A SAMPLING OF MOLECULAR DYNAMICS

By

DANIEL JON SINDHIKARA

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2009

To my family

ACKNOWLEDGMENTS

Upon completion of my studies at University of Florida, I primarily thank my advisor, Professor Adrian Roitberg. Throughout the years I have learned a tremendous amount from him via our scientific conversations, and especially our lively arguments. He has supported me scientifically, academically, and personally more than could ever be asked. I also thank "my second advisor", Professor Kenneth Merz Jr. for exceptional scientific and professional guidance. I appreciate participation by all my committee members including Professors Stephen Hagen, Aneta Petkova, and James Dufty.

I am also grateful for significant assistance on theoretical topics from my colleagues, especially: Gustavo Seabra, Yilin Meng, Andrew Taube, Josh McClellan, and Prakash Verma. I appreciate all my friends and colleagues at the Quantum Theory Project, and the Departments of Physics and Chemistry.

I thank the person who inspired me to become a computational physicist, a professor from my undergraduate institution: Dr. Louis DeChiaro. Finally, I thank my parents. They have given me a tremendous amount of support and encouragement throughout graduate school and everything else.

I would like to acknowledge sources of funding by the Japan Society for the Promotion of Science (JSPS), the National Science Foundation (NSF) Eastern Asian and Pacific Summer Institute (EAPSI) program, National Institutes of Health grant # GM 066689. Also I have used more than my share of computational resources from the University of Florida High Performance Computing center (UFHPC), and NSF Large Allocations Resource Committee through Grant Nos. TG-MCA05S010 and UT-NTNL0002.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

PPII	Poly-proline II helix
NMR	Nuclear magnetic resonance
FRET	Fluorescence resonance energy transfer
CD	Circular dichroism
CCSDT	Coupled-cluster with singles, doubles, and triples
UA	United atom force field
AMBER	Assisted model building and energy refinement, a simulation package for biomolecules, also a model for an all-atom representation of a molecule
MC	Monte Carlo sampling
MD	Molecular Dynamics
NVE	Constant number of particles, volume, and energy (the microcanonical ensemble)
NVT	Constant number of particles, volume, and temperature
SPC	Single point charge solvent model
TIP3P	Transferrable intermolecular potential three position model (for water)
TIP4P	Transferrable intermolecular potential four position model (for water)
GB	Generalized-Born solvent model
PB	Poisson-Boltzmann
TI	Thermodynamic integration
DOS	Density of states
WHAM	Weighted histogram analysis method
T-WHAM	WHAM for temperature biased ensembles
MUCA-WHAM	WHAM for multicanonical algorithm biased ensembles
REMD	Replica exchange molecular dynamics
PT	Parallel tempering

MUCAREM	Multicanonical algorithm replica exchange method
EAF	Exchange attempt frequency
RMSD	Root mean squared deviation
NEM	Normalized ergodic measure
SANDER	Simulated annealing with NMR-derived energy restraints, an MD engine in the AMBER package
PRNG	Pseudo random number generator
T-REMD	Temperature based replica exchange molecular dynamics, same as parallel tempering
SS	Same seed, simulations utilizing the same pseudorandom number seeds
DS	Different seed, simulations utilizing different pseudorandom number seeds
HFP	Highest fractional population
NikR	A class of nickel regulatory proteins
AS	Apo-straight form of NikR
AT	Apo-trans form of NikR
AC	Apo-cis form of NikR
HS	Holo-straight form of NikR
HT	Holo-trans form of NikR
HC	Holo-cis form of NikR
MBD	Metal binding domain
DBD	DNA binding domain
COM	Center of mass

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

A SAMPLING OF MOLECULAR DYNAMICS

By

Daniel J. Sindhikara

May 2009

Chair: Adrian Roitberg
Major: Physics

The sheer vastness of the number of computations required to simulate a biological molecule puts incredible pressure on algorithms to be efficient while maintaining sufficient accuracy. This dissertation summarizes various projects whose purposes address the large span of types of problems in molecular dynamics simulations of biological systems including: increasing efficiency, measuring convergence, avoiding pitfalls, and an application and analysis of a biological system.

Chapters 3 and 4 deal with an enhanced sampling algorithm called “replica exchange molecular dynamics” which is designed to speed-up molecular dynamics simulations. The optimization of a key parameter of these simulations is analyzed. In these successive projects, it was found conclusively that maximizing “exchange attempt frequency” is the most efficient way to run a replica exchange molecular dynamics simulation.

Chapter 5 describes an enhanced metric for convergence in parallel simulations called the normalized ergodic measure. The metric is applied to several properties for several replica exchange simulations. Advantages of this metric over other methods are described.

Chapter 6 describes the implementation and optimization of an enhanced sampling algorithm similar to replica exchange molecular dynamics called multicanonical algorithm

replica exchange molecular dynamics. The algorithm was implemented into a biomolecular simulation suite called AMBER. Additionally several parameters were analyzed and optimized.

In Chapter 7, a pitfall in molecular dynamics is observed in biological systems that is caused by negligent use of a simulation's "thermostat". It was found that if the same pseudorandom number seed were used for multiple systems, they eventually synchronize. In this project, synchronization was observed in biological molecules. Various negative effects including corruption of data are pointed out.

Chapter 8 describes molecular dynamics simulation of NikR, a homotetrameric nickel regulatory protein whose binding to free Ni^{++} increases its binding affinity for a nickel transporter gene. Three forms of the *Pyrococcus Horikoshii* species of NikR were simulated including two apo-forms and one nickel-bound form. A quantum-mechanics-based force field parameterization was required to accurately represent the four nickel-centers in the holo-form. Extensive analysis of the three 100-ns-long trajectories was performed.

CHAPTER 1
INTRODUCTION

"…it all works because Avogadro's number is closer to infinity than to 10."

R. Baierlein, Am. J. Phys., Vol 46, No. 10, October 1978.

1.1 Biological Macromolecules

Biological macromolecules such as proteins and nucleic acids are the primary components in cellular machinery. Knowledge of the structure, dynamics, and functions of such molecules can greatly enhance the understanding of living creatures great and small. This understanding leads to power over natural phenomena that we formerly had no control over: disease, aging, pain, etc. Although many experiments exist that can be used to determine the function of biological molecules, functional analysis alone cannot suggest physical or chemical behavior inherent in the molecule. Thus it is advantageous to study the structure and dynamics of these molecules in order to gain better insight into their biological function.

1.2 Protein Structure

Proteins are biological macromolecules comprised of a string of amino acid (AA) "residues". An amino acid is comprised of a central carbon atom (C_α) bound to a carboxyl group (COOH) an amine group (NH_2) a hydrogen (H) and one of 20 unique "sidechains" (R). The amino acids are connected in a chain via peptide bonds. The sequence of amino acids in a protein is called the "primary structure" which uniquely determines its behavior in its environment. Figures 1-1A,B show an example of both a zwitterionic amino acid, alanine, and a peptide , $(Ala)_3$. For alanine, the unique sidechain is a methyl group (CH_3).

Three types of atoms are repeated along the chain: amide nitrogen, alpha carbon, and carbonyl carbon. The rotation about the bonds between these "backbone" atoms are called the ϕ, ψ, and ω angles respectively (see Figure 1-1B). At physiological temperatures, both ϕ and ψ are

flexible. ω, which represents the rotation about the peptide bond, is very stiff and tends to stay near 180°, or in rarer cases near 0°. For most proteins, there are three likely secondary structures the right-handed alpha helix, the beta sheet, and the polyproline II region (PPII); each of these structures correspond to a series of residues with similar sets of dihedral angles. The secondary structure can be illustrated in a Ramachandran plot, which is essentially a two-dimensional histogram in backbone dihedral angle (ϕ/ψ) space. In the plot, the most likely structures are easily recognized: the right-handed alpha helix near (ϕ=-57, ψ=-47), the beta-sheet near (ϕ=-125, ψ=150) and the PPII near (ϕ=-75, ψ=+145). Figure 1-2 shows an example of a Ramachandran plot. In this example, the highest probability is in the PPII region (the dark red area).

Furthermore, the tertiary structure of a protein is the relative arrangement secondary structure segments. For example, several alpha-helices bound in parallel form a bundle. The tertiary structure is usually referred to by the "type of fold" the protein takes. Lastly, the quaternary structure is the arrangement of multiple protein chains.

1.3 Simulation vs. Experiment

Biological experiments are fundamental in determining functions of biological molecules. While experiments can acquire accurate and essential structural data, there are limitations. Most experimental methods such as nuclear magnetic resonance (NMR) or fluorescence resonance energy transfer (FRET) can only resolve to a timescale of picoseconds to nanoseconds and are typically performed on several molecules at once. More detailed structural experiments such as x-ray crystallography or neutron scattering are much slower yet. Since conformational changes (changes in molecular structure) often occur at much shorter time scales than experiment, and experiments often observe many molecules simultaneously, only the average structure over this

timescale or these “ensembles” of molecules can be reported. For systems stable over the experimental timescale, the determined structure can be useful, other times it can be unphysical.

Additionally, many experiments can only report a subset of structural properties which act as restraints within which the structure must obey. FRET, for example, reports distances between parts of the molecule that have been dye-labeled. Circular dichroism (CD) reports ellipticity, which relates to a molecules secondary structure composition. NMR experiments gather restraints typically based on distances, and angles. While there is much information attainable via experiment, there is certainly room for exploration of biological structure in the realm of simulations that can often gather much better detail.

Simulations offer a representation of systems to whichever detail is considered necessary. Quantum ab-initio methods such as CCSDT[1] represent every atom of a molecule with accurate knowledge of the instantaneous electronic state. While a coarse-grained model, such as a UNITED-ATOM (UA) model, represents a molecule as a chain of subunits – each subunit consisting of several atoms (usually the hydrogen is subsumed into its adjacent atom). The caveat is that with increasing detail of the system comes increasing computational cost. Thus for large systems, electronic detail is intractable; for even larger systems, neither is atomic detail. For a certain size of system there is a limit, based on computational speed, on how much detail can be given (this will be elaborated further). The degree of accuracy required is usually not obvious, so the level of methodology is often difficult to choose

1.4 The All-Atom Representation of Proteins

The all-atom representation of proteins is between the quantum mechanical description and UA description in both accuracy and speed. In the all-atom method individual atoms interact with each other via both bonded and non-bonded energy terms. Equation 1-1 shows an example of an all-atom representation (the “AMBER Force Field”).

$$V(r^N) = \sum_{bonds} \frac{1}{2} k_b (l - l_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 +$$

$$\sum_{torsions} \frac{1}{2} V_N [1 + \cos(n\omega - \gamma)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \left\{ 4\varepsilon_{i,j} \left[\left(\frac{\sigma_{i,j}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{i,j}}{r_{ij}} \right) \right] + \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \right\} \qquad (1\text{-}1)$$

Equation 1-1 represents the potential energy of a protein (and its environment). Here the first two terms represent bond stretching and rotation as springs. The third term represents dihedral angle torsion. The double sum represents non-bonded interactions: Van der Waals and Coulomb interaction respectively. The potential energy for each molecular system will have a unique set of terms according to Equation 1-1 (called the topology). The parameters for a system can often be generated quickly since most proteins are simple combinations of the same twenty amino acids since those parameters can be easily looked up. However, the obtainment of parameters for an arbitrary system can be nontrivial. Current parameters are assumed to have some error and are constantly being analyzed and improved upon.

1.5 Obstacles in Simulation

While all-atom simulations offer great detail about the structure of biological molecules, there are two main reasons why the results from these simulations can be flawed. The most fundamental reason is if the representation of the energy is significantly inaccurate. The representation can be wrong if the parameter set for the force field is incorrect. For example, if the parameter set estimates a hydrogen atom to have a -1 charge, this would be incorrect, and of course, lead to qualitatively poor results. Another reason why the representation can be wrong is if there is a fundamental difference between the force field and the real system. For example, since the most chemical force fields do not allow covalent bond breaking or forming, it could never properly represent a system that undergoes a chemical reaction.

Supposing that the all-atom representation is correct, simulation results can also be flawed if there is incomplete sampling. This means that the simulation trajectory does not accurately reflect the correct population proportionality between relevant states. Solutions to the sampling problem lie in further understanding of the problem.

1.6 Free Energy and Ergodicity

The free energy difference between two states is directly related to the relative ensemble populations of those states as shown in Equation 1-2. In Equation 1-2, $\Delta G_{1\to 2}$ is the free energy difference between state 2 and 1, k_B represents Boltzmann's constant, T is the temperature, and P is the equilibrium probability of the state in the canonical ensemble.

$$\Delta G_{1\to 2} = -k_B T \ln\left(\frac{P_2}{P_1}\right) \tag{1-2}$$

According to the ergodic hypothesis, these populations can be either obtained by fully sampling phase space or time as shown in Equation 1-3.

$$P(x) = \frac{\int \rho(x) dV}{\int dV} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \delta_{x,x(t)} \tag{1-3}$$

Here $\rho(x)$ is the density of x in phase space and $\delta_{x,x(t)}$ is the Kronecker-delta function.

These populations can be either the state populations among many particles, or the state probability of a single particle at any time. Thus, correct populations can be attained by fully sampling either phase space or time. In addition to relative populations, correct expectation values can be obtained via complete sampling of either phase space or time according to Equation 1-4.

$$\langle A \rangle = \frac{\int A(x)\rho(x) dV}{\int \rho(x) dV} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} A(x_t) \tag{1-4}$$

For canonical ensembles, the un-normalized phase space density of x is $\rho(x) = e^{-E(x)/k_B T}$ where the exponential term is deemed the Boltzmann factor. A mapping of free energy to some number of reaction coordinates is sufficient to describe the nature of the system with respect to those coordinates, and is called a free-energy landscape. An accurate landscape is the ultimate goal in simulation.

1.7 Monte Carlo Sampling

The duality of phase space and time lends itself to two approaches to sampling: Monte Carlo (MC) and Molecular Dynamics (MD). In MC, the phase space sampling approach is utilized. A series of configurational snapshots are collected after extensive propagation via attempting configurational changes based upon the transition probability between the before and after states. The transition probability can be derived from the probability factor and the detailed balance condition. Assuming that the populations of any two states are in equilibrium (a dangerous assumption) requires that Equation 1-5 be true.

$$P(a)\Pi(a \to b) = P(b)\Pi(b \to a) \tag{1-5}$$

Here P represents the population, and Π represents the transition probability between two states. Solving this equation for the transition probability yields Equation 1-6.

$$\frac{\Pi(a \to b)}{\Pi(b \to a)} = \frac{P(b)}{P(a)} \tag{1-6}$$

In the case of canonical sampling, the Boltzmann factor dictates P and one sets the probability to transition back, $\Pi(b \to a)$, to 1. This leads to Equation 1-6.

$$\Pi(a \to b) = e^{-(E(b)-E(a))/k_B T} \tag{1-7}$$

Here, the energy, E, is obtained by the potential function of choice. The transitions in MC are implemented by comparing a random number, υ, between zero and one to Π. If υ is less than or equal to Π, the transition (MC move) is accepted, otherwise the current state is added to the

known to the ensemble, and another MC move is attempted. This is known as the Metropolis criterion. These attempts continue until one is satisfied that phase space has been adequately sampled. There is no formal requirement for the nature of any MC moves other than that the reverse must be possible. Typically a simulator chooses a move type to most efficiently sample phase space for the particular system of interest. An advantage of MC over MD (as will be elaborated later) is that the MC moves may be chosen so that they allow system to move between two states regardless of any energy barriers between them. A particular drawback to the MC approach is the fact that sequential moves are not chronologically linked and thus MC simulations yield no temporal information.

1.8 Molecular Dynamics

The molecular dynamics (MD) approach to simulation utilizes integration over time rather than phase space to determine free-energy landscapes. Rather than sampling space by performing random MC moves, in MD, systems are propagated primarily by Newton's second law.

The force on any particle is determined by the negative gradient of the potential and thus the equation of motion is established and solved numerically (details of this procedure are presented in the methods Section 2.1). Successive moves (aka snapshots) describe a movie-like representation of the system. After sufficient propagation, accurate probabilities can be extracted by integrating over the snapshots according to Equation 1-3.

A particular drawback to the MD approach is sometimes called, "kinetic trapping". That is, if the system is inside a deep free-energy "valley", successive integration will redundantly sample the same space. This is a great practical problem since the calculation of these redundant steps can make the computational cost of a "converged" simulation intractable.

1.9 Protein Folding and the Levinthal Paradox

The free energy landscapes of large molecules such as proteins are complex. There are vast degrees of freedom and a large amount of possible configurations. Yet there is often a unique configuration (or small set of configurations), called the "folded" state that allows a protein to perform its biological function. An average protein consists of on the order of 100 amino acid residues. Each residue can sample its own dihedral angle space in which there are commonly three major states (minima). Thus, for the entire protein there should be approximately 3^{100} or around 10^{50} conformational states available. If the protein were to sequentially sample each conformation at a rate similar to the experimental transition rate (about 1 picosecond per transition), it would take about 10^{38} seconds to completely sample configurational space. To put this in perspective, the age of the universe is approximately 10^{17} seconds. Experimental folding rates on the order of seconds or less and the fact that we are alive contradicts this approximation. Cyrus Levinthal originally introduced this contradiction in 1968 in an attempt to explain that nature, rather than randomly searching configurations, seeks "folding pathways" to find the biologically necessary folded state.[2] That is, proteins, starting from an unfolded configuration, gradually move in the general direction of the folded state due to the local gradient of the potential energy surface. To describe the landscape of a protein, however, either MC or MD simulations must sample an equivalent amount configurational space that nature does within the time scale of folding.

1.10 Enhanced Sampling

The fact that so many energy or force calculations are required for simulation of large systems gives impetus to be as efficient with them as possible. MC or MD simulations can be performed in a manner that circumvents the typical problems of simulation in the canonical ensemble by altering the landscape, then compensating by adjusting probabilities.

Conformational sampling is an essential concern to the study of complex molecular systems such as proteins. A major obstacle for the correct sampling of such systems is the fact that the potential energy surfaces of proteins are very rugged and contain a large number of local energy minima.[3] This feature of complex systems causes kinetic trapping due to low barrier crossing rates in constant temperature molecular dynamics. Many different techniques have been introduced to deal with this problem. Itoh et al. wrote an excellent review on the subject of enhanced sampling.[4] In order to overcome kinetic trapping, generalized-ensemble algorithms including multicanonical algorithm[5], simulated tempering[6,7] and parallel-tempering methods[8-12] are often used. These methods allow the system to perform a random walk in temperature or energy space which allows the system under study to more easily overcome energy barriers and hence reduces the problem of kinetic trapping.

1.11 Algorithmic Parallelization

One of the most significant constraints for simulations is computational cost. Simulators are constantly pushing the envelope for getting as much useful scientific data as is possible on modern computers within a reasonable time scale (on the order of months). For the burgeoning years of simulation, scientists could ride the wave of Moore's law, which states that the transistor count on a CPU doubles every two years.[13] However, due to the inherent increase in power consumption and cooling issues, it is becoming more and more advantageous to utilize *more* processors rather than *faster* ones. As of February 2009, it is quite reasonable for a scientist to access 50 2.8 GHz machines simultaneously, and quite impossible to access a single 140 GHz machine.

Thus there is a strong impetus to design algorithms amenable to parallel computing. However, there are many difficulties. Parallel programs rely on the divisibility of the

computation. Thus a programmer must design programs that can be run separately and independently in order to maximize effectiveness in the current technological environment.

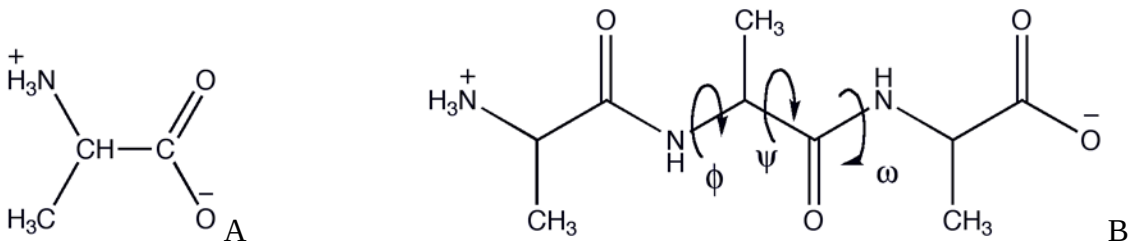


Figure 1-1A,B. Forms of the amino acid Alanine. A) Zwitterionic form. B) Polypeptide form $(Ala)_3$. Backbone dihedral angles, ϕ, ψ, and ω, which determine protein structure, are shown.

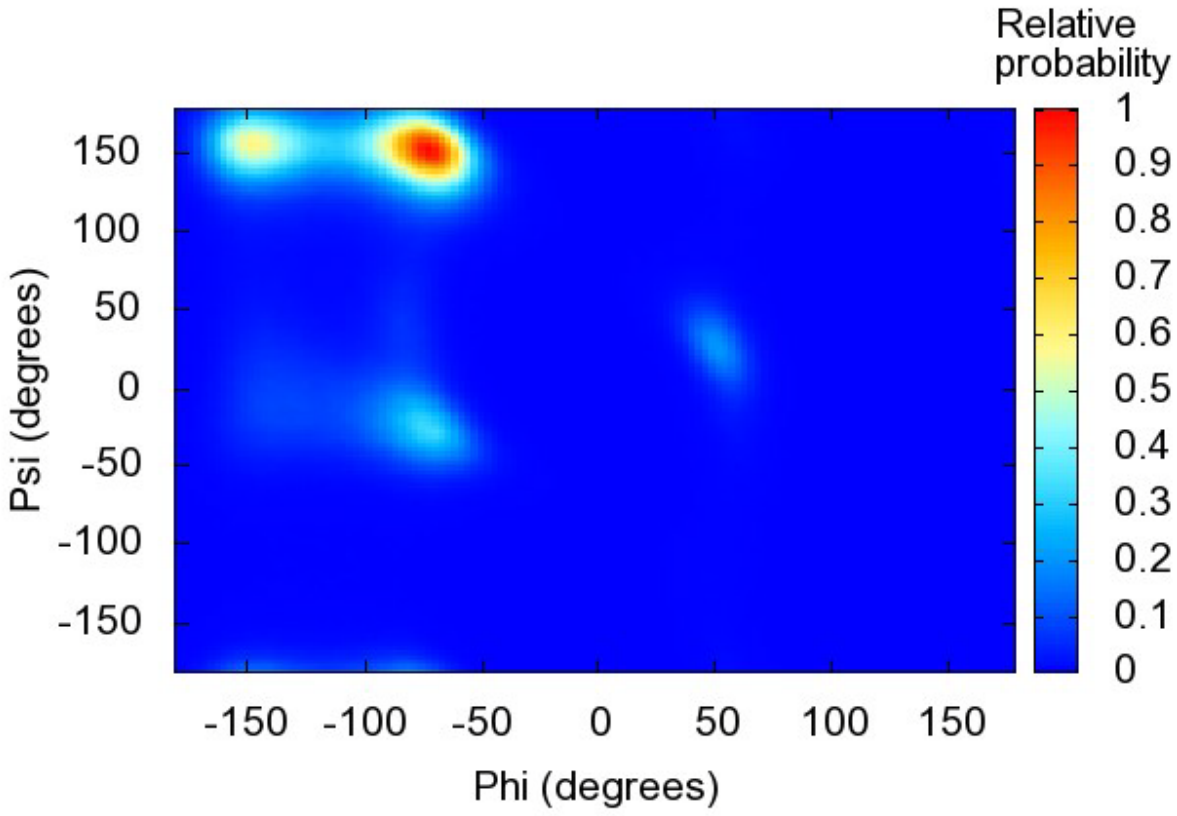


Figure 1-2. An example Ramachandran plot. Backbone dihedral angle histogram data taken from a 50 ns simulation of $(Ala)_5$.

CHAPTER 2
METHODS

"…splintered dreams of unity (our lives are parallel),
so far from reality (our lives are parallel),
independent trajectories (our lives are parallel),
separate terms of equality (our lives are parallel)…"

G. Graffin. "Parallel" Liner notes. The Gray Race:
Music of Bad Religion, Atlantic, 1996

2.1 MD integration

As mentioned previously in this dissertation, in MD, a system is propagated essentially via Newton's 2nd law. Unfortunately, the complexity of the potential function disallows analytical integration. Most commonly, the velocity Verlet algorithm is implemented.[14] Here, we integrate both position and time via a truncated Taylor series expansion (Equations 2-1, 2-2, 2-3, and 2-4).

$$\vec{r}_i(t+\delta t) = \vec{r}_i(t) + \dot{\vec{r}}_i \delta t + \frac{1}{2}\ddot{\vec{r}}_i \delta t^2 \tag{2-1}$$

$$\dot{\vec{r}}_i(t+\delta t/2) = \dot{\vec{r}}_i(t) + \frac{1}{2}\ddot{\vec{r}}_i(t)\delta t \tag{2-2}$$

$$\ddot{\vec{r}}_i(t+\delta t) = -\frac{\nabla V(\vec{r}_i(t))}{m_i} \tag{2-3}$$

$$\dot{\vec{r}}_i(t+\delta t) = \dot{\vec{r}}_i(t+\delta t/2) + \frac{1}{2}\ddot{\vec{r}}_i(t+\delta t)\delta t \tag{2-4}$$

Here $\vec{r}_i$ is the position of the i^{th} atom, $\vec{\nabla} V(\vec{r}_i)$ is the gradient of the potential energy of the i^{th} atom, and δt is the time step. Optimization of the time step is critical to ensure both fast and accurate computation. A time-step that is too small will cause a proportional increase in computational cost (due to more calculations per unit time), a time-step that is too large will result in unphysical results (two atoms that are attracted may find themselves inside each other

before they have a chance to repel). To test if the time-step is too large, one can perform a simulation in the microcanonical (NVE) ensemble and check that the total energy stays constant. For most simulations, we use a time step of 2 fs since it is smaller than the period of the fastest vibrational mode. The smallest step would be shorter than 2 fs if we did not freeze hydrogen bond vibrations via the SHAKE algorithm.[15]

As mentioned in the introduction, we describe our molecular systems via Equation 1-1. Simple integration of this potential would yield a simulation in the microcanonical ensemble (NVE). Our systems of interest behave more like those in the canonical ensemble (NVT) and thus require additional treatment to keep constant temperature. Algorithms that maintain system temperature are thusly called thermostats. Various thermostats occur with certain advantages and disadvantages. Our simulations utilize the Langevin thermostat.[16]

Langevin dynamics modifies the definition of the force on an individual atom (Equation 2-5).

$$\vec{F}(\vec{r}_i) = -\vec{\nabla}V(\vec{r}_i) - \gamma m_i \dot{\vec{r}}_i + \vec{A}(\gamma, T, \upsilon) \tag{2-5}$$

Here m_i, $\ddot{\vec{r}}_i$, $\dot{\vec{r}}_i$, and $\vec{r}_i$ are the mass, acceleration, velocity, and position of the i^{th} particle respectively. $V(\vec{r}_i)$ is the potential energy determined by the force field. This is essentially Newton's second law with two extra terms: a solvent drag force represented by $\gamma m_i \dot{\vec{r}}_i$ and a random force, $\vec{A}$, which obeys the fluctuation-dissipation theorem (Equation 2-6).

$$\left\langle A_i(t)A_j(t+\Delta t)\right\rangle = \int 2m\gamma k_b T\delta(\Delta t)\delta_{ij}d(\Delta t) \tag{2-6}$$

Here the average, $\langle\ \rangle$, is over time, k_b is the Boltzmann constant, and $\delta(\Delta t)$ and δ_{ij} represent the Dirac and Kronecker delta functions respectively. The magnitude and direction of *A*

are based on a pseudorandom number, υ, and a probability distribution based on the temperature and heat bath coupling strength, γ, also known as the collision frequency. The theorem essentially says that a system's response to a perturbing force is identical to a canonical equilibrium fluctuation. Chapter 7 will describe a peculiar effect of using Langevin dynamics.

2.2 Solvent Models

2.2.1 Solvent Models Introduction

Since the biological environment of proteins is generally aqueous (for non membrane proteins), it is advantageous to find accurate and efficient representation of water. Several water models exist, but can generally be classified into two main categories: explicit and implicit. Explicit water models utilize countable extensive water molecules. Typically these models are divided into subclasses based on the number of points required to represent each molecule. For example the SPC, TIP3P, and TIP4P utilize one, three, and four points respectively for each molecule.[17] Each additional point representation makes the single point energy calculation longer, and thus explicit solvent models are generally more computationally expensive than implicit. Section 2.2.2 will describe the TIP3P model in more detail.

Implicit solvent models treat the water as a variable dielectric with which a choice of electrostatics based energy equations can be solved. Two popular classes of implicit solvent models are Generalized Born (GB) and Poisson-Boltzmann (PB). Section 2.2.3 will describe the GB model in more detail.

2.2.2 TIP3P water

The TIP3P solvent model uses three points per water molecule, one at each atomic center of each atom in water. The potential energy for each molecule is calculated as a sum of bonded and non-bonded terms exactly as in the AMBER force field (Equation 1-1). Since only there are only three points per molecule, the only bonded terms allowed are two bond stretching and one

angle bending terms. Typically hydrogen-stretching degrees of freedom are effectively frozen. Figure 2-1 shows a representation of a TIP3P water (though not all parameters are included in the figure).

