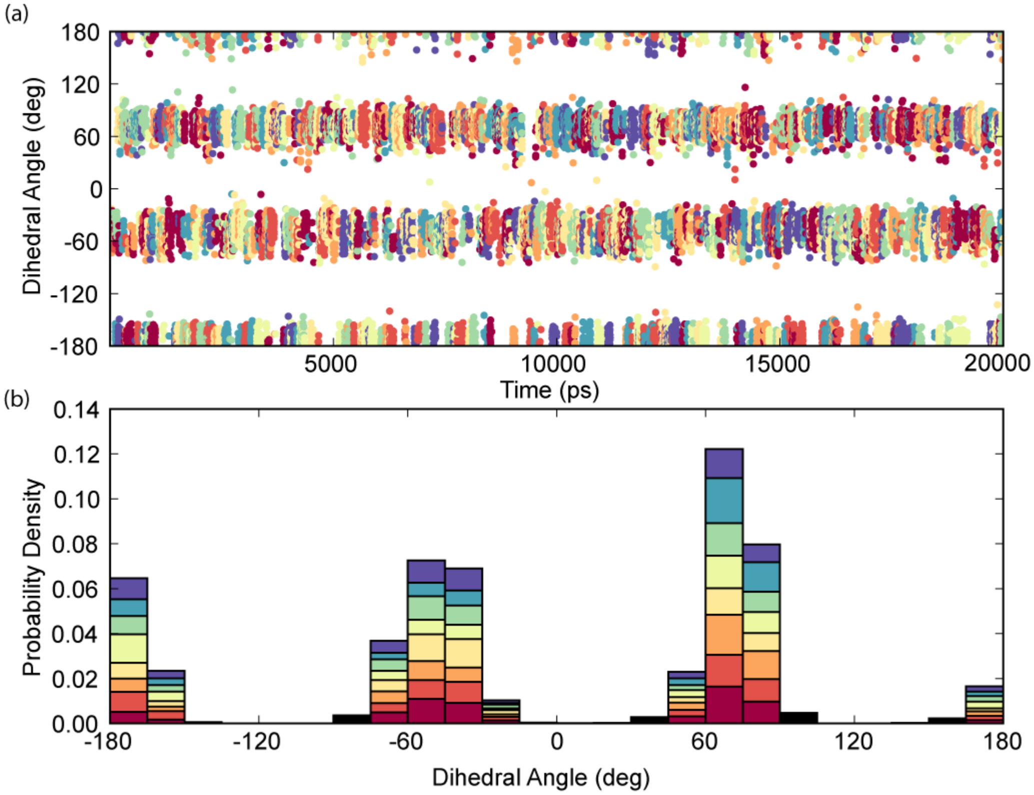

Figure S2.4: A representative example of the conformational sampling of REX-AMDtT for the halogenated dihedral angle of MSB (C2-C3-C4-Cl) at $\lambda = 0.11270$. (a) Time series for the 8 replicas shown in different colors. Conformational transitions very frequent in comparison to REXREG. (b) Histogram of the 8 replicas, with the contribution from each replica shown in different colors. Note the near equal contribution from each replica, indicating ergodic sampling.

Table S2.1: Details for the MSB REXAMDtT simulations. The accelerated molecular dynamics nomenclature follows that in [41]. Units are kcal/mol. bThe effective number of reweighted and uncorrelated instantaneous $dV/d\lambda$ values was calculated using the method of [39]. The total number of effective points from the ten λ 0.11/0.89 and the 0.50 simulation when using the reweighted runs strategy are reported.

State	Boost Parameters				Avg. ΔV		N_{eff}
	E_t	α_t	E_T	α_T	$\lambda = 0.11/0.89$	$\lambda = 0.50$	
s0	n/a	n/a	n/a	n/a	0.0±0.0	0.0±0.0	200000
s1	70	1024	350	2048	29.8±0.7	11.6±0.4	3700; 8340
s2	70	512	350	1024	51.9±1.4	20.9±0.7	1130; 3700
s3	70	256	350	512	81.8±2.7	34.5±1.3	330; 1300
s4	70	154	350	307	105±4	46.5±2.1	160; 530
s5	70	92	350	184	123±7	58.3±3.1	70; 260
s6	70	55	350	111	131±10	67.5±4.2	40; 150
s7	70	33	350	66	129±13	73.0±5.4	30; 100

Chapter 3

Using Multistate Free Energy Techniques to Improve the Efficiency of Replica Exchange Accelerated Molecular Dynamics

Abstract

Replica exchange accelerated molecular dynamics (REXAMD) is a method that enhances conformational sampling while retaining at least one replica on the original potential, thus avoiding the statistical problems of exponential reweighting. In this paper we study three methods that can combine the data from the accelerated replicas to enhance the estimate of properties on the original potential: weighted histogram analysis method (WHAM), pairwise multistate Bennett acceptance ratio (PBAR), and multistate Bennett acceptance ratio (MBAR). We show that the method that makes the most efficient use of equilibrium data from REXAMD simulations is the MBAR method. This observation holds for both alchemical free energy and structural observable prediction. The combination of REXAMD and MBAR should allow for more efficient scaling of the REXAMD method to larger biopolymer systems.

3.1 Introduction

Free energy is the driving force behind biochemical problems of great importance, from drug binding to protein function. A computational approach to free energy calculation affords complete control over the system and method, as well as atomistic detail of the results. The use of computational free energy calculation can theoretically provide free energies for very specific processes that are difficult to isolate in experiment. Among the necessary conditions for accurate free energy prediction are accurate sampling and efficient use of the collected data.

The sampling of computational molecular dynamics has been practically limited due to the topography of the potential energy surface, which requires femtosecond timesteps when propagating the system, and the in the case of biopolymers, sampling many isolated regions of the potential energy. These issues mean that a large number of molecular dynamics steps need to be computed in order to simulate systems with micro- to millisecond relaxation times, but the most heroic brute force approaches have yielded only tens of microseconds. [43, 44] Accurate sampling can also be achieved through methodological developments that increase the sampling efficiency. For example, temperature replica exchange molecular dynamics (TREMD) has enjoyed a surge in popularity following the seminal publication of Sugita and Okamoto. [45] TREMD overcomes local energy barriers by simultaneously simulating the dynamics of a set of non-interacting systems at different temperatures. Conformations sampled at high temperatures are more likely to overcome conformational barriers and exhibit higher sampling efficiency than low temperatures. The sampling efficiency passes between replicas through periodic Metropolis Monte Carlo attempts, which retain the detailed balance of the system. Despite the advantage of conformational exchange between temperatures, TREMD increases the energy of all degrees of freedom instead of activating those specifically important to conformational sampling.

Hamiltonian modification schemes selectively modify the system potential energy and can greatly enhance sampling over TREMD. [46] For example, by applying a harmonic bias along a reaction coordinate of interest, umbrella sampling increases the number of conformations sampled local to the bias minimum [17, 47] Accelerated molecular dynamics (AMD), on the other hand, typically applies a bias along the dihedral degrees of freedom, decreasing the potential barrier between states while retaining the general shape of potential basins. [12] Unlike umbrella sampling AMD does not require prior knowledge of the reaction coordinate, although parameterization of the amount of bias should be done to optimize the acceleration. AMD has been used to reproduce NMR residual dipolar couplings, measures of molecular motion on up to the millisecond time scale, in the third IgG-binding domain of protein G. [38] Despite the promise of AMD to increase the computational efficiency of conformational sampling, each observable must be exponentially reweighted in order to recover the unbiased ensemble average, which Shen and Hamelberg showed yields large statistical uncertainties when applying to highly accelerated simulations. [39]

In order to increase the statistical efficiency of AMD we recently developed a replica exchange accelerated molecular dynamics (REXAMD) scheme, which combines the selective activation of accelerated molecular dynamics and replica exchange. [48] The resulting Hamiltonian replica exchange simultaneously simulates the dynamics on potentials of varying acceleration instead of varying temperature. We previously demonstrated the effectiveness of REXAMD for alchemical free energy calculations of small model systems. [48] These free energy calculations only used the values sampled from the ground, or unaccelerated, state. This analysis strategy makes no use of data generated in the accelerated states and limits the computational efficiency of the method as a result. Furthermore, the computational efficiency will decrease as the system size increases because the number of replicas required for an efficient replica exchange will increase. [49] In order to mitigate this decrease in computational efficiency we will briefly introduce and compare the performance of three methods of recombining multi-state data from REXAMD simulations: WHAM, MBAR, and PBAR.

The most widely used method to combine different biased simulations and produce an estimate of the unbiased result is the weighted histogram method (WHAM). [50] Central to this approach is the definition of an observable of interest, for example $dV/d\lambda$. The bias from each of the multiple states is removed through an exponential reweighting of the biased probability density function of the observable. These unbiased probability density functions, one from each state, are then combined in a weighted sum at each value of the observable subject to the constraints that the weighted variance be minimized and the weights normalized. The WHAM formulation has been extended to REXAMD and applied to combine the (ϕ, ψ) distribution from accelerated potentials during a REXAMD simulation, yielding the unbiased (ϕ, ψ) distribution of various oligopeptides. [40] While making use of the data generated in accelerated states improves the computational efficiency of the REXAMD method, WHAM suffers from an inherent systematic bias that plagues all histogram methods. The bias arises when the probability density of the observable varies greatly over the interval spanned by the bin, an observation rigorously derived by Kobrak. [51] Decreasing the bin width attenuates the problem, but increases the statistical error of the histogram of the simulation data used to approximate the population density. The optimal bin width that balances these two effects should be found for each specific system, but the effects are always present. The pairwise multistate Bennett acceptance ratio method (PBAR) method developed by Maragakis et al. extends the maximum likelihood derivation of the Bennett acceptance ratio method to handle multiple pairs of states simultaneously. [33]

The PBAR method is applicable to both equilibrium and non-equilibrium work data and requires that the work data between different pairs of states be independent. The independence criterion is a practical limitation when applying the PBAR method to the equilibrium samples generated from REMD simulations; the total amount of samples per state N_i must be split into independent samples for each pair of states N_{ij} , thus reducing the statistical quality of the estimates. The PBAR method was validated on a gas-phase alchemical mutation of a capped amino acid simulated using TREMD and showed considerable precision and accuracy from a large data set (360 ns).

Shirts and Chodera developed a different multistate Bennett acceptance ratio method (MBAR). [4] The MBAR approach requires equilibrium samples from each state and the energy of each sampled structure at every state. This lends itself naturally to a REMD approach when the temperature or Hamiltonian modification is straightforward and computing the energy of a structure at the different states is easy. In extreme cases post-processing of the equilibrium samples to compute the energy at different states can be expensive, such as when a soft-core alchemical potential is used. [25, 26] The MBAR analysis uses all of the equilibrium data for each pair of states, and therefore should scale better with the number of states than the PBAR method for equilibrium simulations. Validation was performed by constructing the potential of mean force of the extension of a DNA hairpin in an optical double trap from different constant force trajectories.

In this paper we will first describe the analytical model system and REXAMD simulation scheme. The WHAM, PBAR and MBAR techniques will be discussed in further depth, highlighting the main equations and practical implementation. The performance of WHAM and MBAR will then be compared against the previous ground state approach. The MBAR and PBAR methods will then be compared, and finally the extension of MBAR to computing equilibrium structural properties will be investigated.

3.2 Model System and Methods

A simple, analytically solvable model system provides a wealth of information against which to compare the simulation performance. We selected a linear four-atom molecule (pseudo-butane) with no van der Waals or electrostatics forces, leaving the dihedral angle as the sole degree of freedom. There are two stable conformations at ï£¡ 90 degrees (p-form and m-form, respectively) with different relative depths shown in Figure 3.1. The alchemical change progressed from the dominant p-form to the dominant m-form according to a linear scaling of the potentials (Equation 3.1), and due to the symmetry of the endpoints the alchemical free energy change is zero. The energetic barrier between the two conformations is at least 8 kcal/mol for all , and thus requires aggressive acceleration in order to generate an equilibrium distribution.

$$V(\lambda) = (1 - \lambda) V_0 + \lambda V_1 \tag{3.1}$$

$$\frac{dV}{d\lambda} = V_1 - V_0 \tag{3.2}$$

The REXAMD simulations have four states each, ranging from un-accelerated (s00) to an acceleration that results in a completely flat potential energy surface along the torsional degree of freedom (s03). The exchange rate between these states

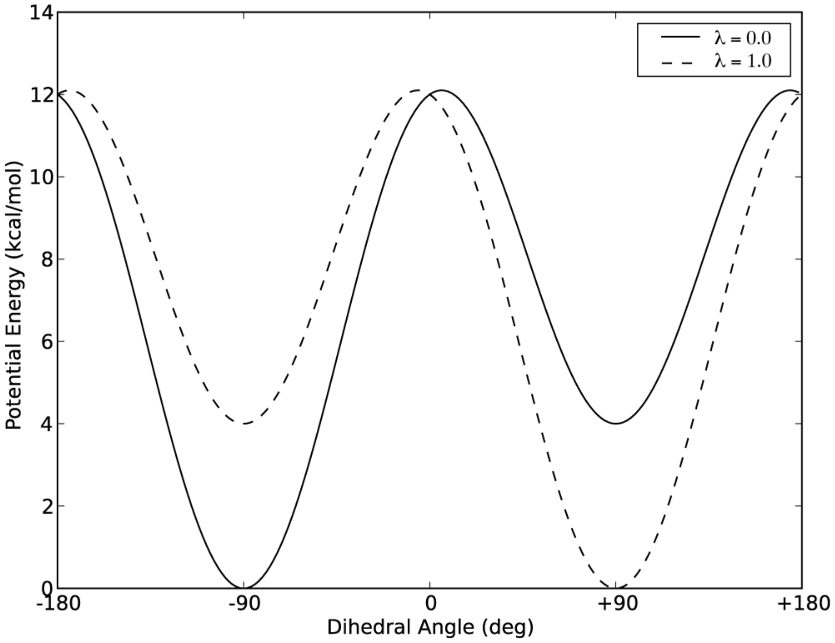

Figure 3.1: Potential energy of the model system.

is higher than 50%, and thus efficient mixing of the replicas occurs on a very short timescale. [3] The exchange period of the REXAMD simulations is 1 ps and the simulations are coupled to a 300K Langevin thermostat with a collision frequency of 50 ps-1.

The instantaneous energy and $dV/d\lambda$ values are taken from the exchanging structures. Systems corresponding to the endpoint lambda values (0.0 and 1.0) as well as the lambda values corresponding to a five-point Gaussian quadrature (0.04691, 0.23077, 0.50000, 0.76923, 0.95309) were simulated for five nanoseconds, and each REXAMD simulation was run four times.

The WHAM method applied to REXAMD relies on unbiasing the probability densities of a specific reaction coordinate for each state j through exponential reweighting (Equation 3.3). Equation 3.3 is exact in the limit of zero bin width, but $P^{j}_{biased}(Q)$ must be approximated by a finite bin width histogram of the observable Q. Kobrak showed analytically that the discretization of a continuous observable Q required by WHAM results in a competition between systematic and statistical error. [51] In the formalism of WHAM for REXAMD the systematic error arises from the approximation of the $P^j_{biased}\left(Q\right)$ and $\left\langle exp\left(\beta V^j_{bias}\right)\right\rangle$ from a finite bin width and increases with increasing bin width. The discretized $P^j_{biased}\left(Q\right)$ and $\left\langle exp\left(\beta V^j_{bias}\right)\right\rangle$ are estimated from a histogram of the sampled data, which introduces a statistical error that decreases with increasing bin width. The $P^j_{unbiased}\left(Q\right)$ from each state are then combined according to Equation 3.4 with a set of weights $w_j\left(Q\right)$ that minimize the variance at each Q. These conditions result in a set of self-consistent equations that can be iterated over until the desired level of precision is achieved (Equation A15 in Reference [40]).

$$P^j_{unbiased}\left(Q\right) = \frac{Z^j_{biased}}{Z_{unbiased}} P_{biased}j\left(Q\right)\left\langle exp\left(\beta V^j_{bias}\right)\right\rangle \tag{3.3}$$

$$P_{unbiased} = \sum_{j=1}^{K} w_j\left(Q\right) P^j_{biased}\left(Q\right) \tag{3.4}$$

The MBAR method is rooted in the identity given in Equation 3.5 (Equation 5 in Reference 14) where $q_i\left(x\right)$ and $q_j\left(x\right)$ designate the un-normalized probability density functions of the configuration x in states i and j respectively, is some arbitrary function, Z_i and Z_j are the configuration integrals from state i and j respectively, and Γ indicates that the integrals are evaluated over all configuration space. Approximating the expectation values as discrete averages of equilibrium data and summing over all states results in a set of K estimating equations (Equation 3.6), where K is the total number of states and N_i is the number of structures sampled at state i. The details of the selection of the function $\alpha_{ij}\left(x\right)$ are outside of the scope of this paper, but the selection exhibits the lowest variance of common reweighting estimators. [4, 52] The solution of Equation 3.6 yields estimates for the partition function Z_i of each state.

$$\int_{\Gamma} q_i\left(z\right)\alpha_{ij}\left(x\right) q_j\left(x\right) dx = Z_i \left\langle \alpha_{ij} q_j\right\rangle_i = Z_j \left\langle \alpha_{ij} q_i\right\rangle_j \tag{3.5}$$

$$\sum_{j=1}^{K} \frac{Z_i}{N_i} \sum_{n=1}^{N_i} \alpha_{ij} q_j\left(x_{in}\right) = \sum_{i=1}^{K} \frac{Z_j}{N_i} \sum_{n=1}^{N_j} \alpha_{ij} q_i\left(x_{jn}\right) \tag{3.6}$$

The Python implementation of MBAR, PyMBAR, was used for all MBAR analysis. The statistical uncertainty of the free energies and expectation values are based on an estimate of the asymptotic covariance matrix of the provided data, which requires an uncorrelated dataset. Correlation can be removed from the molecular dynamics data by subsampling the original data set at an interval greater than or equal to the equilibrium relaxation time of the molecular dynamics system. For this work we used the subsampling technique implemented in PyMBAR. [4]

The PBAR method extends the maximum-likelihood derivation of the Bennett acceptance ratio to multiple states. The log likelihood (Equation 3.7) involves the Fermi function $f(x) = 1/(1 + exp(x))$ of the instantaneous work values between states W_{ij}, the free energy between states ΔF_{ij}, and the constant $M_{ij} = k_B T ln(N_{ij}/N_{ji})$ that accounts for a different number of samples in the forward and reverse direction. The instantaneous work W_{ij} is defined as the difference in potential energies of a specific configuration at states i and j, and these work values must be independent for Equation 3.7 to hold. This requires that a structure x_{in} pulled from an equilibrium simulation of state i can only be used to calculate the work to go to a single other state. The entire set of x_{in} must be separated into K non-overlapping sets of $x_{n_{ij}}$, which greatly reduces the number of data points per state pair ij compared to MBAR. [4] The log likelihood has well defined derivatives and thus any optimization method can be used to find the set of Z_i that maximizes the likelihood function. We implemented the PBAR method in Python using a gradient descent optimization. In order to estimate the statistical uncertainty of the PBAR method the PBAR calculation was repeated multiple times using random subsets of the provided work data and we report the average and standard deviation of the results. [33]

$$ln(L) = \sum_{i}^{K} \sum_{i \neq j}^{K} \sum_{n_{ij}}^{N_{ij}} f\left(-\beta\left[W_{ij}\left(x_{n_{ij}}\right) - \Delta F_{ij} + M_{ij}\right]\right) \tag{3.7}$$

3.3 Combining Multistate Information for TI

The first application of the REXAMD method to alchemical free energy calculations used the instantaneous $dV/d\lambda$ values taken from the un-accelerated, or ground state, to compute $\langle dV/d\lambda \rangle$ for use with Gaussian quadrature thermodynamic integration (Equation 3.8). [48] The ground state only represents $1/N$ of the total data, where N is the number of replicas per REXAMD, and therefore a significant fraction

Table 3.1: Compiled Expectation Values of $dV/d\lambda$

λ_i	Expectation value of $dV/d\lambda$ at λ_i (kcal/mol)				
	TI-GS	TI-WHAM	*Individual Lambda* TI-MBAR	*All Lambdas* TI-MBAR	Analytical
0.04691	+3.904 ± 0.004	+3.138 ± 0.008	+3.922 ± 0.001	+3.922 ± 0.000	+3.923
0.23077	+3.69 ± 0.01	+2.54 ± 0.02	+3.714 ± 0.005	+3.722 ± 0.002	+3.722
0.50000	-0.09 ± 0.08	-0.08 ± 0.06	-0.06 ± 0.04	+0.01 ± 0.02	+0.000
0.76923	-3.703 ± 0.006	-2.48 ± 0.02	-3.710 ± 0.005	-3.721 ± 0.002	-3.722
0.95309	-3.926 ± 0.003	-3.125 ± 0.004	-3.922 ± 0.001	-3.922 ± 0.000	-3.923
ΔF	-0.03 ± 0.02	-0.01 ± 0.02	-0.02 ± 0.01	+0.002 ± 0.006	+0.000

of the data is never used. Multiple REXAMD simulations at each λ_i are then used to estimate the statistical uncertainty of the TI Gaussian quadrature calculation. This was previously referred to as the reweighted runs strategy, but we will refer to this approach as TI-GS (Ground State) in this paper. The five-point Gaussian quadrature $\langle dV/d\lambda \rangle$ from the four combined runs are summarized in Table 3.1. The TI-GS values are not within the estimated statistical uncertainty of the analytical result, so there is room for the multistate methods to show improvement.

$$\Delta F = \int_0^1 \langle dV/d\lambda \rangle_{\lambda'} \, d\lambda' \approx \sum_i^N w_N\left(\lambda_i\right) \langle dV/d\lambda \rangle_{\lambda_i} \tag{3.8}$$

In TI-WHAM the four replicas of a REXAMD simulation at a specific λ_i are combined to estimate $\langle dV/d\lambda \rangle_{\lambda_i}$. The instantaneous $dV/d\lambda_{\lambda_i}$ values were separated into 8000 bins, which gave the optimal estimates of $\langle dV/d\lambda \rangle_{\lambda_i}$. We then calculated the unbiased probability density function of $dV/d\lambda_{\lambda_i}$ and subsequently the $\langle dV/d\lambda \rangle_{\lambda_i}$ shown in Table 3.1. The WHAM estimates of $\langle dV/d\lambda \rangle_{\lambda_i}$ deviate strongly from the analytical results, and are actually worse than the TI-GS estimates. The steep curvature of the population density of $dV/d\lambda_{\lambda_i}$ (not shown) forces a narrow bin width, which in turn increases the statistical error of the histogram approximation leading to a poor estimate of the biased probability density function and eventually $\langle dV/d\lambda \rangle_{\lambda_i}$. The high accuracy of the computed free energy change from TI-WHAM is an artifact of using a symmetric alchemical change, and cannot be expected in realistic systems.