2.2.3 Generalized-Born Implicit Solvent Model

Due to the inherent added cost of including explicit solvent models, it is computationally advantageous, where appropriate, to use implicit solvent. The generalized-Born equation is a widely-used representation for solvent effects on electrostatic interactions as well as solvation free-energy (see Equation 2-7).[18]

$$\Delta G_{elec} = -\left(1-\frac{1}{\varepsilon}\right)\sum_{i=1}^{N}\sum_{j=i+1}^{N}\frac{q_i q_j}{r_{ij}} - \frac{1}{2}\left(1-\frac{1}{\varepsilon}\right)\sum_{i=1}^{N}\frac{q_i^2}{a_i}$$

(2-7)

Here, ε is the dielectric constant (which can vary), q_i is the partial charge of the i^{th} atom, r_{ij} is the distance between the i^{th} and j^{th} atom, and a_i is called the "Born radius". The Born radius of an atom is effectively the radius that would give the correct electrostatic energy if all other particles were uncharged. Though I will not describe it here, it is not trivial to derive the correct dielectric constants, Born radii, or even partial charges in this model.

Equation 2-7 replaces the electrostatic portion of the AMBER force field. It is important to note that this representation is a free-energy, and not a potential energy. This means that the parameters fit for the solvation free energy term are only useful at the temperature for which the parameters were fit. Essentially the solvent entropy is implicit in this model.

2.3 Thermodynamic Integration

Thermodynamic Integration (TI) is a simple method to calculate free energy differences between two states for reasonably small systems. In this method, a reaction coordinate, λ is defined such that at $\lambda = 0$ a system is in its reactant state, and at $\lambda = 1$, its product. The TI

equation (Equation 2-8) is solved by simulating a complete set of simulations at consecutive λ values (λ_i, λ_{i+1}) separated by some decided amount (e.g λ=0, λ=0.1, λ=0.2,…).

$$\Delta G = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial E}{\partial \lambda} \right\rangle_{\lambda} d\lambda \tag{2-8}$$

Each set of simulations at neighboring λ's must be run until $\left\langle \frac{\partial E}{\partial \lambda} \right\rangle_{\lambda}$ converges. Once all simulations are complete, the sum of the $\left\langle \frac{\partial E}{\partial \lambda} \right\rangle_{\lambda}$'s is compiled to find the total change in free energy between states.

TI is an excellent method for free-energy calculation for many small systems, but can be computationally cumbersome when a reaction coordinate cannot be well defined. TI was used in Chapter 8 to determine simulation parameters to match experimental solvation free energies.

2.4 Biased Ensembles and the Weighted Histogram Analysis Method

Often it is advantageous to simulate a system in an alternate or array of alternate ensembles in order to enhance sampling. For example if a system of interest has two stable states separated by a high-energy transition state, kinetic trapping is likely to significantly hinder computational efficiency. However, the landscape can be modified in many ways to increase the sampling in the high-energy regions and thus reduce the trapping. One way to do this is with umbrella sampling.[19] If the position of a barrier is known to some reasonable degree of accuracy an additional potential can be added in the form of a harmonic potential (Equation 2-9) where V_U is the original, unbiased potential, and V_B is the additional biasing potential (Equation 2-10).

$$V = V_U + V_B \tag{2-9}$$

$$V_B = 1/2k\left(x - x_{TS}\right)^2 \tag{2-10}$$

This additional term will encourage sampling near the transition state. Figure 2-2 shows the various terms in a biased ensemble. Typically umbrella sampling simulations require an array of umbrella potentials to amplify sampling throughout reaction-coordinate space. Fortunately these different umbrella potentials, or any alternate landscape are trivially parallelizable – each simulation can be run independently. A plethora of enhanced sampling techniques exist in addition to umbrella sampling, only a few of which will be detailed later in the methods section. Data acquired through enhanced sampling is biased and thus has to be reweighted in order to obtain the correct (unbiased) ensemble.

When more than one ensemble is utilized, the system's ensemble-independent density of states (DOS) should be self-consistent within the array of ensembles. There are numerical methods for solving the self-consistent DOS including the weighted histogram analysis method (WHAM)[20,21], Wang-Landau method[22,23], and the multistate Bennett acceptance ratio (MBAR).[24] Below, I will describe the most popular of these methods, WHAM.[20,21]

The probability of any observable, q_0, in ensemble *i,* is usually written as a normalized integral over phase space (Equation 2-11).

$$P_i(q_0) = \frac{\int \delta\big(q(\tau) - q_0\big)\rho_i(\tau)d\tau}{Z_i} \tag{2-11}$$

Here, τ is the phase-space coordinate (which can be a configuration of the system), ρ_i is the unnormalized probability distribution function, and Z_i is the partition function of the i^{th} ensemble (defined by Equation 2-12).

$$Z_i = \int \rho_i(\tau)d\tau \tag{2-12}$$

Equations 2-11 and 2-12, however, can also be rewritten in terms of the density of q states in phase space, $\Omega(q_0)$ (Equations 2-13, 2-14, respectively).

$$P_i(q_0) = \frac{\Omega(q_0)\rho_i(q_0)}{Z_i} \tag{2-13}$$

$$Z_i = \int \Omega(q)\rho_i(q)dq \tag{2-14}$$

In an ergodic simulation a histogram count of coordinate q_0 of the i^{th} ensemble taken over all snapshots can be approximated as in Equation 2-15.

$$C_i(q_0) \approx P_i(q_0)N_i\Delta q \tag{2-15}$$

Here N_i is the total number of snapshots taken in the i^{th} ensemble and Δq is the bin width of coordinate q.

Combining Equations 2-13 and 2-15, one can solve for the DOS in terms of the histogram count (Equation 2-16).

$$\Omega(q_0) = \frac{C_i(q_0)Z_i}{\rho_i(q_0)N_i\Delta q} \tag{2-16}$$

When a simulation is complete, all elements on the right hand side of Equation 2-16 are known except for the complete set of Z_i's. If the form of ρ_i is known, Equations 2-14 and 2-16 (called the WHAM equations) can be solved iteratively. For example, if temperature biasing is used, the WHAM equations can be written in energy space as in Equations 2-17 and 2-18.

$$\Omega(E) = \frac{\sum_T H_i(E)}{\sum_T N_i \Delta E e^{(f_i - \beta_i E)}} \tag{2-17}$$

$$Z_i \equiv e^{-f_i} = \sum_E \Omega(E)e^{-\beta_i E} \tag{2-18}$$

Here $\beta_i = 1/k_B T_i$ and f_i is effectively the relative unitless free energy of ensemble i. Once the DOS and partition functions converge, these values can be used to calculate probabilities and expectation values of observables from the simulation as in Equations 2-19 and 2-20.

$$P(A_0) = \frac{\sum_E \Omega(E)\delta(A(E) - A_0)e^{-\beta_i E}}{\sum_E \Omega(E)e^{-\beta_i E}} \tag{2-19}$$

$$\langle A \rangle_i = \frac{\sum_E \Omega(E)A(E)e^{-\beta_i E}}{\sum_E \Omega(E)e^{-\beta_i E}} \tag{2-20}$$

$A(E)$ can be found by mapping A to E via looping through the trajectory.

For the purposes of analyzing enhanced sampling simulations, I wrote a C++ program approximately 1100 lines long to perform WHAM analysis and probability and expectation output for both temperature biasing (T-WHAM) and multicanonical biasing (MUCA-WHAM). This program is designed specifically to work with AMBER. Coding suggestions from a review of generalized ensemble algorithms by Okamoto[25] were used in the program.

2.5 Replica Exchange Molecular Dynamics

2.5.1 Background

REMD, the MD version of parallel tempering (PT)[10], is one of the more frequently used generalized-ensemble methods[26-32] aimed at alleviating the problem of kinetic trapping. During a REMD simulation, several non-interacting replicas of the original system are simulated independently and simultaneously at different temperatures using standard molecular dynamics methods. Periodically (after a given time interval in regular MD) an attempt is made to exchange conformations between two temperature-adjacent replicas based on a Metropolis criterion. Successful exchanges then move structures between lower and higher temperatures, passing replicas throughout the temperature spectrum and thus each replica contributes to each canonical ensemble. The replicas continuously heat up and cool down -- walking over energetic barriers. REMD has been proven to drastically increase rates of convergence towards a proper equilibrium distribution.[26-29] Recently, extensions to the original REMD algorithm[12] such as

using non-exponential temperature distribution[28,33] and attempting to exchange among all pairs of replicas[34] have been proposed to optimize its efficiency.

In order to efficiently accept the attempted exchanges between replicas, there must be enough overlap between potential energy distributions of neighboring temperatures.[35-38] A number of replicas are then needed to span the entire desired temperature range. For the conventional REMD algorithm, the number of replicas increases as ~O($f^{1/2}$), where f is the number of degrees of freedom of the system.[39] Unfortunately, this relation severely restricts the ability of the REMD algorithm to simulate large systems such as proteins in explicit solvent due to the very large number of processors and CPU time needed. Reducing the number of replicas but keeping the same sampling accuracy becomes an important issue. Recently, several new methods such as Hamiltonian REMD algorithm (H-REMD)[39-41], hybrid solvent model[42,43], reservoir REMD method (R-REMD),[44-46] replica exchange with dynamical scaling (REDS)[41] and the coupling of multicanonical algorithm with REMD (REMUCA and MUCAREM)[47] have been developed to improve on conventional REMD.

The design of replica exchange is especially amenable to parallelization. Each replica can easily be run on a separate processor. The only communication between replicas that is required is during the exchange attempts and requires only a few numbers to be passed.

The frequency of exchange attempts is an issue of current debate. I will address this issue in Chapters 3 and 4.

2.5.2 Implementation

REMD relies on the assumption that transition rates will increase at higher temperatures. Several simultaneous independent copies (replicas) of a molecule are simulated, each at one rung of a ladder of temperatures (usually exponentially distributed). The temperature range is

typically chosen such that the lowest temperature is below the temperature of interest, and the highest is above the folding temperature. *N* non-interacting copies (replicas) of an original system are simulated at *N* different temperatures (one each). Periodically an exchange of conformation between two adjacent temperatures is attempted. Suppose replica *i* at temperature T_n and replica *j* at temperature T_m are attempting to exchange; Equation 2-21 satisfies the detailed balance condition.

$$P_n(\vec{x}_i)P_m(\vec{x}_j)\Pi(\vec{x}_{i,n} \to \vec{x}_{j,n}; \vec{x}_{j.m} \to \vec{x}_{i,m}) = P_m(\vec{x}_i)P_n(\vec{x}_j)\Pi(\vec{x}_{i,m} \to \vec{x}_{j.m}; \vec{x}_{j,n} \to \vec{x}_{i,n}) \quad (2\text{-}21)$$

If the Metropolis criterion is applied, the exchange probability can be written as in Equation 2-22.[48]

$$\frac{\Pi(\vec{x}_{i,n} \to \vec{x}_{j,n}; \vec{x}_{j.m} \to \vec{x}_{i,m})}{\Pi(\vec{x}_{i,m} \to \vec{x}_{j.m}; \vec{x}_{j,n} \to \vec{x}_{i,n})} = e^{-(\Delta E \Delta \beta)} \quad (2\text{-}22)$$

Here Equations 2-23 and 2-24 hold true.

$$\Delta E = E(\vec{x}_j) - E(\vec{x}_i) \quad (2\text{-}23)$$

$$\Delta \beta = \beta_n - \beta_m \quad (2\text{-}24)$$

Here, velocity can be rescaled by a factor of $\sqrt{\frac{T_n}{T_m}}$ or the reciprocal for the opposite exchange to allow ΔE (the change in total energy) to be reduced to ΔU (the change in potential energy). This rescaling of velocities increases the exchange probability and thus allows for more efficient simulation. If the exchange between two replicas is accepted, the temperatures of two replicas will be swapped (see Figure 2-3). Upon completion of a simulation, data are either collected from a single temperature and compiled into an ensemble average or collected from all temperatures and analyzed using T-WHAM.[21,49] A detailed description of REMD algorithm can be found in the papers of Sugita and Okamoto.[12]

2.6 Multicanonical Algorithm Replica Exchange Method

Multicanonical algorithm replica exchange method (MUCAREM)[47] is an enhanced sampling algorithm similar to REMD. The advantage of MUCAREM over REMD is that it effectively reduces the minimum number of parallel simulations (number of replicas). Since the degree of overlap between potential energy distributions in neighboring temperatures in REMD, many replicas are necessary to span the temperature-space between the highest and lowest temperature. In MUCAREM, the potential energy distributions for each replica is widened, allowing less replicas to span temperature space. In a canonical simulation, the probability of being at energy, E, at inverse temperature, β, obeys Equation 2-25.

$$P(E) = \Omega(E)e^{-E\beta} \tag{2-25}$$

In MUCAREM, the potential energy is modified to flatten the probability distribution in energy space. Setting the right-hand-side of Equation 2-25 to one, the resultant "multicanonical energy" is written as in Equation 2-26.

$$E_{MU} = k_B T_0 \ln(\Omega(E)) \tag{2-26}$$

Here T_0 is any reference temperature. The force resulting from this modified energy related to the canonical force by a multiplicative factor as in Equation 2-27.

$$\vec{F}_{MU} = k_B T_0 \frac{\partial \ln(\Omega(E))}{\partial E} \vec{F} \tag{2-27}$$

Here $\vec{F}$ is the original force originating from the gradient of the potential. During a MUCAREM MD simulation, the typical (canonical) force is calculated then multiplied by the prefactor, $k_B T_0 \frac{\partial \ln(\Omega(E))}{\partial E}$, before being integrated.

Figures 2-4A,B show a comparison between potential energy distributions of the same system (HP1 segment of HP36[50]) using either REMD or MUCAREM. It is apparent that each mode in MUCAREM spans a much wider energy space.

If the DOS were known exactly, a single replica could be used to sample the entire span of the targeted energy subspace. However the DOS can only be obtained by running a converged simulation, which would defeat the purpose of enhanced sampling. Thus multiple replicas are assigned to individual, partially overlapping energy subspaces by defining incremental energy modes as in Expression 2-28. Here, $E_L^{\{m\}}$ and $E_H^{\{m\}}$ are the low and high limits of the potential energy that the multicanonical energy will be used for multicanonical mode *m*.

$$\varepsilon_{MU}^{\{m\}} = \begin{cases} \left.\dfrac{\partial E_{MU}(E:T_{m)}}{\partial E}\right|_{E=E_L^{\{m\}}} \times (E - E_L^{\{m\}}) + E_{MU}(E_L^{\{m\}}:T_m) & E < E_L^{\{m\}} \\ E_{MU}(E;T_m) & E_L^{\{m\}} < E < E_H^{\{m\}} \\ \left.\dfrac{\partial E_{MU}(E:T_{m)}}{\partial E}\right|_{E=E_H^{\{m\}}} \times (E - E_H^{\{m\}}) + E_{MU}(E_H^{\{m\}}:T_m) & E > E_H^{\{m\}} \end{cases} \qquad (2\text{-}28)$$

Thus each mode traverses a relatively flat probability landscape while between $E_L^{\{m\}}$ and $E_H^{\{m\}}$, effectively sampling a canonical landscape when past either extremum. The span of an individual mode's multicanonical space should reflect the relative confidence in the accuracy of the DOS for that region. A poor estimate in the DOS would result in enhanced kinetic trapping. Thus initial MUCAREM simulations utilize a span of zero (which is identical to an REMD simulation). A short REMD simulation can be analyzed via T-WHAM to get a guess at the DOS allowing a subsequent MUCAREM simulation. This process can be iterated as the quality of the DOS improves by subsequent MUCA-WHAM calculations. When the DOS is believed to be accurate enough, it can be used to predict probabilities and expectation (as described in the WHAM section of methods). Exchanges between adjacent multicanonical modes are almost

identical to that in REMD save for that the multicanonical mode energy, $\varepsilon_{MU}^{\{m\}}$, is used rather than the original potential energy.

As the DOS improves, the number of replicas can be reduced, and the potential energy span of each multicanonical mode can be increased in order to maximize sampling efficiency. The optimum number of replicas and span is not known and is a problem addressed in Chapter 6 of this work. Finally, expectation values of a property at a target temperature are calculated by reweighting the probabilities to the correct distribution as in Equation 2-29. Here, M is the number of multicanonical modes, n_m is the number of snapshots taken in that mode, and $\beta_T = 1/k_B T$.

$$\left\langle A \right\rangle_T = \frac{\sum_{m=1}^{M}\sum_{k=1}^{n_m} A(x_m(k)) \dfrac{e^{-\beta_T E(x_m(k))}}{\sum_{l=1}^{M} n_l e^{f_l + \beta_l \varepsilon_{mu}^{\{l\}}(E(x_m(k))}}}{\sum_{m=1}^{M}\sum_{k=1}^{n_m} \dfrac{e^{-\beta_T E(x_m(k))}}{\sum_{l=1}^{M} n_l e^{f_l + \beta_l \varepsilon_{mu}^{\{l\}}(E(x_m(k))}}} \tag{2-29}$$

Theoretically, MUCAREM can only improve on the efficiency of REMD. However the implementation requires much more coding along with a need for a strong interface with WHAM (or any other method to attain DOS). This issue is addressed in Chapter 6 of this work.

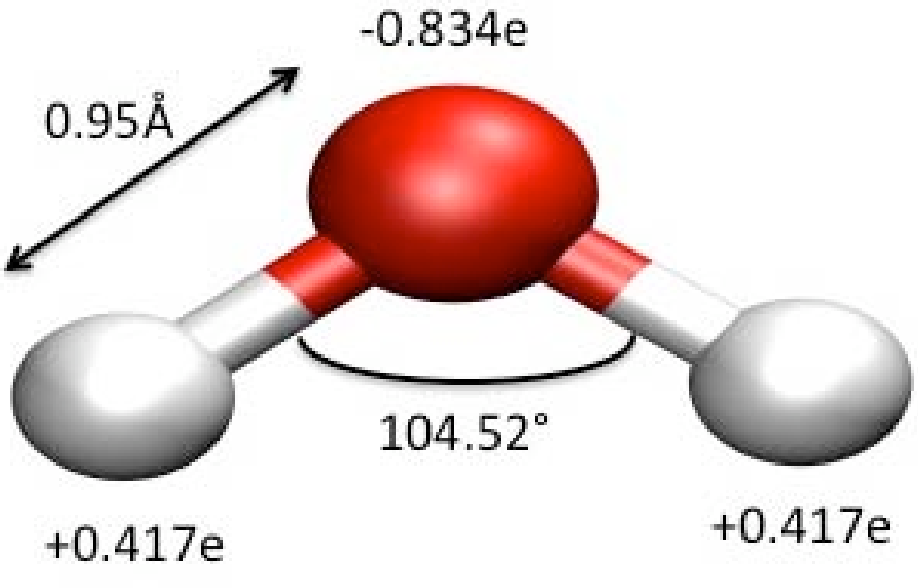


Figure 2-1. TIP3P water.

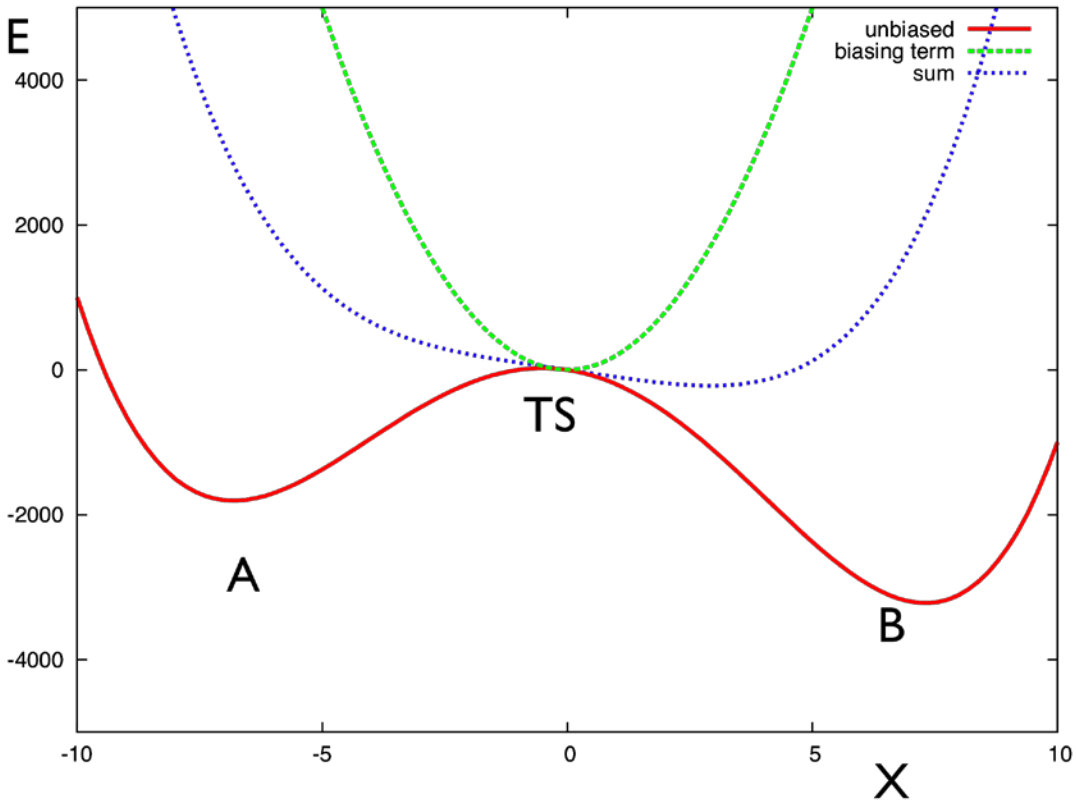


Figure 2-2. Alternate landscapes. The three curves pictured describe three potential energy functions. The red line shows the original function with an energy barrier. The green dotted line shows a biasing term. The blue line shows the sum of the biasing and original terms. Both x and y scales are arbitrary.

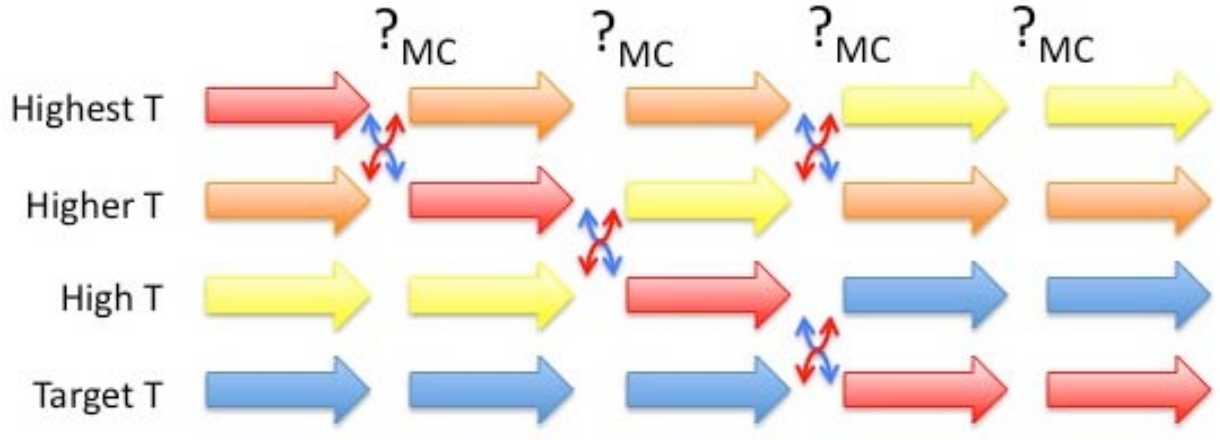
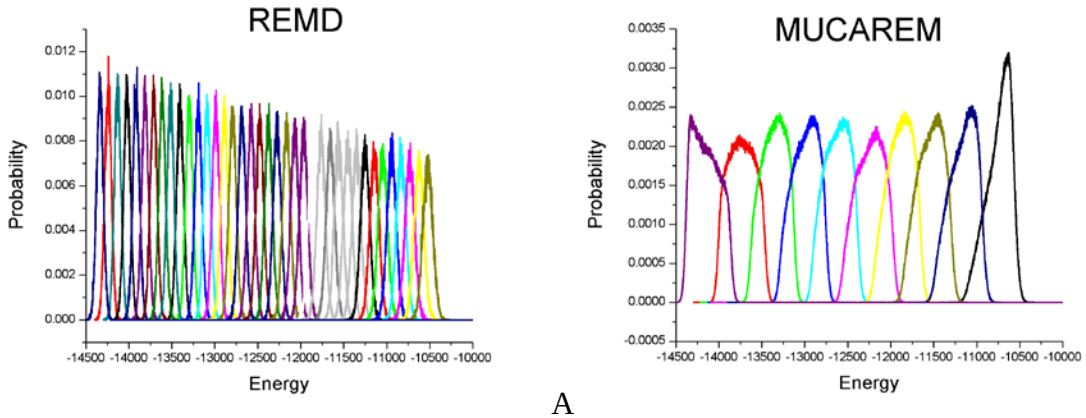


Figure 2-3. Diagram of replica exchange molecular dynamics. Large arrows represent replicas propagating in MD. Changes in level represent changes in simulation temperature.



A B

Figure 2-4A,B. Potential energy distributions of HP1 for various modes in two enhanced sampling algorithms. A) REMD B) MUCAREM

CHAPTER 3
EXCHANGE ATTEMPT FREQUENCY IN REPLICA EXCHANGE MOLECULAR DYNAMICS*

3.1 Introduction

While REMD has been shown to be effective,[26-28] the question of how often to make these exchange attempts has not yet been adequately explored. The exchange attempt frequency (EAF) is often chosen in an ad-hoc manner, using the same values as in previous papers on the subject. The time between exchange attempts can be found in the literature in a wide range.[26-28,30-32] EAF have been reported as large[12] as 100 ps^{-1} and as small[31] as 0.05 ps^{-1}. Duan et al., while studying the 21-residue helical peptide Fs21 (Ace-AAAAA$(AAARA)_3$A-Nme), looked at three values of EAF and found that 1 ps MD interval is an adequate choice for Fs21.[51] Opps and Schofield found that for various forms of PT, exchange attempts should be performed quite frequently.[52] Still, no clear discussion exists of what elements one should use to make a proper decision for choosing an EAF. As will be shown later, substantial improvement in sampling can be accomplished by properly choosing this parameter. Clearly, as the EAF approaches zero, the REMD sampling becomes identical to that of canonical ensemble MD. An argument has been advanced[30,31,51,53] proposing that if exchanges were attempted too often, equilibration to the local replica temperature will not happen and thus the system would sample conformational space improperly. These two arguments suggest both small and large EAFs would provide wrong results, for different reasons. In this dissertation, it is shown that this reasoning is partially flawed and that exchanges should be attempted as often as possible, provided exchanges are done properly.[54]

---

* Adapted with permission from Sindhikara, D., Meng Y., Roitberg, A.E. Exchange Frequency in Replica Exchange Molecular Dynamics, J. Chem. Phys., 2008; 128:024103-024113 Copyright 2008 American Institute of Physics

3.2 Methods

3.2.1 Simulation Details

For our study, four alanine peptides (Ace-A_n-Nme, n=1,3,5,7) and Fs21 (Ace-AAAAA(AAARA)$_3$A-Nme) (A = Alanine, R = Arginine) were simulated. These were chosen to span a range of sizes. Each peptide was blocked by an acetyl group (Ace) at the N-terminus and an N-methylamine (Nme) group at the C-terminus. The temperatures for REMD were distributed exponentially in a manner such that the expected acceptance probability was 20% for the alanine peptides and 15% for Fs21. Each test simulation was run for 10 ns. Sampling was gauged by computing the deviation of conformations between the test runs and REMD runs (see part C of this section). The number of replicas and the temperature of each replica in each reference simulation were the same as those in the corresponding test runs. The reference calculations started from the same structures as the corresponding test simulations and used an EAF of 0.5 ps^{-1}. The 0.5 ps^{-1} EAF (2 ps between exchange attempts) was chosen since time intervals of that order have been highly used and tested.[29,30,55,56] A regular MD simulation would simply take too long to converge. The details of both test and reference simulations such as temperature ranges and reference simulation time can be found in Table 3-1. Thus if the deviation is relatively small, the test simulation approached the behavior of a longer run in less time and hence would have a relatively high convergence rate.

All simulations were done using the AMBER 9 molecular simulation suite[57] with the AMBER ff99SB force field.[58] The SHAKE algorithm[15] was used to constrain the bonds connecting hydrogen and heavy atoms in all the simulations which allowed use of a 2 fs time step. The Generalized Born implicit solvent model GB(OBC)[59] was used to model water environment in all our calculations. Each calculation was performed in the canonical ensemble (NVT) with a Langevin thermostat, using a collision frequency of 1 ps^{-1}.

3.2.2 Conformation Deviation Between Test and Reference Simulations

In our study, the metric used to evaluate conformational error was the root-mean-squared deviation (RMSD) between backbone dihedral angle populations of test and reference simulations. The dihedral angle populations were computed by constructing a 36 by 36 histogram per residue by binning φ and ψ angle pairs resulting in 10-degree by 10-degree bins. The RMSD between the test and reference run histograms was calculated and later averaged over all residues in each peptide.

3.3 Results and Discussion

Figure 3-1 shows the average dihedral angle RMSD per residue versus EAF, for four peptides. For the smaller peptides (data not shown), any trend is within the noise of the measurement, due to the fact that in the 10 ns of the test simulations conformational space has already been adequately sampled. With increased system size, however, a significant behavior becomes apparent. As expected, over the range of small EAFs, deviation decreases with increasing EAF. Contrary to expectation, however, no upswing exists towards large EAFs. These results imply that in REMD, having a large EAF has clear advantages with no negative effect on sampling.

For Fs21 in Figure 3-1, there is no apparent improvement in final deviation by increasing the EAF above 0.05 ps^{-1}. The RMSD for these EAFs appear the same (within the noise). However, these are the final deviations after a fixed length (10 ns) of simulation time.

Figure 3-2 shows the dihedral angle RMSD versus total simulation time. Data are shown for $(Ala)_7$ and Fs21 for various values of EAFs. Deviations were calculated starting from the second nanosecond with time intervals incremented by 100 ps. As expected, the curves decrease with total simulation time. Figure 3-2 suggests that although the final deviations may be similar between mid-range and large EAF simulations, the large EAF simulations clearly converge much

faster. After 1 ns simulation, the RMSD of the Fs21 test simulation at 0.05 ps^{-1} was almost twice as much as that from the test simulation at 125 ps^{-1}. There was also much faster convergence for higher EAF simulations during $(Ala)_7$ simulations. These results indicate that high EAF simulations will achieve convergence in less time than more moderate EAFs.

3.3.1 Toy Model

To verify the fundamental effect of exchange frequency on REMD, a simplified model was created that can explain the qualitative behavior of convergence vs. EAF. The system was constructed with two temperatures and three possible configurational states, degenerate in energy. Convergence of the configurational distribution at the lower temperature (as is typically done for REMD simulations) was monitored. A transition matrix is used to represent the time-evolution operator of an ensemble as it progresses towards the equilibrium distribution. The distribution, after *n* iterations of the evolution operator, *M*, obeys Equation 3-1.

$$v_n = M^n v_0 \tag{3-1}$$

Here v_0 is the initial distribution vector, and v_n is the distribution vector after *n* iterations. The transition matrix was constructed based on Arrhenius-type rates (Equation 3-2).

$$k_{ijT} = A\exp(-E_{ij}/k_B T) \tag{3-2}$$

Here k_{ijT} is the rate of transition between states *i* and *j* at temperature *T*, E_{ij} is the barrier height between the states, k_B is Boltzmann's constant, *T* is the absolute temperature of the system, and *A* is a prefactor that depends on the system and is assumed to be proportional to $T^{1/2}$. At a given temperature, conservation of matter requires that Equation 3-3 must be true.

$$k_{iiT} = 1 - \sum_{i \neq j} k_{ijT} \tag{3-3}$$

We name k_{Tx} the rate of exchange between temperatures (replica exchanges) and in this model it is assumed to be constant. With this implementation, and Equation 3-4 can be written.

$$k_{iiT} = 1 - k_{Tx} - \sum_{i \neq j} k_{ijT} \tag{3-4}$$

In practice, k_{Tx} is the product of r_a, the rate of attempted exchanges (equivalent to our EAF in the previous sections) and P_{acc}, the probability of accepting an exchange. It is assumed that the probability of acceptance is the same as the average probability of acceptance as in Equation 3-5 since the transition matrix represents an averaged time block of simulation.

$$P_{acc} = \overline{P_{acc}} \tag{3-5}$$

$\overline{P_{acc}}$ is actually a nontrivial function of the conformation based on the potential energy distribution and the temperature distribution ($\overline{P_{acc}}$ will be discussed later in this chapter) . Many REMD simulations are reported in the literature using a P_{acc} between 10% and 40%.[26-28,30-32] An optimal P_{acc} of 20% was proposed for REMD to have the smallest deviations in heat capacity calculations.[37]

In order to simplify our system, the barriers between all three configurational states are also made equal such that Equations 3-6 and 3-7 are true.

$$k_{12T} = k_{23T} = k_{31T} = k_{ijT} \tag{3-6}$$

$$k_{11T} = k_{22T} = k_{33T} = k_{iiT} \tag{3-7}$$

The transition matrix was constructed in the basis (1*h*, 2*h*, 3*h*, 1*l*, 2*l*, 3*l*) where 1, 2, 3 are the conformations, *l* is the low temperature, and *h* is the high temperature. For these calculations, $T_h = 2*T_l$. The symmetry of the matrix, *M*, ensures microscopic reversibility as seen in Equation 3-8.

$$M = \begin{pmatrix} k_{iiTh} & k_{ijTh} & k_{ijTh} & k_{Tx} & 0 & 0 \\ k_{ijTh} & k_{iiTh} & k_{ijTh} & 0 & k_{Tx} & 0 \\ k_{ijTh} & k_{ijTh} & k_{ijTh} & 0 & 0 & k_{Tx} \\ k_{Tx} & 0 & 0 & k_{iiTl} & k_{ijTl} & k_{ijTl} \\ 0 & k_{Tx} & 0 & k_{ijTl} & k_{iiTl} & k_{ijTl} \\ 0 & 0 & k_{Tx} & k_{ijTl} & k_{ijTl} & k_{iiTl} \end{pmatrix} \tag{3-8}$$

Only two unique initial ensemble states (initial distribution vectors) are possible in this system due to the high degeneracy. The first is the state where the high and low temperature are in the same configuration called the homo-populated state (e.g.(1,0,0,1,0,0)). The second is the hetero-populated state where the configurations are different (e.g. (1,0,0,0,1,0)). Only results from the homo-populated state are shown since there was no significant difference in behavior for the other case.

Based on the system setup, the converged distribution vector is obvious (Equation 3-9).

$$v_{\infty,i} = \frac{1}{3} \tag{3-9}$$

The difference between the current configurational ensemble after *n* iterations and the limiting, equilibrium population can be quantified as follows. Figure 3-3 shows the population RMSD from the iterative solution versus k_{Tx} (EAF%) with three different unit-less barrier heights, $\frac{E_{ij}}{k_B T_l}$. Final distribution vectors for each k_{Tx} were found by iterating the multiplication of the distribution vector with the corresponding transition matrix ten times. These results agree nicely with our results from actual REMD simulations (Figure 3-1) showing that, indeed, a larger value of EAF accelerates convergence without a turnaround limit. It is also clear that for larger barrier heights, there is an increased benefit of higher EAF.

3.3.2 Replica Exchange Diagnostics at High Exchange Attempt Frequencies

3.3.2.1 Potential energy and conformational distribution

As discussed in the introduction, it has been proposed[30,31,51,53] that if exchanges are made 'too often', thermal equilibration would be inhibited, thus corrupting sampling. If such were the case, there should be a serious problem with potential energy distributions for different EAFs. In order to test for this problem, potential energy distributions were compared between simulations at high and moderate EAF. The moderate EAF was used for comparison since it is commonly used and well tested. The distributions of potential energy for the 1 ps^{-1} and 100 ps^{-1} EAF simulations for Fs21 are shown in Figure 3-4. The difference in noise simply arises from a difference in the data size for different EAFs and does not affect the results.

The distributions for both EAF simulations at given temperature appear the same (within noise), indicating an identical equilibration of potential energy. The similarity can be quantified by computing the overlap between two distributions at the same temperature but different EAFs, computed as in Equation 3-10. Table 3-2 shows that the overlap at each temperature is nearly unity.

$$Overlap = \int_{-\infty}^{\infty} dE \cdot \sqrt{P(E)}_{EAF=1ps^{-1}} \sqrt{P(E)}_{EAF=100ps^{-1}} \tag{3-10}$$

If the potential energy probability distributions follow a Boltzmann distribution, the system's energy distributions at two canonical temperatures where an overlap exists must obey Equation 3-11.[53]

$$\ln\left[\frac{P(E,T_2)}{P(E,T_1)}\right] = \left(\frac{1}{k_B T_1} - \frac{1}{k_B T_2}\right) E + \text{constant} \tag{3-11}$$

The above equation was applied to simulations at both 1 ps^{-1} and 100 ps^{-1} to test canonical equilibration. Figure 3-5 shows a scatter plot of $\ln\left[\frac{P(E,T_2)}{P(E,T_1)}\right]$ vs Energy for all adjacent-

temperature overlaps for two EAF REMD simulations. Though not shown, the plots were fitted linearly each yielding an r^2 of at least 0.985. Between the two EAF simulations there was a maximum error of 2%. Each EAF had a maximum error of 3% against the ideal slope, $\left(\frac{1}{k_B T_1} - \frac{1}{k_B T_2}\right)$. These results suggest that high EAF simulations maintain the Boltzmann distribution.

Thermal equilibration, however, does not necessarily imply correct conformational sampling. In addition to the backbone dihedral deviation data (Figures 3-1 and 3-2), helical data for Fs21 was compared against the reference run for various EAF (Figure 3-6).

Fs21 is known to be partially helical in water[60]. The thermodynamics of Fs21 helix-coil transition and helical properties have been studied experimentally and theoretically by several research groups[51,60-64]. Different force fields and sampling methods and both explicit and implicit water were used in theoretical studies[51,63,64]. The theoretical helical content highly depends on the force field used during simulations, but a discussion of the reliability of our force field is not an objective of this project. In our study, Fs21 helical properties at 300K using various EAFs were also checked in order to observe whether using high EAF would affect peptide thermodynamics. The average helical content of Fs21 of various EAFs at 300K was calculated using the DSSP definitions.[65] The results are shown in Fig. 3-6. The horizontal line in the figure is the helical content given by our reference simulation and the dots are the helical fraction calculated based on test simulations with different EAFs. Uncertainties of average helical fraction at EAF of 0.05 and 100 ps^{-1} were also calculated. The last 9 ns of each test simulation were separated into nine pieces and helical contents of each piece of 1 ns trajectory were calculated. The uncertainties were obtained from those nine 1 ns helical fractions. Combining the averages and the uncertainties that were obtained, the helical fractions are essentially the same at

all EAFs, indicating that the use of a very high EAF does properly sample conformational space. This result is consistent with our potential energy distribution and backbone dihedral angle sampling results.

3.3.2.2 Sampling time

One still might ask: If the system is allowed to exchange very often, and hence each replica potentially spends very little time at a given temperature (continuously), then when is the conformational space sampled? The answer is quite simple: always. There is always a replica at each temperature and each replica is always sampling some conformational space at its own instantaneous temperature. It should be noted that there is no restriction that barriers must be hopped continuously at one temperature. Figure 3-7A shows the temperature excursions for replica 1 for an EAF of 100 ps^{-1}. This behavior is representative of the other replicas. A full coverage of the temperature space is apparent. The behavior of residue 4 ψ angle of $(Ala)_7$ versus time for replica 1 is shown in Figure 3-7B. This backbone dihedral angle trajectory shows that the peptide does sample conformational space while moving across temperature space. Though this does not necessarily indicate that the accumulated MD time before an exchange was accepted is sufficient to sample conformations within the local potential energy basin, as in typical Monte Carlo methods, an entire basin need not be sampled continuously. Since the MC step is inexpensive, taking them more frequently results in more opportunities to get accepted moves with negligible drawback. In another perspective, each replica could be considered a particle that continuously samples multicanonical space. Thus as long as exchanges are proper, maintaining detailed balance, increasing EAF will increase the diffusion rate through multicanonical space.

3.3.2.3 Acceptance of MC moves

Increasing exchange *attempt* frequency does not necessarily imply an increase in the number of accepted moves. The number of accepted moves only increases with an increase in EAF if Equation 3-12 is true.

$$\frac{EAF^{new}}{EAF^{old}} \bullet \frac{P_{acc}^{old}}{P_{acc}^{new}} < 1 \tag{3-12}$$

Therefore, the effect of EAF on P_{acc} is examined. The average probability for acceptance of an exchange in the REMD regime with velocity rescaling is shown in Equation 3-13.[35,37]

$$\overline{P_{acc}}(\beta_1, \beta_2) = \int_1 \int_2 P_1(U_1) P_2(U_2) * \min(1, e^{\Delta\beta\Delta U}) dU_2 dU_1 \tag{3-13}$$

where 1 and 2 represent two temperatures at which an exchange attempt is made.

Therefore, if the energy distributions are independent of EAF, P_{acc} will also be independent of EAF. This criterion is verified in our system as shown in Figure 3-4. The number of accepted MC moves should then increase linearly with EAF as in Equation 3-14.

$$n_{acc} = \overline{P_{acc}} * EAF * time \tag{3-14}$$

This does not mean that the sampling rate will increase linearly with EAF since not all MC moves will necessarily benefit sampling at the same rate.

It is proposed that at high EAF, a phenomenon deemed *exchange trapping* may exist. Exchange trapping would occur when a replica consistently switches back and forth between the same two neighboring temperatures. This might occur due to a shortage of molecular dynamics time to traverse potential energy space. The exchange trapping is calculated as the fraction of exchanges that occur where the new temperature is the same as it was two exchanges before as in Equation 3-15.

$$T_n = T_{n-2} \tag{3-15}$$

The exchange trapping ratios for $(Ala)_7$ and Fs21 can be found in Tables 3-3 and 3-4, respectively. The rest of replicas displayed a similar trend. Based on our calculations, it can be seen that more than 50% of accepted exchanges were trapped when EAFs larger than 1 ps^{-1} were attempted. Comparing $(Ala)_7$ and Fs21, we find that the exchange trapping ratios are bigger for Fs21. This is no surprise since Fs21 was simulated with a smaller acceptance ratio than $(Ala)_7$ (15% and 20% respectively). Though the exchange trapping ratio does increase with EAF, so does the rate at which a replica traverses temperature space (a usual measure of efficiency and convergence). Note that exchange trapping only occurs ~10% more often in the highest EAF simulations than the ones at moderate EAFs. Table 3-3 shows the number of round-trips between temperature extrema. Still, the exchange trapping does not visibly inhibit the sampling according to the results presented earlier in this dissertation.

3.3.3 High EAF and Current Program Architecture

An important point of caution must be made regarding high EAF simulations on some computer program architectures. Though theoretically, implementing higher exchange attempts should not significantly increase computer time, certain designs may need to be revamped to allow for rapid exchange attempts. The calculation for exchanges requires only some basic arithmetic and very little information passing (new temperatures are passed, not coordinates or velocities). However, some implementations of REMD[66] have taken a regular MD program and simply wrapped it with an outer shell that acts mostly as a script, handling communications and exchange calculations. Under the assumption that the majority of computational time is spent on the MD steps and the exchanges simply add a very small overhead, this architecture holds. However, very large EAFs as those recommended here, violate that assumption. Some codes might have to be re-written to handle communication within the main program instead of through a shell.

3.4 Conclusions

The effect of various values of the exchange attempt frequency parameter of replica exchange molecular dynamics was closely examined as applied to peptides of different sizes. Initially it was expected that the backbone dihedral deviation from the correct ensemble would be affected by two major factors functionally dependent on exchange attempt frequency. The beneficial term, that lowers the deviation, comes from the fact that with increased exchange attempts, molecules have more opportunity to "see" the broader conformational space made available by sampling at higher temperatures. The hindering term was suggested as due to insufficient sampling from the lack of equilibration.

It was made clear through use of a toy model that there is a fundamental benefit of increased exchange frequency. It was also shown that remarkably consistent potential energy distribution is achieved across EAFs as seen in the overlap calculated between high and medium EAFs. The conformational distributions and thermodynamic results obtained at equilibrium are independent of EAF. The potential energy distribution, conformational distribution and thermodynamic results suggested that using high EAFs does not change the canonical ensemble one has chosen. So increasing EAF will increase the rate of barrier-hopping at lower temperatures, which can lead to faster convergence, due to the fact that the ensembles as well as the sampling rates at the highest temperature are the same for different EAFs. This should hold true for any REMD analogue where exchanges are done properly. It is then recommended that one raise the standard EAF to every few time steps for maximum efficiency.

Table 3-1. Details of the simulated systems and parameters used

Peptide	EAF's (ps^{-1})	Reference Simulation Time (ns)	Replicas (count)	Temperature Range (K)
$(Ala)_1$	0.001,0.005,0.01,0.05,0.1,0.5,1,5,10,50,125,250	50	6	190.1-594.7
$(Ala)_3$	0.001,0.005,0.01,0.05,0.1,0.5,1,5,10,50,125,250	50	8	217.2-672.6
$(Ala)_5$	0.001,0.005,0.01,0.05,0.1,0.5,1,5,10,50,125,250	50	10	230.4-755.2
$(Ala)_7$	0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 125	50	10	238.7-667.4
FS21	0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 125	250	14	280.1-683.0

Table 3-2. Overlap between fitted potential energy probability distribution at EAF of 1 ps^{-1} and 100 ps^{-1} respectively.

Temp.(K)	280.1	300.0	321.3	344.1	358.5	394.7	422.7
Overlap	0.997	0.998	0.996	0.998	0.997	0.997	0.994
Temp.(K)	452.7	484.8	519.2	556.0	595.5	637.7	683.0
Overlap	0.996	0.994	0.995	0.994	0.995	0.994	0.995

Table 3-3. Exchange trapping ratio of selected replicas in $(Ala)_7$ simulation. EAF has the unit of ps^{-1}.

Replica/EAF	0.005	1	5	10	100	125
1	0.46	0.48	0.55	0.58	0.57	0.58
5	0.35	0.48	0.55	0.58	0.57	0.57
10	0.33	0.50	0.54	0.57	0.57	0.57

Table 3-4. Exchange trapping ratio of selected replicas in Fs21 simulation

Replica/EAF	0.005	1	5	10	100	125
1	0.44	0.51	0.56	0.59	0.58	0.62
7	0.46	0.49	0.58	0.57	0.59	0.61
14	0.39	0.55	0.57	0.61	0.59	0.60

Table 3-5. Number of trips between the two temperature extrema for selected replicas in Fs21 simulation.

Replica/EAF	0.01	0.05	1	5	10	100	125
1	0	0	4	15	22	53	53
4	0	1	12	19	24	34	65
7	0	0	8	13	35	52	72
10	0	1	10	26	28	55	56
14	0	2	12	24	20	50	50

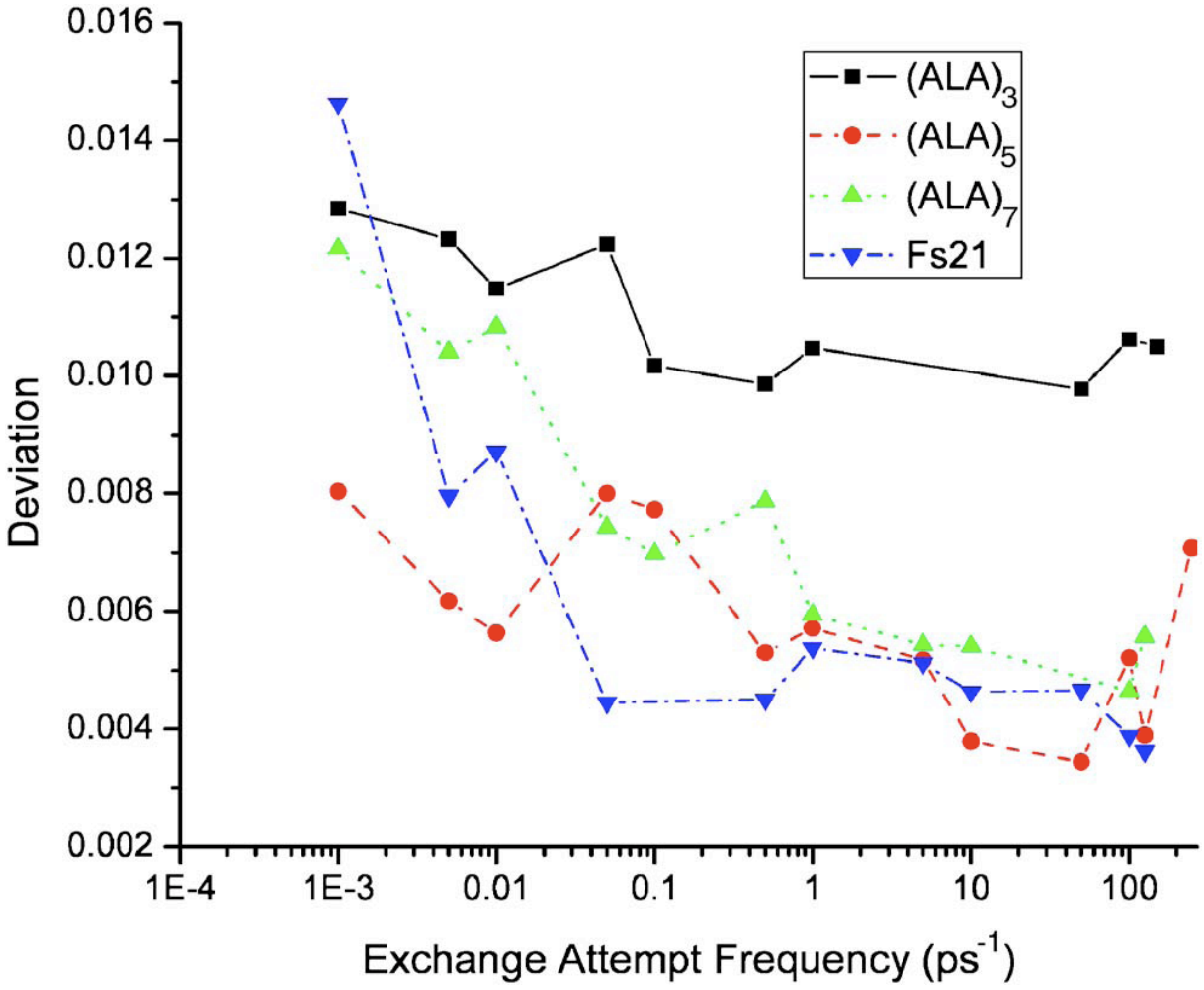


Figure 3-1. Deviation vs. EAF for four peptides.

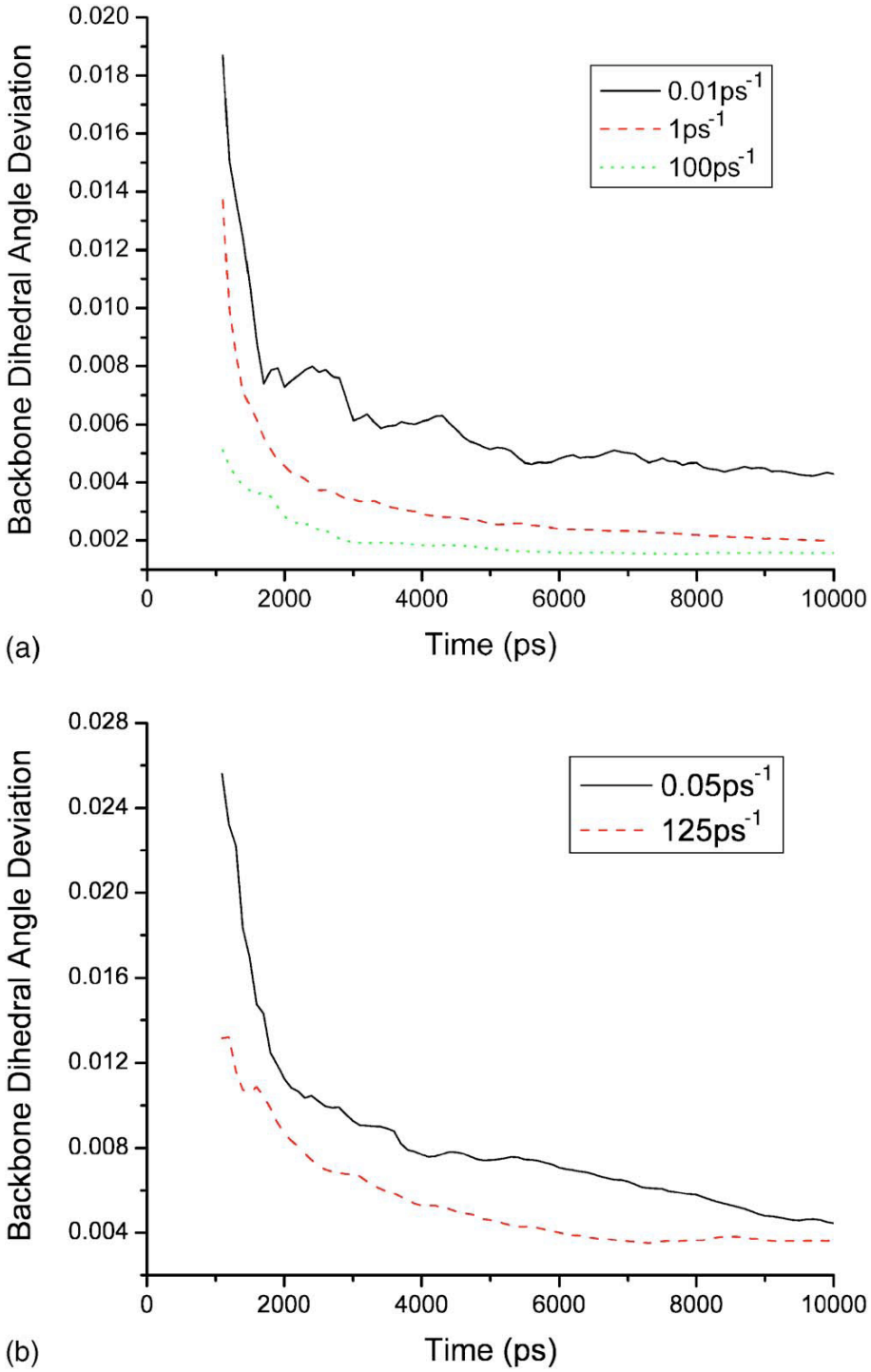


Figure 3-2. Deviation vs time for different EAFs. A) $(Ala)_7$. B) FS21.

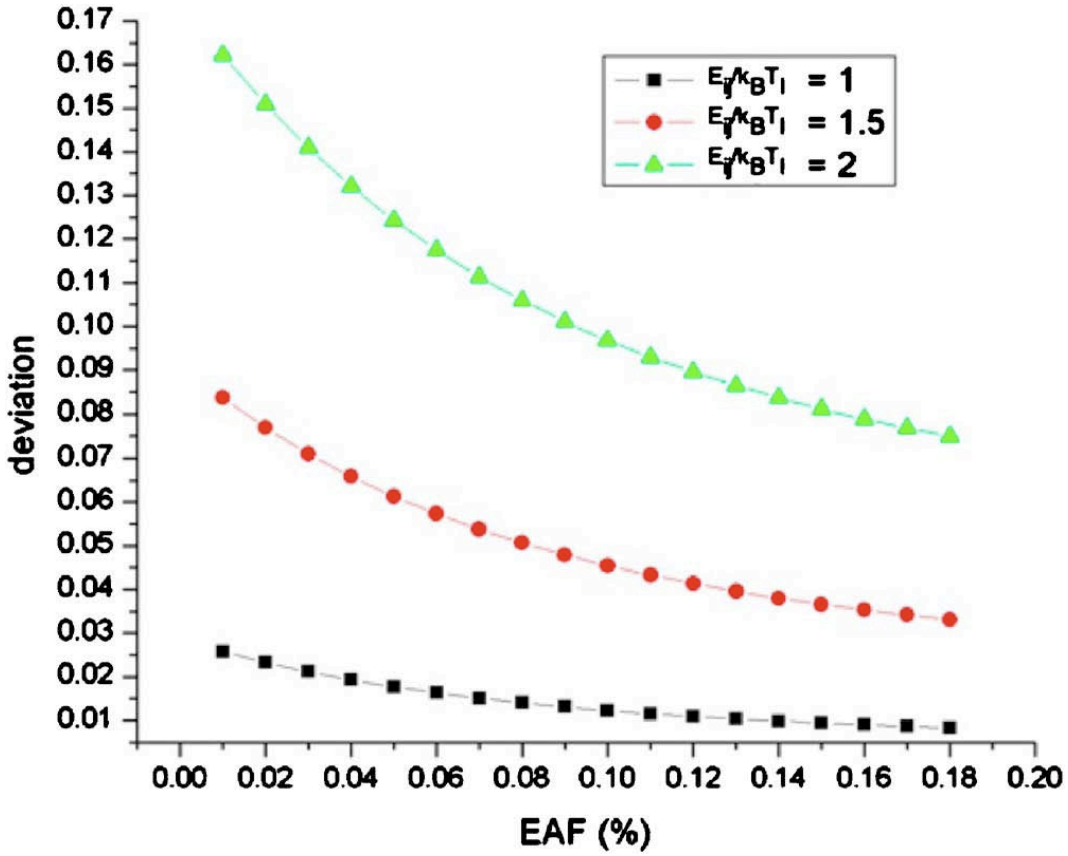
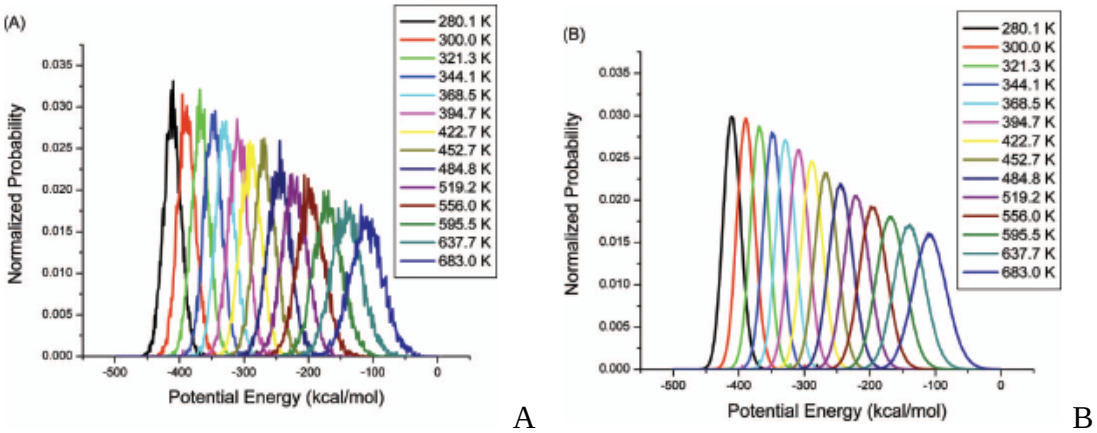


Figure 3-3. RMSD vs correct ensemble vs EAF using the toy model. Black squares, red circles, and green triangles represent solutions with barrier heights of 1, 1.5, and 2 respectively.