The *Individual Lamda* TI-MBAR also uses the four REXAMD replicas for each λ_i to compute $\langle dV/d\lambda \rangle_{\lambda_i}$. This is exactly equivalent to the TI-WHAM method in the limit of zero bin width and should exhibit less bias. [4] The $\langle dV/d\lambda \rangle_{\lambda_i}$ in Table 3.1

Table 3.2: Comparison of TI and TI-MBAR methods. The combined TI-GS results report the average and standard error from the four individual runs. The individual lambda method only uses information from a specific lambda value to compute $\langle dV/d\lambda\rangle_{\lambda_i}$ at each λ_i. The all lambdas method uses information from all the lambda values to compute $\langle dV/d\lambda\rangle_{\lambda_i}$ at each λ_i.

Run	Alchemical Free Energy (kcal/mol)		
	TI-GS	*Individual Lambda* TI-MBAR	*All Lambdas* TI-MBAR
01	-0.055	-0.04 ± 0.02	+0.02 ± 0.01
02	-0.051	-0.07 ± 0.03	-0.02 ± 0.01
03	+0.040	+0.04 ± 0.03	+0.02 ± 0.01
04	-0.067	-0.05 ± 0.03	-0.02 ± 0.01
Combined	-0.03 ± 0.021	-0.02 ± 0.01	+0.002 ± 0.006

show improvement over both the TI-GS and TI-WHAM results, and are very close to the analytical results. The combined TI-GS result is calculated with four duplicate simulations of five nanoseconds each, resulting in a total of 20,000 structures per $\langle dV/d\lambda\rangle_{\lambda_i}$. The *Individual Lambda* TI-MBAR runs utilize the four REXAMD states of five nanoseconds each for a similar 20,000 structures per $\langle dV/d\lambda\rangle_{\lambda_i}$, and yield results that are quite comparable to the combined TI-GS result (Table 3.2). In other words, these results imply that *Individual Lambda* TI-MBAR is able to calculate $\langle dV/d\lambda\rangle_{\lambda_i}$ to a comparable accuracy as TI-GS with only a quarter of the data required by TI-GS.

A more efficient use of the data is to simultaneously use all four states from all five λ_i and thus utilize 100,000 structures per $\langle dV/d\lambda\rangle_{\lambda_i}$. This *All Lambdas* TI-MBAR performs very well when considering the four individual five nanosecond runs, but really shines when all four runs are combined (Table II). It is noteworthy that the MBAR estimate of the statistical uncertainty for a relatively low number of data points, namely the individual runs of *All Lambdas* TI-MBAR, is too low to account for the offset of the calculated free energy from the analytical result. An increase in the number of samples used does correct this, but this effect should be studied further to gain confidence in the MBAR uncertainty estimate as it applies to larger systems.

Table 3.3: Comparison of TI-MBAR to MBAR and PBAR. The uncertainty of the TI-MBAR and MBAR comes from the asymptotic covariance matrix estimator from Shirts and Chodera. [4] The uncertainty in the MBAR Subsets results is estimated from the standard deviation from 1000 different subsets of 185 random reduced potential values per state for the individual runs, and 1000 different subsets of 740 random reduced potential values for the combined result. The uncertainty in the PBAR results is estimated from the standard deviation from 1000 different subsets of 185 random work values per pair of states for the individual runs, and 1000 different subsets of 740 random work values for the combined result.

Run	Free Energy (kcal/mol)			
	TI-MBAR	MBAR	MBAR Subsets	PBAR
01	+0.01 ± 0.01	+0.005 ± 0.009	-0.00 ± 0.03	-0.01 ± 0.02
02	-0.02 ± 0.01	-0.016 ± 0.009	-0.01 ± 0.03	-0.02 ± 0.02
03	+0.00 ± 0.01	+0.000 ± 0.009	+0.01 ± 0.03	+0.00 ± 0.02
04	+0.00 ± 0.01	+0.001 ± 0.009	-0.01 ± 0.03	-0.01 ± 0.02
Combined	-0.002 ± 0.005	-0.001± 0.005	-0.00 ± 0.01	-0.007 ± 0.008

3.4 Direct Multistate Free Energy Estimates

The TI-MBAR method was helpful in comparing the advantage of MBAR over TI-WHAM, but the PBAR and MBAR methods were developed to directly estimate the free energy difference between states. The λ_i endpoints (0.0 and 1.0) need to be simulated in order for a direct estimate of the free energy difference from MBAR and PBAR, and the results in Table 3.3 include data from all seven λ_i. The addition of the endpoint data improves the combined runs TI-MBAR result from $+0.002 \pm 0.006$ to -0.002 ± 0.005 kcal/mol. The direct MBAR method shows slightly better accuracy and precision than TI-MBAR for all of the runs as well as the combined data set. The performance gain of MBAR relative to TI-MBAR is expected to increase with the system size as the bias inherent the Gaussian quadrature process will be more evident in complex, non-symmetric alchemical changes.

Interestingly, the direct MBAR method also outperforms the PBAR results. Shirts and Chodera predicted that because the PBAR method requires independent sets of work values between each pair of states MBAR would make better use of equilibrium data. [4] To get independent sets of work values the N_i samples per state must be distributed into N_{ij} sets that will be used to calculate the instantaneous work to go from state i to state j. With K total states and N_i structures per state each N_{ij} can have at most N_i/K structures, and for large values of K the decrease in the number of available data points relative to the equilibrium number N_i is large. For example, the four replicas and seven λ_i used in the model system PBAR calculation reduces the 5000 N_i structures to only 185 N_{ij} structures. In order to better use the equilibrium data set Maragakis et al. suggested repeating the PBAR analysis with different random subsets and reporting the average and standard deviation. [33] The average should be recovered with this method, but due to the small number of in each PBAR calculation the standard deviation can be expected to be significantly higher in PBAR than MBAR, which is observed in Table 3.3.

The MBAR Subsets results follow the procedure outlined above for sampling random subsets of the total structures and reporting the average and standard deviation. The size of for each MBAR Subset was limited to same size required by PBAR (185 for the individual runs, 740 for the combined runs), although for MBAR to work properly it must use the same 185 structures from N_i for each pair of states ij. The MBAR Subset results are comparable to the PBAR results, showing that the increased precision of the direct MBAR method is indeed due to the more efficient use of equilibrium data.

3.5 Equilibrium Conformations

The REXAMD method is not limited to free energy calculations, and has been used for determining equilibrium structural properties in conjunction with the WHAM method. [40] The calculation of expectation values using MBAR instead of WHAM should avoid the bias introduced by discrete binning during the reconstruction process. [51, 4] The MBAR method naturally computes equilibrium expectation values, and this can be extended to computing population histograms and potential of mean force for observables. Figure 3.2 illustrates the distribution of the model systemï£¡s dihedral angle from a single simulation at where the potential energy surface is symmetric. The five thousand structures from the ground state method (Figure 3.2) show an overpopulation of the p-form. All four REXAMD states at can be used

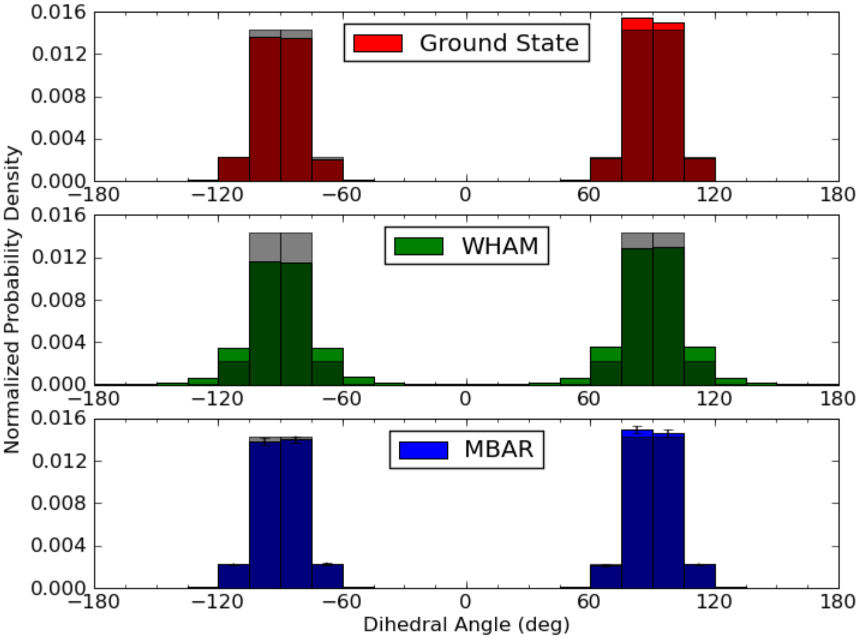

Figure 3.2: Comparison of the unbiased dihedral histogram for $\lambda_i = 0.5$ from 5 nanoseconds and 7 λ_i. The shaded histogram in each plot is the analytical result for the model system. The Ground State result shows the distribution of the $\lambda_i = 0.5$ ground state replica. The WHAM result shows the WHAM reconstruction from the four REXAMD states at $\lambda_i = 0.5$. The MBAR result shows the histogram derived from the four REXAMD states at all 7 λ_i.

to generate the unbiased probability density histogram using WHAM. The result does show a higher degree of symmetry than the ground state method, but the systematic error introduced by the large bin widths causes the result to deviate strongly from the analytical result. Decreasing the bin width does reduce this effect (Supplemental), but finding the optimal bin width to balance the systematic and statistical errors can be difficult. [51] If instead all four states of all seven are used by MBAR to compute the population histogram the probability density the analytical result is completely recovered (blue in Figure 3.2). This finding reiterates the benefit of using MBAR over WHAM when combining equilibrium distributions from different biases.

Many important equilibrium conformational analysis techniques can be expressed as expectation values, and thus would greatly benefit form the combination of

REXAMD and MBAR. For example, the covariance matrix and the related principal components analysis and quasi-harmonic entropy (for the mass-weighted covariance matrix), which are very sensitive to sampling, are defined in terms of expectation values (Equation 3.9). The root mean squared deviation and the root mean squared fluctuation are also expectation values that are frequently used in conformational analysis.

$$Cov\left(X_i, X_j\right) = \left\langle \left(X_i - \left\langle X_i \right\rangle\right)\left(X_j - \left\langle X_j \right\rangle\right) \right\rangle = \left\langle X_i X_j \right\rangle - \left\langle X_i \right\rangle \left\langle X_j \right\rangle \tag{3.9}$$

3.6 Conclusion

All REMD methods require an increasing number of replicas as the system size increases, and it becomes important to make efficient use of all of the replica data and not just the data at the desired temperature or acceleration. We set out to determine the best way to combine the biased data from REXAMD simulations to improve the accuracy of free energy calculations and structural analysis. The performance of WHAM and MBAR at computing the $\left\langle dV/d\lambda \right\rangle_{\lambda_i}$ from multiple acceleration states at specific λ_i was compared to the $\left\langle dV/d\lambda \right\rangle_{\lambda_i}$ from the ground state (no acceleration). TI-WHAM performed worse than the TI-GS, and this result was discussed in light of the competing errors when selecting the WHAM bin width. The Individual Lambda TI-MBAR, which is comparable to TI-WHAM in terms of number of structures used to compute each $\left\langle dV/d\lambda \right\rangle_{\lambda_i}$, was very close to the analytical result. The asymptotic covariance matrix estimator was not able to cover the offset from the analytical results and indicates that MBAR will underestimate the statistical uncertainty when used with a relatively low number of samples. The All Lambdas TI-MBAR represented the most efficient method of calculating $\left\langle dV/d\lambda \right\rangle_{\lambda_i}$ with MBAR given the five intermediate λ_i and with this amount of data the estimates of $\left\langle dV/d\lambda \right\rangle_{\lambda_i}$ were both precise and accurate. The results were comparable to using four times as much ground state data, and show that the computational gain by combining the replica information from REXAMD is excellent.

The MBAR and PBAR methods were then compared against each other, which required simulating the alchemical endpoints. The MBAR method was approximately an order of magnitude more precise than the PBAR method. The inefficiency of PBAR relative to MBAR was shown to be due to the requirement of independent data sets for PBAR, which reduced the amount of data available for each pair of states during the analysis. The most efficient way of combining equilibrium samples of REXAMD data is conclusively MBAR. We then demonstrated the usefulness of MBAR in combining multiple states to generate unbiased structural quantities. The combination of REXAMD and MBAR should allow large system sizes to be efficiently sampled and analyzed.

Acknowledgement

Chapter 3, in full, is a minimally modified reprint of the material as it appears in M. Fajer, R. Swift, J. A. McCammon, âĂIJUsing Multistate Free Energy Techniques to Improve the Efficiency of Replica Exchange Accelerated Molecular Dynamics.âĂİ J. Comp. Chem. 30, 1719-1725 (2009). The dissertation author was the principal investigator and first author of this paper.

Chapter 4

Intrinsic free energy of the conformational transition of the KcsA signature peptide from conducting to non-conducting state

Abstract

We explore a conformational transition of the TATTVGYG signature peptide of the KcsA ion selectivity filter and its GYG to AYA mutant from the conducting α-strand state into the non-conducting pII-like state using a novel technique for multidimensional optimization of transition path ensembles and free energy calculations. We find that the wild type peptide, unlike the mutant, intrinsically favors the conducting state due to G77 backbone propensities and additional hydrophobic interaction between the V76 and Y78 sidechains in water. The molecular mechanical free energy profiles in explicit water are in very good agreement with the corresponding adiabatic energies from the Generalized Born Molecular Volume (GBMV) implicit solvent model. However comparisons of the energies to higher level B3LYP/6-31G(d) Density Functional Theory calculations with Polarizable Continuum Model (PCM) suggest that the non-conducting state might be more favorable than predicted by molecular mechanics simulations. By extrapolating the single peptide results to the

tetrameric channel, we propose a novel hypothesis for the ion selectivity mechanism.

4.1 Introduction

Organisms transmit electric impulses by means of cellular membrane polarization that critically depends on the work of ion channels. These channels permit passage of specific ion types across the membrane. Ion channels selective for potassium such as KcsA [53, 54] are particularly interesting as they solve a non-trivial problem of selecting larger K+ over smaller Na+ ions. Despite the wealth of information derived from both experimental [55, 56, 54, 53] and computational [57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70] studies of potassium channels, the mechanism of selectivity in these biological machines remains too difficult to tackle as it requires probing the multi-ion permeation transition states. [56, 55, 57, 59, 60, 63, 64, 65, 66, 67, 68, 69, 70] From the computational perspective, this task demands computing multi-dimensional potentials of mean force (PMFs) for which efficient tools have been lacking. [71, 72, 73, 21, 74, 75, 76, 77, 78, 79, 80, 81, 82, 9, 83, 84, 85, 86]

Recently, we have developed and the generalized gradient-augmented Harmonic Fourier Beads (ggaHFB) method [73, 21, 74] that allows studying rare events in complex molecular systems by extending FukuiâĂŹs intrinsic reaction coordinate (IRC) approach [87, 88] with the help of the multi-dimensional free-energy gradient. [21, 74, 89, 90]

In the present paper we apply the ggaHFB methodology to study an important functional transition of the signature peptide TATTVGYG of the KcsA selectivity filter that pinches the filter shut by flipping its V76 carbonyl group away from the channel axis coupled with the V76 sidechain rotation in response to lowering the K+ concentration. [53, 56, 55] The V76 carbonyl group flip in the KcsA channel is associated with the αL to pII backbone conformational transition at the G77 residue of the signature peptide, and is believed to switch the selectivity filter from a conducting (αL) to a non-conducting (pII) state. This transition has been alluded to by X-ray crystallography that detected a partial flip of the V76 backbone carbonyls in the wild type KcsA upon lowering K+ concentration, [53, 54] and recently a more pronounced flip in the E71A mutant. [91] Similar transitions have been observed in numerous molecular dynamics (MD) simulations of KcsA [59, 60] and other related channels. [92, 93, 94] Interestingly, X-ray crystallographic studies indicate that the αL to pII backbone transition is accompanied by rotation of the V76 sidechain. However, to the best of our knowledge, previous MD simulations of the KcsA and related potassium channels did not report such a rotation. Furthermore, while the carbonyl flip into pII state observed by X-ray crystallography preserved the four-fold symmetry of the channel, the MD simulations reported only a single strand out of four identical strands to undergo the αL to pII transition, thus breaking the symmetry of the channel.

It is possible that averaging over the four strands of the filter might artificially diminish the extent of the transition seen by the X-ray crystallography, thus masking the symmetry breaking. However, unambiguous demonstration of the symmetry breaking requires assessing the free energy of the conformational transition in the full tetrameric channel. Although possible to accomplish with the help of the ggaHFB method, this task is computationally intensive as it requires free energy optimization of a transition path ensemble for a relatively large system. On the other hand, exploring the same transition using a single peptide might provide useful insights into the function of the tetrameric channel with reduced computational burden. In particular, the intrinsic free energy profile should provide relative free energies of the αL and pII states along with the corresponding free energy barrier outside the channel environment, and thus suggest whether multiple transitions inside the channel are likely.

We define the intrinsic free energy profile of the peptide as that of a single peptide in water. Our choice of water medium has been motivated by the following observations. The distributions of the Ramachandran dihedral angles of various residues in the existing protein structures resemble those from the corresponding adiabatic maps in water, but differ markedly from those in gas phase. [95, 96, 97, 98, 99] Even though KcsA is a trans-membrane protein, when fully assembled and in conducting state, water molecules can access the back of the selectivity filter, where they participate in hydrogen bonding with E71 and D80 residues (not present in our model). [53, 54, 67, 100] Additional water molecules reach behind the selectivity filter to interact with other residues of the signature peptide in the non-conducting state. [54, 56] Furthermore, the filter is known to conduct water with and without the ions and hence has a water accessible interior. [53, 54, 56, 101] Therefore, we feel that the study of the behavior of a single selectivity peptide in water will provide useful insights for understanding the behavior of the same peptide in the tetrameric channel.

This paper is organized as follows. First, we review the ggaHFB methodology for finding minimum adiabatic potential energy paths and minimum free energy transition path ensembles, and computing corresponding energy profiles. Combining the ggaHFB transition path ensemble optimization and free-energy evaluation capabilities with the available X-ray structural information, we then explore the intrinsic free energy profile of the signature peptide underlying the flip of the V76 carbonyl from conducting into the non-conducting state. [102, 103, 104, **?**, 106] Furthermore, we evaluate the effect of the V76 sidechain rotation on the backbone transition. To derive additional support for the functional importance of the specified transition to the KcsA channel, we compare the free energy profile of the wild type peptide to that of the GYG to AYA mutant. Note that a closely related G77A mutant either abolishes the selectivity [107] or abrogates the activity of the channel. [108] To diffuse any doubts regarding the choice of the water environment for our study, we examine the changes to the functional transition upon removing the peptide from water and placing it into gas phase. Here we fully utilize the ggaHFB capabilities in finding minimum adiabatic potential energy pathways and computing the corresponding energy profiles via the generalized line integral formalism. Finally, we provide some benchmarks to lend credence to the computed energy profiles in water. In particular, we gauge the molecular mechanical (MM) CHARMM22 forcefield [109, 110] against a popular Quantum Mechanical (QM) Density Functional Theory model, namely B3LYP [111, 112, 113] with 6-31G(d) basis set. To account for the solvent contribution, we employ the Generalized Born Molecular Volume (GBMV) [97, 114] and Polarizable Continuum Model (PCM) [115, 116, 117, 118] with the MM and QM energy functions, respectively.

4.2 Methodology

Given the novelty of the employed transition path and path ensemble optimization technique, the generalized gradient augmented Harmonic Fourier Beads method, that makes this study possible, we briefly describe the main points of the method in the following paragraphs.

4.2.1 Reactive coordinate space (RCS) and biasing potential

The generalized gradient-augmented Harmonic Fourier Beads (ggaHFB) method considers an arbitrary system of N atoms described by $3N$ generalized coordinates $\bar{Q} = (q_1, ..., q_{3N})$, and, equivalently, by $3N$ Cartesian coordinates $\bar{X} = (x_1, ..., x_{3N})$. The method derives the gradient of either adiabatic potential energy or the free energy of the system with respect to a selected subset of $S \leq 3N$ coordinates $\bar{q} = (q_1, ..., q_S)$ that comprise the reactive coordinate space (RCS) by employing either biased optimization or biased molecular dynamics (MD) or Monte Carlo (MC) simulations, correspondingly. The remaining $3N - S$ degrees of freedom $\bar{r} = (q_{S+1}, ..., q_{3N})$ comprise the spectator coordinate space (SCS) and do not contribute explicitly to the energy gradient.