A B

Figure 3-4. Potential energy probability distributions at all simulated temperatures. A) EAF=1 ps^{-1}. B) EAF=100 ps^{-1}.

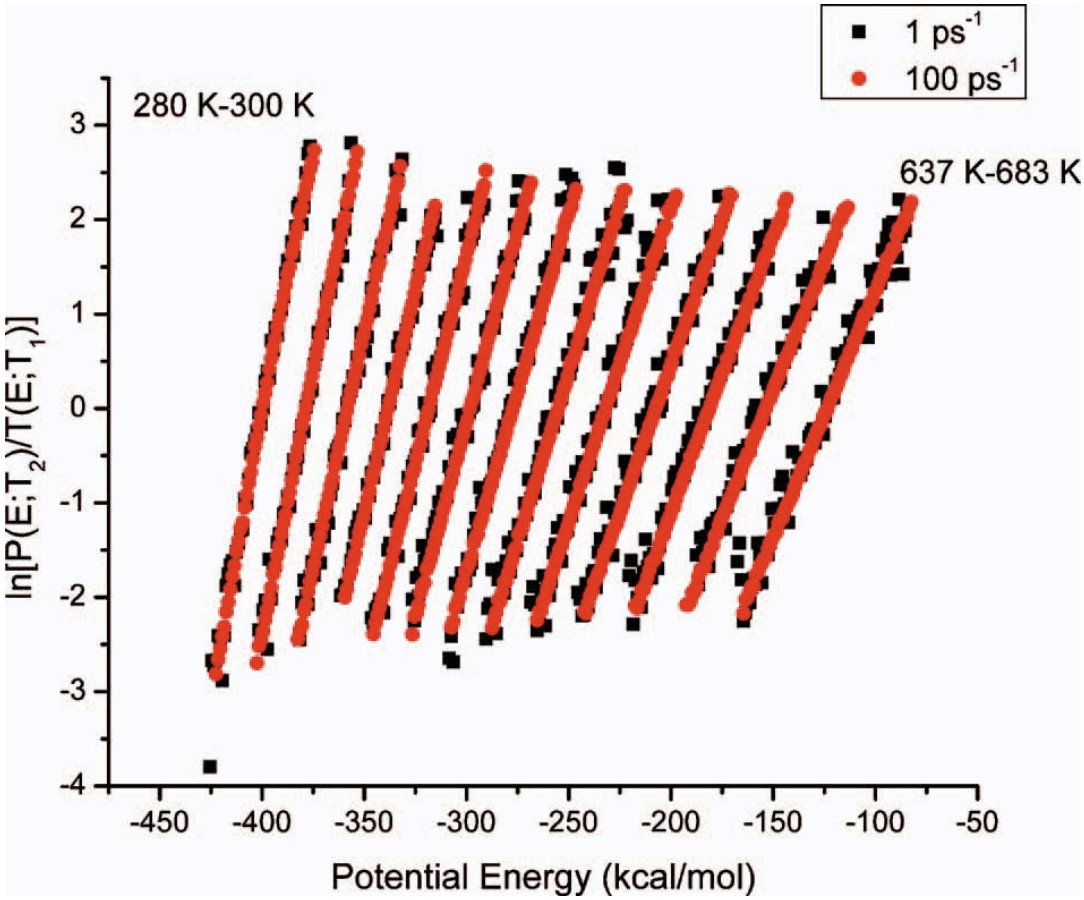


Figure 3-5. Logarithm of energy population for overlap between adjacent temperatures. Data are shown for both 1 and 100 ps^{-1} simulations.

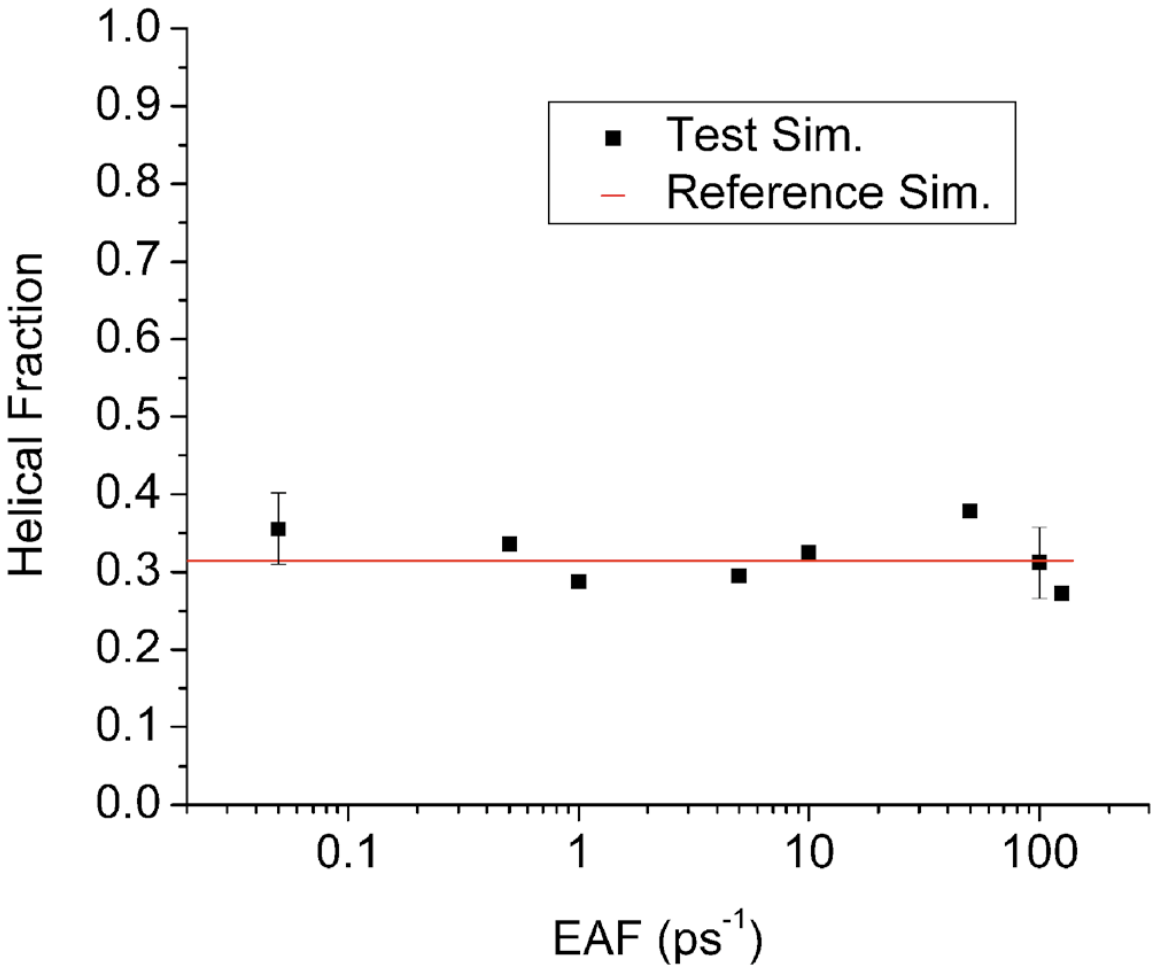
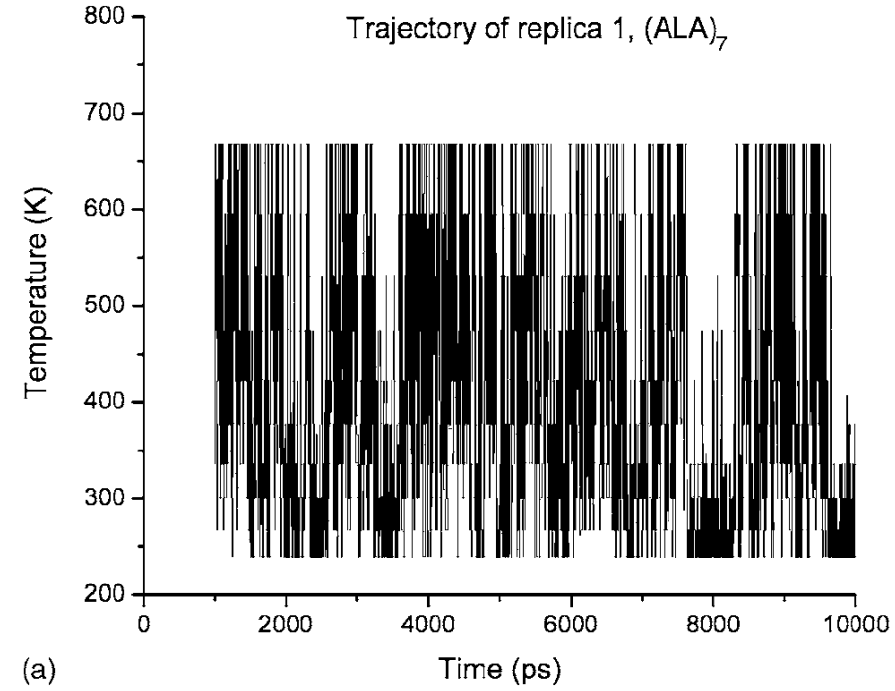
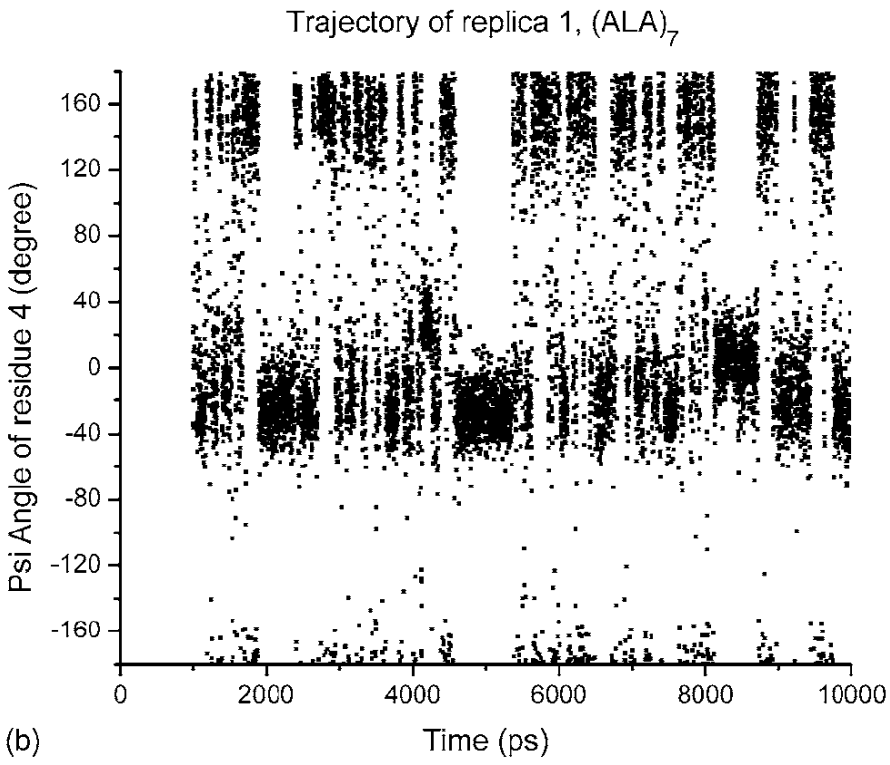


Figure 3-6. Average helical content of Fs21 for various EAF simulations at 300K. The solid line indicates the average helical content of the reference simulation.





Figures 3-7A,B. Time evolutions of replica 1 from the $(Ala)_7$ simulation at EAF=100 ps^{-1}. A) Evolution of replica 1 in temperature space. B) Evolution of φ angle of residue 4.

CHAPTER 4
CONFIRMATION OF HIGH EXCHANGE ATTEMPT FREQUENCY BEHAVIOR OF REPLICA EXCHANGE MOLECULAR DYNAMICS IN EXPLICIT SOLVENT

4.1 Introduction

As mentioned in the previous chapter, the optimum exchange attempt frequency (EAF) in REMD is a topic of debate. Since the publication of the work arguing for the maximization of EAF by Sindhikara et al.[54], some still question whether the results are generally applicable. For example, Abraham et al. suggested that results would be different for explicit solvent simulations despite showing no contradictory evidence.[67] To confirm the results of the original study, a similar study using explicit solvent systems was performed. The conclusions are identical with the original – maximizing EAF maximizes efficiency in REMD.

4.2 Methods

For this explicit solvent study, two polyalanine peptides (Ace-A_n-Nme, for n=3,7) were simulated. Each peptide was capped by an acetyl group (Ace) at the N-terminus and an N-methylamine (Nme) group at the C-terminus. All simulations were performed using AMBER 10 molecular simulation package with the AMBER ff99SB parameter set.[57,58] The SHAKE algorithm in which all bonds involving hydrogen were constrained giving a 2 fs time step.[15] Calculations were performed in a canonical ensemble with the Langevin thermostat (collision frequency of 1 ps^{-1}). Each molecule was solvated in an octahedral box of TIP3P water molecules. 1301 water molecules were added to $(Ala)_7$ (total system size 3985 atoms) and 761 water molecules were added to $(Ala)_3$ (total size of 2325 atoms). Each system was run with REMD with temperatures exponentially distributed with a theoretical acceptance probability of 20%. For $(Ala)_3$, 24 temperatures were used ranging from 293.2 K to 495.5 K; for $(Ala)_7$, 28 temperatures were used ranging from 294.9 K to 467.9 K. For each system, 20 ns-long test runs of varying EAFs were compared to a reference 100 ns-long REMD run. The reference run

utilized a "moderate" EAF of 0.1 ps^{-1}. Each test run was compared to the reference run to judge its convergence to the "correct" value.

4.3 Results and Discussion

4.3.1 Thermal Equilibration at High EAF

To check if EAF affects thermal equilibration, energetic properties were compared between a moderate EAF of 1 ps^{-1} and a high EAF of 100 ps^{-1} for the $(Ala)_7$ simulations. The potential energy overlap was compared between them as in Chapter 3, using Equation 3-11. For all 28 temperatures, the lowest overlap between distributions at the two EAFs is 0.996. This indicates that the distributions are nearly identical. Figure 4-1 shows the logarithm of the ratio of energy distributions between neighboring temperatures for the $(Ala)_7$ simulations at both 1 ps^{-1} and 100 ps^{-1}. Since the points for either EAF lie on the same line, it is likely that the Boltzmann distribution is maintained for the high EAF simulation.

4.3.2 Effect of EAF on Sampling Efficiency

Figures 4-2A,B show the RMSD of the dihedral population for both $(Ala)_3$ and $(Ala)_7$. Consistent with the results from the implicit solvent study[54], the trend is that the highest EAF simulations have the lowest error. It is noted that the reference runs are only 5 times longer than the test runs and thus if any of the test runs were around 5 or times more efficient than the reference run, the deviation would not be indicative of sampling efficiency.

Another indicator of sampling efficiency in REMD is the total number of times a replica visits both the lowest and the highest temperature. This "round trip number" for a fixed simulation time represents the speed of diffusion of a replica in temperature space and is thus indicative of the freedom of the simulation to sample. Figures 4-3A,B show the round trip number for both the $(Ala)_3$ and $(Ala)_7$ simulations. The trend is clearly that the higher EAF simulations move faster through temperature space.

4.4 Conclusions

Despite strong evidence in prior studies that REMD using high EAFs are optimal for fastest sampling, skepticism remained about the consistency of this result for explicit solvent systems. Here, a similar study to the original, implicit solvent study[54] shows that the trend is the same for explicit solvent systems. This new study should resolve many remaining questions about the effect of EAF on REMD simulations. Combined with the extensive analysis of the study, we strongly conclude that for any system where REMD exchanges are performed properly, maximum sampling efficiency can be obtained by exchanging every few MD steps.

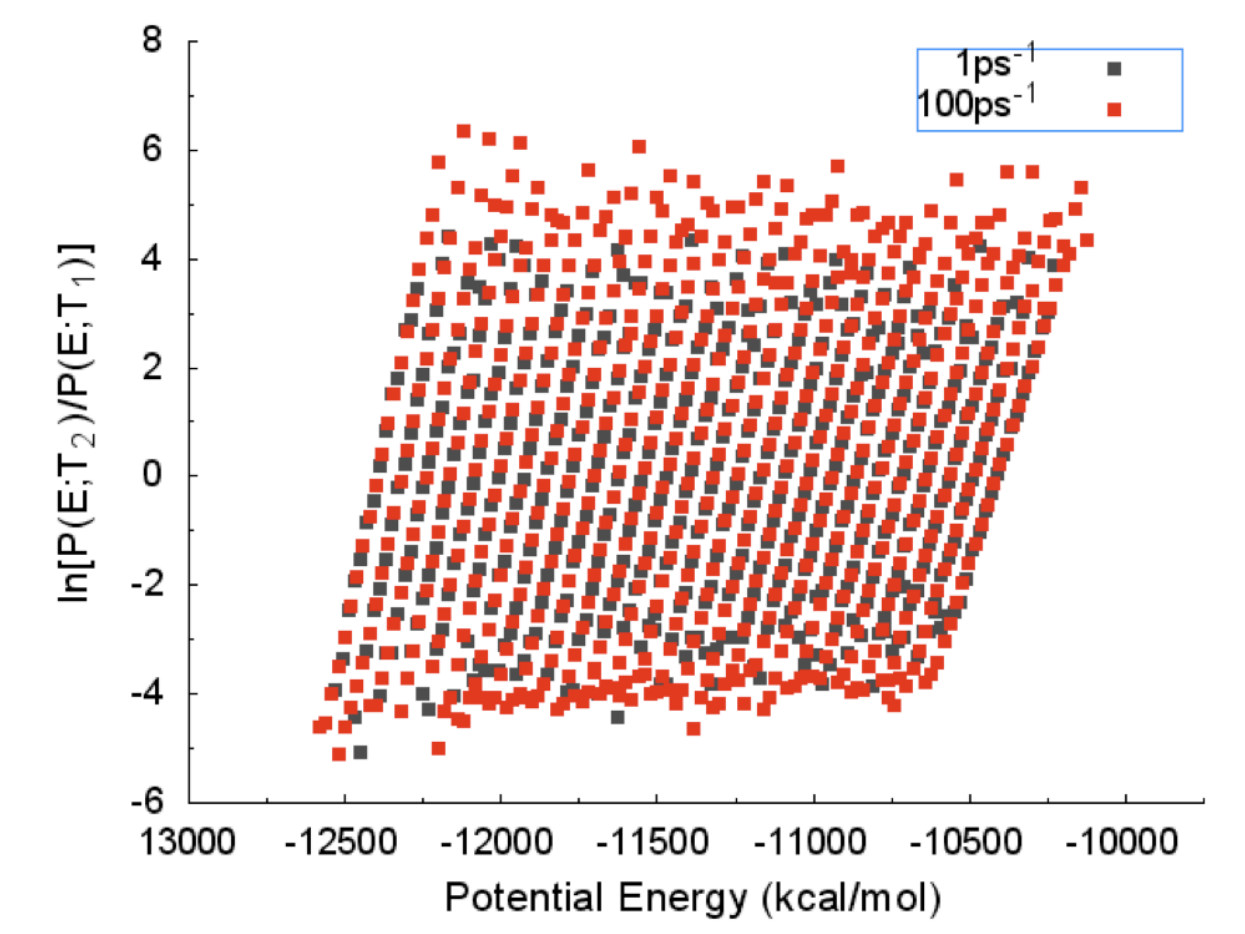


Figure 4-1. Logarithm of energy population for overlap between adjacent temperatures for $(Ala)_7$ simulations. Data are shown for both 1 and 100 ps^{-1} simulations.

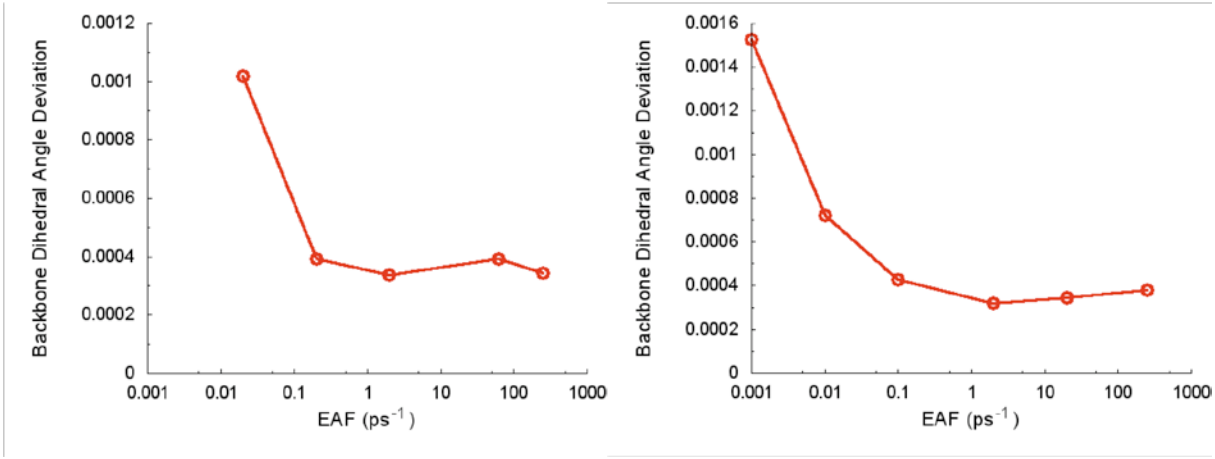


Figure 4-2A,B. RMSD between backbone dihedral angle populations of test and reference runs for both $(Ala)_3$ and $(Ala)_7$.

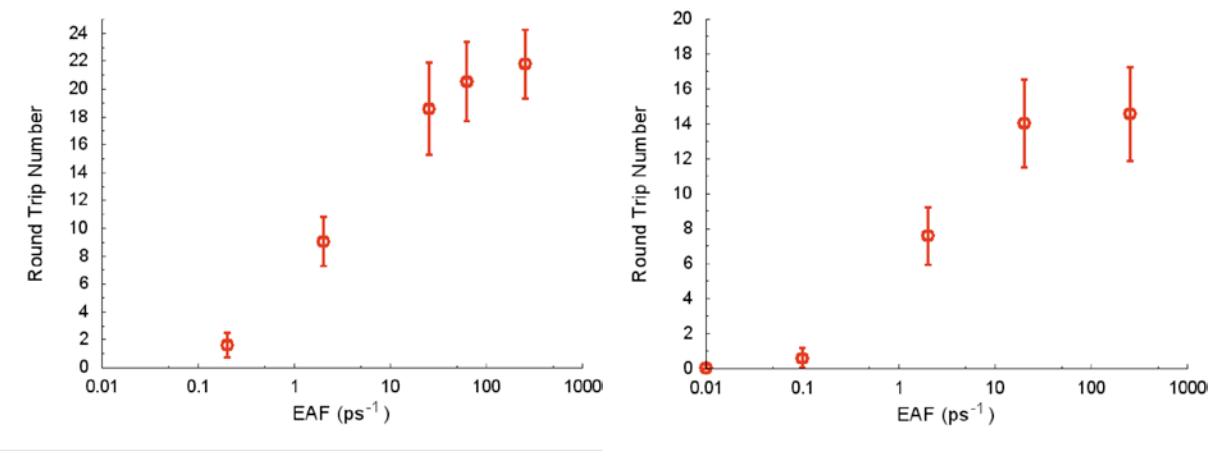


Figure 4-3A,B. Round trip number for various EAFS for both $(Ala)_3$ and $(Ala)_7$ simulations.

CHAPTER 5
THE NORMALIZED ERGODIC MEASURE FOR PARALLEL SAMPLING ALGORITHMS

5.1 Introduction

As mentioned in Chapter 1 of this dissertation, there are two sources of inaccuracy in simulation results: a poor energy representation or poor sampling. The accuracy of an energy representation (force-field) can only be tested in a simulation with sufficient sampling. Unfortunately, it can never be determined if simulation has reached convergence (of sampling). According to the ergodic theorem, only the equivalent of an infinitely long simulation has completely sampled phase space. There is, however, hope, for mortals. Only the phase space sampled well by nature in physiological timescales is necessary to be sampled by simulations.

5.1.1 Measuring Convergence

There are practical measures that simulators use to gauge the sampling of their systems. Often, a rule-of-thumb time scale for adequate sampling is used. That is, if the simulator believes the slowest conformational transition takes on average time, τ, then, depending on the number of times one must witness this transition to consider it "adequately sampled", n, the simulator would simulate for a length of time, $n \cdot \tau$. This method for choosing simulation length is both practical and convenient, but can easily be mishandled.

More quantitative measures rely on the fact that phase space is unique to a system. Any data set deemed "adequately sampled" should have identical average properties to that of another adequately sampled data set. This principle is commonly exploited in two ways. The first way is to run a "long" simulation and subsequently break this trajectory into pieces. Similarity between these pieces supports the likelihood of adequate sampling. There are more advanced versions of this method called "resampling techniques" such as "bootstrapping" and "jackknifing" where error can be estimated by comparing sets of data randomly and possibly redundantly extracted

from the long trajectory. The second way to exploit the uniqueness of phase space is to run separate simulations of the same system under different initial conditions and subsequently compare each simulation.

In addition to the choice of the manner in which properties are compared, one must decide which properties to compare. The convergence of different properties occurs at different rates and will have different values. Average potential energy, for example, converges very quickly but is a poor descriptor of the convergence of structure. Convergence of some measure of structure might be more strict than necessary for a simulation. The property whose convergence is measured should be representative of the property or properties to be reported.

5.1.2 Metrics Used for Parallel Algorithms

There are many types of parallel algorithms in common practice. The use of multiple simulations of the same system separately is the most common example. All of the "replica-type" algorithms are also parallel (e.g. REMD[12] (PT[10]), MUCAREM[47], HREMD[68], REST[69]). Parallel algorithms lend themselves to comparative convergence analysis since each replica of the system exists in the same phase space.

In 1989, Thirumalai et al. utilized a simple convergence metric for parallel simulations called the "energy metric" which compares average potential energies as in Equation 5-1.[70,71] Here, *a* and *b* represent two distinct initial states of the system, *N* is the number of particles, and $\overline{V}(t)$ is the average potential energy up to time *t*. This simple metric, which compares the deviation in the running average of the potential energy between simulations, should go to zero at infinite time according to ergodic theory.

$$d(t) = \frac{1}{N}\sum_{j=1}^{N}\left[\overline{V}_{aj}(t) - \overline{V}_{bj}(t)\right]^2 \qquad (5\text{-}1)$$

In 2000, Neirotti et al. applied a modified version of this metric to a more modern parallel simulation scheme (PT). In that work, a more general expression for multiple initial states was used as in Equation 5-2. In this expression, *M* represents the total number of parallel simulations, similar to *a* and *b* in Equation 5-1. Equation 5-2 was utilized as a qualitative metric of convergence by separately by both Calvo[34] and Rick[41] to compare enhanced versions of REMD to the original.

$$d(t) = \frac{2}{M(M-1)} \sum_{i=2}^{M} \sum_{j}^{i-1} \left[\bar{V}_i(t) - \bar{V}_j(t) \right]^2 \tag{5-2}$$

Both Equations 5-1 and 5-2 pose two major problems. First, the magnitude of the metric is essentially meaningless since one cannot decide based on the value of the metric what the level of convergence is (the dimension of the metric is E^2). Second, the quantity measured in the metrics, potential energy, is not reflective of the structural convergence of the simulation. Since the free energy surface along some structural coordinate system is often the goal of a simulation study, it is appropriate that the convergence of the structural coordinate should be measured. Here, an enhanced metric for convergence in parallel algorithms is presented that is both quantitatively meaningful and structurally meaningful.