The biasing potential is a linear combination of relatively stiff harmonic restraints and applies only to the RCS degrees of freedom centered at a reference configuration $\bar{q}^{b,\mathbf{ref}} = \left(q_1^{b,\mathbf{ref}}, ..., q_S^{b,\mathbf{ref}}\right)$: [21, 74]

$$V^b \left(q_1, ..., q_S; q_1^{b,\mathbf{ref}}, ..., q_S^{b,\mathbf{ref}}\right) = \sum_{i=1}^{S} k_i^b \left(q_i - q_i^{b,\mathbf{ref}}\right)^2 \tag{4.1}$$

Here superscript b indicates the bias, and k_i^b is the $i^{\mathbf{th}}$ coordinate bias force constant. This biasing potential allows deriving the desired energy gradients using a very simple idea described in the following section.

4.2.2 Adiabatic potential energy gradient from biased optimization

The key idea for computing the energy gradients is most clearly demonstrated on the example of the adiabatic potential energy. Let us add the biasing potential (1.1) to the total energy of the system $U\left(\bar{Q}\right) = U\left(\bar{q}, \bar{r}\right)$ and then perform potential energy optimization on the modified potential energy surface. Such optimization should reach an equilibrium point at which the forces from the biasing potential that apply only to the S degrees of freedom balance those from the potential energy. Because the forces on the remaining $3N - S$ degrees of freedom become identically zero due to optimization, the equilibrium point provides the gradient not of the full potential energy, but instead of the adiabatic potential energy. Therefore, the biased optimization yields the gradient of the adiabatic potential energy in RCS. The following equations summarize the above.

$$\left.\frac{\partial U\left(\bar{q},\bar{r}\right)}{\partial q_i}\right|_{\bar{q}=[\bar{q}]^b,\bar{r}=[\bar{r}]} = \left.\frac{\partial V\left(\bar{q}\right)}{\partial q_i}\right|_{\bar{q}=[\bar{q}]^b} = -2k_i^b\left([q_i]^b - q_i^{b,\mathbf{ref}}\right) \tag{4.2}$$

The square brackets indicate the local minimum on the modified potential energy surface. This procedure effectively reduces the full potential energy surface of $3N$ degrees of freedom to the adiabatic potential energy surface of $S \leq 3N$ degrees of freedom. It is worth noting that in order to compute the adiabatic potential energy gradient on steep slopes in the vicinity of transition states one has to use somewhat stiff springs. Otherwise the minimum on the modified energy surface will slide downhill close to the corresponding minimum on the full energy surface providing little or no information about the transition state region. This remark also applies to the free energy gradient discussed in the next paragraph.

4.2.3 Free energy gradient from biased simulations

The idea used to derive the gradient of the adiabatic potential energy can be applied to derive the gradient of the free energy from biased simulations. For the proof of this statement we refer the reader to the previous work [21, 74, 89, 90] and only summarize the results here. It has been demonstrated that for somewhat stiff Cartesian restraint (1.1) with reference configuration $\bar{x}^{b,\mathbf{ref}}$ in RCS, one can compute the corresponding Cartesian free energy gradient via equation (1.3).

$$\left.\frac{\partial W^u\left(\bar{x}\right)}{\partial x_i}\right|_{\bar{x}=\langle\bar{x}\rangle^b} \approx -2k_i^b\left(\langle x_i\rangle^b - x_i^{b,\mathbf{ref}}\right) \tag{4.3}$$

Similarly, for the restraint (1.1) in generalized coordinates centered at $\bar{q}^{b,\mathrm{ref}}$ the corresponding free energy gradient is given by equation (1.4). [74]

$$\left.\frac{\partial W^u(\bar{q})}{\partial q_i}\right|_{\bar{q}=\langle\bar{q}\rangle^b} \approx -2k_i^b\left(\langle q_i\rangle^b - q_i^{b,\mathbf{ref}}\right) + k_B T \left.\frac{\partial \ln|J(\bar{q})|}{\partial q_i}\right|_{\bar{q}=\langle\bar{q}\rangle^b} \tag{4.4}$$

Here W^u is the unbiased free energy, k_B is the Boltzmann constant, T is the simulation temperature, and $|J(\bar{q})|$ is the ensemble-reduced Jacobian for the transformation from Cartesian to the generalized coordinates. Note that equation (1.4) is practically identical to equation (1.2) for the adiabatic potential energy gradient, where the biased ensemble average $\langle\bar{q}\rangle^b = \left(\langle q_1\rangle^b, ..., \langle q_S\rangle^b\right)$ replaces the local minimum $[\bar{q}]^b$ configuration. The additional logarithmic Jacobian term on the right hand side of the generalized gradient expression (1.4) is the consequence of using Cartesian MD or MC propagators with the nonlinear restraints. [74, 47] Unlike the case for the adiabatic potential energy gradient, the free energy gradient expression is approximate.

The quality of the free energy gradient depends on the stiffness of the harmonic restraint [21, 74] and on the quality of the corresponding configuration averages. To achieve the highest quality, one can either run a single very long simulation or run several short simulations and then combine the results into the cumulative average. We prefer the latter approach for accurate free energy calculations as it allows monitoring convergence of the gradient. Specifically, running P batches of short MD or MC simulations of equal length subject to the restraint (1.1) provides P sets of averaged coordinates or âĂIJevolved beadsâĂİ $\langle\bar{q}\rangle = \left(\langle q_1\rangle^{b,j}, ..., \langle q_S\rangle^{b,j}\right)$ for a given reference bead, where j is the batch number. These averages could then be easily combined to yield the higher quality cumulative average:

$$\langle\bar{q}\rangle^b = \frac{1}{P}\sum_{j=1}^{P}\langle\bar{q}\rangle^{b,j} \tag{4.5}$$

Importantly, the averaged configuration provides the complete free energy gradient in RCS and not just one of its components:

$$\nabla W^u(\bar{q})\big|_{\langle\bar{q}\rangle^b} = \left(\left.\frac{\partial W^u(\bar{q})}{\partial q_1}\right|_{\langle\bar{q}\rangle^b}, ..., \left.\frac{\partial W^u(\bar{q})}{\partial q_S}\right|_{\langle\bar{q}\rangle^b}\right) \tag{4.6}$$

This property of the ggaHFB method is a great advantage over the histogram-based free energy estimates that require much larger arrays of simulations to populate multidimensional histograms. [21, 74, 89, 90, 50, 119] Therefore, the ggaHFB method offers a practical alternative to the conventional umbrella sampling simulations with weighted histogram analysis method (WHAM). [50, 119]

The ability to compute the free energy gradient efficiently makes it possible to perform gradient-driven optimization on free energy surfaces and ultimately to find minimum free energy transition path ensembles.

4.2.4 Minimum adiabatic energy transition path

The ggaHFB method as a path finding tool belongs to the class of double-ended reaction path methods that require a reactant and a product state to describe a transition of interest. [71, 72, 73, 21, 74, 75, 76, 77, 120, 121, 122, 123, 124, 125, 126, 127] Importantly, the ggaHFB method finds reaction or transition paths that are invariant with respect to coordinate transformations. The concept of invariant reaction paths, called âĂIJintrinsic reaction coordinateâĂİ (IRC), has been developed by Fukui for the full potential energy surfaces [87, 88] and has been further elaborated by many authors since. [71, 77, 123, 128, 129, 130, 131, 132] In simple terms IRC represents the center curve of the reaction path region that follows the invariant energy gradient.

In particular, in Cartesian coordinates the IRC curve satisfies the following simple condition:

$$\nabla_{\perp} U\left(\bar{X}\right) = \nabla U\left(\bar{X}\right) - \vec{n}\left(\bar{X}\right) \frac{\vec{n}\left(\bar{X}\right) \cdot \nabla U\left(\bar{X}\right)}{\vec{n}\left(\bar{X}\right) \cdot \vec{n}\left(\bar{X}\right)} = \vec{0} \tag{4.7}$$

where $\vec{n}\left(\bar{X}\right)$ is the curve tangent and $\vec{0}$ is the null vector.

Importantly, for nonlinear coordinates the direction of the gradient vector has to be corrected using the corresponding contravariant metric tensor G that potentially depends on all $3N$ degrees of freedom:

$$G = (g_{ij}) = \left(\sum_{k=1}^{3N} \frac{\partial q_i}{\partial x_k} \frac{\partial q_j}{\partial x_k} \right) \tag{4.8}$$

otherwise different nonlinear coordinate systems will yield different reaction paths for the same stationary points. [71, 87, 88, 128, 129, 130, 131] Thus, to be invariant the transition path curve in nonlinear coordinates must satisfy the following more complicated condition:

$$\left(G\nabla U\left(\bar{Q}\right)\right)_{\perp}=\left(G\nabla U\left(\bar{Q}\right)\right)-\vec{n}\left(\bar{Q}\right)\frac{\vec{n}\left(\bar{Q}\right)\cdot\left(G\nabla U\left(\bar{Q}\right)\right)}{\vec{n}\left(\bar{Q}\right)\cdot\vec{n}\left(\bar{Q}\right)}=\vec{0} \qquad (4.9)$$

where $\vec{n}\left(\bar{Q}\right)$ is the curve tangent in the nonlinear coordinates.

Both equations (1.7) and (1.9) apply also to the adiabatic energy surfaces. Because the system of equations (1.9) is somewhat complicated by the need to compute the metric tensor, the ggaHFB method employs Cartesian coordinates for the path curve optimization instead of the generalized coordinates.

4.2.5 Minimum free energy transition path ensemble

Using the free energy gradient, the ggaHFB method generalizes the concept of the FukuiâĂŹs IRC [87, 88] to free energy surfaces. In deriving the free energy gradient the SCS degrees of freedom orthogonal to RCS are averaged over, which results in each point in the RCS representing an ensemble. Thus, the ggaHFB method finds continuous curves that connect the provided reactant and product ensembles through a series of transition and intermediate state ensembles. These curves must satisfy the condition that the invariant free energy gradient be tangential to the path curve at any point. In particular, the ggaHFB method uses the straightforward generalization of equation (1.7) to free energy surfaces in Cartesian coordinates:

$$\nabla_{\perp}W\left(\langle\bar{x}\rangle^{b}\right)=\nabla W\left(\langle\bar{x}\rangle^{b}\right)-\vec{n}\left(\bar{x}\right)\frac{\vec{n}\left(\bar{x}\right)\cdot\nabla W\left(\langle\bar{x}\rangle^{b}\right)}{\vec{n}\left(\bar{x}\right)\cdot\vec{n}\left(\bar{x}\right)}=\vec{0} \qquad (4.10)$$

As noted above, working with nonlinear coordinates requires computing logarithmic Jacobian corrections to the free energy gradient. Furthermore, finding invariant paths requires additional metric tensor corrections. [71, 122] No such complications arise in Cartesian coordinates, which is why the ggaHFB method employs these coordinates to optimize transition path ensembles.

4.2.6 Transition path optimization

To optimize a transition path in Cartesian coordinates, we take K unique configurations $\left\{\bar{Q}_k\right\}_{k=1,K}$ that gradually progress from the reactant to the product and assign them to a uniform grid $\left\{\alpha_k = \frac{k-1}{K-1}\right\}_{k=1,K}$ with mesh size of $\delta\alpha = \frac{1}{K-1}$. If initial configurations $\bar{Q}_k = \bar{Q}\left(\alpha_k\right)$ are unavailable they could be derived via a linear interpolation or by the activated evolution procedure [73] that is similar to the growing string method. [60] Using these K configurations, we obtain up to K corresponding Fourier amplitudes for each degree of freedom by applying the standard Fourier transform integration with the trapezoidal rule on the grid: [133]

$$b_n^i = \sum_{k=1}^{K-1} \left(f_n^{i,k} + f_n^{1,k+1}\right) \delta\alpha \tag{4.11}$$

where $f_n^{i,k} = \left[q_i\left(\alpha_k\right) - q_i\left(0\right) - \left(q_i\left(1\right) - q_i\left(0\right)\right)\alpha_k\right] \sin\left(n\pi\alpha_k\right)$.

This procedure globally interpolates between all the K points, yielding a continuous Fourier curve [73, 134] which is an analytical function of a progress variable $\alpha \in [0; 1]$:

$$q_i\left(\alpha\right) = q_i\left(0\right) + \left(q_i\left(1\right) - q_i\left(0\right)\right)\alpha + \sum_{n=1}^{K} b_n^i \sin\left(n\pi\alpha\right) \tag{4.12}$$

We then redistribute the K beads along the path curve such that they conform to a particular metric. Usually, we reposition the beads to make the arc lengths between adjacent beads of equal length in the RCS.

The newly redistributed beads serve as reference beads to compute the corresponding adiabatic potential energy gradients or the free energy gradients via the evolution procedures described in previous sections. Thus, for each reference bead $\bar{q}_k^{\mathbf{ref}} = \bar{q}\left(\alpha_k^{\mathbf{ref}}\right)$, the evolution returns either the minimized $\left[\bar{q}\right]_k^b = \left(\left[q_1\right]_k^b, ..., \left[q_S\right]_k^b\right)$ or the average bead $\left\langle\bar{q}\right\rangle_k^b = \left(\left\langle q_1\right\rangle_k^b, ..., \left\langle q_S\right\rangle_k^b\right)$, also called the *raw evolved bead.*

The ggaHFB method borrows the idea of re-distributing beads along the curve, and re-parametrizing the curve given the re-distributed beads from the string method. [121, 122, 123, 135] All the other essential ingredients of the ggaHFB method, such as the multidimensional energy gradient derived on the fly from the harmonic biasing potential, the Fourier representation of both the path and of the corresponding energy gradient (see below) employed in the energy profile integration via generalized line integral, as well as optimization strategies have been obtained from sources independent of the string method. [73, 21, 74, 127, 134]

In the following discussion, we omit the complementary SCS coordinates for clarity. These coordinates are assumed to be either completely minimized or averaged over and do not explicitly affect either the path or its energy. Optimization implies that the SCS coordinates are passed along either through dynamics restart files or through the complete coordinate files. In addition, it is assumed that the changes in the SCS coordinates between the beads are continuous.

For brevity, we only discuss how to drive optimization of the transition path ensembles and compute the corresponding free energy profiles. The same strategies apply to finding the transition paths on adiabatic potential energy surfaces and computing the corresponding energy profiles. In this case, the adiabatic potential energies could also be calculated exactly for all the points along the path and compared to those computed using the ggaHFBâĂŹs generalized line integral formalism.

Substituting the raw evolved beads into equation (1.3) gives estimates of the free energy gradients for each bead. These gradients are then used in the steepest descent step to generate the *enhanced evolved beads*:

$$\bar{q}_k^{\mathbf{SD}} = \langle \bar{q} \rangle_k^b + \gamma_k \nabla W^u \left(\langle \bar{q} \rangle_k^b \right) \tag{4.13}$$

Here γ_k is the parameter that controls the SD step size for the $k^{\mathbf{th}}$ bead. In the present paper we use the uniform step size parameter γ for all the beads for simplicity.

Following the Fourier transform of the enhanced beads to obtain new Fourier amplitudes, redistribution of the beads along the resulting curve provides new reference beads. These reference beads are re-aligned to maintain the coordinate system. For this purpose we invoke a mass-weighted best-fit procedure in a suitable space, usually RCS, to enforce the Eckart conditions on the beads. [21, 127, 136, 137, 138] In cases where only a few coordinates are available for the best fit or if their geometric arrangement breaks down the standard best fit procedure, simpler alignment methods could be used. The final re-aligned beads then replace the previous reference beads in the next round of evolution. This procedure is repeated until convergence of the path, i.e. until the path curve changes cease. The final optimized curve represents an invariant minimum free energy transition path ensemble that satisfies the Fukui's IRC criteria. [87, 88]

The convergence rate of the ggaHFB method depends to some extent on the employed bias force constant and step size parameter. Therefore, devising an optimization strategy to achieve the fastest convergence possible is desirable and is an active area of research in our lab.

4.2.7 Computing the free energy along the Fourier path

Given a Fourier path in the generalized multi-dimensional coordinate space and the corresponding free energy gradients, we can compute the free energy profile along that path via the generalized line integral formalism. To achieve the highest accuracy, we Fourier transform both the evolved beads (1.5) and the corresponding free energy gradients (1.6) along the path. With the continuous Fourier representations of the forces and the path, we could then analytically evaluate the corresponding reversible work line integral passing through the evolved beads:

$$W^u(\alpha) = \sum_{i=1}^{S} \int_0^{\alpha} \left[\frac{\partial W^u(\alpha)}{\partial q_i} q_i'(\alpha) \right] d\alpha \tag{4.14}$$

In practice, we evaluate the generalized line integral of the second order in equation (1.14) on a fine uniform grid with $L >> K$ quadrature points.

This procedure provides the free energy or the potential of mean force (PMF) profile as an analytical function of the progress variable. Unlike umbrella sampling with WHAM, the interpolation-based ggaHFB free energy integration procedure does not require overlap between the windows. Furthermore, the ggaHFB integration procedure allows natural decomposition of the free energy into contributions from individual coordinates.

The analytical form of the energy profile and that of the corresponding path provided by the ggaHFB method renders pinpointing the energy extrema and their accurate RCS coordinates particularly trivial. One can easily find the values of the progress variable α corresponding to extrema on the energy profile and then substitute these values into equation (1.12) to get the matching structures.

4.2.8 Summary of the ggaHFB methodology

In summary, the ggaHFB method finds the Fukui's IRC curves on the adiabatic potential energy surfaces and further generalizes this approach to Cartesian free energy surfaces. Thus, the ggaHFB method finds either minimum adiabatic potential energy paths or minimum free energy transition path ensembles via a gradient driven optimization procedure. Optimizing the transition paths and path ensembles in Cartesian coordinates bypasses the need to calculate the corresponding metric tensors. The optimized transition paths provide structural and energetic information about all the intermediates and transition states connecting given reactants and products at once. Furthermore, the global Fourier representation of the path and the forces provide useful means to control various aspects of the path optimization and ultimately makes the ggaHFB optimization extremely robust.

Independent from the path optimization, the ggaHFB method is a practical alternative to the conventional approach to free energy calculations via umbrella sampling with WHAM. Advantageously, the ggaHFB method is histogram-free, which makes it applicable to cases with arbitrary many dimensions. Even though ggaHFB uses somewhat stiff springs, it does not require the overlap between the windows to integrate the free energy profile. Additionally, the Cartesian version of the ggaHFB method avoids the need to compute the logarithmic Jacobian correction that is required if either WHAM or ggaHFB is used with nonlinear coordinates such as bond distances, angles, dihedrals etc to compute free energy profiles. [74, 139] Finally, the energy profiles can be straightforwardly decomposed into contributions from the individual degrees of freedom that could be useful for analysis and design purposes.

4.2.9 Results

4.2.10 Minimum free energy transition path ensembles

To explore the free energy of the αL to pII backbone transition of a TATTVGYG signature peptide in water and the effect of the V76 sidechain rotation, we use the ggaHFB method with two reactive coordinate spaces (RCSs) of different dimensionalities. Specifically, we include all heavy atoms of the peptide into RCS1, and derive RCS2 from RCS1 by excluding sidechain atoms. The RCS1 surface provides the free energy of the backbone configuration subject to a particular sidechain orientation. In contrast, the RCS2 surface provides the free energy of the backbone configuration irrespective of the sidechains. Unless otherwise stated, throughout this work we employ molecular dynamics in the isothermal isobaric NPT ensemble at 298 K and 1 atm using CHARMM22 molecular mechanical force field [109, 110] with the CHARMM-modified TIP3P explicit water model [140, 141, 142, 143, 144] to derive all the required free energy gradients.

Our preliminary free-energy optimization runs revealed that both the αL and pII states at position 77 are local free-energy minima of the isolated peptide in water. Interestingly, the partially flipped, non-conducting conformation observed by X-ray crystallography at low K+ concentration (PDB code 1R3K) is unstable by itself in water despite the rotation of the V76 sidechain away from the high K+ concentration, conducting conformation (PDB code 1R3J). During optimization of the peptide from the partially flipped state its backbone, but not the V76 sidechain, collapses to the conducting state conformation.

Therefore, to study the full range of the peptide flip, we have constructed an initial path that includes both αL and pII states of G77 backbone. To assess the effect of the V76 sidechain rotation, we have included two such rotations by requiring that the endpoints have the same V76 sidechain orientation, matching that of the conducting state. Furthermore, we have inserted the crystallographic non-conducting state with partially flipped backbone and rotated V76 in the middle of the path (refer to Supporting Information for details).

Performing thorough optimization on the RCS1 free energy surface (see Supporting Information), we obtain an intrinsic transition path ensemble for the G77 backbone conformational transition from αL to pII state in the TATTVGYG peptide in water. Using the RCS1-optimized path ensemble as the reference we then compute the final free-energy profile, both coupled with (RCS1) and uncoupled from (RCS2) the V76 sidechain rotation. To elaborate on the energetics of the backbone transition further, we compute an analogous optimal path and the free-energy information for a GYG to AYA mutant. The resulting cumulative PMFs at different collection times are depicted in Figure 1.1 and the representative structures of the wild type peptide are shown in Figure 1.2.

The PMFs at the RCS1 level display two events involving the V76 rotation as two sharp peaks with barrier heights ranging from 5 to 8 kcal/mol in both directions. Interestingly, the V76 sidechain rotation from the conducting into non-conducting orientation destabilizes αL state by 2 to 3 kcal/mol, indicating that the hydrophobic interaction between the V76 and Y78 sidechains provides additional stabilization of the αL state in the conducting conformation. The αL to pII backbone transition at position 77 follows the second, restoring V76 sidechain rotation. In the wild type, the free energy barrier for the backbone transition given a specific orientation of the sidechains (the RCS1 PMF) has a forward barrier of 6.0 kcal/mol. The pII state is 2.2 kcal/mol less favorable than the αL state, and converts back with a barrier of 3.8 kcal/mol. In sharp contrast, in the mutant the forward barrier is only 0.7 kcal/mol and the pII state is 6.3 kcal/mol more favorable than the α-strand. Restoring the α-strand state in the mutant requires surmounting a high 7.0 kcal/mol free-energy barrier.

Setting the sidechains free (the RCS2 PMF) permits evaluating the free energy of the backbone transition alone. As is seen from Figure 1.1, switching to RCS2 space collapses the sharp peaks (labeled with arrows) corresponding to the V76 sidechain rotations, but leaves the portion of the PMF underlying the backbone transition from the αL to the pII state virtually unchanged. In particular, for the wild type peptide the forward activation barrier is 5.9 kcal/mol, and the pII state is still less stable than the αL by slightly smaller 1.7 kcal/mol. Restoring the conducting state requires overcoming a slightly higher barrier of 4.2 kcal/mol. In contrast, the mutant exhibits a forward barrier of 0.9 kcal/mol and the relative pII state stabilization energy of 7.0 kcal/mol that makes the reverse barrier increase to 7.9 kcal/mol.

Minimum adiabatic potential energy paths

To explicitly evaluate the effect water has on the conformational transitions of the signature peptide and to further demonstrate the capabilities of the ggaHFB method, we have computed the minimum adiabatic potential energy paths for the wild

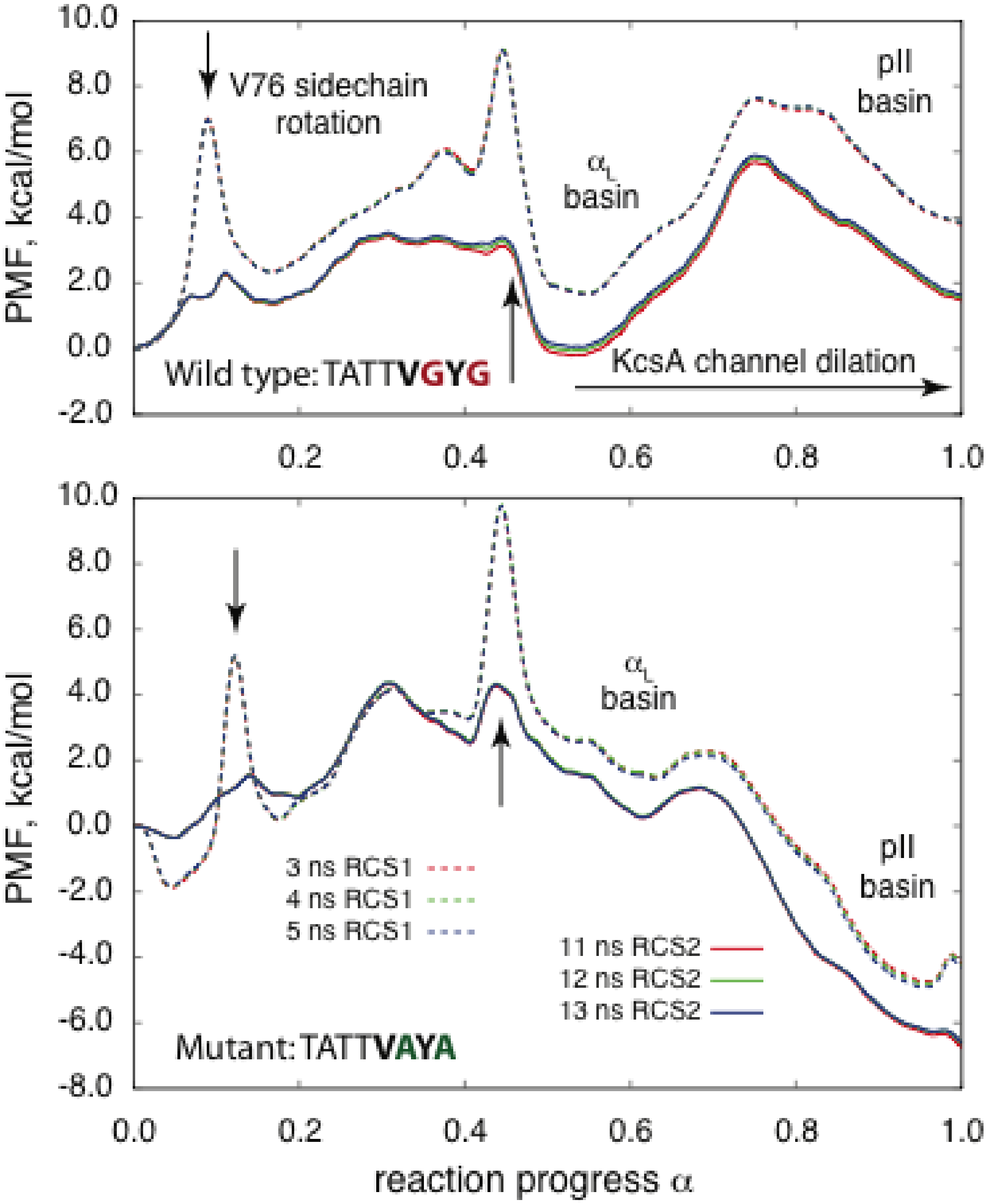

Figure 4.1: Cumulative PMFs for the conformational transition of the signature TATTVGYG KcsA peptide and its AYA mutant from the αL to pII state in explicit water on RCS1 and RCS2 free-energy surfaces at different collection times. The ggaHFB method employed 89 beads to integrate the free energy profile. Arrows point to the V76 sidechain rotations.

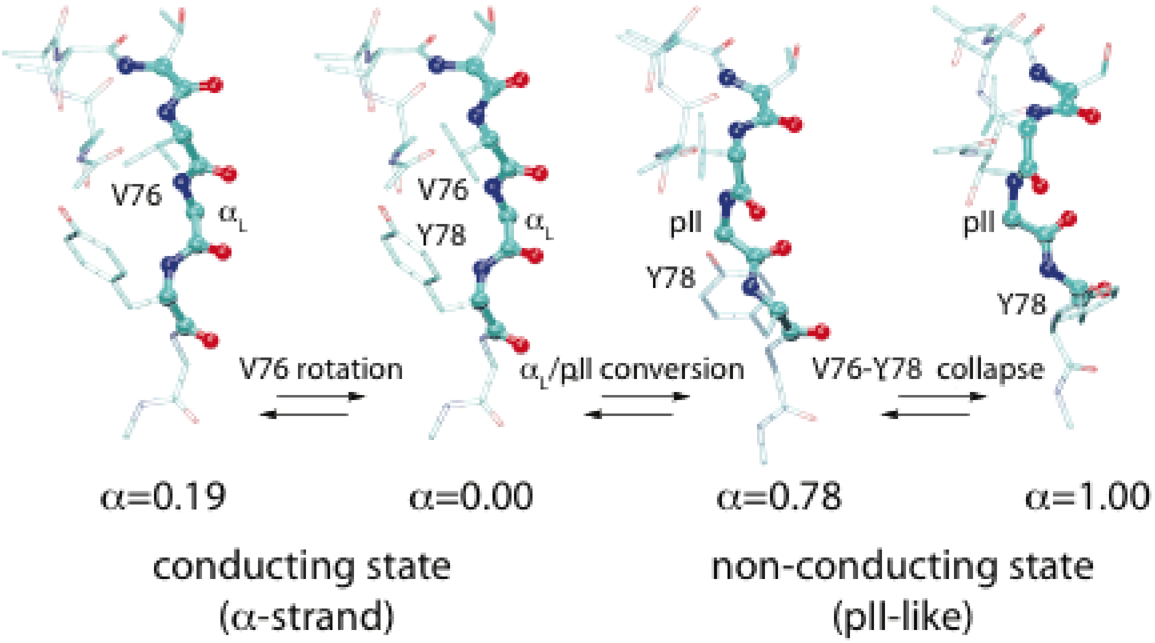

Figure 4.2: Representative structures from the free energy transition path ensemble of the wild type TATTVGYG signature peptide of the KcsA selectivity filter in explicit water. The values of the progress variable α provided relate structures to the free energy profile of the wild type peptide in Figure 1.1.

type TATTVGYG peptide and its GYG to AYA mutant in gas phase. Both peptides have three threonine and one tyrosine residues with rotatable OH bonds that were averaged over in the minimum free energy transition path ensembles computed in water. The orientation of these hydrogens significantly perturbs the overall potential energy, therefore we include these four hydrogen atoms into the reactive coordinate space. Thus, by adding the polar hydrogen atoms to the all-heavy-atom RCS1 we derive the RCS1h for adiabatic potential energy path optimization.

We have to assign some initial values to the tyrosine and threonine OH groups, which have two and three rotameric states, respectively, in order to compute the adiabatic potential energy paths. The total number of possible initial path configurations is therefore 2^1x$3^3 = 54$. To control the configurations, we follow the Protein Data Bank atom naming convention and use dihedral angles Cϵ1-Cς-Oη-Hη and Cα-Cβ-Oγ-Hγ for the tyrosine and the threonines, respectively. Here, we arbitrarily choose dihedral angles of 180, 180, -30 and 0 degrees for the T72, T74, T75 and Y78, respectively, as the initial conditions for the path optimization. To prepare the initial path with these conditions, we fix the RCS1 coordinates and apply stiff harmonic restraints of 1000 kcal/(mol·rad^2) on the corresponding dihedral angles during an optimization of the hydrogen positions.

Because the optimization on adiabatic potential energy surfaces with the bare CHARMM22 molecular mechanical forcefield is relatively inexpensive, we have initiated the ggaHFB optimization using 89 beads. Using 89 beads is sufficient to integrate the adiabatic potential energy, given the initial orientation of the four OH bonds. Nevertheless, optimization of the OH groups requires increasing the number of beads further to correctly integrate the adiabatic potential energy. The increase reflects the fact that rotations of the OH groups correspond to small changes in the RCS1h, resulting in very sharp transitions along the path. Although, proper integration could be achieved by locally increasing the number of beads at the sharp transitions leading to non-uniform bead distributions, [74] in the present work we use uniform grid for simplicity. Thorough path optimization increases the overall path length dramatically, in the end requiring 705 beads to properly integrate the adiabatic potential energy along the path.

The final paths in the gas phase have little if any resemblance with the paths optimized in water and exhibit a greater number of local minima and transition states. For the wild type peptide, the α-strand disappears almost completely. First, the G79 residue spontaneously flips into the C5 conformation and then converts into the C7ax conformation. In the flipped configuration on the reactant side of the path G79 forms a hydrogen bond with the OH group of T71 using its carbonyl oxygen. The G79 residue flip significantly perturbs the rest of the α-strand, which quickly collapses further residue-by-residue along the path.

In the mutant, the α-strand is annihilated completely in the reactant basin, where the A79 along with the A77 residues flip into the C7ax conformation. The four residues V76, A77, Y78 and A79 surround the T71 residue like a belt, with alternating axial and equatorial configurations, namely C7ax, C7eq, C7ax and C7eq, respectively. During the optimization the mutant pathway deviates substantially from that of the wild type.

Figure 1.3 depicts the corresponding adiabatic energy profiles that underscore the complexity of the changes in the gas phase. It also provides the benchmarks for the adiabatic potential energy integration via the generalized line integral formalism. In particular, comparison of the line integral energies with the exact adiabatic potential energies from the CHARMM22 forcefield show the accumulated errors of 0.07 and 0.12 kcal/mol for the wild type and mutant adiabatic energy profiles, respectively. We consider this a very good agreement between the generalized line integral energy and the exact energy profiles.

The V76 sidechain rotations (labeled with arrows) have been preserved in both wild type and mutant paths, although in some cases have been coupled with other structural rearrangement as seen in Figure 1.3. The forward and reverse barrier heights for the V76 sidechain rotation vary, but are similar to those in water.

Overall the gas phase structures are more compact than the ones in water and establish as many intra-molecular hydrogen bonds as possible. Given the complexity of the adiabatic paths and their divergence from the structures obtained by either the X-ray crystallography or by the free energy optimization in water, we omit detailed description of the structural changes along the path and simply provide the corresponding trajectories in Supporting Information.

4.2.11 Comparison of the MM and QM energy profiles

Gas Phase Because the present paper investigates an important conformational transition of the TATTVGYG signature peptide from the KcsA potassium channel, it might be useful to assess the molecular mechanical (MM) forcefield employed. Of particular interest is evaluating the energetics of the signature peptide and its mutant along the path optimized in water. To establish useful benchmarks, we first compute the gas phase adiabatic energy profiles along the minimum free energy transition path ensembles in the RCS1. In particular, we compare the MM energy profiles with one of the most popular density functional theory models, namely B3LYP, with 6-31G(d) basis set as a high-level quantum mechanical (QM) model (see Supporting Information for details). This model is not expected to produce accurate energy profiles when it comes to dispersion interactions between the Y78 residue and the V76 sidechains and hence should be used with caution. [145, 146, 147, 148]

Figure 1.4 shows the corresponding adiabatic potential energy profiles for the

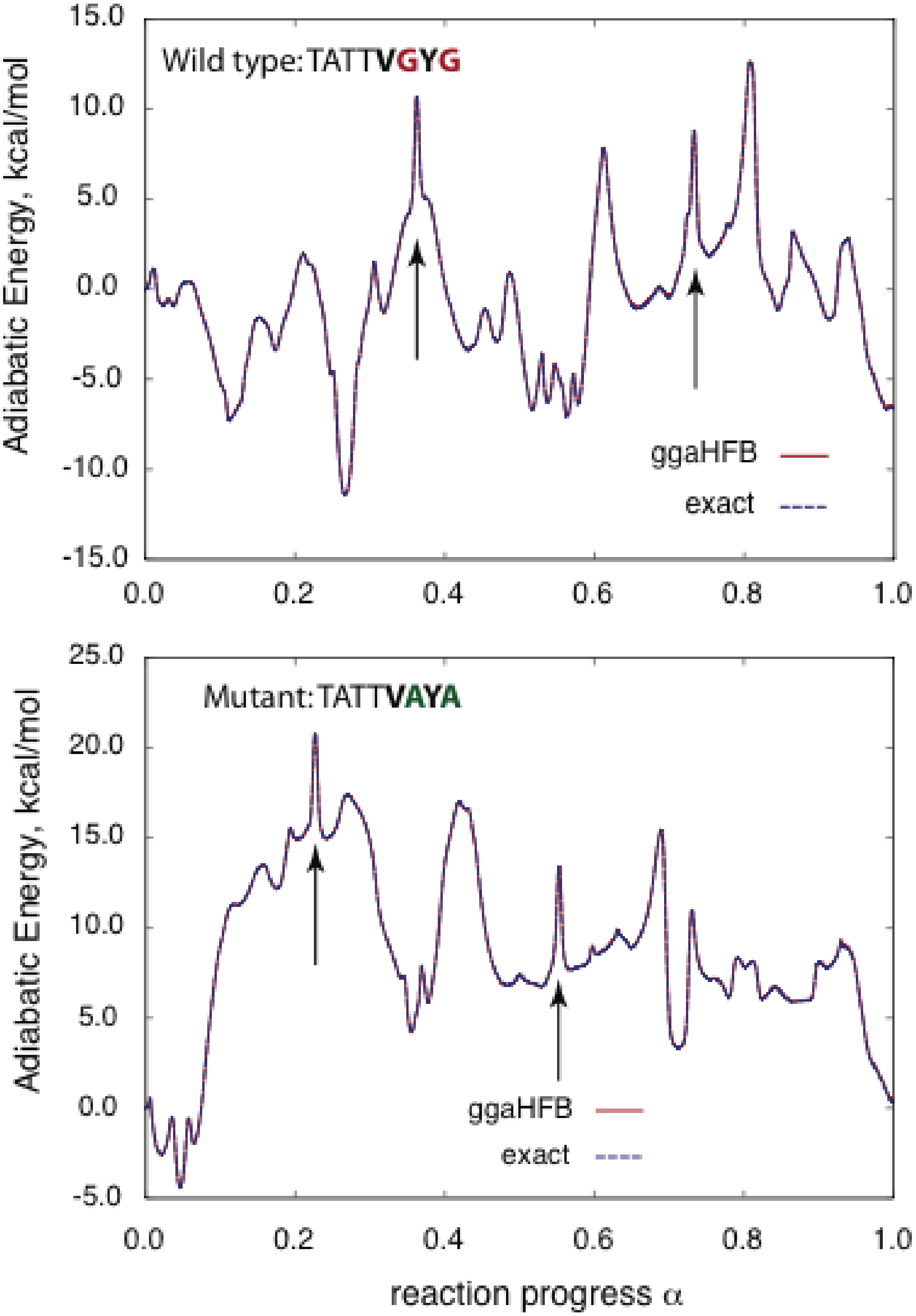

Figure 4.3: Adiabatic potential energy profiles along the optimized reaction paths of the signature TATTVGYG KcsA peptide and its AYA mutant in gas phase. The ggaHFB method employed 705 beads to integrate the potential energy with the RCS1h (see text for details) âĂŞ solid lines; the CHARMM22 exact energies are shown in dashed lines. Arrows point to the rotations of the V76 sidechain.

peptides with the rotatable OH bonds fixed at the conformation used as initial condition for the path re-optimization in gas phase. Interestingly, the energy profiles obtained with MM and the QM models for the same path differ substantially. The V76 sidechain rotation barriers appear reduced in the QM model.

Both the MM and QM models favor the pII state over the α-strand. The QM model predicts the α-strand to be much less stable than the pII state in the wild type, but relatively more stable in the mutant peptide. In contrast, the MM model suggests that the α-strand is much more stable in the wild type peptide, not the mutant. It is likely that the gas phase adiabatic energy surfaces of the MM and QM models are significantly different in the gas phase and such single point energy profile comparisons should be taken with caution.

Implicit Solvent As mentioned above, we are primarily interested in the energetics of the peptides in water and not in the gas phase. After all, the transition path ensembles for the functional conformational transitions of the peptides have been optimized using the MM model in the explicit water. To compare the MM with QM models in water, we choose the Generalized Born Molecular Volume (GBMV) [97, 114] and Polarizable Continuum Model (PCM) [115, 116, 117, 118] implemented in Gaussian 03 [149] implicit solvent models, respectively. Using the free-energy optimized transition path ensembles in explicit water as reference profiles, we have performed optimization of all the degrees of freedom orthogonal to the RCS1 and subject to the same dihedral restraints on the rotatable OH bonds as discussed above with MM-GBMV model and then computed single point QM-PCM energies for all the beads along the path (see Supporting Information for details). The results are provided in Figure 1.5.

The MM-GBMV model yields the free-energy profiles that are in very good agreement with the explicit solvent calculations (Figure 1.1), thus further validating the intrinsic free energy profiles. We cannot expect better agreement between the two profiles given the fixed conformations of the rotatable OH bonds necessary to compute the adiabatic energy profile with GBMV.

The QM-PCM model produces energy profiles very different from those of the

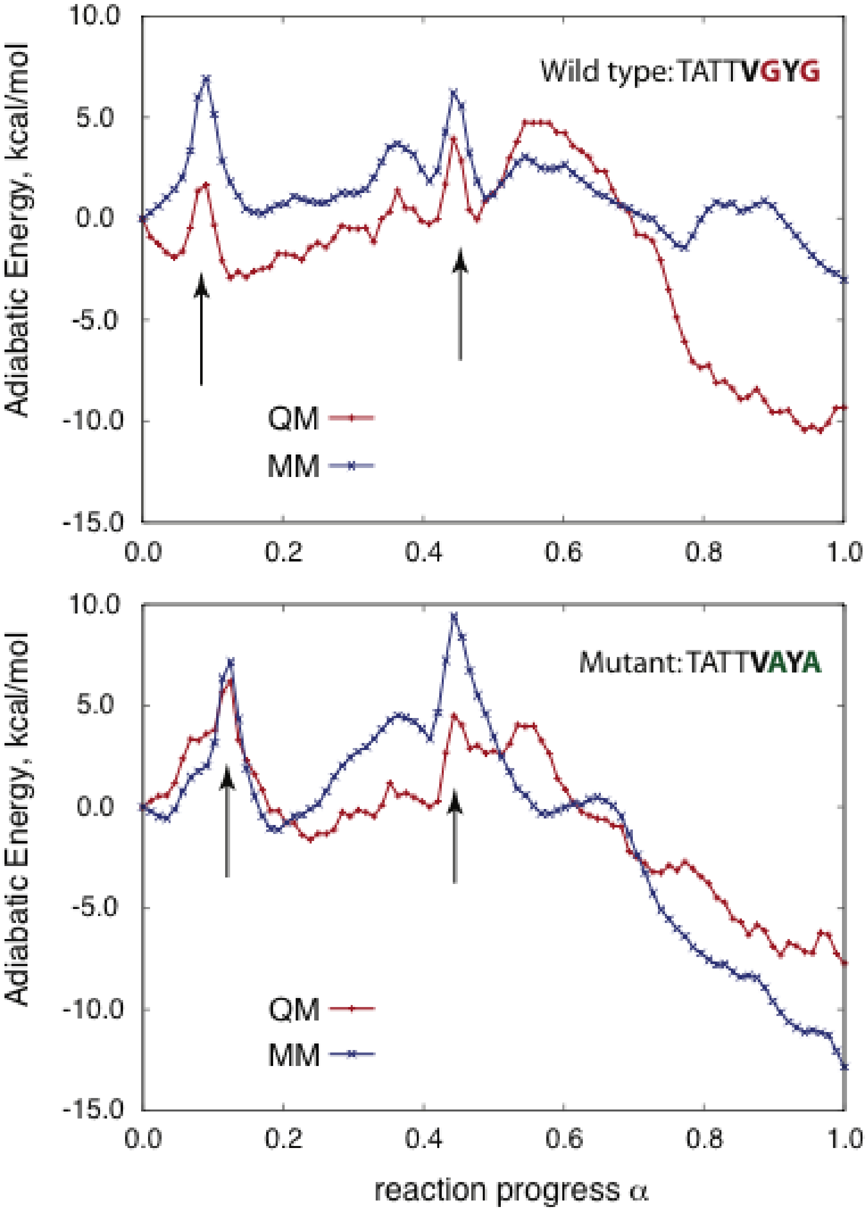

Figure 4.4: Gas phase single point energy profile along the α-strand to pII state conformational transition of the signature TATTVGYG KcsA peptide and its AYA mutant in water. MM is the CHARMM22 forcefield, and QM is the B3LYP/6-31G(d) Density Functional Theory model. Arrows point to the rotations of the V76 sidechain.