5.2 Theory and Methods

5.2.1 The Normalized Ergodic Measure

Equation 5-2 can very simply be generalized to compare structural properties rather than energetic ones as in Equation 5-3. Here, the normalization is dropped and the potential energy is generalized to any property the simulator decides. The quantity $\bar{A}_m^i$ is the running average of property *A* for replica *i* in mode *m* (each mode corresponds to a different ensemble such as a temperature or Hamiltonian).

$$\mathrm{Y}^{U}_{A,m}(t) = \sum_{i=2}^{M}\sum_{j}^{i-1}\left[\bar{A}^{i}_{m}(t) - \bar{A}^{j}_{m}(t)\right]^{2}$$

(5-3)

The dimension of this metric is A^2. Simply dividing by the maximum value can normalize the metric as in Equation 5-4.

$$\mathrm{Y}_{A,m}(t) = \frac{\mathrm{Y}^{U}_{A,m}(t)}{\max(\mathrm{Y}^{U}_{A,m})} \tag{5-4}$$

The value of $\mathrm{Y}_{A,m}$, the normalized ergodic measure (NEM), is dimensionless and should decay to zero at infinite time scales according to the ergodic hypothesis. The normalization factor will vary even within the same system. This makes it somewhat subjective. However, the greatest value of $\mathrm{Y}^{U}_{A,m}$ is an excellent indicator for the deviation a certain property can have within the length scale of the simulation. Since $\mathrm{Y}^{U}_{A,m}$ tends to decay with time, the maximum value usually occurs very early within the simulation and thus will not change as simulations continue. If the simulator has knowledge of the range of some property, they may start separate parallel simulations from configurations near the largest range. This will guarantee that the normalization is optimal. If the simulator does not know the widest range, then $\mathrm{Y}_{A,m}$ may converge prematurely, then jump when a new state is sampled. This premature convergence can be avoided if the simulator is knowledgeable for the length scale of transitions for their system.

5.2.2 Simulation Details

To utilize the diagnostic abilities of the metric, eleven total REMD simulations of two polypeptide systems were performed. The parameters for the simulations are given in Table 5-1. $(Ala)_7$ was simulated in GB(OBC)[59] implicit solvent, and in explicit TIP3P solvent with five different exchange attempt frequencies (EAFs). $(Ala)_3$ was simulated in explicit TIP3P solvent

with five different EAFs. The effect of EAF on REMD simulation is discussed in detail in Chapters 3 and 4 of this dissertation as well as the work of Sindhikara et al.[54]

5.3 Results and Discussion

The convergence of two quantities was measured: Potential energy and the ${}^3J_{H^N H^\alpha}$ J-coupling of the first residue. The J-coupling was used since it is a measurable by NMR experiments. The equation used to calculate this value is given in Equation 5-5 as parameterized by Hu et al.[72]

$${}^3J_{H^N H^\alpha} = 7.09\cos^2(\phi - 60°) - 1.42\cos(\phi - 60°) + 1.55 \quad (5\text{-}5)$$

Figure 5-1 shows the potential energy versus time of three replicas from the $(Ala)_3$ simulation with an EAF of 2 ps^{-1}. Each replica spans an energy range of about 1500 kcal/mol. Figure 5-2 shows the running average of the potential energy at 300K for ten replicas of the same simulation. The gradual narrowing of the spread indicates that the potential energy at 300K is converging for this system. The relative size of this spread is the quantity that is measured by Equation 5-4. Note that a replica only updates its average at a specified temperature when it is sampling that temperature.

Figures 5-3A,B show $Y_{V,300K}$, the normalized ergodic measure for potential energy at 300 K, for the $(Ala)_3$ and $(Ala)_5$ simulations respectively. For either molecule, the slowest EAF is unconverged by the end of the simulation. For the faster EAFs, the 20 ns simulation is ample time to converge 300K potential energy. In $(Ala)_3$, three out of four simulations drop below 0.1 in less than 5 ns. In $(Ala)_7$, only the two fastest EAFs seem well-converged in under 5 ns.

Figures 5-4A,B show $Y_{J,300K}$, the normalized ergodic measure for ${}^3J_{H^N H^\alpha}$ of the first residue at 300K. The asymptotic decay in these figures is much slower than in Figures 5-2A,B. This indicates, unsurprisingly, that ${}^3J_{H^N H^\alpha}$, a structural property, converges much slower than

the potential energy. Interestingly, this property behaves very similarly for both molecules. The three fastest EAFs perform similarly, achieving value below 0.1 at about 15 ns. Despite the fact that the 2 ps^{-1} $(Ala)_7$ simulation (blue line in Figures 5-3B and 5-4B) converges energy noticeably slower than the faster EAFs, the convergence of $^3J_{H^N H^\alpha}$ is very similar.

One might note that since the normalization factor in the NEM is taken from the largest deviation within a single simulation, that this might vary between simulations and thus skew the results. If this were true then the similarity of the NEM for the simulations with the three highest EAFs might be deceitful. To check this, the final average value of $^3J_{H^N H^\alpha}$ was compared in Table 5-2. All the simulations with a final NEM under 0.1 are consistent with each other for both $(Ala)_3$ and $(Ala)_7$. Thus is NEM is a reasonable predictor of convergence in these cases.

5.4 Conclusions

An enhanced measure for convergence in parallel simulations is proposed and tested. The normalized ergodic measure was applied to both a structural and energetic property of various REMD simulations. The results show energetic convergence occurring faster than for structural properties. The simulations with the slowest energetic convergence also converged the slowest in structure. The other three simulations for each molecule converged energy at different rates, yet they converged structure at about the same rate. This might be a result of the loss of the kinetic trapping bottleneck. If barrer-hopping is no longer a problem, the rate-limiting step might then become conformational searching.

Comparing final results of simulations with low NEM values was done to test the value of the normalized ergodic measure. Each simulation that had a low final NEM had consistent properties with each other. It is likely that the NEM is a sufficient descriptor of self-consistency and thus convergence in most cases.

Table 5-1. Details of the systems simulated.

Molecule	Solvent	Temperature range (K)	Number of Replicas	Simulation time (ns)	EAF (ps^{-1})
$(Ala)_3$	TIP3P	293-495	24	20	250, 25, 2 0.2
$(Ala)_7$	TIP3P	295-467	28	20	250, 20, 2, 0.1

Table 5-2. Relationship between normalized ergodic measure and consistency.

Molecule	$(Ala)_3$	$(Ala)_3$	$(Ala)_3$	$(Ala)_3$	$(Ala)_7$	$(Ala)_7$	$(Ala)_7$	$(Ala)_7$
EAF (ps^{-1})	250	25	2	0.2	250	20	2	0.1
$\left\langle {}^{3}J_{H^N H^\alpha} \right\rangle$	2.40	2.39	2.37	2.28	2.36	2.33	2.35	2.5
$Y_{J,300K}$ (20 ns)	0.042	0.059	0.042	0.87	0.042	0.027	0.056	0.81

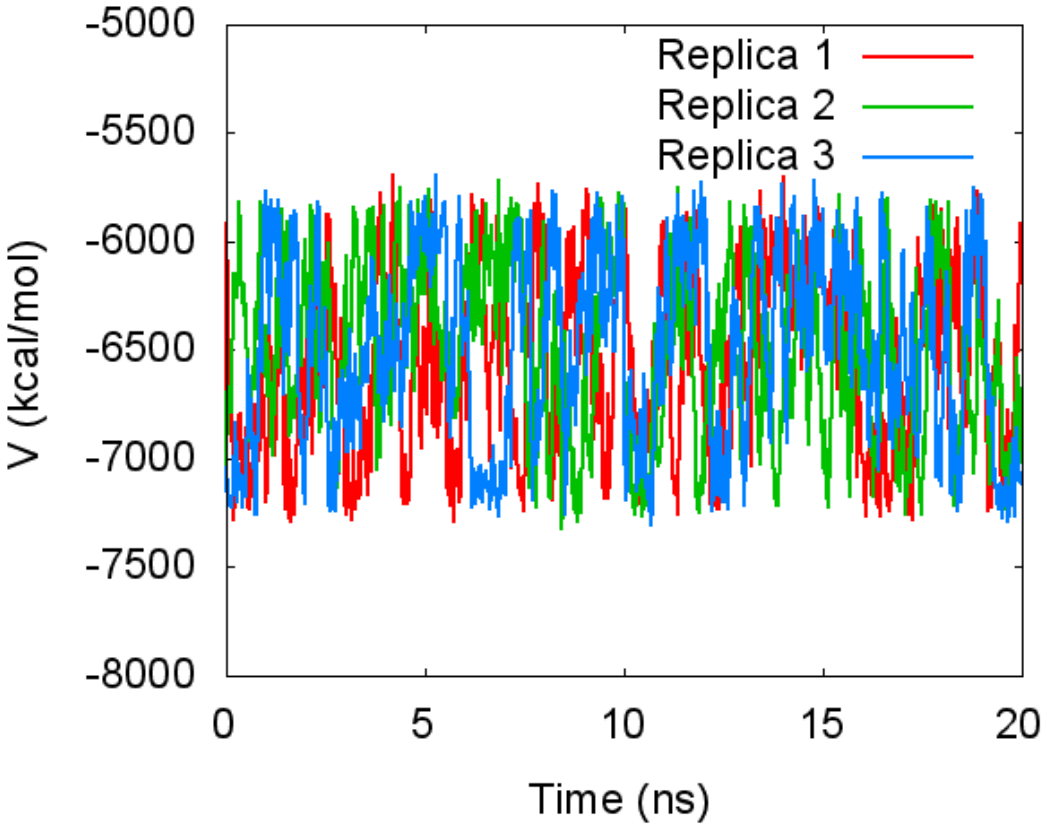


Figure 5-1. Potential energy vs time for three replicas in an $(Ala)_3$ REMD simulation performed with an EAF of of 2 ps^{-1}.

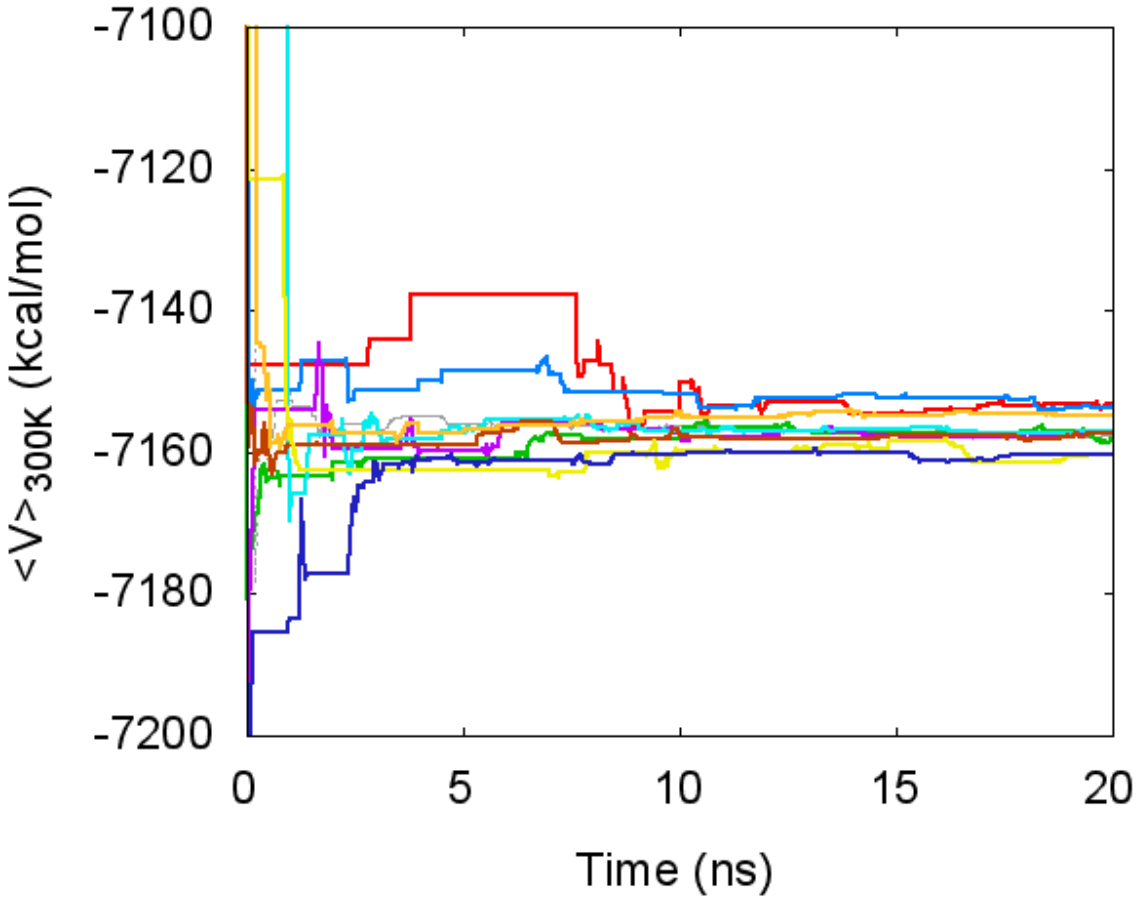
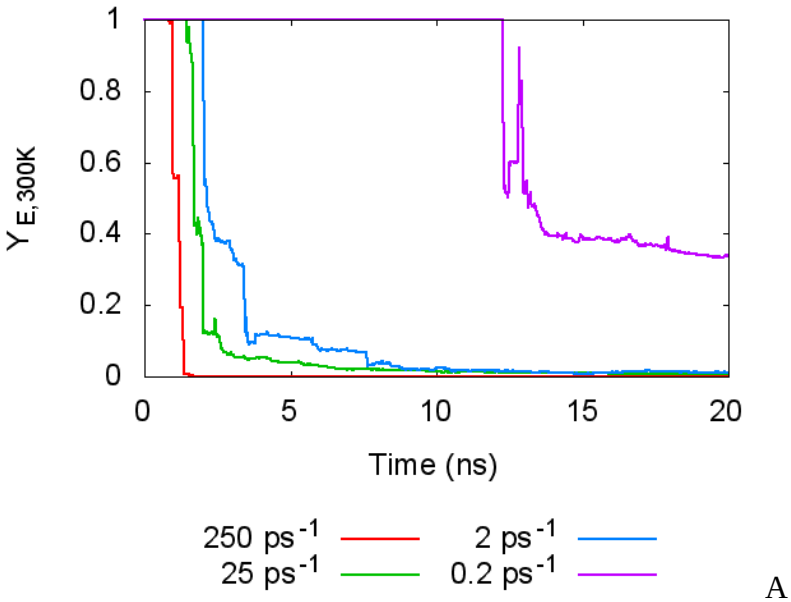
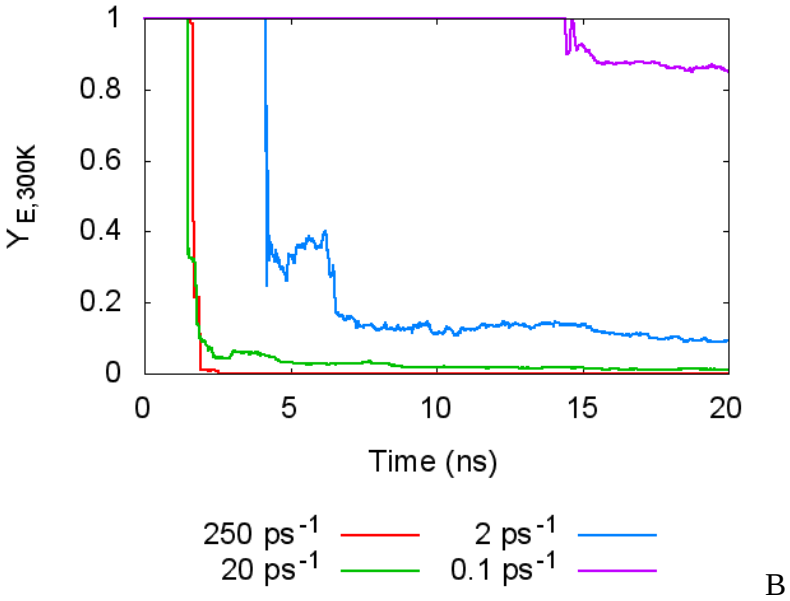


Figure 5-2. Running average potential energy at 300 K vs time for ten replicas in an $(Ala)_3$ REMD simulation performed with an EAF of of 2 ps^{-1}.

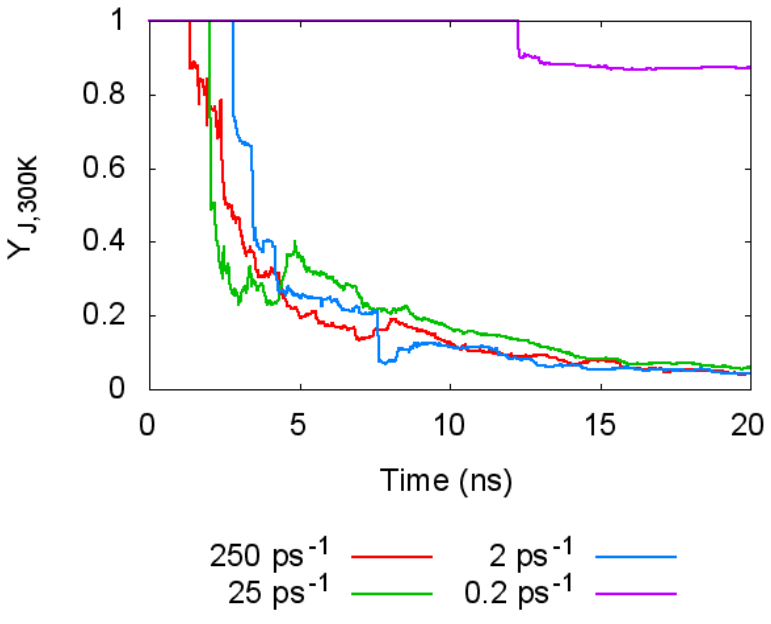


A

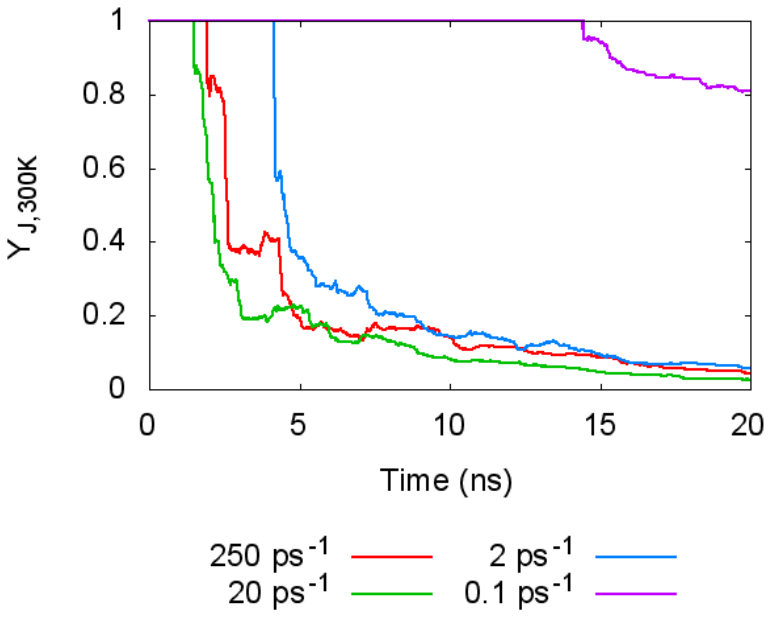


B

Figures 5-3A,B. Normalized ergodic measure for potential energy at 300 K at various EAFs. A) $(Ala)_3$ simulations. B) $(Ala)_7$ simulations



A



B

Figures 5-4A,B. Normalized Ergodic Measure for a J-coupling at 300 K at Various EAFs. A) $(Ala)_3$ Simulations. B) $(Ala)_7$ Simulations

CHAPTER 6
ADVANCING MULTICANONICAL ALGORITHM REPLICA EXCHANGE METHOD

6.1 Introduction

Multicanonical Algorithm Replica Exchange Method (MUCAREM[47]) is an enhanced sampling algorithm similar to REMD but with increased efficiency due to greater exchange acceptance rates (see Methods Sections 2.4 and 2.5). Despite the fact that MUCAREM is more efficient, REMD is much more popular. This is most likely because of the simplicity of coding of REMD that only requires programming of the calculation of the exchange and standard MD simulations at arbitrary temperatures. While it is possible for individual simulator to program his or her own version of MUCAREM by hand, implementation into a large software suite would vastly increase the impact of the method. In this project, MUCAREM was implemented into the AMBER[57] simulation suite, a comprehensive WHAM[20,21] program was written.

Additionally, several venues for optimization were explored. In MUCAREM, unlike REMD, it is much more difficult to guess the number of replicas to use. Ideally if the density of states (DOS) were known, only one replica would be necessary. But due to the initial poor quality of the density of states, a large number of replicas are necessary in order to maximize sampling. For each replica mode, one must also choose the "multicanonical limits" which are the extrema in energy space for which a replica is treated as multicanonical (see Methods Section 2.5). In this work, both number of replicas and multicanonical energy spans are tested and optimized.

Finally, it is problematic that each MUCAREM simulation usually begins with a REMD simulation in order to get a first guess at the DOS. This means that at least for the beginning of the simulation, the simulation is hindered by the same problem of REMD as having to use many replicas. To alleviate the problem, I propose that a guess of the DOS can be proposed for the first

MUCAREM run. This means that the expensive REMD step can be skipped. A simple guess at the DOS is tested with promising results.

6.2 Theory and Methods

6.2.1 MUCAREM General Procedure

MUCAREM is usually performed in an iterative manner. Typically a short REMD simulation is first run on the system for some period then the output energetics are analyzed with WHAM ("T-WHAM" since replicas are temperature biased). The DOS is obtained from this which allows MUCAREM to be run (thus DOS information must be input into the simulation). When the MUCAREM run is finished, the new energetics are analyzed with WHAM ("MUCA-WHAM" in this case). If the simulator is satisfied that the results are converged, trajectory information can be utilized with WHAM to predict expectation values and free energy surfaces in whichever conformational space he or she can extract. If the results appear to be unconverged, the data can be used to create a better guess at the DOS and MUCAREM can be run again. This procedure is iterated until the simulator is satisfied that convergence has been achieved according to whichever criterion was chosen. Figure 6-1 shows a flowchart for the procedure to run MUCAREM.

6.2.2 AMBER Code Modification

Several AMBER subroutines under the global module name "SANDER" were modified to handle special input variables, calculate modified forces, and calculate new exchange probabilities. Although specific details of the code are not necessary for understanding how to use MUCAREM, key parameters including the energetic limits, and the input style of the DOS are. Figure 6-2 shows an example input file for a single mode of a MUCAREM simulation. As can be seen in Expression 2-27 (and as explained in section 2.6), the multicanonical energy is piecewise defined for each mode. The energetic limits of each mode, $E_L^{\{m\}}$ and $E_H^{\{m\}}$, must be

specified, and are deemed "*mucael*" and "*mucaeh*" in the input as seen in Figure 6-2. Additionally, the program must know the DOS to calculate the multicanonical energy. The DOS is input as a polynomial fit to $\ln(\Omega(E))$. In Figure 6-2, "*mucap(n)*" represents the n^{th} order polynomial parameter for the fit to $\ln(\Omega(E))$.

6.2.3 WHAM code customized for AMBER, REMD, and MUCAREM

The information of the DOS required for a MUCAREM simulation typically comes from the WHAM analysis of a prior REMD or MUCAREM simulation. WHAM code was written not only to attain DOS information from REMD (T-WHAM) and MUCAREM (MUCA-WHAM), but also to output expectation values, probability distributions, free energy surfaces, and create AMBER MUCAREM input files. The latest version of this C++ program is over 1100 lines and can be found at http://www.qtp.ufl.edu/~sindhika/scripts/WHAMdan5.cpp.

6.2.4 Optimization of MUCAREM Parameters

To examine the effect of number of replicas and energy spans of multicanonical modes in MUCAREM, simulations were run on two peptide systems: $(Ala)_{10}$ and $(Ala)_{20}$. Each system was simulated with a modified version of AMBER9[57] utilizing implicit solvent GB$^{(OBC)}$[59] model with SHAKE constraints. The Langevin thermostat was utilized with a friction coefficient of 1 ps^{-1}. The molecules were simulated using the ff94 parameter set, known to overstrengthen helices, in order to ensure an energetically significant phase transition. A fast exchange attempt frequency (50 ps^{-1}) was used as recommended by Sindhikara et al.[54] Initial DOS data were obtained for $(Ala)_{10}$ from an initial 1 ns REMD simulation using 12 replicas exponentially spaced in temperature between 247 K and 721 K. For $(Ala)_{20}$, the initial REMD simulation utilized 14 replicas ranging in temperature from 261 K to 651 K. Four subsequent test MUCAREM simulations were performed with different numbers of replicas (4 or 8) and spans (r=0.4, 0.8). The value of r is the ratio of potential energy overlap with the adjacent replica, and

is directly related with the potential energy span, *w*, of each multicanonical mode as in Equation 6-1.

$$w = \frac{E_{\max} - E_{\min}}{(1-r/2)(N-2) + (1-r)/2 + 0.5} \tag{6-1}$$

Here, $E_{\max}$ and $E_{\min}$ are the global energy extrema, and *N* is the number of replicas. The description of the energy span in Equation 6-1 allows for constant replica density in energy space when the local mode energy extrema obey Equations 6-2 and 6-3.

$$E_L^{\{m\}} = \begin{cases} E_{\min} & m = 1 \\ E_{\min} + w\left((1-r/2)(m-2) - (1-r)/2\right) & m > 1 \end{cases} \tag{6-2}$$

$$E_H^{\{m\}} = \begin{cases} E_{\min} + w/2 & m = 1 \\ E_{\min} + w\left((1-r/2)(m-2) - (1-r)/2 + 1\right) & 1 < m < N \\ E_{\min} + w\left((1-r/2)(m-2) - (1-r)/2 + 0.5\right) & m = n \end{cases} \tag{6-3}$$

Figure 6-3 shows a pictorial representation of MUCAREM modes with different values of *r*. Each MUCAREM test simulation was run for 10 ns.

6.2.5 Skipping REMD in MUCAREM

The initial expensive REMD step in MUCAREM can be skipped if one can approximate the DOS to a similar level of accuracy as could be extracted from the REMD step. A simplistic approach would be to assume the system has a fixed heat capacity at constant volume, C_V. Kofke derived a simple expression for the DOS when studying acceptance rates in REMC (replica exchange Monte Carlo) as shown in Equation 6-4.[35]

$$\Omega(U) = \left(1 + \frac{1}{C}\beta_r(U - U_r)\right)^C \Omega(U_r) \tag{6-4}$$

Here, *C* is the unitless heat capacity, C_V / k_B, and the subscript *r* refers to some reference state where the average potential energy, and DOS is known. One can approximate the system to be an ideal gas giving an expression for the heat capacity in Equation 6-5.

$$C_V = \frac{3}{2} N k_B \tag{6-5}$$

Here, N is the number of particles. For any enhanced sampling simulation, an initial relaxation is required. The final or average potential energy from this short step can be used as the reference state in Equation 6-4. Also, the reference DOS, $\Omega(U_r)$, can be set to 1 since the force calculation in MUCAREM only requires the derivative of the DOS (see Equation 2-27). Thus one can completely skip the REMD run by using this rough approximation for the DOS. This technique was applied to the same systems as in Section 6.2.5.

6.3 Results and Discussion

6.3.1 MUCAREM parameter optimization

To test the performance of different parameters in MUCAREM the unsigned energy-space replica flux was calculated for each system. This measurement, which counts the number of times a replica passes though a point in energy space, is quantifies the freedom of replicas to move through energy space. The more free they can move, the faster they can barrier-hop and sample. Figures 6-4A,B show the unsigned replica flux for $(Ala)_{10}$ and $(Ala)_{20}$. Any dips in the flux are bottlenecks in the energetic sampling of the simulation. As expected, doubling of the number of replicas leads to smoother sampling. However, since the number of replicas is proportionate to the number of processors required in the simulation, it is not necessarily more computationally efficient to implement more replicas. Secondly, for the simulations with only four modes, the increase in overlap from r of 0.4 to 0.8 smoothes out the bottlenecks. Thus it is recommended that simulators utilize MUCAREM with as many processors as is available, and for them to utilize a larger span rather than a smaller.

6.3.2 Approximating the DOS

As seen in Equation 2-27, the multicanonical force is dependent on $\frac{\partial \ln(\Omega(E))}{\partial E}$. The approximation of the DOS from Equation 6-4 in the form of the unitless force prefactor at 300K is compared to that of the DOS extracted from REMD as shown in Figure 6-5. The force prefactor is very similar for both the REMD simulation and the approximation. This suggests that the information acquired from a short REMD run is not a major improvement over the approximation of constant heat capacity.

MUCAREM simulations were run using this guess for both $(Ala)_{10}$ (4 modes, and 0.8 overlap) and $(Ala)_{20}$ (8 modes and 0.8 overlap). Figures 6-6A,B show the unsigned flux for the MUCAREM simulations utilizing the approximation for the DOS along with the originals from the REMD DOS. The simulations using the approximation perform worse, in general, than those utilizing the DOS from REMD as is apparent by the dip in flux at the higher end of the energy spectrum. The performance decrease is substantial, but outweighed by the decrease in initial computational cost. By skipping REMD, the initial minimum number of CPUs goes from 12 to 4 for $(Ala)_{10}$ and from 14 to 8 for $(Ala)_{10}$. This decrease in computational cost is significant especially for simulators whose resources may be not quite enough to run a full REMD simulation.

6.4 Conclusions

MUCAREM, a superior enhanced sampling algorithm, was implemented into a large simulation suite, AMBER. A comprehensive analysis package, including WHAM was written as a complement to the MUCAREM engine.

Simulations were performed on two polypeptides, $(Ala)_{10}$ and $(Ala)_{20}$ to test optimization of parameters in MUCAREM. It was found that increasing the number of modes significantly

smoothes out bottlenecks, but not always enough to justify the additional computational expense. A definition of the energy span of multicanonical modes was made that ensures constant replica density in a manner that quantifiably depends on the overlap with neighboring modes. Consistently, the results showed that higher overlap helps mollify sampling bottlenecks.