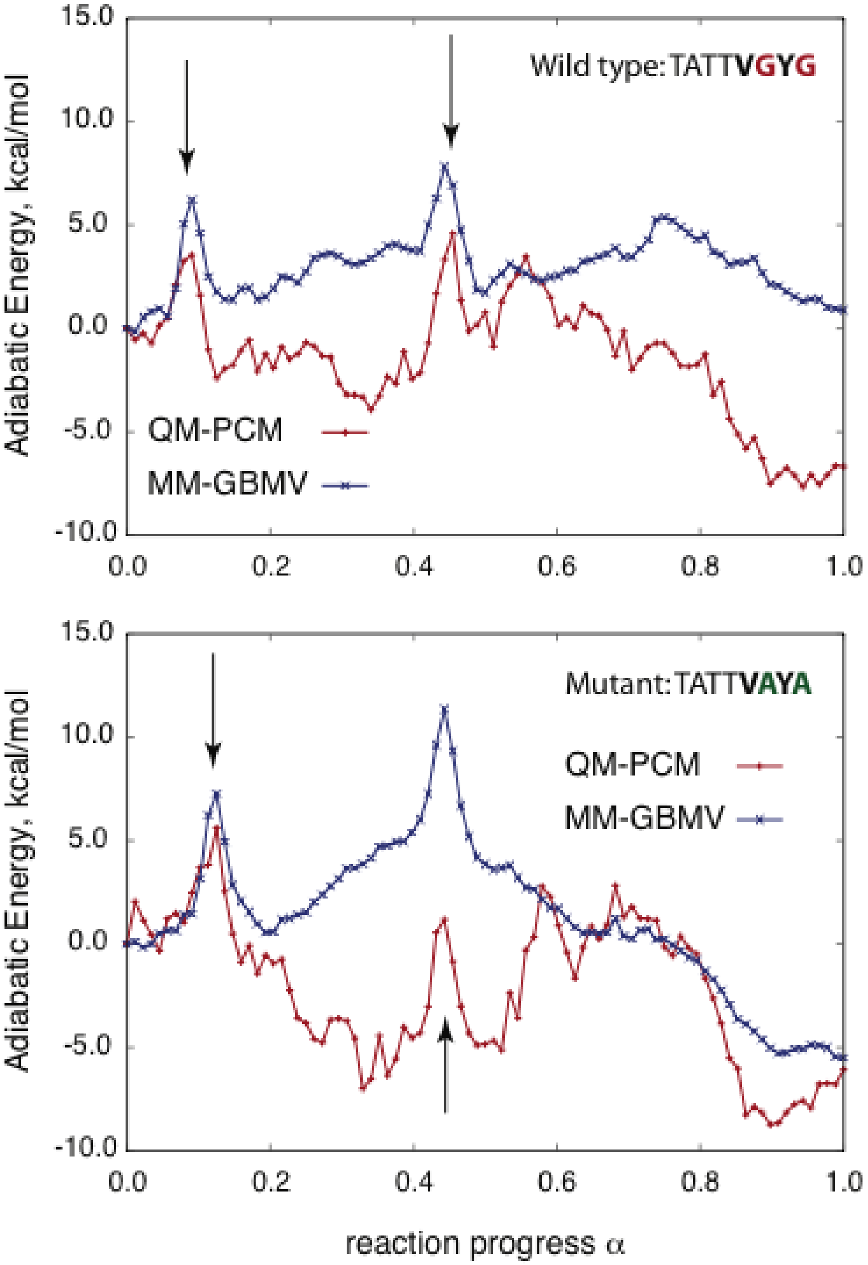

Figure 4.5: Implicit solvent single point energy profile along the α-strand to pII state conformational transition of the signature TATTVGYG KcsA peptide and its AYA mutant in water. MM-GBMV âĂŞ uses the CHARMM22 forcefield with the Generalized Born Molecular Volume implicit solvent model, QM-PCM âĂŞ B3LYP/6-31G(d) Density Functional Theory with the Polarizable Continuum Model implicit solvent model. Arrows point to the rotations of the V76 sidechain.

MM-GBMV model, and most importantly does not favor the αL state over the pII state in the wild-type peptide. In the mutant, on the other hand, the αL state remains unstable with the QM-PCM model.

Because we have performed optimization on the water modified RCS1 free energy surfaces of the peptides with the MM model and explicit water any comparison with other surfaces that have not been optimized do not warrant good agreement, unless the surfaces are exactly the same. This conflict could in principle be resolved by the QM-PCM optimization of the product, reactant, and a few key intermediates, which unfortunately presents a significant challenge at present.

4.3 Discussion

The finding that the signature peptide taken from the partially flipped non-conducting state (PDB code 1R3K) collapses back into the conducting state in water despite the V76 sidechain rotation shows the intrinsic width of the peptide αL basin. Furthermore, it suggests that either the channel provides additional interactions to stabilize partially flipped backbone structure or that only one of the four strands of the tetrameric channel undergoes the full transition. If the latter symmetry breaking were to occur, the apparent configuration observed by X-ray crystallography would correspond to the average over four strands, thus artificially reducing the extent of the transition in a single peptide.

The fact that the free energy profiles for the αL to the pII transition are relatively insensitive to the position of the sidechains (see Figure 1) reflects the robustness of the backbone transition. The forward and reverse free energy activation barriers in the wild type peptide are 5.9 and 4.2 kcal/mol. Interestingly, previous calculations of an even lower dimensional PMF for a similar transition inside the wild type KcsA channel in the presence of two ions gave a rough estimate of the free energy barrier between 0.5 and 4.0 kcal/mol.8 In sharp contrast, the mutant exhibits a forward barrier of 0.9 kcal/mol and the reverse barrier of 7.9 kcal/mol.

The width of the αL basin in the intrinsic free energy profile of a single peptide might determine the range of a local dilation/contraction of the tetrameric KcsA channel at the V76 carbonyl ring. If the V76 carbonyl were pushed away from the channel axis beyond the limits of the αL basin, the peptide would go over the transition barrier and into the non-conducting pII state. We emphasize that by local dilation/contraction of the channel we imply the change in the distance between the V76 carbonyls associated with the backbone motions within the bounds of the αL basin, and not with the transition from αL to pII or back.

The full αL to pII transition has been demonstrated to be unnecessary for the ion selectivity, at least in a synthetic channel with the D-Ala residue in place of G77. [56] Note that the wild type KcsA channel, in addition to K+, permits ions of larger size, namely Cs+ and Rb+, which are expected to pass the V76 carbonyl ring without triggering the transition from αL to pII. [55] Such a wide range of dilation/contraction would not have been possible if G77 was substituted for regular Ala, as the width and the depth of the αL basin would have been dramatically reduced as seen from the PMFs for the mutant depicted in Figure 1.1. Note however, that the AYA mutant would be sterically prevented from forming the conducting α-strand conformation in the tetrameric channel. [56]

In an effort to validate the results obtained with the MM force field in explicit water we have profiled the energetics along the paths using MM and QM methods both in gas phase and in implicit solvent. The results of these calculations are summarized in Figures 1.4 and 1.5 that highlight the stark disagreement between the MM and QM models. Although QM models usually have higher fidelity than MM models, the particular DFT method used in this work, namely the B3LYP functional, is well known to fail to account for short-range dispersion interactions necessary to properly describe the energetics of the hydrophobic interactions such as those between V76 and Y78 sidechains. [145, 146, 147, 148] Higher level, more expensive ab initio methods that properly account for the dispersion suggest that the interactions between the CH bonds of the V76 and the phenol ring of the Y78 could favor the α-strand by about 1 kcal/mol.101, 103 Additional discrepancies between the MM and QM in this work may arise due to the fact that no optimization has been performed at the QM level of theory. Therefore, the differences between the QM and MM models should be interpreted with caution.

It appears that the stability of the pII state in the wild type peptide might be overestimated by the QM model with implicit water, because it would require at least

20 kcal/mol to assume the conducting conformation in the tetrameric channel. On the other hand the MM model with implicit water predicts the αL and pII configurations to be nearly degenerate. If the QM-PCM model more accurately reproduced the energetics of the solvated peptide even without optimization, the ground or resting state of the wild type KcsA channel would be the non-conducting pII state. Thus the channel would have to be activated by a conformational change from the pII resting state into the α-strand state to conduct ions. This would only be possible due to a strong perturbation such as strong attraction of the ions in the lumen of the tetrameric channel to its carbonyl oxygens.

The switching between the non-conducting and conducting state and the functional contraction/dilation of the tetrameric KcsA channel would require a certain balance between the electrostatic repulsion of the V76 carbonyls and the free energy of the backbone rotation of the residue at position 77. Because the electrostatic repulsion can be relaxed by transiently flipping one or more of the carbonyls out from the αL into pII state, the filter must also ensure to favor the αL over the pII state at least in the presence of ions in the lumen of the filter. The potassium channel seems to have achieved the α-strand stabilization by using a G residue that has high propensity for the αL configuration at position 77, and in addition by the hydrophobic interaction between V76 and Y78 sidechains. The importance of the hydrophobic interaction is supported by the experimental observation that the V76A mutant abrogates tetrameric assembly of the channel. [108] Taking the above into consideration, it appears that the free energy profiles computed with the MM model in explicit water agree better with the proposal than the corresponding single point energy profiles obtained with B3LYP/6-31G(d)-PCM QM model.

In the absence of the actual tetrameric channel in our model, the bulk water better reproduces the environment of the KcsA selectivity filter than the gas phase and therefore provides useful insights into the channel function. In particular, the differences between the adiabatic energy maps of the peptide residues in water and gas phase suggest that the gas phase transition pathways must deviate strongly from those of the transition path ensembles optimized in water. This is particularly true of the αL region that is forbidden in gas phase. [98, 99] The ggaHFB optimization of the adiabatic paths in gas phase explicitly demonstrates that water plays active and important role in defining the intrinsic path and the energetics of the peptide backbone transition.

The outcome of the gas phase optimization can be predicted based on the previous studies of the glycine and alanine dipeptides. [98, 99] In particular, the referred work demonstrated that the αL configurations collapse into the C7ax, while pII configurations collapse into the C7eq. [98, 99] These are the exact changes we observe upon the adiabatic potential energy optimization in gas phase. The final optimized paths in gas phase are rather complex (see Figure 1.4) and seem irrelevant for the functional transition of the selectivity peptide in the KcsA channel. On the other hand the pathways in water show very good qualitative agreement with the peptide conformations observed in the tetrameric channel.

Finally, based on the present findings, we are able to propose a novel hypothesis for the mechanism of ion selectivity in the tetrameric KcsA channel. Specifically, we conjecture that in its conducting α-strand state the carbonyl rings should contract around the ion entering the channel and that this contraction would propagate to the nearby carbonyl rings along the channel axis (see Figure S1.1 in Supporting Information for an illustration). Because ions are believed to pass the KcsA filter stripped of all but two water molecules that co-translate with the ion while hydrogen bonding to the carbonyls, these water molecules will experience greater difficulty to pass neighboring carbonyl rings due to the contraction, in turn impeding the ion movements along the channel.

This hypothesis could explain why the channel selects larger K+ over smaller Na+ ions. Specifically, we anticipate that smaller Na+ ions would contract the carbonyl rings to a greater extent than the larger K+ ions thus impeding the passage of the co-translating water molecules to a greater degree. With water passage impeded the ions themselves must in turn slow down.

Our hypothesis suggests that in the absence of ions the KcsA channel should stay open to water permeation unless one of the four V76 carbonyls flips out pinch-shutting the channel. Indeed, the KcsA channel has been experimentally demonstrated to conduct water in the absence of permeating ions. [101] The partial flipping of the carbonyls (while still within the ÎśL basin) might serve selectivity purpose, whereas complete flip (transition into the pII basin) can be used to gate the channel.

[60] To provide further support of this hypothesis we are currently performing an optimization of several transition path ensembles for ion-water co-permeation through the tetrameric KcsA selectivity filter. The results of this work will be reported in a forthcoming publication.

To conclude, we have explored an intrinsic free energy landscape of an important functional transition of the signature peptide from the KcsA selectivity filter that is responsible for locally dilating/contracting the channel at the V76 carbonyl ring, in addition to switching the channel between conducting and non-conducting states. We have found that the wild type peptide intrinsically favors the conducting state due to the combination of the high G77 backbone propensity for the ÎśL configuration and the stabilizing hydrophobic interaction between V76 and Y78 sidechains. In sharp contrast, the mutant strongly favors the non-conducting state. However, additional steric effects in the tetrameric channel that are absent in the present study are expected to prevent formation of the conducting conformation in the mutant.

We have found the ÎśL to pII transition to be exceptionally robust and intrinsically funneled toward the conducting state in the wild type KcsA peptide at the MM level with explicit water. Although the intrinsic free energy profiles have been validated using the MM with implicit water model, efforts to gauge performance of the MM model against QM model indicated that our results should be interpreted with caution. Based on the QM-PCM model it may be possible that the ground state of the channel in the absence of ions could in fact be the non-conducting state and that the conducting state would only form upon ion entrance into the lumen of the channel. Nevertheless, the present study has allowed us to propose a novel hypothesis for the ion selectivity within the KcsA channel in which local contraction of the channel interior in response to the ion presence regulates co-permeation of water through the channel to a degree that inversly proportional to the ion size. Work is currently underway in our lab to test the proposed hypothesis in the tetrameric KcsA channel model. We hope that the present work will stimulate future transition path ensemble studies of rare events in complex molecular systems.

Acknowledgement

Chapter 1, in full, is a minimally modified reprint of the material as it appears in I. Khavrutskii, M. Fajer, J. A. McCammon, âĂIJIntrinsic Free Energy of the Conformational Transition of the KcsA Signature Peptide from Conducting to Nonconducting StateâĂİ J. Chem. Theory Comput. 4, 1541-1554 (2008). The dissertation author was the second author of this paper.

Supporting Information

Model setup

The model of the wild type ion selectivity signature peptide was derived from the KcsA structures obtained in the presence of Tl^+ ions in place of K^+ ions. There are two such crystallographic structures: 1R3J that corresponds to high Tl^+ concentration, and 1R3K that corresponds to low Tl^+ concentration. [53] We extracted the following sequence from the monomer: THR^{72}(1)-ALA^{73}(2)-THR^{74}(3)-THR^{75}(4)-VAL^{76}(5)-GLY^{77}(6)-TYR^{78}(7)-GLY^{79}(8). Superscript numbering corresponds to the original KcsA numbering in the PDB files, whereas numbers in parenthesis correspond to the model. In a single letter aminoacid code the sequence is T72-A73-T74-T75-V76-G77-Y78-G79 or TATTVGYG for short.

The corresponding double ala-mutant was generated by replacing the two glycine residues namely GLY77 and GLY79 with alanine, resulting in THR(1)-ALA(2)-THR(3)-THR(4)-VAL(5)-ALA(6)-TYR(7)-ALA(8) sequence or TATTVAYA for short.

Because the peptide model was derived by truncating the KcsA peptide we mended the ends of the chain with the standard neutral capping groups, namely Acetyl and N-methyl amide for the N- and C-termini, respectively. In the following we use the standard definitions of atom types from the CHARMM22 forcefield.

Methods

Molecular Dynamics

We use CHARMM22 TIP3P water model (CTIP3P) and peptide parameters. [109] Initial simulations were performed using GBMV implicit solvent model to expedite path optimization. The following parameters of the GBMV model were used (see charm c32b1 documentation for explanation of the parameters):

With the GBMV implicit solvent model we employed the following MD protocol. The electrostatic and vdW non-bonded interactions were truncated by switching functions between 12 and 13 Å. Langevin dynamics (LD) was performed with leap frog integrator using a 1.5 fs time step at 298 K and with a friction coefficient of 10 ps-1 for all heavy atoms. All bonds involving hydrogen atoms were constrained using SHAKE [150, 131, 151] with tolerance of 10-8 Å. Each evolution step involved 5,000 equilibration steps and 25,000 production steps. Coordinates from production runs were recorded every 50 steps for subsequent averaging.

The results of the GBMV simulations have been refined using explicit CTIP3P solvent model. The explicit water simulations were performed in the NPT ensemble using truncated octahedron box. For the wild type peptide the water box contained 1155 water molecules, whereas for the mutant the box contained 1222 water molecules. The nearest image distance was approximately 36 and 37 Åfor the wild type and mutant peptides, respectively. The temperature was maintained with the Nose-Hoover thermostat at 298 K using a thermal piston of mass 250 (kcal/mol) · ps, [152, 153] whereas the pressure was maintained at 1 atm by the Langevin piston method with the piston mass of 250 amu and Langevin collision frequency of 20 ps-1. [154, 155] Electrostatic interactions were computed using particle mesh Ewald method7 with 12 Åreal space cutoff. The vdW non-bonded interactions were truncated by switching functions between 10 and 12 Å. The covalent bonds between the hydrogen and the heavy atom were constrained using the SHAKE [150, 131, 151] algorithm with a tolerance in the bond length deviations of 10-10 Å. The MD integration step size was 2 fs. For the evolution runs we performed short MD runs including 5,000 steps of equilibration and 25,000 steps of production. The solute coordinates from the production

Table S4.1: GBMV Parameters

BETA	-20
EPSILON	80
DN	1.0
WATR	1.4
GEOM	
TOL	1e-8
BUFR	0.5
MEM	10
CUTA	20
HSX1	-0.125
HSX2	0.25
ALFRQ	1
EMP	1.5
P4	0.0
P6	8.0
P3	0.70
ONX	1.9
OFFX	2.1
WTYP	2
NPHI	38
SHIFT	-0.102
SLOPE	0.9085
CORR	1

runs were recorded for subsequent averaging every 50 MD steps.

Transition Path Ensemble Optimization with ggaHFB

Initial path for optimization was prepared as follows. Starting from the X-ray structure corresponding to high Tl^+ concentrations (PDB code 1R3J), we gradually rotate its atoms into pII-like state producing three more states. Specifically, we rotate all the atoms of residues 1 through 5 along with the HN atom of residue 6 by -35.0 degrees about the N-CA bond of residue 6 (ÏŢ(6) angle). In addition, we rotate atoms of residues 7 through 8 along with atom O of residue 6 by -20.0 degrees about the CA-C bond of residue 6 (ÏĹ(6) angle). Overall we perform three such rotations to produce intermediates int1, int2 and int3.

We then take the low Tl^+ concentration structure (PDB code 1R3K) as yet another intermediate to seed the path produced by the rotation described above. This structure introduces the VAL sidechain rotation. The reason this structure cannot be used as the product is that its backbone atoms collapse during the free energy optimization back to the high Tl^+ concentration structure, while leaving the VAL sidechain in the conformation found in 1R3K structure. Therefore, the 1r3k structure is placed right after the 1r3j but before the int1. Thus, the following sequence of structures 1r3j, 1r3k, int1, int2, int3 are fed to the ggaHFB interpolation procedure to generate 12 beads using the Fourier series truncation parameter of 4.

In constructing the Reactive Coordinate Space (RCS), we setup the following subspaces or levels:

Only heavy (non hydrogen) atoms of the peptide are included into the RCS.

Using these levels we define RCS1 and RCS2 as follows. RCS1 combines levels 1 through 7, whereas RCS2 only includes the base level 1.

Throughout the path optimization we used uniform step size parameter beta. Slightly different protocols were used for the wild type and mutant optimization with the details provided in the following tables.

Table S4.2: RCS vectors for different levels

Level	Atom Types	Comments
1	C, N, CA, NT, CY, CAY and CAT	base of the polypeptide chain
2	additionally O, CB and OY	base + one bond
3	additionally CG, CG1, CG2 and OG1	base + two bonds
4	additionally CD1 and CD2	base + three bonds
5	additionally CE1 and CE2	base + four bonds
6	additionally CZ	base + five bonds
7	additionally OH	base + six bonds

Table S4.3: The optimization protocol for the wild type.