A hypothesis that the density of states of a system can be guessed with reasonable accuracy was tested. Results showed that a MUCAREM simulation performed utilizing a guess of the density of states performs adequately, and allows for a significant decrease in the minimum computational cost (by skipping the typical initial REMD step).

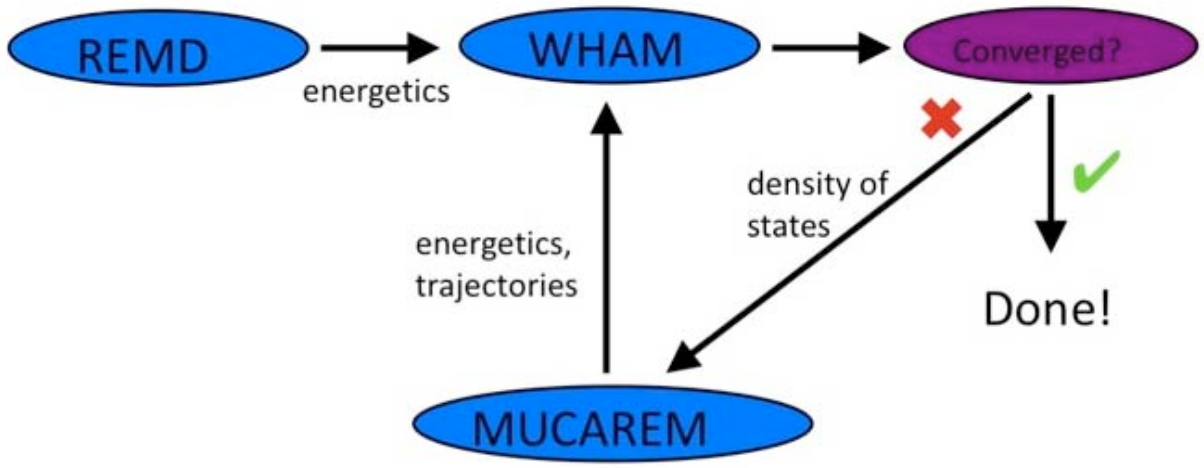


Figure 6-1. A flowchart for MUCAREM.

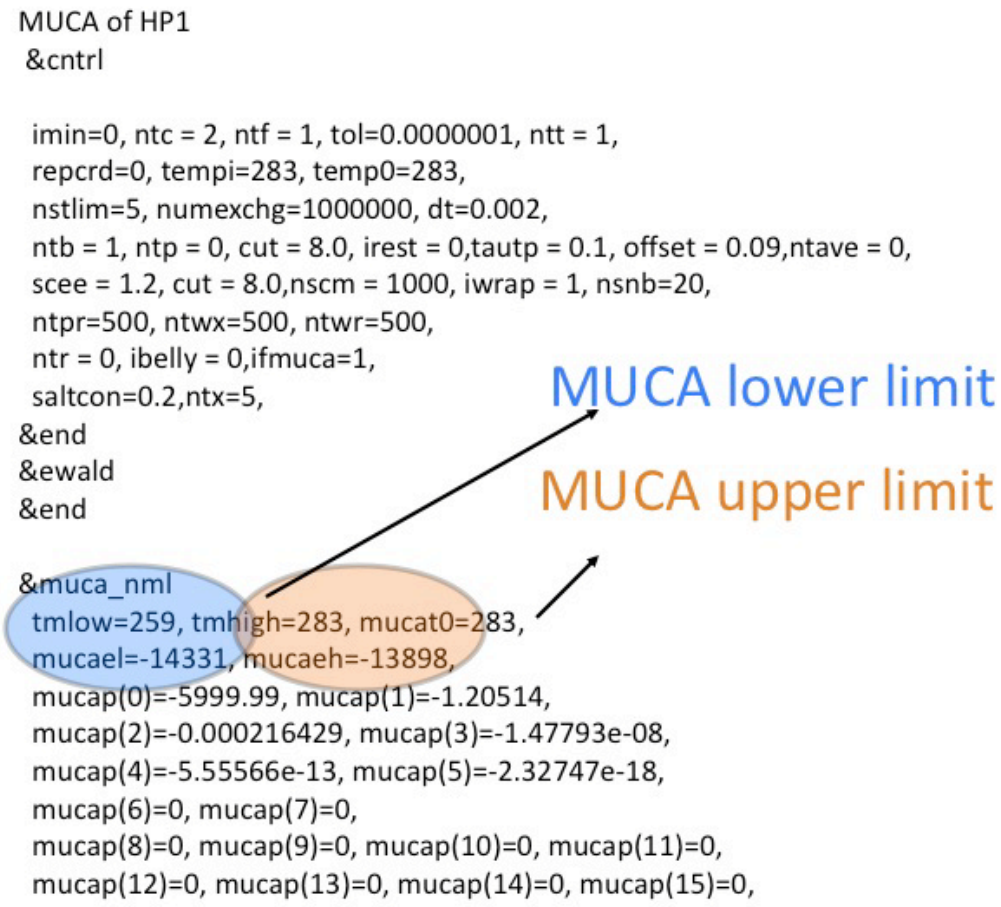


Figure 6-2. An highlighted example input file for a MUCAREM simulation.

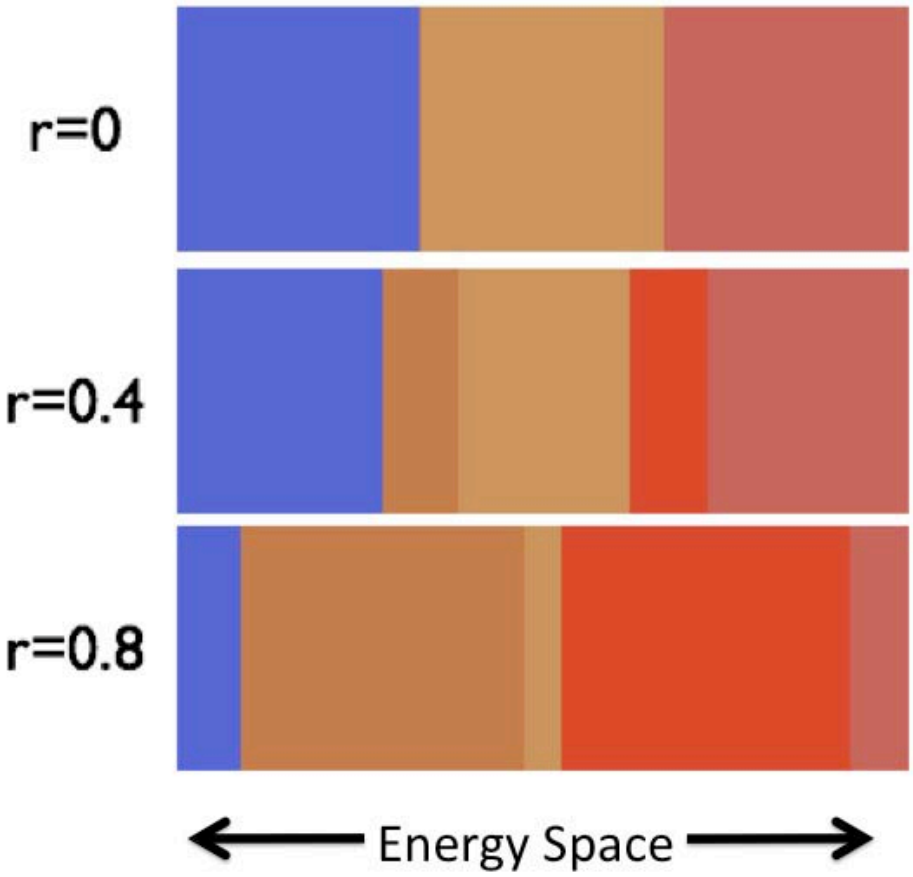
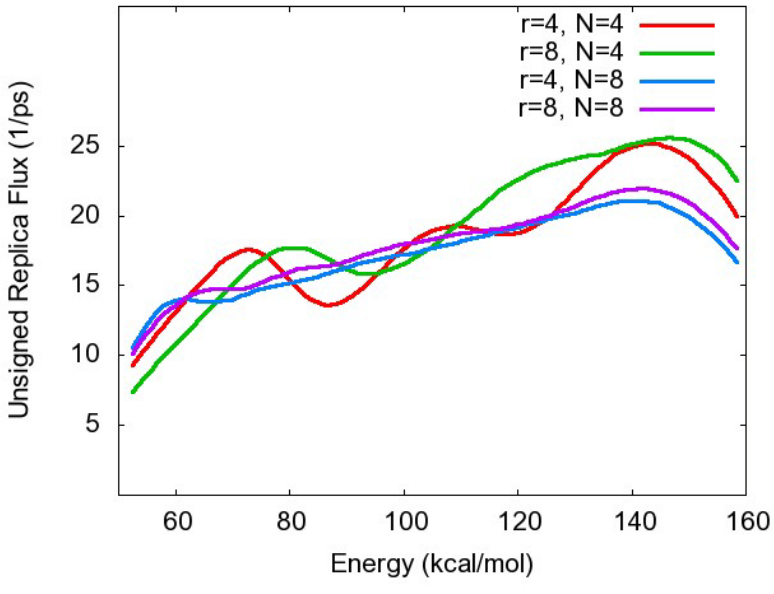
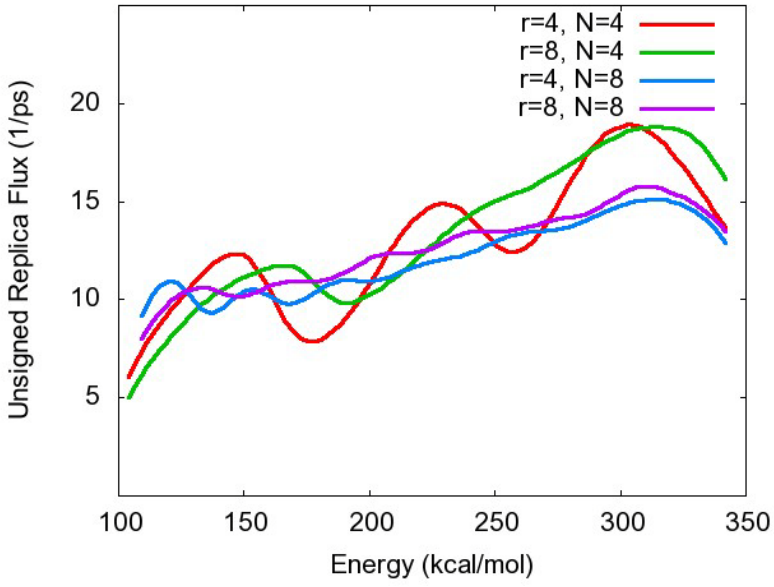


Figure 6-3. Pictorial representation of different overlap ratios for MUCAREM modes. Blue, orange and red represent successive modes.



A



B

Figures 6-4A,B. Unsigned Replica Flux for MUCAREM simulations with different number of replicas and energy spans. A) $(Ala)_{10}$. B) $(Ala)_{20}$

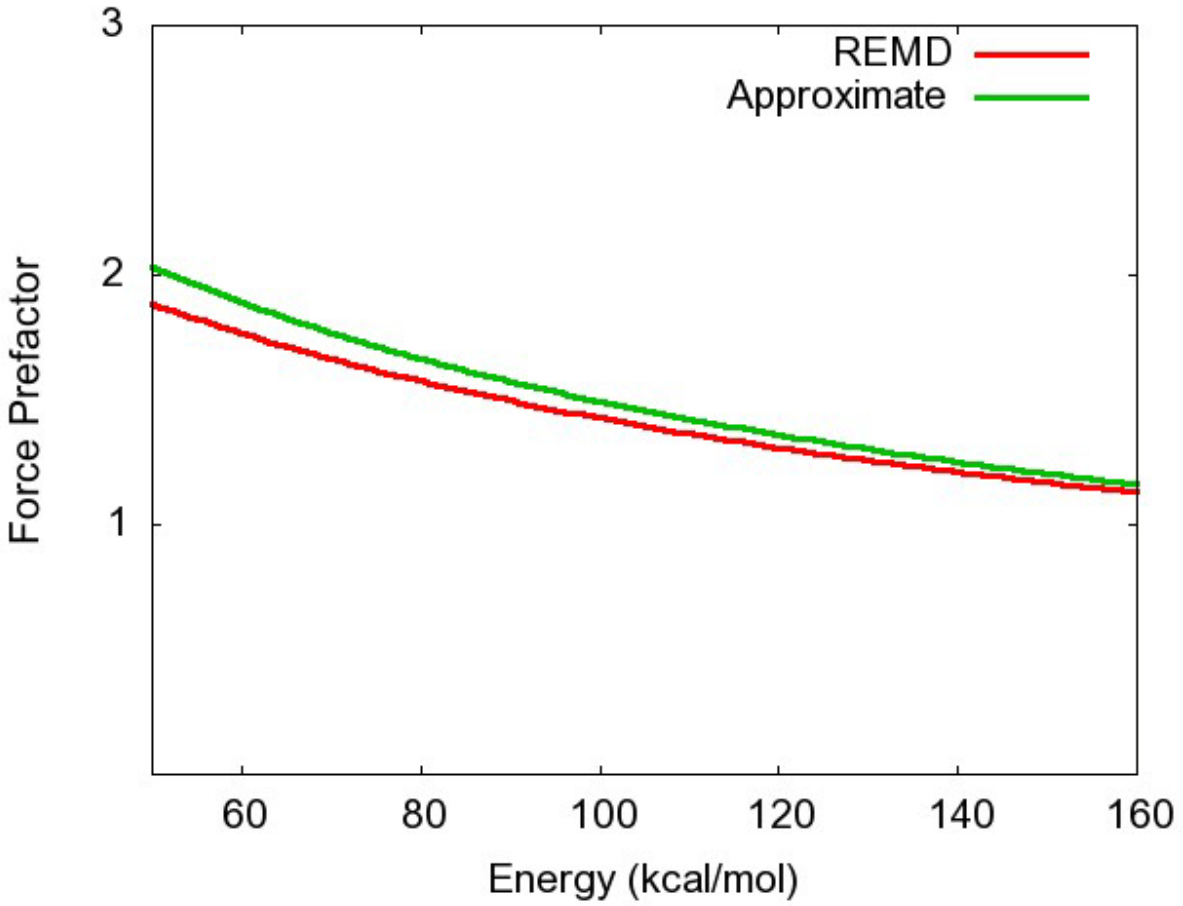
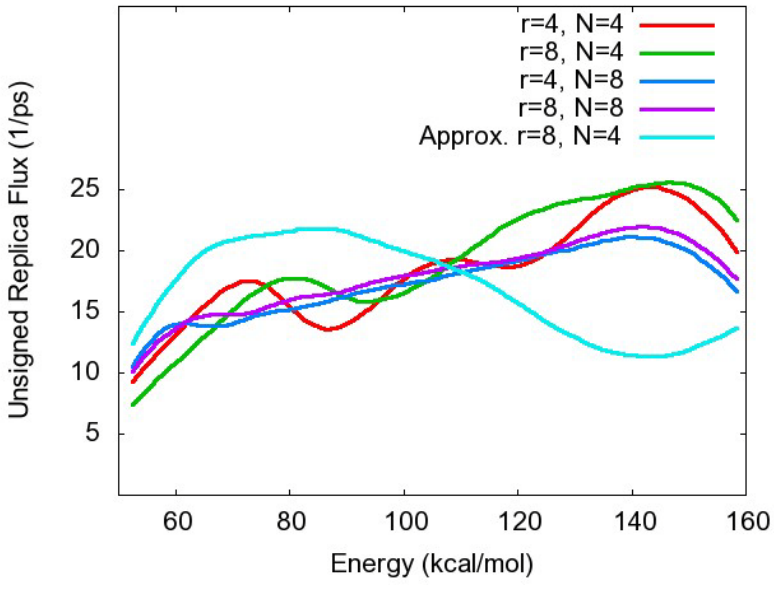
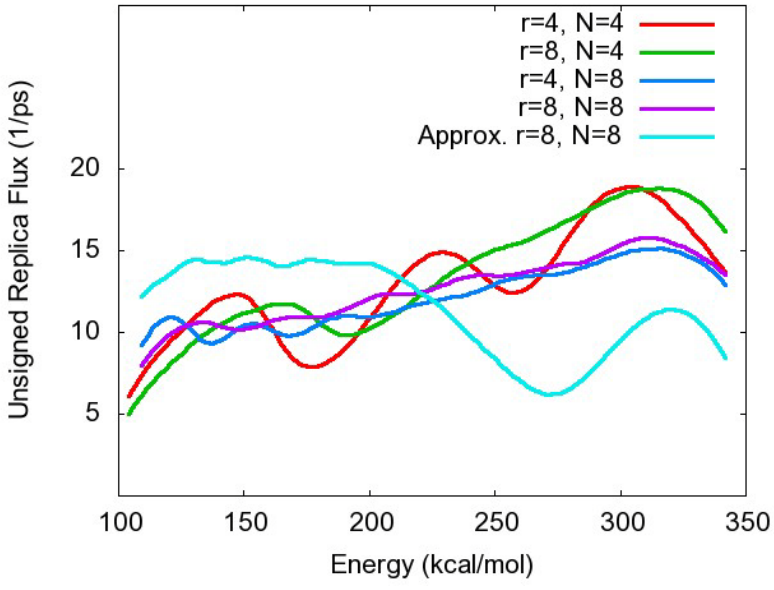


Figure 6-5. Unitless multicanonical force prefactor at 300K for two approximations of the DOS. A) From the 1 ns REMD simulation. B) From the approximation in Equation 6-4.



A



B

Figures 6-6A,B. Unsigned Replica Flux for MUCAREM simulations including approximation of DOS. A) $(Ala)_{10}$. B) $(Ala)_{20}$.

CHAPTER 7
BAD SEEDS SPROUT PERILOUS DYNAMICS: STOCHASTIC THERMOSTAT INDUCED TRAJECTORY SYNCHRONIZATION IN BIOMOLECULES[§]

7.1 Introduction

The use of molecular-dynamics (MD) simulations is widespread across various fields[73]. It is often useful to perform MD simulations in the canonical ensemble (NVT) in order to compare with experimental processes. In such circumstances, a thermostat is used to regulate temperature. Many types of thermostats are commonly employed, including Berendsen[74], Nose-Hoover[75-77], Andersen[78], and Langevin[16]. Andersen and Langevin are stochastic in nature -- including random forces to mimic the effect of solvent collisions. Both can be proven[79,80] to give true canonical sampling. The characteristics of simulations using these stochastic thermostats, especially the commonly used Langevin thermostat, are the primary focus of this project. Hunenberger wrote an excellent review of characteristics of various thermostats.[80]

Biomolecular simulations run at constant energy or constant temperature using a nonstochastic thermostat, such as Berendsen or Nose-Hoover, are considered to be chaotic and thus extremely sensitive to initial conditions. Braxenthaler et al. found that for a peptide system, root-mean-square deviations between two simulations can grow from only 0.001 Å to roughly 1 Å after only one or two picoseconds.[81] One might then expect that a stochastic thermostat, due to its use of random forces, would increase this divergent behavior. We will show in this article that under certain conditions this assumption is untrue, and failure to recognize this can lead to incorrect simulation results.

[§] Adapted with permission from Sindhikara, D., Kim S., Voter, A. F., Roitberg, A.E. Bad Seeds Sprout Perilous Dynamics: Stochastic Thermostat Induced Trajectory Synchronization in Biomolecules, J. Chem. Theory, 2009; *in press* Copyright 2009 American Chemical Society

Stochastic thermostats use sequences of pseudorandom numbers to mimic the random solvent impacts. Pseudorandom number generators (PRNGs) are deterministic; given an initial 'seed', they always produce the same sequence of numbers. This trait is useful in that it allows for reproducibility of results when needed. Thus, if multiple simulations are run with different initial conditions, $(\vec{x}, \vec{v})$, but identical random seeds, their random forces will remain the same for all simulations for their full length. Uberuaga et al. recently showed[82] that for the case of dynamics in a simple harmonic potential basin, Langevin (or Andersen) trajectories with identical seeds are *driven* to synchronize – the difference in both $\vec{x}$ and $\vec{v}$ between two trajectories decay exponentially, ultimately leading to a single trajectory path, no matter how different the initial conditions were. More generally, they argued that for a convex potential basin or even more general confining potentials a similar synchronization effect should occur. This behavior is consistent with rigorous mathematical results stating that under fairly general conditions, when the largest Lyapunov exponent is negative, trajectories starting from any ensemble of initial conditions are attracted to "random sinks"[83]. Other groups have also observed this synchronization effect in model systems[84-88]. Cerutti et al. recently described[89] a situation where rapid restarts with the same seed of a single MD simulation in Langevin or Andersen Dynamics would result in a residual (nonzero average) stochastic force. We note, though, that this residual force is not the same as the synchronization effect.

In this chapter, I will show that synchronization can also occur in the much more complex potential energy landscapes of biomolecular systems. The potential energy surfaces for these systems typically consist of a complex network of many local minima separated by negatively curved saddle regions. Nonetheless, we observe that use of the same random number seed for different trajectories leads to strongly biased behavior due to partial synchronization occurring

on the typical simulation time scale. We show this for Langevin dynamics of small peptides (Alanine trimer and pentamer) and a simulation of a temperature-jump experiment for peptide folding of a 14-residue peptide.

We also explore the possibility that trajectories with identical seeds at different temperatures can synchronize, extending the harmonic-well derivation of Uberuaga et al. to the case in which the two trajectories have different temperatures. It will be shown that there is a well-defined synchronization of the coordinates and hence a strong correlation between the trajectories exists. This multiple temperature synchronization has important implications for the method of replica exchange molecular dynamics among temperatures (T-REMD)[10,12] and variants thereof.[41,90,91] If the same random number seed is used for all the replicas, correlations among the different trajectories will contaminate the statistics of the study.

We first review the derivation of the driven synchronization for a pair of trajectories in a harmonic oscillator, then extend it to the case of two trajectories at different temperatures. We then present results from various peptide simulations in which the synchronization effect causes a bias in the results, culminating with the case of the T-REMD simulations of Alanine trimer. We close with a discussion of the importance of understanding, and avoiding, the statistical contamination that can be caused by this synchronization effect in biomolecular simulations, and we identify and explain several common situations in which a simulator may unknowingly initiate multiple trajectories with the same random number seed, including neglecting to distribute random seeds for simultaneous simulations or sequential restart simulations, or using programs that do not enforce distributions of random seeds.

7.2 Theory

7.2.1 Single Temperature Langevin Synchronization

In Langevin dynamics, particles are propagated based on the Langevin equation of motion as shown in Equation 7-1.

$$m_i \ddot{\vec{r}}_i = -\vec{\nabla} V(\vec{r}_i) - \gamma m_i \dot{\vec{r}}_i + \vec{A}(\gamma, T, \upsilon) \tag{7-1}$$

Here m_i, $\ddot{\vec{r}}_i$, $\dot{\vec{r}}_i$, and $\vec{r}_i$ are the mass, acceleration, velocity, and position of the i^{th} particle respectively. $V(\vec{r}_i)$ is the potential energy determined by the force field. Equation 7-1 is essentially Newton's second law with two extra terms: a solvent drag force represented by $\gamma m_i \dot{\vec{r}}_i$ and a random force, $\vec{A}$, which obeys the fluctuation-dissipation theorem as shown in Equation 7-2.

$$\left\langle A_i(t) A_j(t+\Delta t) \right\rangle = 2m\gamma k_b T \delta(\Delta t) \delta_{ij} \tag{7-2}$$

Here the average, $\langle\ \rangle$, is over time, k_b is the Boltzmann constant, and $\delta(\Delta t)$ and δ_{ij} represent the Dirac and Kronecker delta functions respectively. The magnitude and direction of *A* are based on a pseudorandom number, υ, and a probability distribution based on the temperature and heat bath coupling strength, γ, also known as the collision frequency, or friction. It has been shown[82] that for two particles in the same harmonic well with the same random number sequence, their trajectories are driven to synchronize. That is, for the i^{th} degree of freedom in a single dimension, as seen in Equation 7-3, the difference between two trajectories *a* and *b*, tends to zero as time increases.

$$\Delta x_i = x_i^a - x_i^b \tag{7-3}$$

Let us first consider the differences (between trajectories *a* and *b*) in the instantaneous accelerations on each degree of freedom in the Langevin regime as seen in Equation 7-4.

$$\Delta\ddot{x} = -\left(\partial V / \partial x\big|_{x_a} - \partial V / \partial x\big|_{x_b}\right) / m - \gamma\Delta\dot{x} + \left(A_a - A_b\right) / m \tag{7-4}$$

Here, Equations 7-5 and 7-6 are true.

$$\Delta\ddot{x} = \ddot{x}_a - \ddot{x}_b \tag{7-5}$$

$$\Delta\dot{x} = \dot{x}_a - \dot{x}_b \tag{7-6}$$

If the approximation is made that the local potential region to be a harmonic well as in Equation 7-7, the difference in accelerations results in Equation 7-8.

$$\partial V / \partial x = m\omega^2 x \tag{7-7}$$

$$\Delta\ddot{x} = -\omega^2\Delta x - \gamma\Delta\dot{x} + \left(A_a - A_b\right) / m \tag{7-8}$$

If the same pseudorandom number initial seed is used for both simulations, the difference in random forces becomes zero at every step. What is then left is Equation 7-9.

$$\Delta\ddot{x} = -\omega^2\Delta x - \gamma\Delta\dot{x} \tag{7-9}$$

Equation 7-9 is that of a damped harmonic oscillator. For long times, the difference in the coordinates becomes zero; i.e. the trajectories 'synchronize'.

Realistic systems are more complicated than a simple harmonic oscillator. For these systems, synchronization rates are disrupted by passage of particles through regions of negative curvature[82]. Despite this, any bound system must inevitably exist in some greater basin. Thus, synchronization must eventually occur for almost any simulated system unless a different initial seed for the PRNG is used.

Even before complete synchronization occurs for these many-minima systems, partial synchronization may occur for particles in basins of similar shape (this is where the shift from

the identical harmonic well solution is small). As we will show, partial synchronization between trajectories does indeed take place on the time-scale of realistic simulations of peptide systems and has an effect strong enough to corrupt the results.

7.2.2 Multiple Temperature Langevin Synchronization

Though constant temperature simulations are both useful and commonplace, it is often necessary to use advanced simulation algorithms that utilize multiple temperatures for better sampling. One such enhanced sampling method that employs multiple-temperature simulation is parallel tempering[10] (PT), also known as replica exchange molecular dynamics among temperatures[12] (T-REMD). In this approach, replicas of the same molecule are simulated in parallel at different temperatures, most of which are above physiological temperatures. Periodically, a Metropolis-style Monte Carlo swap is attempted between conformations at different temperatures. A discussion of many generalized ensemble algorithms for enhanced sampling including T-REMD can be found in a review by Okamoto.[25]

The analytical derivation shown in the previous section suggests that single temperature simulations should synchronize. One might expect that the added complication of multiple temperatures might diminish synchronization. We show that even when using multiple temperatures, this effect is present and of consequence. We follow the same scheme and notation as we did in the single temperature derivation with two trajectories *a* and *b*, employing the same sequence of random numbers, but at two temperatures, T_a and T_b. Following the fluctuation dissipation theorem, we see that the stochastic forces are related by a simple scaling factor as in Equation 7-10, where Equation 7-11 is true.

$$A_t^b(t) = cA_t^a(t) \tag{7-10}$$

$$c = \sqrt{T_b / T_a} \tag{7-11}$$

The equation for the difference in the i^{th} degree of freedom between the two runs is then given by Equation 7-12, where Equation 7-13 is true.

$$\Delta\ddot{x} = -\left(\partial V / \partial x\Big|_{x_a} - \partial V / \partial x\Big|_{x_b}\right) / m - \gamma\Delta\dot{x} + (1-c)A_a / m$$

(7-12)

$$\Delta\ddot{x} = \ddot{x}_a - \ddot{x}_b \tag{7-13}$$

As before, for a single harmonic potential with one degree of freedom this equation results in Equation 7-14.

$$\Delta\ddot{x} = -\omega^2\Delta x - \gamma\Delta\dot{x} + (1-c)A_a / m \tag{7-14}$$

When the two temperatures are the same, c=1, and this simplifies to the damped harmonic oscillator equation. When the temperatures differ by a small amount, c is close to unity, and the equation of motion for the difference between the two trajectories, Δx, becomes a Langevin equation with only a small noise term.

We can be more specific about how the two trajectories differ for any two temperatures by using a simple rescaling argument. Equation 7-4 can be modified to elucidate this behavior as in Equation 7-15.

$$c\ddot{x}_a - \ddot{x}_b = -\left(c\partial V / \partial x\Big|_{x_b} - \partial V / \partial x\Big|_{x_b}\right) / m - \gamma(c\dot{x}_a - \dot{x}_b) + (c-c)A_a / m$$

(7-15)

The last term (the noise) vanishes, and for a harmonic oscillator, the linearity allows us to simplify this to Equation 7-16.

$$c\ddot{x}_a - \ddot{x}_b = -\omega^2\left(cx_a - x_b\right) - \gamma(c\dot{x}_a - \dot{x}_b) \tag{7-16}$$

Using the definition of y in Equation 7-17, a damped harmonic oscillator equation can be formed as in Equation 7-18.

$$y = cx_a - x_b \tag{7-17}$$

$$\ddot{y} = -\omega^2 y - \gamma \dot{y} \tag{7-18}$$

Thus the trajectory for cx_a synchronizes to the trajectory for x_b; i.e. the trajectories at the two temperatures are related by a simple rescaling by *c*. For this single-well case, knowing the trajectory at any temperature is sufficient to specify exactly what the trajectory at any other temperature will be, once they have run long enough to be synchronized. Regardless of the potential, the trajectories are will be strongly correlated, and the effective sampling will be greatly diminished.