Beads	ggaHFB Steps	Force Constant	Beta	Trunc
GBMV Implicit				
12	1-100	10.0	0.000	8
	101-199	5.0	0.004	8
23	200-201	1.0	0.000	18
	202-499	5.0	0.005	18
	500-599	10.0	0.002	18
CTIP3P Explicit				
23	600-609	10.0	0.000	18
	610-1099	10.0	0.002-0.004	18
45	1100-1199	10.0	0.0035	42
89	1200-1209	10.0	0.0035	86

Table S4.4: The optimization protocol for the mutant.

Beads	ggaHFB Steps	Force Constant	Beta	Trunc
GBMV Implicit				
12	1-9	10.0	0.000	8
	10-20	10.0	0.001	8
	21-199	10.0	0.002	8
23	200-209	1.0	0.000	18
	210-599	5.0	0.005	18
	600-799	10.0	0.003	18
CTIP3P Explicit				
23	800-804	10.0	0.000	18
	805-1199	10.0	0.002-0.004	18
45	1200-1299	10.0	0.0035	42
89	1300-1309	10.0	0.0035	86

PMF Integration with ggaHFB

Once the optimization was completed as determined by cessation of the changes in the coordinates of the RCS atoms a collection procedure was initiated. Note, that we only performed path optimizations on the RCS1 free energy surface and not RCS2. However, we used the subset of the atoms from the optimized RCS1 transition path ensemble as a reference to compute the PMF for the RCS2 level.

The final PMFs were collected using the averaged positions of the RCS atoms from 1 ns long batches of MD simulations using the final 89âĂŞbead path optimized at the RCS1 level as a reference. To restrain the atoms to the reference positions we used the force constant of 10.0 $\mathrm{kcal} \cdot \mathrm{g}^{-1} \cdot \mathring{\mathrm{A}}^{-2}$. To compute the RCS1 level PMFs only five batches were necessary to achieve high convergence, whereas for the RCS2 level we needed 13 batches. The averages were combined into the final cumulative averages over the whole simulation time using the standard procedure described earlier.

During the final integration procedure we used the Fourier series truncation parameter of 88 with 1024 quadrature points for the reversible work line integral.

Comments on the PMF Validity

Keeping the RCS1 path and releasing the sidechains for the RCS2 PMF calculations substantially expands the SCS space. Unfortunately, proper averaging over all possible configurations of the sidechains necessitates overcoming barriers as high as 5-8 kcal/mol as we have seen on the V76 example. Therefore, a complete averaging cannot be achieved for the RCS2 PMF within the limited simulation time of regular MD. We could improve sampling by using parallel tempering or replica exchange MD, [156, 157, 45, 158] but that would increase the cost of the computed PMFs by a factor equal to the number of replicas. Therefore, in this paper we limit ourselves to regular MD simulations. We still get apparently well-converged PMFs without having properly sampled alternative configurations of the sidechains. As expected the RCS2 PMFs take significantly longer to converge to the accuracy comparable to that of the RCS1 PMFs. Nevertheless, the RCS2 ensemble contains the discontinuities in the regions of the V76 transitions inherited from the preceding RCS1 path even after 13 ns. In particular, the sharp peaks corresponding to the V76 rotations at RCS1 level collapse at the RCS2 level creating the discontinuities in the SCS space. Strictly speaking such discontinuities invalidate the portion of the RCS2 PMFs in that region. Therefore, these PMFs should only be considered as tentative until more extensive sampling of the sidechain conformations is achieved. Work along those lines is currently in progress in our lab.

Nevertheless, the portion of the PMF underlying the important backbone transition from the α-strand to the pII state remains virtually unchanged upon going from RCS1 to RCS2 level. Thus, for the wild type peptide the forward activation barrier is 5.9 kcal/mol and the pII state is still less stable than the α-strand by slightly smaller 1.7 kcal/mol. Restoring the conducting state requires overcoming a slightly higher barrier of 4.2 kcal/mol. For the mutant we find the forward barrier of 0.9 kcal/mol and the relative pII state stabilization energy of 7.0 kcal/mol that makes the reverse barrier increase slightly to 7.9 kcal/mol. The differences between RCS1 and RCS2 PMFs are quite small in the backbone transition region. Thus, we conclude that structural details of the sidechains have very little effect on this functionally important backbone transition making it extremely robust.

The PMFs for the wild type peptide indicate that the free energy landscape is funneled toward the α-strand, whereas the mutation changes direction of transition opposing the α-strand formation. Changes in the PMFs due to mutation are consistent with the Hammond postulate. [159] Further examination of the optimized transition path ensembles reveals that formation of the α-strand is coupled with the hydrophobic collapse between the V76 and Y78 residues. Such hydrophobic interactions are often considered the driving force in protein folding. Nevertheless, the strength of this particular interaction is not sufficient to stabilize the α-strand in the mutant.

Minimum Adiabatic Potential Energy Transition Path Optimization with ggaHFB

To optimize the paths in gas phase using the bare CHARMM22 forcefield we start from the minimum free energy transition path ensembles optimized on RCS1 surface in water. These paths have 89 beads. We first rebuild all the hydrogen positions by potential energy optimization with the fixed RCS1 atoms for all the beads. Because tyrosine and threonine OH groups have two and three rotameric states, we have to initialize their dihedral angles Cϵ1-Cς-Oη-Hη and Cα-Cβ-Oγ-Hγ, respectively. Thus, we assign 180, 180, -30 and 0 degrees for the T72, T74, T75 and Y78, dihedral angles, correspondingly. Adding tyrosine and threonine hydrogen atoms from the OH groups to the RCS1 we create the new reactive coordinate space RCS1h that is sufficient to integrate the adiabatic potential energy along the path.

Initially we performed ggaHFB optimization using 89 beads. With 89 beads we used 76 basis functions, the force constant of 20.0 kcal $\cdot$ g^{-1} $\cdot$ $^{-2}$ and step size parameter of 0.0025. We did not attempt to find the optimal optimization conditions and executed on the order of 2000 ggaHFB optimization steps. For each bead we used the mass weighted harmonic restraints in Cartesian coordinates, and tandem Steepest Descent/Adaptive Basis Newton Raphson optimizer with up to 200/1000 optimization steps. Because bead optimizations are very fast they were done for each bead consecutively within a single CHARMM input script, which was run on a single CPU. The optimization was set to exit once the average gradient change was less than

10^{-5}kcal $\cdot$ mol^{-1-1}. Throughout this work we truncated the electrostatic interactions with 16 Åcutoff, switching the interactions off between 16 and 18 Å. The non-bonded list cutoff was 21 Å.

As the ggaHFB optimization continues the path length increases and the OH groups rotate to different optimal positions. Soon the forces between the beads become discontinuous and adiabatic potential energy profile integration via the generalized line integral formalism no longer gives correct result. Despite that the optimization can be continued further and the progress of the path optimization can be followed by the path RMSD or by plotting a two-dimensional projection of the path onto the reactant and product vectors (not shown).

We started the optimization with 89 beads and then subsequently increased the number of beads to 177, 353, and finally 705 beads until the energy profile could be integrated precisely again. With 177 beads we performed on the order of 1000 ggaHFB optimization steps using up to 168 basis functions and the step size parameter of 0.0025. With 353 beads we performed on the order of 100 ggaHFB steps with up to 324 basis functions and we also changed the force constant of the harmonic restraint from 20.0 to 40.0 kcal $\cdot$ g^{-1} $\cdot$ $^{-2}$, and turned of the steepest descent optimization boost by setting the step size parameter to 0.0000. Finally with 705 beads we performed on the order of 100 ggaHFB steps with up to 695 basis functions keeping the rest of the parameters the same as in the case of 353 beads.

Computing the Final Adiabatic Potential Energy Profile with ggaHFB

In the end of optimization we performed a single ggaHFB step with 705 beads and 705 basis functions (for the highest accuracy) in the Fourier series, using the force constant of 40.0 kcal $\cdot$ g^{-1} $\cdot$ $^{-2}$. To evaluate the generalized line integral we used 2820 quadrature points. Finally, we have computed the exact energies along the structures at each of the 2820 quadrature points. The highest deviations must be observed at the progress variable value of 1.0. We find the accumulated errors with respect to the exact CHARMM22 energies of 0.07 and 0.12 kcal/mol for the wild type and mutant adiabatic energy profiles, respectively. These numbers could be improved even further with additional beads, because the onsets of the deviations between the exact and the ggaHFB profiles appear at the sharp peaks.

Energy Profiling of the Minimum Free-energy Transition Path Ensembles

Gas Phase Energy

We used Density Functional Theory model, namely B3LYP with 6-31G(d) basis set to compute the single point Quantum Mechanical energies along the path. To do that we optimized all the degrees of freedom orthogonal to RCS1 except for the four dihedral angles mentioned above that were kept restrained at the 180, 180, -30 and 0 degrees for the T72, T74, T75 and Y78, respectively using the CHARMM22 Molecular Mechanical forcefield. The B3LYP/6-31G(d) energies were then compared to the corresponding CHARMM22 energies.

Implicit Solvent Energy

To compare the Quantum Mechanical and Molecular Mechanical energies in water, we employed Polarizable Continuum Model. For the QM-PCM model we used B3LYP/6-31G(d) with scrf=(pcm,solvent=water,read) keywords in the Gaussian G03 input file with additional parameters pcmdoc, radii=uaks, scfvac, ofac=0.8, rmin=0.5. For the MM we used GBMV implicit solvent model with the exact same parameters as has been described above. Prior to computing single point QM-PCM energies we optimized all the degrees of freedom orthogonal to the RCS1 except the four dihedral angles mentioned above at the MM-GBMV level.

Conclusion

The present work demonstrates the utility of the novel ggaHFB method in studying complex processes on multidimensional free energy surfaces on the example of the important functional transition of the selectivity filter of KcsA ion channel. Other important questions that require exploring multidimensional free-energy sur-

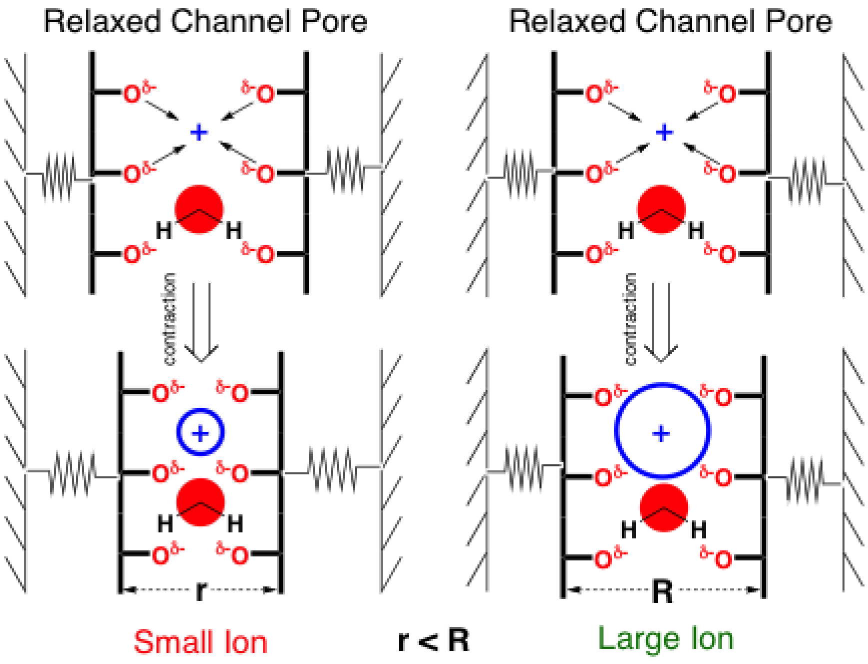

Figure S4.1: An illustration of the ion selectivity hypothesis. Thick black lines represent the signature peptide of the selectivity filter connected with springs to the outer barrel of the channel. In this model, smaller ions will impede co-translating water passage to a greater degree than larger ions by contracting the carbonyl rings of the channel. The ion sizes are not drawn to scale, but exaggerated to demonstrate the point.

faces can now be addressed. Work is now in progress in our lab to verify the hypothesis for the ion selectivity (see Figure S1.1) within the tetrameric KcsA channel.

Chapter 5

Calculation of Absolute Binding Free Energies by Thermodynamic Integration in AMBER12

Abstract

5.1 Introduction

Prior to AMBER12 the implementations of thermodynamic integration in the *sander* program were intended for relative binding free energy computation only. [160, 161] Two states must be well defined in order to compute a free energy difference between them, and in relative binding free energy calculations this is generally trivial since only unique components of the ligands are altered. The unperturbed portions of the ligand are assumed to keep the ligand in the correct *bound* pose. Absolute binding free energies must define the bound state more rigorously [162], for example by applying translational and sometimes rotational restraints to the ligand. [23, 163]

This communication outlines an update to the *sander* program in AMBER12 that correctly handles restraints during thermodynamic integration. We also implemented the ability to run at arbitrary lambda points in the range of $0 \leq \lambda \leq 1$. This feature becomes important when calculating absolute binding free energies, or when no preliminary knowledge of the $\langle \frac{\partial U}{\partial \lambda} \rangle_\lambda$ vs. λ curve is available. The new features of these implementations are presented and described through an illustrative example.

5.2 Theory

Absolute binding free energies for receptor-ligand binding can be estimated using the *double decoupling* method, outlined in figure 1.1. [23] Restraints are used to define the bound state of the ligand in the ΔG_c leg of the alchemical cycle. [164, 165, 35] In order to properly treat the restraint forces on perturbed atoms we must decompose the perturbed potential energy function $V\left(\mathbf{q}, \lambda\right)$, where $\mathbf{q}$ is a vector containing the atomic positions of the system and λ is the alchemical coupling parameter. For each leg of the alchemical cycle the two endpoints are denoted with subscripts of 0 or 1. The endpoints are further divided into perturbed and unperturbed portions, with p and u subscripts, respectively. A final separation between the covalent plus restraint and non-bonded energy terms lets us completely describe which energy terms are scaled, as shown in equation 1.1. The $V_{u\leftrightarrow u}$ term is encompasses all of the interactions including only unperturbed atoms. The $V_{0,u\leftrightarrow p,cov+rest}$ and $V_{1,u\leftrightarrow p,cov+rest}$ terms are the covalent and restraint forces between the unperturbed and perturbed atoms for the initial and final endpoints, respectively. The $V_{0,p\leftrightarrow p}$ and $V_{1,p\leftrightarrow p}$ terms encompass all of the interactions including only perturbed atoms for the initial and final endpoints, respectively. The only terms that are scaled by λ are $V_{0,u\leftrightarrow p,nb}$ and $V_{1,u\leftrightarrow p,nb}$ terms, which are the non-bonded interactions between the non-perturbed atoms and the perturbed atoms for the initial and final endpoints, respectively.

$$\begin{aligned} V\left(\mathbf{q}, \lambda\right) = & V_{u\leftrightarrow u}\left(\mathbf{q}\right) + V_{0,u\leftrightarrow p,cov+rest}\left(\mathbf{q}\right) + V_{1,u\leftrightarrow p,cov+rest}\left(\mathbf{q}\right) \\ & + V_{0,p\leftrightarrow p}\left(\mathbf{q}\right) + V_{1,p\leftrightarrow p}\left(\mathbf{q}\right) + \left(1-\lambda\right) V_{0,u\leftrightarrow p,nb}\left(\mathbf{q}, \lambda\right) \\ & + \lambda V_{1,u\leftrightarrow p,nb}\left(\mathbf{q}, \lambda\right) \end{aligned} \tag{5.1}$$

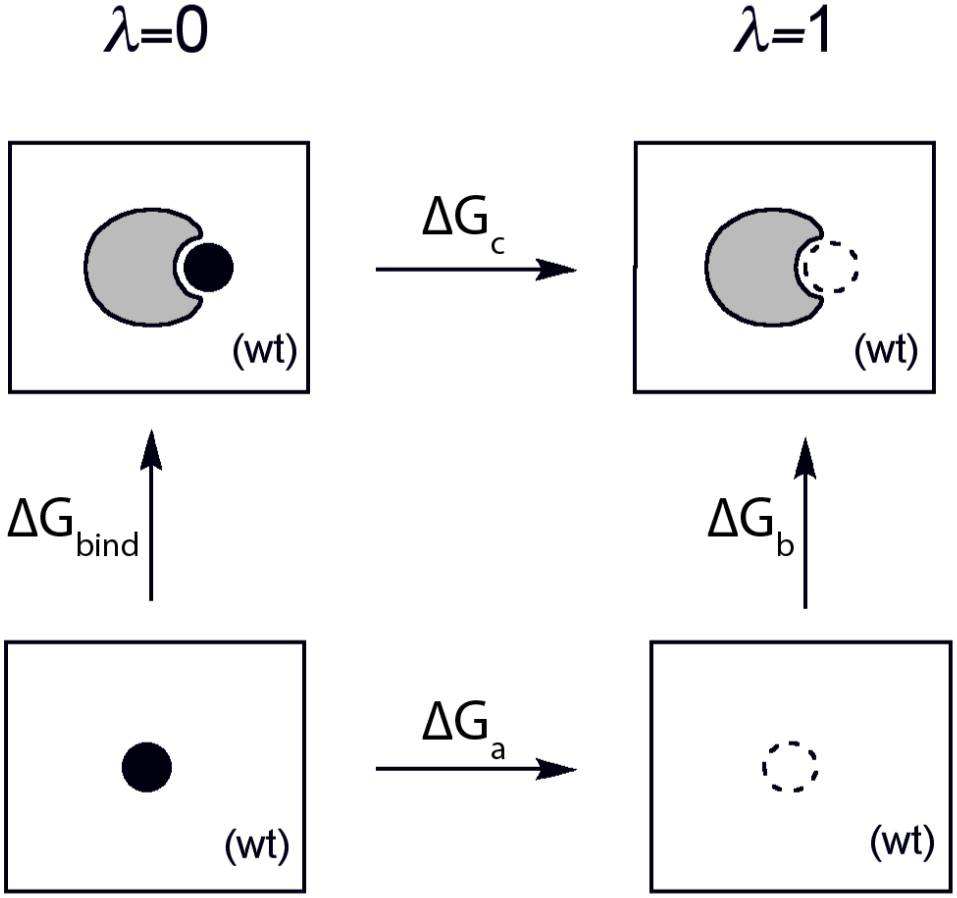

Figure 5.1: Schematic of the double-decoupling method taken from [2]. The closed and dashed circles refer to the coupled and decoupled ligand, respectively, and (wt) refers to the water solvent. The alchemical free energy to decouple the ligand from bulk solvent (ΔG_a) and from the protein (ΔG_c) are computed through simulations. The contribution from moving the decoupled ligand from bulk solvent to the protein (ΔG_b) is zero.

5.3 Methods

The bicyclo[2.2.2]octane compounds are putative binders to the seven-unit cucurbitural host system. [166] We parameterized 1,4-dimethyl alcohol ginbicyclo[2.2.2]octane, referred to as B2, using the AMBER GAFF parameters and the RESP charge fitting procedure.

Simulations were run using the *sander* program from AMBER10 [167] and a pre-release version of AMBER12. In order to emphasize the modifications we only looked at the soft-core Lennard-Jones decoupling [25] of an uncharged B2 molecule in a 36x35x36 Å box of TIP4P water [143]. A Langevin thermostat with a target temperature of 300 Kelvin and a collision frequency of 20 ps^{-1} and weak-coupling pressure control with a relaxation time of 0.5 ps was used to enforce a NPT ensemble. [168] The conformational space of the ligand was restricted by a harmonic force applied to one of the core atoms with a force constant of 2.5 kcal/mol/Å^2. The simulations were run for 5 ns each, and the first nanosecond discarded as equilibration. The λ endpoints had to be 0.01 and 0.99 for AMBER10. The AMBER12 endpoints were also run at 0.01 and 0.99 for comparison with AMBER10, and additional runs at λ of 0.00 and 1.00 to determine the numerical integration error.

5.4 Results

The first thing to establish is that the restraint forces are properly treated during the Hamiltonian perturbation. The displacement of the restrained atom from its average position at $\lambda = 0.99$ is shown Figure 1.2. The restraint force in the AMBER10 simulation is scaled improperly and samples a larger volume than in the AMBER12 simulation. The average displacement during the AMBER12 simulation is also independent of λ (Figure S1.1), showing that the state is well defined during the alchemical perturbation.

The $\langle \frac{\partial U}{\partial \lambda} \rangle_\lambda$ vs. λ curves for the AMBER10 and AMBER12 codes are shown in Figure 1.3. The scaling of the restraint forces in the AMBER10 simulation creates the offset between the two curves. The contribution to $\partial V/\partial\lambda$ from the restraint can be removed from the TI curve, but there is another problem that cannot be so easily addressed. The soft-core potential allows atoms to overlap with an energy penalty that decreases inversely proportional to λ. In the case of a protein-ligand system there is a value of λ at which the unrestrained ligand can begin overlapping with the protein atoms, and eventually travel through the protein into bulk solvent. The $\partial V/\partial\lambda$ term will change as the environment changes, and thus if the ligand is not properly restrained to the binding site the $\langle\partial V/\partial\lambda\rangle$ will have a complex offset. To emphasize the role of the restraint potential we presented a simple example of the perturbation of B2 in water. There is an example of the dependence of $\partial V/\partial\lambda$ on the environment in the supplementary materials.

5.5 Conclusion

The theoretical and practical importance of restraining a ligand to a well-defined state during the course of absolute binding free energy computation was summarized. We presented the results of modifications to the AMBER12 *sander* program that addressed the proper handling of restraints. The changes were validated with a Cartesian restraint and showed that the ligand can be restrained to a well-defined volume as is necessary in absolute binding free energy calculations.

Acknowledgement

Chapter 1, in full, is a minimally modified manuscript that is being submitted. The dissertation author was the principal investigator and the first author of this paper.