7.3 Methods

7.3.1 Single Temperature Simulations

Single temperature MD simulations were performed on three peptides: tri-alanine (Ac-AAA-NH_2, $(Ala)_3$), penta-alanine (Ac-AAAAA-NH_2, $(Ala)_5$) and a 14-residue peptide (Ac-YGSPEAAAKAAAA-r-NH_2, where r represents d-Arg). All simulations were performed using the AMBER 9 molecular simulation suite[57] with Langevin Dynamics in generalized-Born implicit solvent. All simulations were performed in AMBER 9 with the AMBER ff99SB force field[58] and the Generalized Born implicit solvent model $GB^{(OBC)}$ was used to model the water environment in all our calculations.[59] The SHAKE algorithm[15] was used to constrain the bonds connecting hydrogen and heavy atoms in all the simulations. For the poly-Alanine peptides, a 1 fs integration timestep was used and each calculation was performed in the canonical ensemble (NVT) with a Langevin thermostat, using collision frequencies, γ, of 1 ps^{-1} or 50 ps^{-1} (as specified). For the 14-residue peptide, a 2 fs timestep was used with Langevin collision frequency of 5 ps^{-1}. For the 14-residue peptide, 1200 initial coordinate sets were taken from

previously equilibrated run for the DS and SS production runs, which were run at an increased temperature of 372 K in order to simulate a Temperature-jump (T-jump) experiment

To demonstrate single temperature trajectory synchronization, multiple simulations were run, all with different initial coordinates, using either the same initial random seed (SS) or different seeds (DS). 100 simulations each were run for 1 ns for poly-alanines; 1200 simulations for 5 ns each for the 14-residue peptide.

7.3.2 Multiple Temperature Simulations

Synchronization across multiple temperatures is demonstrated by use of T-REMD simulation. Both a DS and SS 100-ns T-REMD simulations were performed using the AMBER9 package with Langevin dynamics in implicit solvent GB model. A Langevin thermostat was used with a collision frequency of 50 ps^{-1}. The SHAKE algorithm was employed allowing use of a 2 fs timestep. Both systems utilized 6 replicas and started from the same initial configurations. The replica temperatures were spaced geometrically: 251.8K, 300.0K, 357.5K, 426.0K, 507.6K, and 604.8K. Exchanges were attempted every 500 steps (1 ps). The T-REMD code was altered to keep the random number sequences synchronized for all replicas for the SS simulation. Snapshots were recorded every 25 ps.

7.4 Results & Discussion

7.4.1 Single Temperature Synchronization

For both the $(Ala)_3$ and $(Ala)_5$ simulations, the dihedral angle of the second residue (ϕ_2) was measured versus time as an internal unit. We could have chosen any other set of coordinates to illustrate the synchronization effect.

Figures 7-1A and 7-1B show probability distributions of ϕ_2 across the sets of 100 simulations for $(Ala)_3$ for a collision frequency of 1 ps^{-1} or 50 ps^{-1} respectively. According to the harmonic theory[82], in the low collision frequency regime, increasing the frequency, γ, should

yield faster synchronization of trajectories. Histograms are shown at arbitrary intervals of 1 ps, 333 ps, and 666 ps into the trajectory. For comparison, probability distributions across the entire trajectories are shown in thin lines though since nearly identical, they are virtually indistinguishable.

Regardless of the random seeds (SS or DS), after a very long time, the distributions of ϕ_2 angles (very thin lines) are the same in each case. If synchronization is not present (as in the case of DS), the distribution of ϕ_2 among the 100 trajectories at a given time should be similar to the longer time average population. Conversely, if the same seeds (SS) are used for all 100 trajectories, the system behaves very differently. For instance, at 666 ps, for both figures 1a and 1b, a large number of trajectories have very similar values of ϕ_2, as represented by a sharply peaked histogram. Figure 7-1 clearly shows that even in complex systems, the effect of synchronization is observable. The behavior of a coordinate for a set of SS simulations is similar to a swarm that expands and tightens as if compelled to come together. A movie of the Ramachandran plot (in beta/ppII region) of the first residue of $(Ala)_3$ with $\gamma = 50\ ps^{-1}$ demonstrates this behavior (http://www.qtp.ufl.edu/~sindhika/sync/ALA3-rama-movie.mov).

To quantify synchronization among the entire set of 100 simulations, we used a measure of how many of them were similar to each other at any particular time. This was done by histogramming a physical observable (again ϕ_2 in our case) and counting how many of the 100 simulations reside in the histogram bin with maximum population. This is equivalent to the maximum height in Figure 7-1. If all 100 systems were perfectly synchronized, the highest fractional population (HFP) would be exactly one. Conversely, for completely unsynchronized systems, the HFP should stay relatively constant (and small for small bin sizes). For our simulations, the frame-by-frame ϕ_2 population was binned in narrow 2-degree windows. Figures

7-2A and 7-2B display the time series of the fractional population of the most popular bin (HFP) for both $(Ala)_3$ and $(Ala)_5$, versus time. The sideplots show the probability distributions of those HFP time series'. Included in the SI are additional HFP time series' and probability distributions.

From the figures it is clear that the DS simulations have a small and relatively constant HFP. This is expected since the trajectories evolve independently from each other (there are not many simulations where the ϕ_2 angles are the same). In contrast, the SS simulations HFPs are much larger than for the DS case and in some cases achieve extremely high values. For instance, at 628 ps, a HFP value of 0.78 (Figure 7-2A) means that 78 of the 100 simulations have the same value of ϕ_2 (to within 2 degrees). The HFP difference between DS and SS is significantly smaller for the $(Ala)_5$ possibly because the synchronization is slower for larger systems. However, large homogeneous systems, such as those with explicit solvent, may still synchronize quickly.

To visualize the system as a whole, a time snapshot from the $(Ala)_3$ simulations was chosen and shown with all 100 simulation frames superimposed (see Figure 7-3, created in VMD[92]). In this figure, the red and blue spheres represent atom locations in the SS and DS simulations respectively. A stick representation is shown in gray as a visual aid. There is the same number of red spheres as blue (100 per atom). The figure shows fluctuation among the DS snapshots is much greater than that for SS; since the SS simulations are partially synchronized, the atomic positions are more condensed than they should be otherwise.

When simulating a complex system, it is often useful to utilize many simulations to reduce the error. The average over many simulations of a property, *A*, at time *t*, $\langle A \rangle_{sim}(t)$, is likely to be closer to the true average, $\overline{A}$, than the value of a property of a single simulation, $A(t)$ since value of *A* will fluctuate naturally in time. If the simulations are uncorrelated with each other, it can be shown that the standard deviation over time of these averages over simulations,

$\sigma_{time}(\langle A \rangle_{sim})$, is less than the standard deviation over time of a single simulation $\sigma_{time}(A)$ (which is caused by the natural fluctuations) (see Equation 7-19).

$$\sigma_{time}(\langle A \rangle_{sim}) = \sigma_{time}(A)/\sqrt{N_{sim}} \tag{7-19}$$

Here, N_{sim} is the number of simulations. However, if the simulations are correlated with each other, the average over simulations $\langle A \rangle_{sim}$ will fluctuate in time with greater amplitude, similar to that of a single simulation, $A(t)$. This has the same effect as reducing N_{sim}. Thus, according to Equation 7-20, correlated simulations will have a higher standard deviation over time of the average over simulations, $\sigma_{time}(\langle A \rangle_{sim})$. We have presented above some arguments and results showing that many simulations run with the same initial random number generator seed will become somewhat synchronized over time. This effect will cause correlations between different simulations.

We present here a striking example of this effect demonstrated in a simulation of temperature jump folding for a 14-residue peptide. This peptide was chosen since the T-jump kinetics were recently measured experimentally.[93] We have previously published a protocol for the simulation of that type of experiment.[94] In physical T-jump experiments, proteins are heated rapidly by a laser to observe folding events. Typically, a spectroscopic measure such as Trp-fluorescence or IR absorbance is used to follow the subsequent population relaxation.[95,96] Unfortunately, in simulations, these phenomena are difficult to estimate. The expected CD spectra, rather, can be estimated in simulations based on the structure of the system. Since the CD signal at 222 nm is sometimes used to measure the average ellipticity of molecules, we focused on this measure to observe T-jump kinetics.

We computed ellipticity at 222 nm vs time averaged over 1200 simulations, using the method introduced by Sreerama and Woody.[97] Figure 7-4 shows the ellipticity vs time averaged

over all 1200 simulations for SS (Red line) and DS (Black line) respectively. As can be seen in Figure 7-4, the signal to noise ratio is dramatically worse for the SS simulations.

The standard deviation for the last 2.5 ns of the T-Jump simulation is 689 and 104 deg*cm^{2}*dM^{-1} for the SS and DS simulations respectively. Thus, according to Equation 7-20, the effective number of simulations is 44 ($\sqrt{689/104}$) times smaller for the SS than the DS. According to our preceding explanation, this means that the single seed runs act not like 1200 runs, but as if only 27 (1200/44) truly independent runs. Thus, the average over 27 DS simulations (1200/44) should have similar standard deviation to that of the 1200 SS simulations (thin green line in Figure 7-4). This effect is clearly shown in Figure 7-4 as a thin green line. We clarify that the single seed runs are not 'wrong', but that they produce overall fluctuations that are equivalent to a much smaller number of independent runs.

7.4.2 Multiple Temperature Synchronization

In the T-REMD simulations of Alanine trimer, only six simultaneous simulations could be compared – one for each replica (as opposed to the 100 or 1200 simulations from the single temperature simulations above). Figure 5 shows the histogram of the highest fractional population of ϕ_2 bins (2 degree bins) for SS and DS simulations. The DS simulations (black bars) have a higher probability to have an HFP of 1/6, that is, that no two replicas have a ϕ_2 angle within 2 degrees of each other. The SS simulations are more likely to have two or three replicas with the same ϕ_2 angle. This indicates that some synchronization does occur between replicas. Such a simulation would not only be biased, but also additional consequences for exchange probabilities may exist. Although T-REMD was used as an example, we expect synchronization to occur for any set of multiple temperature simulations.

7.4.3 Relevance of Synchronization

As evidenced by the results, thermostat induced trajectory synchronization biases results and should be avoided. Depending on the severity of the synchronization, the bias may or may not be obvious to the researcher. It is thus important to understand the nature of synchronization to be aware of situations where it might occur.

Synchronization occurs when there is an overlap of pseudorandom number sequences, and this is typically caused by using the same initial seed for multiple runs. This can happen inadvertently for many reasons. Some simulation programs use a default random seed. AMBER, for example, uses a default random seed if none is specified. Others may use a time-seeded PRNG, which, depending on the implementation, may give a high risk of giving identical seeds. For example, if the program uses a time seed connected with a clock that is discretized to the nearest second, then if many simulations are initiated simultaneously, there is a high probability that many or all will receive the same seed.

Quite often simulators restart simulations. If one restarts with the same parameters (including initial seed), then the simulations could become self-synchronized. Cerutti et al. recently reported[89] a different negative consequence of repeatedly restarting Langevin or Andersen MD runs with the same initial seed -- trajectory corruption caused by a nonzero average stochastic force. Additionally, coders of new methods, such as T-REMD, which string together MD segments, might unknowingly build a code that uses the same seed. As we have shown, even Langevin MD runs at different temperatures can become synchronized.

Furthermore, since PRNGs have an inherent period, MD runs which call the PRNG more than this amount will naturally repeat the sequence. Although advanced PRNGs such as the Marsaglia algorithm[98] have extremely long periods (2^{144} for AMBERs implementation), older

PRNGs have much shorter periods. We highly recommend that simulators take note of the PRNG period of the program they are running.

7.5 Conclusion

It has been shown here that identical-noise synchronization effects, previously observed for relatively simple systems under the influence of a stochastic thermostat, can also occur in the much more complex systems typical of biomolecular simulations. Even in the case of trajectories at different temperatures, harmonic analysis shows a special scaled synchronization will occur. It was indeed found evidence of synchronization bias in a replica-exchange simulation. In a simulation study, this synchronization tendency, even if weak, will corrupt the statistical quality of the results, and may even lead to incorrect conclusions about the qualitative behavior of the system. Using modern biomolecular simulation programs and methods, many ways exist in which one can inadvertently initiate trajectories with identical seeds. It is possible that many papers have already been published with data that is biased by synchronization. We advise that great care be taken to avoid this situation by meticulous preparation of seeds, and we suggest that authors may wish to state specifically whether different initial seeds have been used when their results are based on multiple trajectories with a stochastic thermostat.

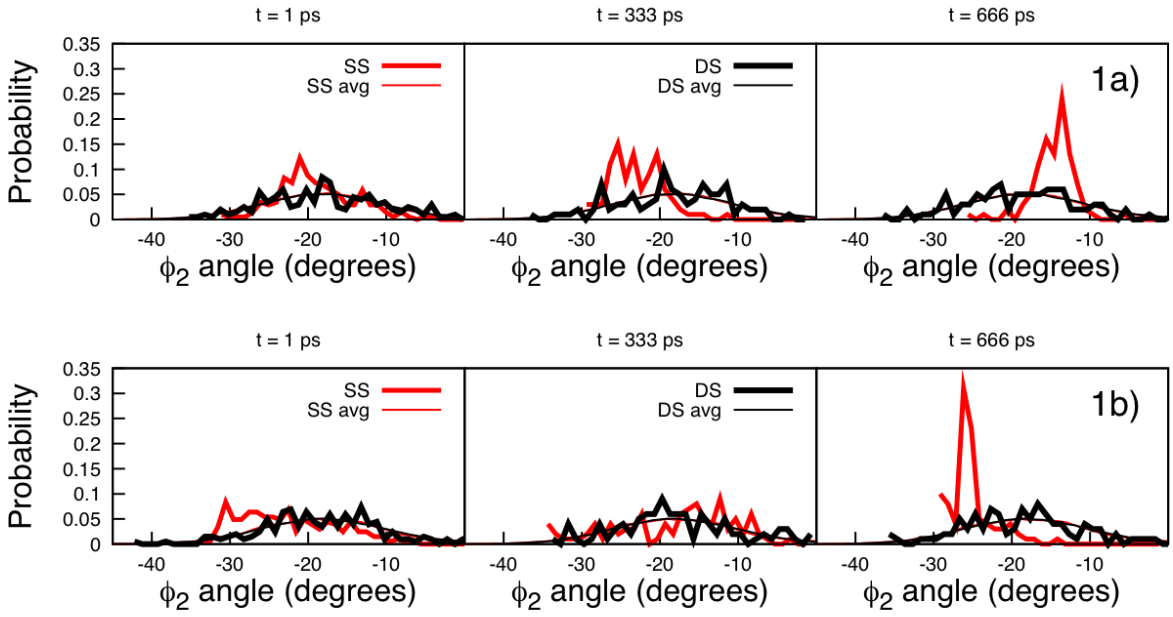
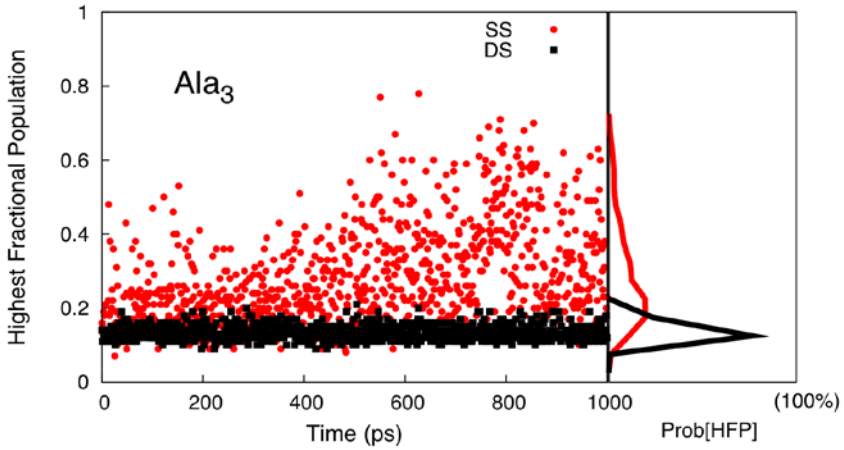
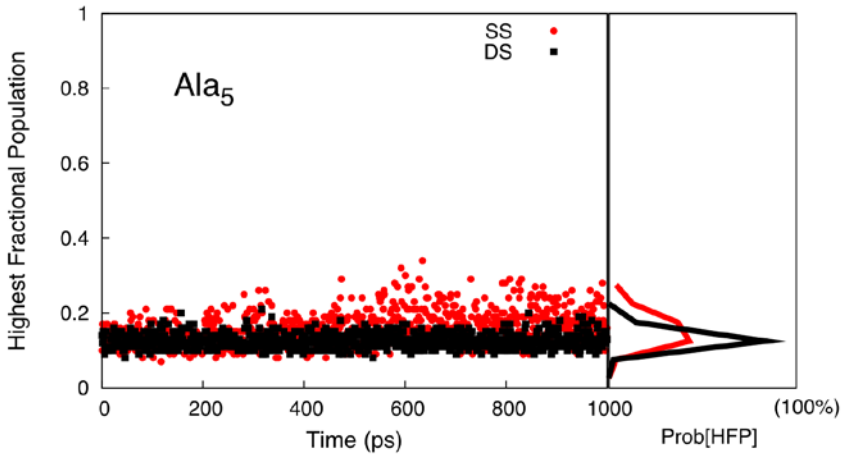


Figure 7-1A,B. Probability distributions of the dihedral angle φ_2 across sets of 100 simulations for Alanine polymer simulations. Data from SS simulations are shown in red, DS in black. Figures 1a and 1b show (Ala)3 with $\gamma = 1$ ps^{-1} and $\gamma = 50$ ps^{-1} respectively.



A



B

Figure 7-2A,B. Highest fractional population (HFP) for φ_2 in Alanine polymer simulations with a collision frequency, γ, of 50 ps^{-1}. Figures 2A and 2B show HFP for $(Ala)_3$ and $(Ala)_5$ respectively. Sideplots are the appropriate probability distribution of HFPs.

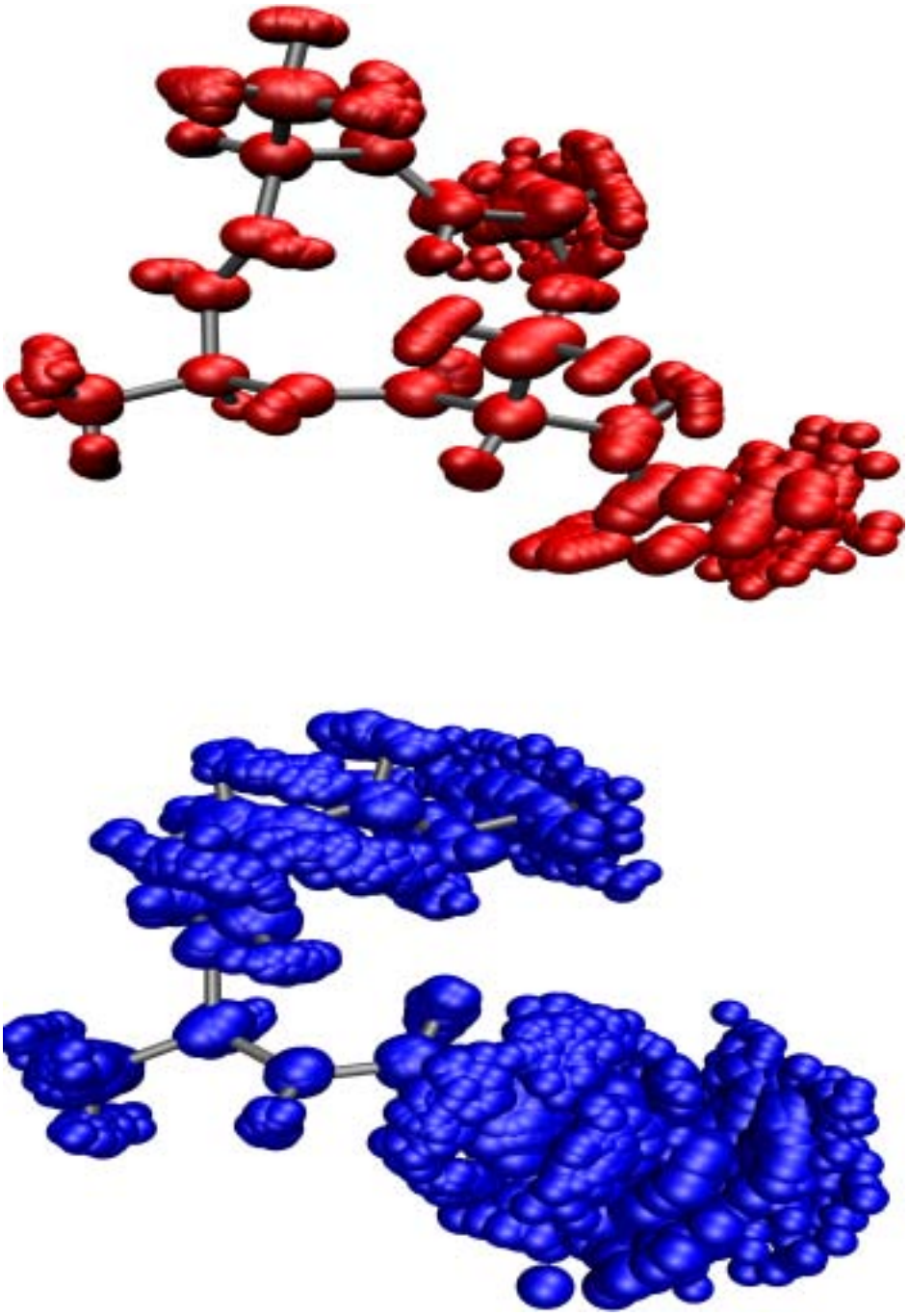


Figure 7-3. Sphere representations of simultaneous frames of 100 simulations at 836th ps of $(Ala)_3$. Red spheres represent atoms in SS simulations; Blue spheres represent atoms in DS simulation.

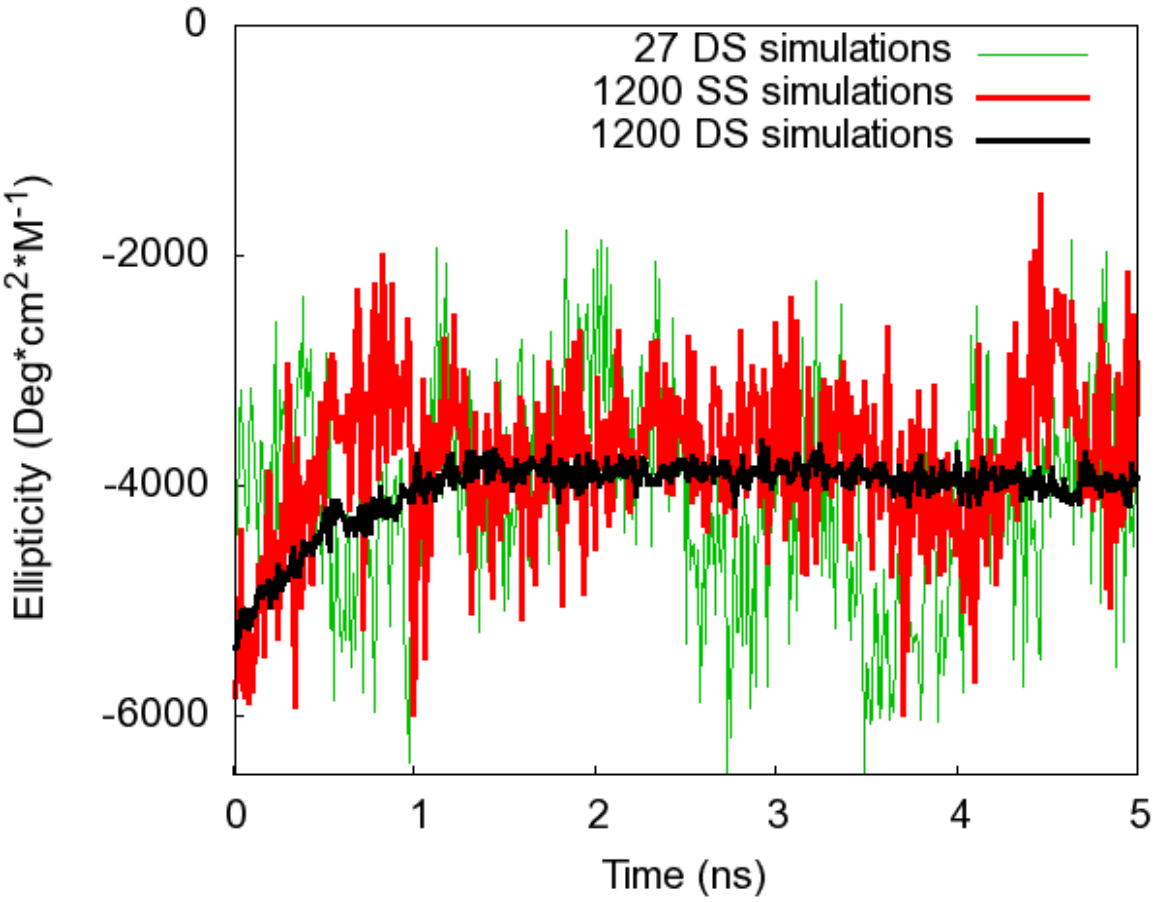


Figure 7-4. Average ellipticity vs time for a simulated T-jump experiment averaging over 1200 trajectories. SS simulations are shown in red, DS in black. The average of only 27 DS simulations is shown in green.

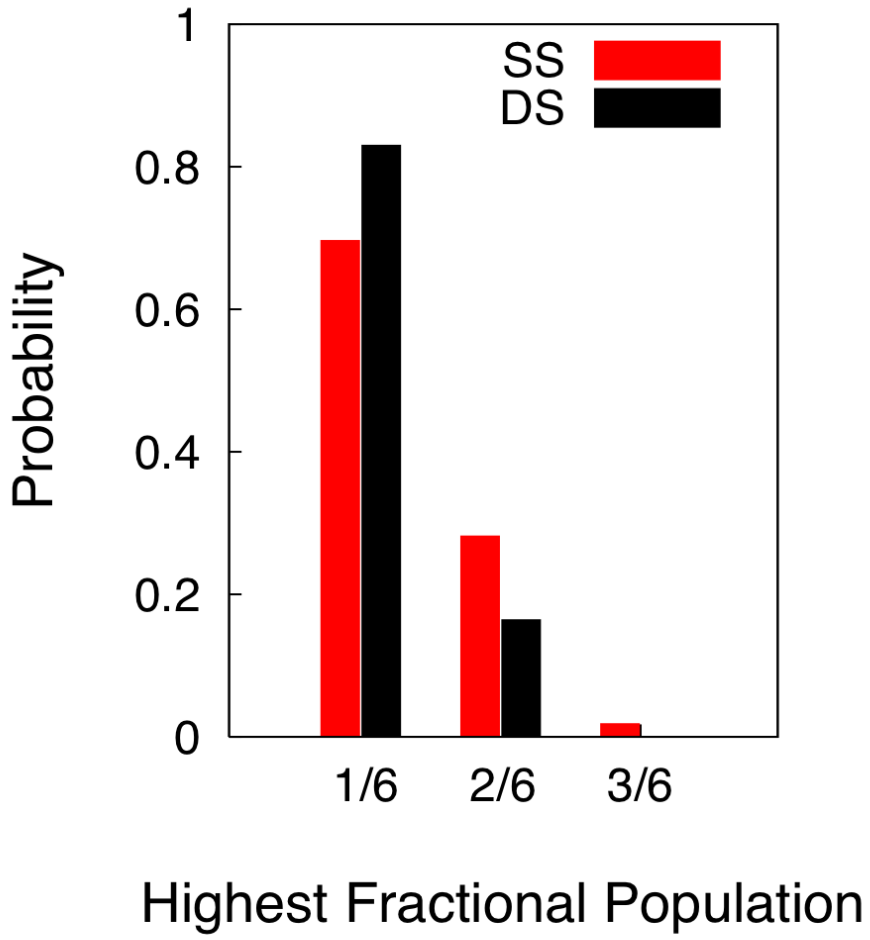


Figure 7-5. Histogram of highest fractional population for φ_2 angle in 2-degree bins for $(Ala)_3$ T-REMD simulation. SS shown in red, DS shown in black.