Supplementary Material

Displacement during Alchemical Perturbation

The bound state is well defined during the entire absolute binding alchemical perturbation for AMBER12, which was not the case in AMBER10. This is shown

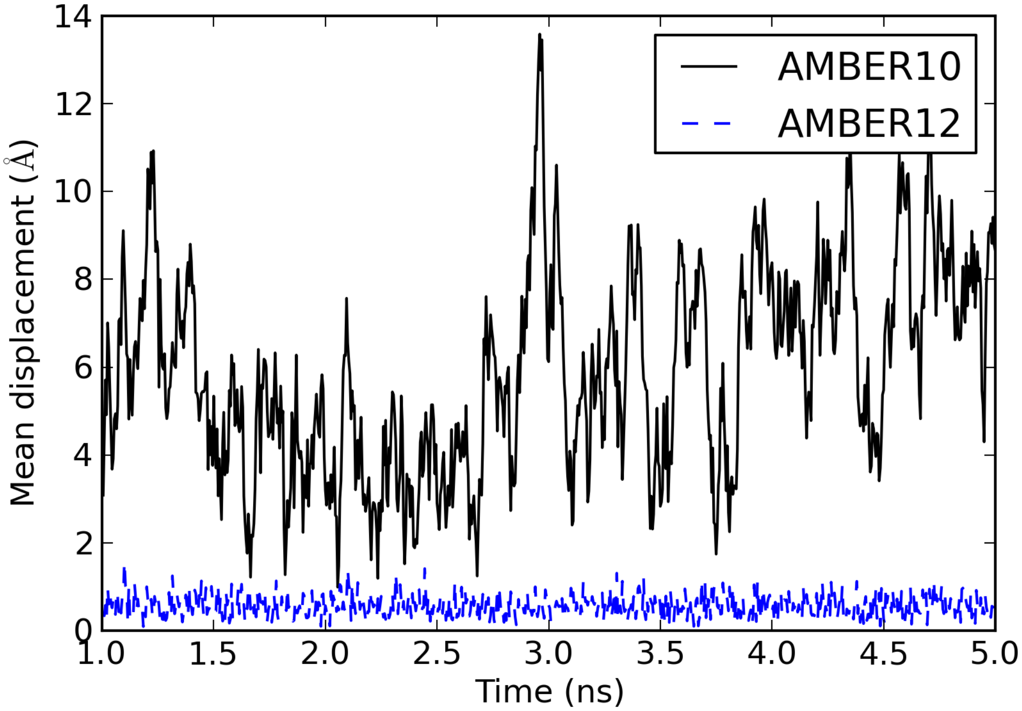

Figure 5.2: The mean displacement (Å) of the restraint atom from its average position. The displacement corresponding to $3/2 \cdot k_B T$ at 300K is 0.85 Å.

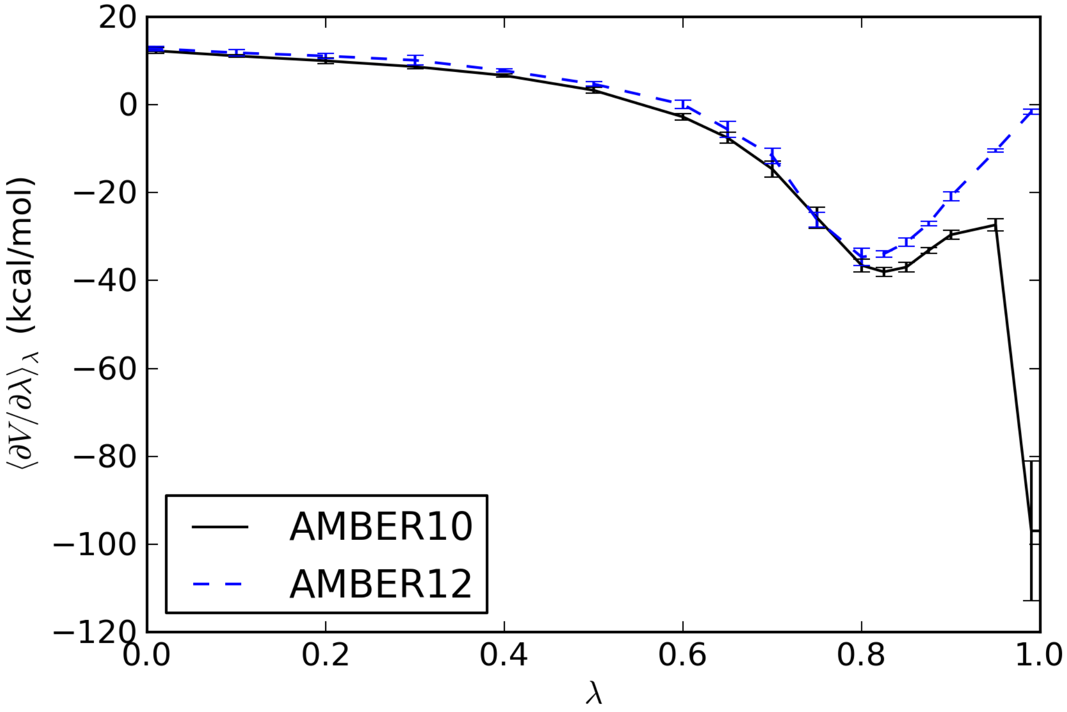

Figure 5.3: Thermodynamic integration curve for the perturbation of van der Waals interactions between B2 and water.

with the mean displacement of the B2 restraint atom as a function of λ in Figure S1.1).

Dependence of $\partial V/\partial\lambda$ on Environment

The complex of B2 with CB[7] has a more restricted conformational space than most protein-ligand complexes. The decoupled $\lambda = 0.99$ linear-mixing soft-core Hamiltonian used the same simulation parameters outlined in the paper. The major difference was to replace the Cartesian restraint with a center-of-mass distance restraint applied between CB[7] and B2. In order to show the dependence of $\partial V/\partial\lambda$ on the center-of-mass distance between ligand and host a series of six 100-ps simulations were performed with flat-bottomed harmonic potentials. The flat portions of the simulations were 0-2, 2-4, 4-6, 6-8, 8-10, and 10-12 Å for each of the simulations, and the force constant was 2.5 kcal/mol/Å^2. The simulations were started consecutively starting from 0-2, and each simulation starting from the last structure of its predecessor.

Figure S1.2 shows the correlation between $\partial V/\partial\lambda$ and the center-of-mass displacement of CB[7] and B2. This highlights how important it is to properly define and restrain the ligand to the bound state.

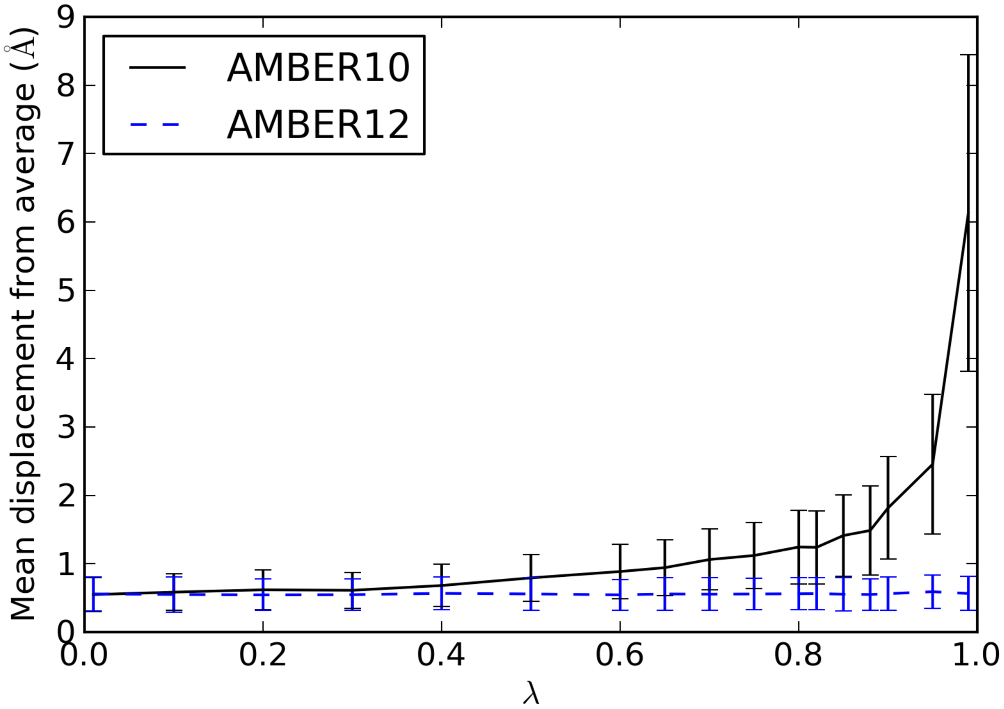

Figure S5.1: The mean displacement (Å) of the restraint atom as a function of λ.

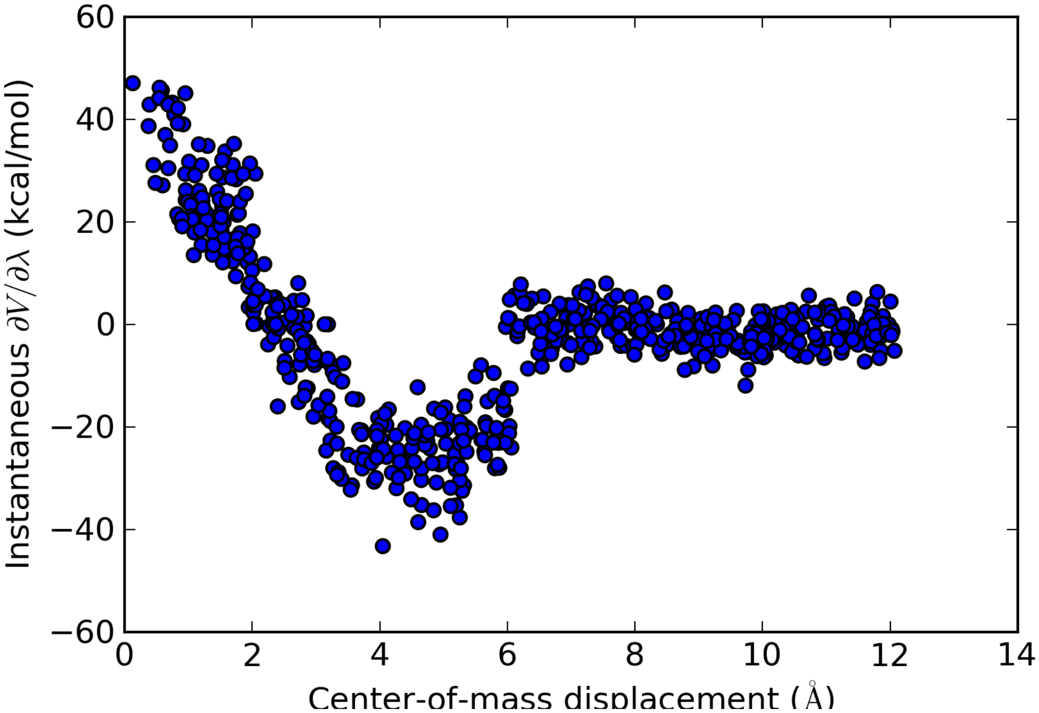

Figure S5.2: The instantaneous $\partial V/\partial\lambda$ at various intermolecular displacements.

Bibliography

[1] Alessandro Laio and Francesco L. Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports On Progress In Physics*, 71(12), 2008.

[2] Morgan Lawrenz, Riccardo Baron, and J. Andrew McCammon. Independent-trajectories thermodynamic-integration free-energy changes for biomolecular systems: Determinants of h5n1 avian influenza virus neuraminidase inhibition by peramivir. *Journal of Chemical Theory and Computation*, 5(4):1106–1116, 2009.

[3] Mark J. Abraham and Jill E. Gready. Ensuring mixing efficiency of replica-exchange molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 4(7):1119–1128, 2008.

[4] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129(12):124105, 2008.

[5] W. T. Astbury and A. Street. X-ray studies of the structure of hair, wool, and related fibres. i. general. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 230:pp. 75–101, 1932.

[6] Linus Pauling, Robert B. Corey, and H. R. Branson. The structure of proteins. *Proceedings of the National Academy of Sciences*, 37(4):205–211, 1951.

[7] Linus Pauling and Robert B. Corey. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. *Proc Natl Acad Sci USA*, 37(11):729–740, 1951.

[8] Tamar Schlick. *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Springer, 2010.

[9] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proc Natl Acad Sci U S A*, 99(20):12562–12566, 2002.

[10] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov. New approach to monte carlo calculation of the free energy: Method of expanded ensembles. 96(3):1776–1783, 1992.

[11] Arthur F. Voter. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.*, 78(20):3908–3911, 1997.

[12] Donald Hamelberg, John Mongan, and J Andrew McCammon. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J Chem Phys*, 120(24):11919–29, 2004.

[13] Hongzhi Li, Mikolai Fajer, and Wei Yang. Simulated scaling method for localized enhanced sampling and simultaneous "alchemical" free energy simulations: a general method for molecular mechanical, quantum mechanical, and quantum mechanical/molecular mechanical simulations. *J Chem Phys*, 126(2):024106, 2007.

[14] Hongzhi Li, Donghong Min, Yusong Liu, and Wei Yang. Essential energy space random walk via energy space metadynamics method to accelerate molecular dynamics simulations. *The Journal of Chemical Physics*, 127(9):094101, 2007.

[15] David A. Pearlman and Peter A. Kollman. The lag between the hamiltonian and the system configuration in free energy perturbation calculations. *The Journal of Chemical Physics*, 91(12):7831–7839, 1989.

[16] Huan-Xiang Zhou and Michael K. Gilson. Theory of free energy and entropy in noncovalent binding. *Chemical Reviews*, 109(9):4092–4107, 2009.

[17] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187 – 199, 1977.

[18] Davide Branduardi, Francesco Luigi Gervasio, Andrea Cavalli, Maurizio Recanatini, and Michele Parrinello. The role of the peripheral anionic site and cationâĹŠÏĂ interactions in the ligand penetration of the human ache gorge. *Journal of the American Chemical Society*, 127(25):9147–9155, 2005. PMID: 15969593.

[19] Justin R. Gullingsrud, Rosemary Braun, and Klaus Schulten. Reconstructing potentials of mean force through time series analysis of steered molecular dynamics simulations. *Journal of Computational Physics*, 151(1):190 – 211, 1999.

[20] Graeme Henkelman, Blas P. Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. 113(22):9901–9904, 2000.

[21] Ilja V. Khavrutskii and J. Andrew McCammon. Generalized gradient-augmented harmonic fourier beads method with multiple atomic and/or center-of-mass positional restraints. *The Journal of Chemical Physics*, 127(12):124901, 2007.

[22] A.C. Pan, D. Sezer, and B. Roux. Finding transition pathways using the string method with swarms of trajectories. *Journal of Physical Chemistry B*, 112(11):3432–3440, 2008.

[23] M.K. Gilson, J.A. Given, B.L. Bush, and J.A. McCammon. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical Journal*, 72(3):1047 – 1069, 1997.

[24] H. Lee Woodcock, Benjamin T. Miller, Milan Hodoscek, Asim Okur, Joseph D. Larkin, Jay W. Ponder, and Bernard R. Brooks. Mscale: A general utility for multiscale modeling. *Journal of Chemical Theory and Computation*, 7(4):1208–1219, 2011.

[25] M. Zacharias, T. P. Straatsma, and J. A. McCammon. Separation-shifted scaling, a new scaling method for lennard-jones interactions in thermodynamic integration. *The Journal of Chemical Physics*, 100(12):9025–9031, 1994.

[26] Thomas C. Beutler, Alan E. Mark, Rene C. van Schaik, Paul R. Gerber, and Wilfred F. van Gunsteren. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters*, 222(6):529 – 539, 1994.

[27] Thomas Steinbrecher, David L Mobley, and David A Case. Nonlinear scaling schemes for lennard-jones interactions in free energy calculations. *J Chem Phys*, 127(21):214108, 2007.

[28] Tri T. Pham and Michael R. Shirts. Identifying low variance pathways for free energy calculations of molecular transformations in solution phase. *The Journal of Chemical Physics*, 135(3):034114, 2011.

[29] John G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313, 1935.

[30] Robert W. Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *The Journal of Chemical Physics*, 22(8):1420–1426, 1954.

[31] Michael R Shirts and Vijay S Pande. Comparison of efficiency and bias of free energies computed by exponential averaging, the bennett acceptance ratio, and thermodynamic integration. *J Chem Phys*, 122(14):144107, 2005.

[32] Charles H. Bennett. Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.

[33] Paul Maragakis, Martin Spichty, and Martin Karplus. Optimal estimates of free energies from multistate nonequilibrium work data. *Phys Rev Lett*, 96(10):100602, 2006.

[34] Hideaki Fujitani, Yoshiaki Tanida, Masakatsu Ito, Guha Jayachandran, Christopher D. Snow, Michael R. Shirts, Eric J. Sorin, and Vijay S. Pande. Direct calculation of the binding free energies of fkbp ligands. *The Journal of Chemical Physics*, 123(8):084108, 2005.

[35] D. Hamelberg and J.A. McCammon. Standard free energy of releasing a localized water molecule from the binding pockets of proteins: Double-decoupling method. *Journal of the American Chemical Society*, 126(24):7683–7689, 2004.

[36] Jiyao Wang, Yuqing Deng, and Benoit Roux. Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophys. J.*, 91(8):2798–2814, 2006.

[37] C.J. Woods, J.W. Essex, and M.A. King. The development of replica-exchange-based free-energy methods. *Journal of Physical Chemistry B*, 107(49):13703–13710, 2003.

[38] P.R.L. Markwick, G. Bouvignies, and M. Blackledge. Exploring multiple timescale motions in protein gb3 using accelerated molecular dynamics and nmr spectroscopy. *J. Am. Chem. Soc.*, 129(15):4724 – 4730, 2007.

[39] Tongye Shen and Donald Hamelberg. A statistical analysis of the precision of reweighting-based simulations. *J Chem Phys*, 129(3):034103, 2008.

[40] Chao Xu, Jun Wang, and Haiyan Liu. A hamiltonian replica exchange approach and its application to the study of side-chain type and neighbor effects on peptide backbone conformations. *Journal of Chemical Theory and Computation*, 4(8):1348–1359, 2008.

[41] Donald Hamelberg, César Augusto F de Oliveira, and J Andrew McCammon. Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. *J Chem Phys*, 127(15):155102, 2007.

[42] David A. Pearlman, David A. Case, James W. Caldwell, Wilson S. Ross, Thomas E. Cheatham, Steve DeBolt, David Ferguson, George Seibel, and Peter Kollman. Amber, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1-3):1 – 41, 1995.

[43] Paul Maragakis, Kresten Lindorff-Larsen, Michael P. Eastwood, Ron O. Dror, John L. Klepeis, Isaiah T. Arkin, Morten O. Jensen, Huafeng Xu, Nikola Trbovic, Richard A. Friesner, Arthur G. Palmer, and David E. Shaw. Microsecond

molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins. *The Journal of Physical Chemistry B*, 112(19):6155–6158, 2008.

[44] Cameron Mura and J Andrew McCammon. Molecular dynamics of a kappab dna element: base flipping via cross-strand intercalative stacking in a microsecond-scale simulation. *Nucleic Acids Res*, 36(15):4941–4955, 2008.

[45] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2):141–151, 1999.

[46] Jose D Faraldo-Gomez and Benoit Roux. Characterization of conformational equilibria through hamiltonian and temperature replica-exchange simulations: assessing entropic and environmental effects. *J Comput Chem*, 28(10):1634–1647, 2007.

[47] Giovanni Ciccotti, Mauro Ferrario, James T. Hynes, and Raymond Kapral. Constrained molecular dynamics and the mean potential for an ion pair in a polar solvent. *Chemical Physics*, 129(2):241–251, 1989.

[48] Mikolai Fajer, Donald Hamelberg, and J. Andrew McCammon. Replica-exchange accelerated molecular dynamics (rexamd) applied to thermodynamic integration. *Journal of Chemical Theory and Computation*, 4(10):1565–1569, 2008.

[49] Hiroaki Fukunishi, Osamu Watanabe, and Shoji Takada. On the hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *The Journal of Chemical Physics*, 116(20):9058–9067, 2002.

[50] Shankar Kumar, Djamal Bouzida, Robert H. Swendsen, Peter A. Kollman, and John M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules. i: The method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[51] Mark N Kobrak. Systematic and statistical error in histogram-based free energy calculations. *J Comput Chem*, 24(12):1437–1446, 2003.

[52] Zhiqiang Tan. On a likelihood approach for monte carlo integration. *Journal of the American Statistical Association*, 99(468):1027–1036, 2004.

[53] Yufeng Zhou and Roderick MacKinnon. The occupancy of ions in the k+ selectivity filter: Charge balance and coupling of ion binding to a protein conformational change underlie high conduction ratesncy of ions in the k+ selectivity filter: Charge balance and coupling of ion binding to a protein conformational change underlie high conduction rates. *Journal of Molecular Biology*, 333(5):965–975, 2003.

[54] Yufeng Zhou, Joao H. Morais-Cabral, Amelia Kaufman, and Roderick MacKinnon. Chemistry of ion coordination and hydration revealed by a k+ channel-fab complex at 2.0å resolution. *Nature*, 414(6859):43–48, 2001.

[55] Steve W Lockless, Ming Zhou, and Roderick MacKinnon. Structural and thermodynamic properties of selective ion binding in a k+ channel. *PLoS Biol*, 5(5):e121, 2007.

[56] Francis I. Valiyaveetil, Manuel Leonetti, Tom W. Muir, and Roderick MacKinnon. Ion selectivity in a semisynthetic k+ channel locked in the conductive conformation. *Science*, 314(5801):1004–1007, 2006.

[57] Simon Berneche and Benoit Roux. Molecular dynamics of the kcsa k+ channel in a bilayer membrane. *Biophysical Journal*, 78(6):2900–2917, 2000.

[58] Johan Aqvist and Victor Luzhkov. Ion permeation mechanism of the potassium channel. *Nature*, 404(6780):881–884, 2000.

[59] Simon Berneche and Benoit Roux. Energetics of ion conduction through the k+ channel. *Nature*, 414(6859):73–77, 2001.

[60] Simon Berneche and Benoit Roux. A gate in the selectivity filter of potassium channels. *Structure*, 13(4):591–600, 2005.

[61] R. Jay Mashl, Yuzhou Tang, Jim Schnitzer, and Eric Jakobsson. Hierarchical approach to predicting permeation in ion channels. *Biophysical Journal*, 81(5):2473 – 2483, 2001.

[62] Leonardo Guidoni and Paolo Carloni. Potassium permeation through the kcsa channel: a density functional study. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1563(1-2):1 – 6, 2002.

[63] M. Compoint, C. Boiteux, P. Huetz, C. Ramseyer, and C. Girardet. Role of water molecules in the kcsa protein channel by molecular dynamics calculations. *Phys. Chem. Chem. Phys.*, 7:4138–4145, 2005.

[64] Hendrick W. de Haan, Igor S. Tolokh, C. G. Gray, and Saul Goldman. Nonequilibrium molecular dynamics calculation of the conductance of the kcsa potassium ion channel. *Phys. Rev. E*, 74:030905, 2006.

[65] Michael Grabe, Delphine Bichet, Xiang Qian, Yuh Nung Jan, and Lily Yeh Jan. K+ channel selectivity depends on kinetic as well as thermodynamic factors. *Proceedings of the National Academy of Sciences*, 103(39):14361–14366, 2006.

[66] Sebastian Kraszewski, Celine Boiteux, Marek Langner, and Christophe Ramseyer. Insight into the origins of the barrier-less knock-on conduction in the kcsa channel: molecular dynamics simulations and ab initio calculations. *Phys. Chem. Chem. Phys.*, 9:1219–1225, 2007.

[67] Carmen Domene, Satyavani Vemparala, Simone Furini, Kim Sharp, and Michael L. Klein. The role of conformation in ion permeation in a k+ channel. *Journal of the American Chemical Society*, 130(11):3389–3398, 2008.

[68] Jean-Fang Gwan and Artur Baumgaertner. Cooperative transport in a potassium ion channel. *The Journal of Chemical Physics*, 127(4):045103, 2007.

[69] Sergei Yu. Noskov, Simon Berneche, and Benoit Roux. Control of ion selectivity in potassium channels by electrostatic and dynamic properties of carbonyl ligands. *Nature*, 431(7010):830–834, 2004.

[70] Sergei Yu. Noskov and Benoit Roux. Ion selectivity in potassium channels. *Biophysical Chemistry*, 124(3):279 – 291, 2006. Ion Hydration Special Issue.

[71] Luca Maragliano and Eric Vanden-Eijnden. On-the-fly string method for minimum free energy paths calculation. *Chemical Physics Letters*, 446(1-3):182–190, 2007.

[72] Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *The Journal of Chemical Physics*, 126(16):164103, 2007.

[73] Ilja V. Khavrutskii, Karunesh Arora, and Charles L. Brooks. Harmonic fourier beads method for studying rare events on rugged energy surfaces. *The Journal of Chemical Physics*, 125(17):174108, 2006.

[74] Ilja V. Khavrutskii, Joachim Dzubiella, and J. Andrew McCammon. Computing accurate potentials of mean force in electrolyte solutions with the generalized gradient-augmented harmonic fourier beads method. *The Journal of Chemical Physics*, 128(4):044106, 2008.

[75] Hao Hu, Zhenyu Lu, Jerry M. Parks, Steven K. Burger, and Weitao Yang. Quantum mechanics/molecular mechanics minimum free-energy path for accurate reaction energetics in solution and enzymes: Sequential sampling and optimization on the potential of mean force surface. *The Journal of Chemical Physics*, 128(3):034105, 2008.

[76] Steven K. Burger and Weitao Yang. Sequential quadratic programming method for determining the minimum energy path. *The Journal of Chemical Physics*, 127(16):164107, 2007.

[77] Steven K. Burger and Weitao Yang. Quadratic string method for determining the minimum-energy path based on multiobjective optimization. *The Journal of Chemical Physics*, 124(5):054109, 2006.

[78] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.*, 100(2):020603, 2008.

[79] Davide Branduardi, Francesco Luigi Gervasio, and Michele Parrinello. From a to b in free energy space. *The Journal of Chemical Physics*, 126(5):054103, 2007.

[80] A. Laio and M. Parrinello. Computing free energies and accelerating rare events with metadynamics. In Mauro Ferrario, Giovanni Ciccotti, and Kurt Binder, editors, *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1*, volume 703 of *Lecture Notes in Physics*, pages 315–347. Springer Berlin / Heidelberg, 2006.

[81] G Bussi, A Laio, and M Parrinello. Equilibrium free energies from nonequilibrium metadynamics. *Physical Review Letters*, 96(9), 2006.

[82] Alessandro Laio, Antonio Rodriguez-Fortea, Francesco Luigi Gervasio, Matteo Ceccarelli, and Michele Parrinello. Assessing the accuracy of metadynamics. *J Phys Chem B*, 109(14):6714–6721, 2005.

[83] Arjan van der Vaart and Martin Karplus. Minimum free energy pathways and free energy profiles for conformational transitions based on atomistic molecular dynamics simulations. *The Journal of Chemical Physics*, 126(16):164106, 2007.

[84] Volodymyr Babin, Christopher Roland, Thomas A. Darden, and Celeste Sagui. The free energy landscape of small peptides as obtained from metadynamics with umbrella sampling corrections. *The Journal of Chemical Physics*, 125(20):204909, 2006.

[85] Volodymyr Babin, Christopher Roland, and Celeste Sagui. Adaptively biased molecular dynamics for free energy calculations. *The Journal of Chemical Physics*, 128(13):134101, 2008.

[86] 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, 1994.

[87] Kenichi Fukui. The path of chemical reactions - the irc approach. *Accounts of Chemical Research*, 14(12):363–368, 1981.

[88] Kenichi Fukui, Shigeki Kato, and Hiroshi Fujimoto. Constituent analysis of the potential gradient along a reaction coordinate. method and an application to methane + tritium reaction. *Journal of the American Chemical Society*, 97(1):1–7, 1975.

[89] G. Hummer and A. Szabo. Free energy surfaces from single-molecule force spectroscopy. *Accounts of Chemical Research*, 38(7):504–513, 2005.

[90] Johannes Kastner and Walter Thiel. Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "umbrella integration". *The Journal of Chemical Physics*, 123(14):144104, 2005.

[91] Julio F Cordero-Morales, Luis G Cuello, Yanxiang Zhao, Vishwanath Jogini, D Marien Cortes, Benoit Roux, and Eduardo Perozo. Molecular determinants of gating at the potassium-channel selectivity filter. *Nat Struct Mol Biol*, 13(4):311–318, 2006.

[92] Carmen Domene, Alessandro Grottesi, and Mark S.P. Sansom. Filter flexibility and distortion in a bacterial inward rectifier k+ channel: Simulation studies of kirbac1.1. *Biophysical Journal*, 87(1):256 – 267, 2004.

[93] Charlotte E. Capener, Peter Proks, Frances M. Ashcroft, and Mark S.P. Sansom. Filter flexibility in a mammalian k channel: Models and simulations of kir6.2 mutants. *Biophysical Journal*, 84(4):2345 – 2356, 2003.

[94] Fatemeh Khalili-Araghi, Emad Tajkhorshid, and Klaus Schulten. Dynamics of k+ ion conduction through kv1.2. *Biophysical Journal*, 91(6):L72 – L74, 2006.

[95] Sven Hovmoller, Tuping Zhou, and Tomas Ohlson. Conformations of amino acids in proteins. *Acta Crystallographica Section D*, 58(5):768–776, 2002.

[96] Alexander D. Mackerell, Michael Feig, and Charles L. Brooks. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry*, 25(11):1400–1415, 2004.

[97] Michael S. Lee, Michael Feig, Freddie R. Salsbury, and Charles L. Brooks. New analytic approximation to the standard molecular volume definition and its application to generalized born calculations. *Journal of Computational Chemistry*, 24(11):1348–1356, 2003.

[98] B. Montgomery Pettitt and Martin Karplus. The potential of mean force surface for the alanine dipeptide in aqueous solution: a theoretical approach. *Chemical Physics Letters*, 121(3):194 – 201, 1985.

[99] Wan F. Lau and B. Montgomery Pettitt. Conformations of the glycine dipeptide. *Biopolymers*, 26(11):1817–1831, 1987.

[100] Stephen B. Long, Xiao Tao, Ernest B. Campbell, and Roderick MacKinnon. Atomic structure of a voltage-dependent k+ channel in a lipid membrane-like environment. *Nature*, 450(7168):376–382, 2007.

[101] Sapar M. Saparov and Peter Pohl. Beyond the diffusion limit: Water flow through the empty bacterial potassium channel. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14):4805–4809, 2004.

[102] James E. Milner-White, James D. Watson, Guoying Qi, and Steven Hayward. Amyloid formation may involve Îś- to Îš sheet interconversion via peptide plane flipping. *Structure*, 14(9):1369 – 1376, 2006.

[103] Roger S. Armen, Mari L. DeMarco, Darwin O. V. Alonso, and Valerie Daggett. Pauling and corey's Îś-pleated sheet structure may define the prefibrillar amyloidogenic intermediate in amyloid disease. *Proceedings of the National Academy of Sciences of the United States of America*, 101(32):11622–11627, 2004.

[104] Roger S. Armen, Darwin O. V. Alonso, and Valerie Daggett. Anatomy of an amyloidogenic intermediate: Conversion of Îš-sheet to Îś-sheet structure in transthyretin at acidic ph. *Structure*, 12(10):1847 – 1863, 2004.

[105] Renee D. JiJi, Gurusamy Balakrishnan, Ying Hu, and Thomas G. Spiro. Intermediacy of poly(l-proline) ii and Îš-strand conformations in poly(l-lysine) Îš-sheet formation probed by temperature-jump/uv resonance raman spectroscopyâĂă. *Biochemistry*, 45(1):34–41, 2006.

[106] Sanford A. Asher, Alexander V. Mikhonin, and Sergei Bykov. Uv raman demonstrates that Îś-helical polyalanine peptides melt to polyproline ii conformations. *Journal of the American Chemical Society*, 126(27):8433–8440, 2004.

[107] L. Heginbotham, Z. Lu, T. Abramson, and R. MacKinnon. Mutations in the k+ channel signature sequence. *Biophysical Journal*, 66(4):1061 – 1067, 1994.

[108] Heiner Splitt, Dirk Meuser, Ilya Borovok, Michael Betzler, and Hildgund Schrempf. Pore mutations affecting tetrameric assembly and functioning of the potassium channel kcsa from streptomyces lividans. *FEBS Letters*, 472(1):83 – 87, 2000.

[109] A. D. MacKerell, D. Bashford, Bellott, R. L. Dunbrack, 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, 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 proteinsâĂă. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998.

[110] Jr. MacKerell, A. D., B. R. Brooks, III Brooks, C. L., L. Nilsson, B. Roux, Y. Won, and M. Karplus. *The Encyclopedia of Computational Chemistry*, chapter CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. John Wiley and Sons: Chichester, 1998.

[111] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098–3100, 1988.

[112] Axel D. Becke. Density-functional thermochemistry. iii. the role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 1993.

[113] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785–789, 1988.

[114] Michael S. Lee, Freddie R. Salsbury, and Charles L. Brooks. Novel generalized born methods. *The Journal of Chemical Physics*, 116(24):10606–10614, 2002.

[115] Maurizio Cossi, Vincenzo Barone, Benedetta Mennucci, and Jacopo Tomasi. Ab initio study of ionic solutions by a polarizable continuum dielectric model. *Chemical Physics Letters*, 286(3-4):253 – 260, 1998.

[116] E. Cancès, B. Mennucci, and J. Tomasi. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *The Journal of Chemical Physics*, 107(8):3032–3041, 1997.

[117] Benedetta Mennucci and Jacopo Tomasi. Continuum solvation models: A new approach to the problem of solute's charge distribution and cavity boundaries. *The Journal of Chemical Physics*, 106(12):5151–5158, 1997.

[118] Maurizio Cossi, Giovanni Scalmani, Nadia Rega, and Vincenzo Barone. New developments in the polarizable continuum model for quantum mechanical and classical calculations on molecules in solution. *The Journal of Chemical Physics*, 117(1):43–54, 2002.

[119] Erik M. Boczko and Charles L. Brooks. Constant-temperature free energy surfaces for physical and chemical processes. *The Journal of Physical Chemistry*, 97(17):4509–4513, 1993.

[120] Elena F. Koslover and David J. Wales. Comparison of double-ended transition state search methods. 127(13):134102, 2007.

[121] Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. Finite temperature string method for the study of rare events. *The Journal of Physical Chemistry B*, 109(14):6688–6693, 2005.

[122] Luca Maragliano, Alexander Fischer, Eric Vanden-Eijnden, and Giovanni Ciccotti. String method in collective variables: Minimum free energy paths and isocommittor surfaces. *The Journal of Chemical Physics*, 125(2):024106, 2006.

[123] Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Phys. Rev. B*, 66:052301, 2002.

[124] Ron Elber. Long-timescale simulation methods. *Current Opinion in Structural Biology*, 15(2):151 – 156, 2005. Theory and simulation/Macromolecular assemblages.

[125] Ron Elber, Alfredo Cárdenas, Avijit Ghosh, and Harry A. Stern. *Advances in Chemical Physics*, chapter Bridging the Gap between Long Time Trajectories and Reaction Pathways, pages 93–129. John Wiley and Sons, Inc., 2003.

[126] Roberto Olender and Ron Elber. Calculation of classical trajectories with a very large time step: Formalism and numerical examples. *The Journal of Chemical Physics*, 105(20):9299–9315, 1996.

[127] Ilja V. Khavrutskii, Richard H. Byrd, and Charles L. Brooks. A line integral reaction path approximation for large systems via nonlinear constrained optimization: Application to alanine dipeptide and the Îš hairpin of protein g. *The Journal of Chemical Physics*, 124(19):194903, 2006.

[128] Wolfgang Quapp and Dietmar Heidrich. Analysis of the concept of minimum energy path on the potential energy surface of chemically reacting systems. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 66:245–260, 1984.

[129] Michel Sana, Georges Reckinger, and Georges Leroy. An internal coordinate invariant reaction pathway. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 58:145–153, 1981.

[130] M. V. Basilevsky. Modern development of the reaction coordinate concept. *Journal of Molecular Structure: THEOCHEM*, 103:139 – 152, 1983.

[131] Themis Lazaridis, Douglas J. Tobias, Charles L. Brooks, and Michael E. Paulaitis. Reaction paths and free energy profiles for conformational transitions: An internal coordinate approach. 95(10):7612–7625, 1991.

[132] Michael Hirsch and Wolfgang Quapp. Reaction pathways and convexity of the potential energy surface: Application of newton trajectories. *Journal of Mathematical Chemistry*, 36:307–340, 2004.

[133] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in Fortran 77: The Art of Scientific Computing.*, volume 1. Cambridge University Press: Port Chester, NY, 2001.

[134] A.E. Cho, J.D. Doll, and D.L. Freeman. The construction of double-ended classical trajectories. *Chemical Physics Letters*, 229(3):218 – 224, 1994.

[135] Baron Peters, Andreas Heyden, Alexis T. Bell, and Arup Chakraborty. A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. *The Journal of Chemical Physics*, 120(17):7877–7886, 2004.

[136] Carl Eckart. Some studies concerning rotating axes and polyatomic molecules. *Phys. Rev.*, 47:552–558, 1935.

[137] Konstantin N. Kudin and Anatoly Y. Dymarsky. Eckart axis conditions and the minimization of the root-mean-square deviation: Two closely related problems. *The Journal of Chemical Physics*, 122(22):224105, 2005.

[138] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976.

[139] Daniel Trzesniak, Anna-Pitschna E. Kunz, and Wilfred F. van Gunsteren. A comparison of methods to compute the potential of mean force. *ChemPhysChem*, 8(1):162–169, 2007.

[140] Hsiang-Ai Yu, Benoit Roux, and Martin Karplus. Solvation thermodynamics: An approach from analytic temperature derivatives. *The Journal of Chemical Physics*, 92(8):5020–5033, 1990.

[141] Pekka Mark and Lennart Nilsson. Structure and dynamics of liquid water with different long-range interaction truncation and temperature control methods in molecular dynamics simulations. *Journal of Computational Chemistry*, 23(13):1211–1219, 2002.

[142] Pekka Mark and Lennart Nilsson. Structure and dynamics of the tip3p, spc, and spc/e water models at 298 k. *The Journal of Physical Chemistry A*, 105(43):9954–9960, 2001.

[143] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.

[144] Guillaume Lamoureux, Alexander D. MacKerell, and Benoît Roux. A simple polarizable model of water based on classical drude oscillators. *The Journal of Chemical Physics*, 119(10):5185–5197, 2003.

[145] Kazuya Ishimura, Peter Pulay, and Shigeru Nagase. A new parallel algorithm of mp2 energy calculations. *Journal of Computational Chemistry*, 27(4):407–413, 2006.

[146] Axel D. Becke and Erin R. Johnson. A density-functional model of the dispersion interaction. *The Journal of Chemical Physics*, 123(15):154101, 2005.

[147] Axel D. Becke and Erin R. Johnson. A unified density-functional treatment of dynamical, nondynamical, and dispersion correlations. *The Journal of Chemical Physics*, 127(12):124108, 2007.

[148] Erin R. Johnson and Axel D. Becke. A unified density-functional treatment of dynamical, nondynamical, and dispersion correlations. ii. thermochemical and kinetic benchmarks. *The Journal of Chemical Physics*, 128(12):124105, 2008.

[149] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. J. A. Montgomery, T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara,

K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. *Gaussian 03, Revision B.1.* Gaussian, Inc., Pittsburgh PA, 2003.

[150] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327 – 341, 1977.

[151] Douglas J. Tobias and Charles L. Brooks. Molecular dynamics with internal coordinate constraints. *The Journal of Chemical Physics*, 89(8):5115–5127, 1988.

[152] William G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31(3):1695–1697, 1985.

[153] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, 81(1):511–519, 1984.

[154] Hans C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393, 1980.

[155] Scott E. Feller, Yuhong Zhang, Richard W. Pastor, and Bernard R. Brooks. Constant pressure molecular dynamics simulation: The langevin piston method. *The Journal of Chemical Physics*, 103(11):4613–4621, 1995.

[156] Ulrich H.E. and Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281(1-3):140 – 150, 1997.

[157] Robert H. Swendsen and Jian-Sheng Wang. Replica monte carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57:2607–2609, 1986.

[158] Adrian E Roitberg, Asim Okur, and Carlos Simmerling. Coupling of replica exchange simulations to a non-boltzmann structure reservoir. *J Phys Chem B*, 111(10):2415–2418, 2007.

[159] George S. Hammond. A correlation of reaction rates. *Journal of the American Chemical Society*, 77(2):334–338, 1955.

[160] Thomas Steinbrecher, David A. Case, and Andreas Labahn. A multistep approach to structure-based drug design:âĂĽ studying ligand binding at the human neutrophil elastaseâĂă. *Journal of Medicinal Chemistry*, 49(6):1837–1844, 2006.

[161] Thomas Steinbrecher, Andrea Hrenn, Korinna L. Dormann, Irmgard Merfort, and Andreas Labahn. Bornyl (3,4,5-trihydroxy)-cinnamate - an optimized human neutrophil elastase inhibitor designed by free energy calculations. *Bioorganic and Medicinal Chemistry*, 16(5):2385 – 2390, 2008.

[162] Ignacio J. General. A note on the standard state's binding free energy. *Journal of Chemical Theory and Computation*, 6(8):2520–2524, 2010.

[163] Y. Deng and B. Roux. Calculation of standard binding free energies: Aromatic molecules in the t4 lysozyme l99a mutant. *Journal of Chemical Theory and Computation*, 2(5):1255–1273, 2006.

[164] Jan Hermans and Shankar Subramaniam. The Free Energy of Xenon Binding to Myoglobin from Molecular Dynamics Simulation. *Israel Journal of Chemistry*, 27, 1986.

[165] B. Roux, M. Nina, R. Pomes, and J.C. Smith. Thermodynamic stability of water molecules in the bacteriorhodopsin proton channel: a molecular dynamics free energy perturbation study. *Biophysical Journal*, 71(2):670 – 681, 1996.

[166] Sarvin Moghaddam, Yoshihisa Inoue, and Michael K. Gilson. HostâĹŠguest complexes with proteinâĹŠligand-like affinities: Computational analysis and design. *Journal of the American Chemical Society*, 131(11):4012–4021, 2009.

[167] D.A. Case, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, M. Crowley, R.C. Walker, W. Zhang, K.M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvary, K.F. Wong, F. Paesani, J. Vanicek, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, and P.A. Kollman. *AMBER10*. University of California, San Francisco, 2008.

[168] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984.