CHAPTER 8
MOLECULAR DYNAMICS ANALYSIS OF APO AND NICKEL-BOUND FORMS OF THE PYROCOCCUS HORIKOSHII SPECIES OF THE METALLOREGULATORY PROTEIN: NIKR

8.1 Introduction

8.1.1 Biological Function of NikR

The understanding of metal ion homeostasis at the molecular-level is critical to our understanding of how toxic metal ions are delivered safely and specifically to the appropriate location within cells.[99-101] There is a significant structural, biochemical and medical literature on the homeostasis of Cu, for example, due to its role in several neurodegenerative disorders (Alzheimer's and amyotrophic lateral sclerosis (Lou Gehrig' disease), for example).[99,100,102-106] Despite its inherent toxicity, many microorganisms require cellular nickel as an enzyme co-factor.[107,108] *E. coli* and *H. pylori*, for example, require nickel for [NiFe]-hydrogenases and urease which are necessary for survival. NikR is a DNA-binding metalloregulatory protein that utilizes nickel-responsive allostery to modulate cellular nickel concentrations in various ways.[101,109,110] In *E. coli*, NikR detects cellular nickel levels and selectively binds and represses the operon which encodes NikABCDE, an ABC-type membrane transporter[109,111] effectively reducing nickel import. In *H. pylori*, NikR, activated by high nickel levels, represses expression of nickel-dependent proteins, such as the HoxN type permease (NixA) which imports nickel, while stimulating activation of other proteins.[112] Hence, NikR senses intracellular levels of Ni and shuts downs the production of the transporter when the level of Ni is appropriate for the cell to function.[109]

8.1.2 Allosteric Mechanism

It has been found that NikR's affinity for DNA is dependent on nickel concentration[113]. NikR contains both low (30 μM) and high-affinity (2 pM) nickel-binding sites; when the high

(low) affinity nickel sites become occupied, DNA affinity increases to nanomolar (picomolar) levels[113]. Carrington *et al.* found that nickel binds to the high affinity sites in a unique square-planar geometry, but changes to 6-coordinate when bound to DNA.[114] Several species of NikR have been crystallized in various forms by Drennan and co-workers (*E. coli* NikR – EcNikR)[115,116] Terradot and co-workers (*H. pylori* NikR - HpNikR)[117] and Chivers and Tahirov (*P. horikoshii* NikR - PhNikR).[118] The overall structure consistently shows NikR to have a central tetrameric metal-binding domain (MBD) constructed of a βαββαβ ferredoxin-like fold binding nickel in square-planar geometry surrounded by two dimeric ribbon-helix-helix (βαα) wing-like DNA binding domains (DBDs) (see Figure 8-1).

There are, however, some inconsistencies between different species' crystal structures. The relatively stable MBD of NikR and wing-like DBDs of NikR can adapt roughly three different main conformations (straight, trans or cis) whose stability seems to be related to both nickel and DNA binding. We will refer to the different forms by the presence of nickel, A (apo) for nickel-free, H (holo) for nickel-bound, and the global conformation, T for trans-like, C for cis-like, and S for straight. Figure 8-2 shows sketches of various forms of NikR. Schreiter *et al.* crystallized EcNikR in AS, HS and HC (DNA-bound) forms.[115,116,119] They found that only the nickel-bound forms had a well-ordered α_3 helix (a helix containing one of the four nickel-binding residues). This lead to the proposition that the stabilization of the α_3 helix assists in DNA binding, a hypothesis supported by comparative metal binding studies.[120] Chivers and Tahirov crystallized five forms of *P. horikoshii* including AS, HS, and three forms of HT: one with the four high affinity nickel-binding sites occupied and PhNikRs unique additional four auxilliary high affinity sites bound, one with all high and low sites bound, and one with all sites bound in the presence of free phosphate ions (to simulate DNA).[118] The AS form, similar to EcNikR, was found to have

a highly mobile and disordered α_3-helix. Interestingly the form with phosphate ions does not fit into the corresponding DNA operon, but does if one half (any two chains that share a wing) is rotated by 180 degrees about the center axis (turning it from a trans to a cis). The crystal forms found by Dian *et al*. of HpNikR were the most unique set of the three species crystallized.[117] They found a unique AT structure, which roughly resembles the HT form of PhNikR but lacks dimeric symmetry. Also, HpNikR was never found with all four high-affinity nickel-binding sites occupied, but rather with a single high-affinity nickel site bound per dimer, an intermediate site, or an external site.

These various permutations of conformations and nickel and DNA bound states both provide insight and raise questions about NikR's allosteric mechanism. What role does nickel binding have in the structural preference of NikR? Does nickel-binding preferentially stabilize one conformation over another? Or does nickel-binding stimulate conformational changes?

To elucidate the role of nickel binding on conformational states and dynamics of NikR we simulated various forms using all-atom molecular dynamics (MD) simulations. MD simulations of biological molecules have often been utilized to elucidate biological mechanisms. Few computational studies of any NikR have been reported. Cui and Merz utilized a course-grained Gaussian-network-model to find natural global fluctuations in PhNikR.[121] They found that natural fluctuations would lead to transitions between the major conformational states found in NikR crystal structures (between straight, trans, and cis). This coarse-grained study supports the pre-existing equilibrium hypothesis – that all the available conformational states are available to each bound state. Bradley et al. utilized[122] an 80 ns all-atom explicit solvent simulation of EcNikR starting from the AT state (PDB code 1Q5V). In that work, a potential interaction network between the metal-binding and DNA binding residues was highlighted using a novel

correlation-matrix-based residue clustering technique.[122] However, the metal binding states we not examined.

Herein, we examine the dynamics of three 100 ns simulations beginning from three forms of PhNikR. The Ph species was chosen because the crystal structures available had the least amount of missing atoms. We chose the AS, HT (with only four nickels bound) and AT forms (PDB codes: 2BJ3, 2BJ7 (with 4 primary nickel ions), 2BJ7 (without nickel ions)). Our goal is to highlight the features of the dynamics of each form to compare and contrast any apparent differences between them. The analyses reported herein include: RMSD analysis to look at conformational changes of various regions of interest, center-of-mass angular analysis to measure global conformational change, correlation analysis to examine inter-residue communication, and fluctuation analysis to quantify residue mobility.

8.2 Methods

8.2.1 Simulation details

We chose to simulate three forms of PhNikR[118]: the apo-straight form, taken from PDB ID 2BJ3 (PhNikR-AS), the holo-trans form with only four nickels simulated, taken from PDB ID 2BJ7 (PhNikR-HT), and an apo-trans from where all nickels were stripped from the original pdb structure 2BJ7 (PhNikR-AT). An implicit generalized Born solvation model ($GB^{(OBC)}$)[59] was chosen to reduce the system size as well as the solvent viscosity to allow for faster kinetics.[123] Recently, Hornak et al. successfully utilized implicit solvation to witness flap-motion in a simulation of HIV-1 protease.[124,125] In our simulations, each form was taken from the appropriate PDB structure with missing residues built by the TLEAP program which is part of the AMBER suite.[57] The PhNikR-HT form required additional modeling in order to properly represent the nickel centers (see the "Force Field Parameters" section). The proteins were simulated in AMBER using the ff99SB parameter set.[58] Each form was minimized in AMBER10 using the

SANDER module, then relaxed at 100K for 10 ps. The production runs were 100 ns long at 300K for each form using the PMEMD module. Analysis of the trajectories was performed using the PTRAJ module.

8.2.2 Force Field Parameters

All 138x4 of the amino acid residues in Ph apo-NikR occur naturally and thus force field parameters are available in various parameter sets (ff99SB in our case[58]). For the apo forms, all Histidine protonation states were chosen to agree with the most likely protonation state at the experimental pH of 7.5 and are thus delta-protonated. Since in the apo form no Histidines are hydrogen bound any other residues, the difference between the delta and epsilon (the second most likely state) protonation will have a limited impact, if at all, to global dynamics.

For the nickel-bound form, there are several approaches to metal modeling that we could have chosen. These approaches can roughly be classified into three categories: quantum mechanical, non-bonded, and bonded. Quantum mechanical treatment or a mixed QM/MM treatment of metals has the most potential for accuracy, but its huge computational cost deters usage for large systems. A non-bonded approach would treat the nickel as a free ion whose interaction with the protein is strictly through Van der Waals and Coulomb forces. This approach is delicate to parameterize, and does not enforce a particular geometry.[126 127-130] Finally, a bonded approach utilizes all terms in the typical all-atom force field to treat the metal ion. Parameterization of this approach can be done by combination of experimental and quantum mechanical methods. This method is the most restrictive geometrically, but is useful when modeling stable metal sites. We chose the bonded approach since NikR is too large to treat quantum mechanically, and we know precisely what the Ni coordination state is in NikR.

New force field parameters were created and used for both the bound nickel ions, and the ligating residues. The MTK++ package was used to expedite the force field parameterization

process.[131] First, a single copy of one of the nickels plus the four ligating sidechains were isolated from the holo structure (pdb code: 2BJ7) then capped with CH_3 groups (in the position of the original C_α). Figures 8-3A,B show representations of this "small" metal-cluster model. The notation H87A, for example, indicates a histidine (H) residue that is sequentially 87th for chain A.

The Histidine protonation states were taken from experiment: H89A and H78D are epsilon-protonated, and H87A is delta-protonated with the remaining sites occupied by the Nickel ion. The resultant cluster was then minimized using the TPSSTPSS/TZVP level of DFT theory with Gaussian.[132] This method and basis has been shown to model nickel compounds well.[133] The minimization did not disrupt the square-planar geometry (though use of smaller basis sets did, data not shown). Following the minimization, force constant calculations were performed using the same methods.

To obtain the partial charges on all the atoms, a larger cluster was selected from the pdb file, this time including backbone atoms, but capped with Acetyl and N-Methyl groups in appropriate positions. Figures 8-4A,B show representations of this "large" cluster.

Charges were obtained using the RESP algorithm[134-141] on this model cluster. Van der Waals parameters were not adjusted for the ligating residues. For the Nickel, thermodynamic-integration simulations were performed for the alchemical reaction (desolvation): $Ni^{2+} + 2Cl^- + nH_2O \rightarrow nH_2O$ for various ε and *r* values for nickel. The combination of *r* and ε that most closely resulted in the experimental free energy of hydration of the ions was used for nickel.[142] The new set of parameters for the nickel ions and ligating residues were then mapped to the four primary nickel binding sites in the holo form of NikR. The ff99SB parameter set[58] was used for all other residues. Note that although no restraints were used to maintain the planarity of the

Nickel center the planarity was maintained throughout the course of the simulations described below (data not shown).

Table 8-1 shows the partial charges of the nickel center along with those of the nickel-bound atoms. For the nickel, we found a van der Waals radius of 1.10Å and a well-depth of 0.013 kcal/mol. The full set of parameters are listed in the appendix.

8.3 Results and Discussion

8.3.1 Energetic Analysis

To evaluate the energetic relaxation of the simulations, the potential energy of each system was examined. Figure 8-5 shows a block-averaged time-series of the potential energy. AS refers to the straight nickel-free structure (pdb ID 2BJ3), AT the trans nickel-free structure (pdb ID 2BJ7), and HT is the nickel-bound trans structure (pdb ID 2BJ7).

The potential energy for all simulations had standard deviations below 80 kcal/mol for the last 90% of the simulation. This is consistent with the fluctuation in the simulation by Bradley *et al.*[122] whose potential energy fluctuated by about 200 kcal/mol. The potential energy drops off slightly for the first 10 ns. After that, all three simulations fluctuate about each of their respective averages. Thus for the final 90 ns, each simulation is likely in energetic equilibrium.

8.3.2 Domain Center-of-Mass Angle Motion

In order to get a general sense of the global conformational changes, (between straight, trans, and cis forms), we looked at the angles between centers of mass between the centers of mass (COMs) of the four major domains: DBD from chains A and B, MBD from AB, MBD from CD, and DBD from CD. The following figures show the angles between both DBD-AB, MBD-AB, and MBD-CD (see Figure 8-6A) and MBD-AB, MBD-CD, and DBDCD (see Figure 8-6B) for all three forms. The AS form, clearly maintains angles just below 180 degrees, indicating that the COMs of the four domains are roughly linear. Conversely, the AT and HT

forms maintain an angle of 150 degrees or less indicating a bend in the structure. While fluctuations exist, no major conformational changes are evident from this analysis other than the HT CD wing, which drifts near 160 degrees. This could indicate that this wing is "straightening out".

To get a sense of the simulations dynamics between "cis" and "trans" states, the dihedral angle between all four major domains was measured. An angle near 180 corresponds to a "trans-like" state, near zero is "cis-like". The following plot shows the dihedral angle between these four points for each simulation. Unsurprisingly, the AS domain-COM dihedral angle fluctuates wildly across all angles (since it is nearly linear and thus small motions can cause large angular changes). Conversely, the AT and HT, both starting at less than -150 degrees, slowly drift up throughout the simulation. This could be simply relaxation, or more significantly, a slow conformational change to the cis form.

8.3.3 RMSD Analysis

In order to get a qualitative glimpse into the motion of all three simulations, a RMSD analysis was performed. RMSD analysis provides insight into global structural changes and the relative stability of a molecular trajectory. Figure 8-8 shows the C_α RMSD against the initial structure *versus* time for all three simulations (AS, AT, and HT). All three simulations deviate more than 2Å within the first nanosecond. Though none of the simulations are stable over the entire 100 ns of simulation, they each maintain a steady RMSD between transitions. Both of the simulations whose initial structures came from forms found *via* crystallography (AS and HT) maintain states for nearly 60 ns. The AT simulation only stays steady for ~40 ns and experiences the highest RMSD change. The magnitude of fluctuations is similar to that of the 80 ns explicit solvent *E. Coli* simulation of Bradley et al.[122] whose backbone RMSD fluctuated between 2.5

and 5Å after relaxation. The greater fluctuation observed in the simulations herein is likely due to the lack of explicit solvent molecules in the GB simulation.

Figure 8-9 shows the C_α RMSD against the first frame for only the metal binding domain of the protein (Residues 57 to 131 of each chain). Both the AS and HT simulations jump initially and stay in a steady state between 1.5 and 2Å from the initial structure. The AT simulation jumps initially to 3Å then rises to 4Å between the 40^{th} and 60^{th} ns. This larger jump may be due to the instability of the initial nickel-free AT structure. It is likely that rearrangement mostly occurs to adapt to the nickel-free form.

Figures 8-10A,B show the RMSD vs initial structure for each of the two DBDs: the AB side and the CD side. We refer to chains A, B, C, and D such that the A and B chains are separated from the C and D chains by a plane through the center of the MBD (the plane essentially contains all four nickel ions). Here the RMSD is minimized with respect to the MBD in order to isolate the DBD motion. In each simulation, RMSDs above 19Å is achieved. Most strikingly, there is no apparent correlation between the behaviors of either "wing" to its dimeric counterpart. Though not shown the RMSD was measured for the wings with a fit to themselves. The RMSD was less than 5Å for all simulations save for the CD wing of the AT simulation that went just above 5Å at around 50ns and then remained steady for the rest of the simulation. This dramatic difference in RMSDs depending on the fit suggests that the wings largely exhibit rigid body (wing-like) motion.

Figures 8-11A-D show RMSD against initial structure of the C terminal domains of each chain. X-ray data show that each symmetry-related half contains one free loop (A and D chains) and one ion-stabilized loop (B and C). Interestingly, in the AS form, the loop is stabilized by a negative chloride ion, and in the holo – trans form, it is stabilized by a positively charged nickel

ion. For the simulations, neither ion was included in the model. The only loop that fluctuates more than 2Å is the A-loop for the AS, and AT simulations. There seems to be very little correlation between either the different loops of the same simulation or the same loop for different simulations. This suggests that the loop motion is independent of conformation and the presence of the four primary nickel ions.

There is experimental evidence that the stabilization of the α_3-helix upon nickel ion binding leads to an increase in DNA affinity.[115,116,119,143] This helix, part of the MBD, contains one residue that binds a primary nickel (see Figure 8-1). There was little resolution in this region in x-ray crystal structures whenever the nickel ion was not bound. We performed RMSD analysis on the α_3-helix to see how the difference in stability between the different forms (see Figure 8-12). Both the AS and HT simulations maintain relatively low RMSD, with holo-trans being the most stable. The AT simulation jumps to around 3.5Å and stays relatively stable throughout the simulation. Contrary to the hypothesis of nickel stabilization of the α_3-helix, both the AS and HT forms behave similarly. More extended simulations may reveal fluctuation of the helix in the AS form at a time-scale larger than 100 ns.

Schreiter *et al.* also found loops in the MBD that make noncovalent contacts with DNA.[116] The RMSD of these loops (residues 64-67) are shown in Figures 8-13A-D. Similar to the C-terminal loop, in the absence of DNA and ions other than the primary nickel ions, no correlation can be seen between the three forms in this region of the protein.

8.3.4 Correlation Analysis

To elucidate inter-residue dynamical relationships, correlation analysis was performed on all C_α atoms for each simulation with the ptraj module from AMBER10 (see Equation 8-1).

$$C(i,j) = \frac{\left\langle \Delta x_i(t_k) \cdot \Delta x_i(t_k) \right\rangle}{\left\langle \Delta x_i(t_k)^2 \right\rangle^{1/2} \left\langle \Delta x_j(t_k)^2 \right\rangle^{1/2}} \quad (8\text{-}1)$$

Figures 8-14A-C show the correlation matrices with the averages performed over the entire simulation for each form. Qualitatively, the appearance of small (~50 residue) medium (~75) and large (~140) residue-large squares shows correlation within the DBD, MBD and entire chains respectively. Off-block-diagonal terms represent relations between different chains.

To evaluate the stability of the correlations, the correlation of the 10^{th}-45^{th} ns was subtracted from the 45^{th} to 80^{th} ns of the same simulation (each correlation matrix element was subtracted from its equivalent counterpart. These correlation differences are shown in Figures 8-15A-C on a scale (-2:2) double the original (-1:1). A value of +2 indicates that the interaction went from being anti-correlated to correlated. The graphs indicate no significant change in correlation between the first and second halves of the simulation. For the AS, AT, and HT simulations there were 2, 14, and 0 changes in correlation greater in magnitude than 1.

To compare the correlation between different forms of NikR, the correlation matrix elements were subtracted from each other in all three possible unique permutations as shown in the following figures. Figures 16A-C show the correlation differences, $C_{ij}^{AS} - C_{ij}^{AT}$, $C_{ij}^{AS} - C_{ij}^{HT}$, and $C_{ij}^{HT} - C_{ij}^{AT}$ respectively. For each comparison, the magnitude of the correlation difference was greater than 1 a total of 128, 150, and 31 times for the $C_{ij}^{AS} - C_{ij}^{AT}$, $C_{ij}^{AS} - C_{ij}^{HT}$, and $C_{ij}^{HT} - C_{ij}^{AT}$ respectively. Interestingly this suggests that despite the difference in behavior between HT and AT (which began from the same structure) based on RMSD analysis, the inter-residue correlations are very similar.

In order to rank the relevance of specific residues, we used a measure which sums the absolute value of the correlation along either a row or column for each residue:

$$REL_i = \sum_{j=1}^{N} \left| c_{i,j} \right| \tag{8-2}$$

Here $c_{i,j}$ is the correlation coefficient between the i^{th} and j^{th} residue and N is the total number of residues. This is plotted in Figure 8-17A. To get a better qualitative assessment of the importance of residues, the next plot is the same but summed over chain id as in Equation 8-3.

$$cREL_i = \sum_{chains} \sum_{j=1}^{M} \left| c_{i,j} \right| \tag{8-3}$$

Here *M* is the number of residues per chain. Figure 8-17B shows the correlation relevance summed over chains. The colored area just above the x-axis indicates biological information. The light gray region is the DBD, the dark gray is the MBD. Bright and faded yellow bars are completely and partially conserved residues among different NikR species respectively. The purple bars represent residues known to bind directly to the major groove of DNA. The pink bars are the residues that bind directly to nickel. The graphs indicate that each form exhibits similar behavior, residue-to-residue, and that the residues of AS and HT forms are much more coupled. Interestingly, for most of the DBD (approximately first 38 residues) for these two forms, the chain-independent correlation relevance are nearly identical. In the MBD, the AS is more correlated than the HT form (and even more so than the AT).

Information can be extracted for not only which residues affect the whole structure, but which affect other domains as well (Relevance for DBD residues affecting MBD residues and vice-versa). This "cross-domain" relevance can qualitatively determine "allosterically important" residues. The cross domain relevance can be represented by Equations 8-4 and 8-5.

$$xDREL_i = \sum_{j=1}^{N} \left| c_{i,j} \right| \cdot \frac{\xi(i,j)}{n(D(j))} \tag{8-4}$$

$$\xi(i,j)=\begin{Bmatrix}0 & D_i = D_j \\ 1 & otherwise\end{Bmatrix} \tag{8-5}$$

Here, D_i is the domain of residue *i* (DBD or MBD) and $n(D(j))$ is the number of residues in domain type D. These values were plotted either by residue or summed over chains in Figures 8-18A,B with linking residues removed. Similar to the previous chain-independent graph, the AS and HT have nearly identical behavior in the DBD. The MBD regions act more similarly between AS and HT forms. This could indicate that although the MBD of AS form is much more coupled than HT form, the cross-domain interactions are similar. It appears as though both the nickel-binding and DNA-binding residues have strong interdomain interactions. This is could be a result of the allosteric mechanism. These residues, marked purple and pink on the x-axis, have the highest cross-domain correlation.

It is likely that in this allosteric system, the "biologically important" residues indicated by the colored bars at the bottom of Figures 8-17A,B and 8-18A,B will have a strong impact on cross-domain correlation. Table 8-2 shows the averaged cross-domain correlation relevance for each form for different sets of biologically important residues. As the criterion for biological importance becomes more strict, the average cross-domain correlation increases.

8.3.5 RMS Fluctuations

To observe the stability of the protein, RMS fluctuations of the C_α atom of each residue of the protein were measured for each form. Figures 8-19A-D show the fluctuations by residue of each chain (A,B,C,D). The figures indicate, not unexpectedly, that the DBDs (wings) fluctuate much more wildly than the MBDs. As seen in the RMSD plots, the two experimentally observed forms (AS and HT) behave similarly, especially in the MBD. The AT form seems much less stable, especially in the MBD. The AB wing (DBD) of the AS form seems to fluctuate much more wildly than its dimeric counterpart (the CD wing).

8.4 Conclusions

NikR regulates cellular nickel concentration by responding to high nickel concentrations by suppressing nickel transport protein expression. Experimental data gives clues to its mechanism but without conclusive detail. We have performed three simulations to help elucidate the dynamics of NikR. The modeling of this system required special parameterization to accurately represent the nickel and its ligating residues. Our analyses of these simulations provide additional insight into the dynamical preference that different initial conformations and nickel-binding states have on the protein. While not observing global conformational changes on the computational time-scale available, several distinct features of NikR were observed. The wing-like DBDs move independently of each other. The AT form is less stable than the other two, especially in the region of the MBD. Surprisingly, and contrary to prior belief, the alpha-3 helix is not dramatically more stable in the nickel-bound form (HT) than the apo (AS) form. Biologically relevant residues tend to influence motion across domains much more strongly than others residues. It was found that the fluctuation of AS and HT are very similar in the MBD, but significantly different in the DBD.

With current computational capacity, simulatating a global conformational change in a system like NikR is a significant challenge. Beyond brute force simulations, it is possible that advanced sampling algorithms could be applied to not only establish a meaningful reaction coordinate between NikR states, but also to calculate the free-energy change between states.

Table 8-1. Partial charges for the nickel center along with each ligating atom. Though not shown in the main text, the partial charges were created for each entire ligating residue (see Appendix).

Ni^{2+}	Cys 97 S	H87 NE (HID)	H89 ND (HIE)	H78D NE (HID)
0.66275	-0.52432	-0.66101	0.00690	-0.16857

Table 8-2. Average Cross-domain Correlation relevance for various forms of NikR. The column marked all considers all residues for which the cross-domain relevance was considered. B is for either nickel or DNA binding residues, F is for fully conserved residues, and M is for mostly conserved residues.

	All	B,F,M	B,F	B
AS	2.22	2.59	2.65	3.32
AT	1.44	1.62	1.67	2.00
HT	1.97	2.35	2.38	2.83

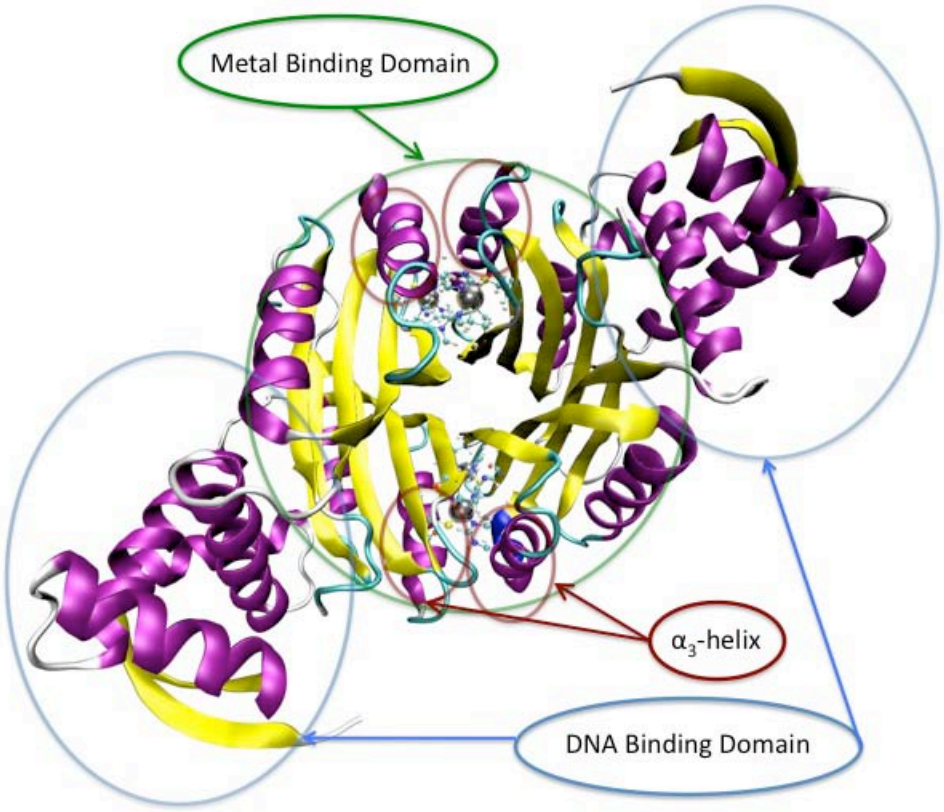


Figure 8-1. A rendering of PhNikR. The DNA binding domains surround the central metal binding domain. Metallic spheres represent bound nickels.

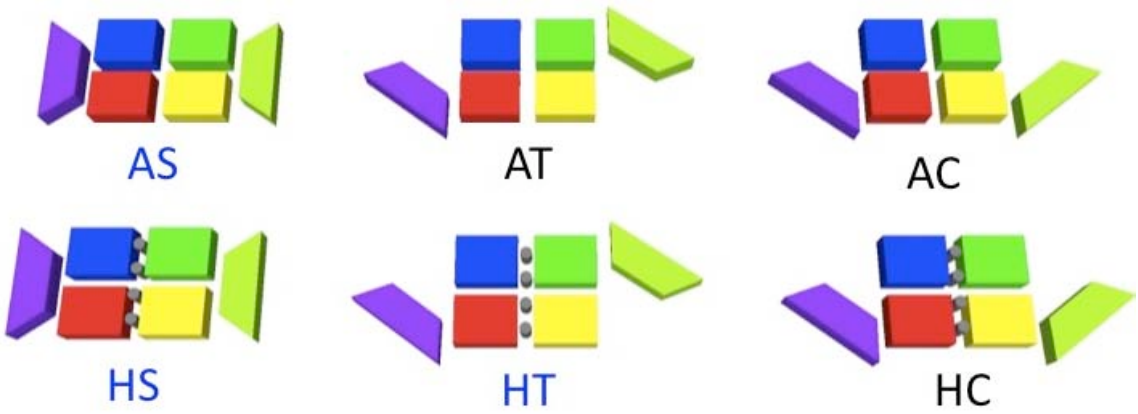
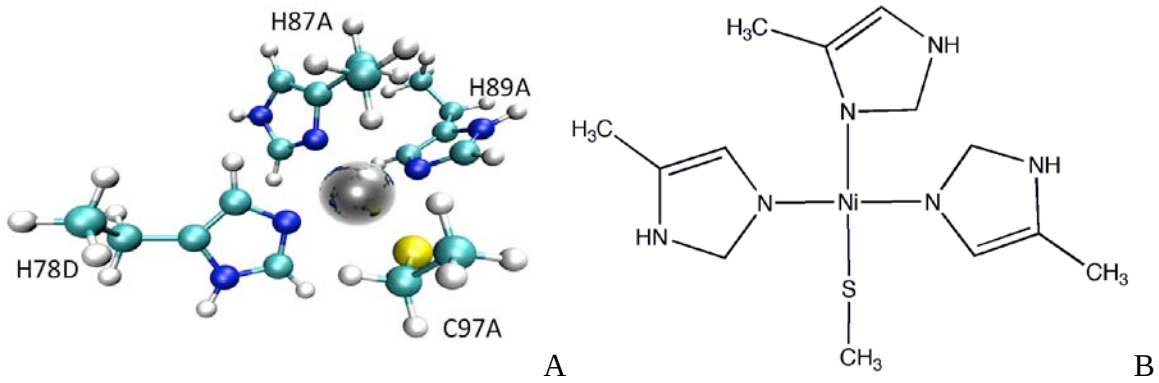
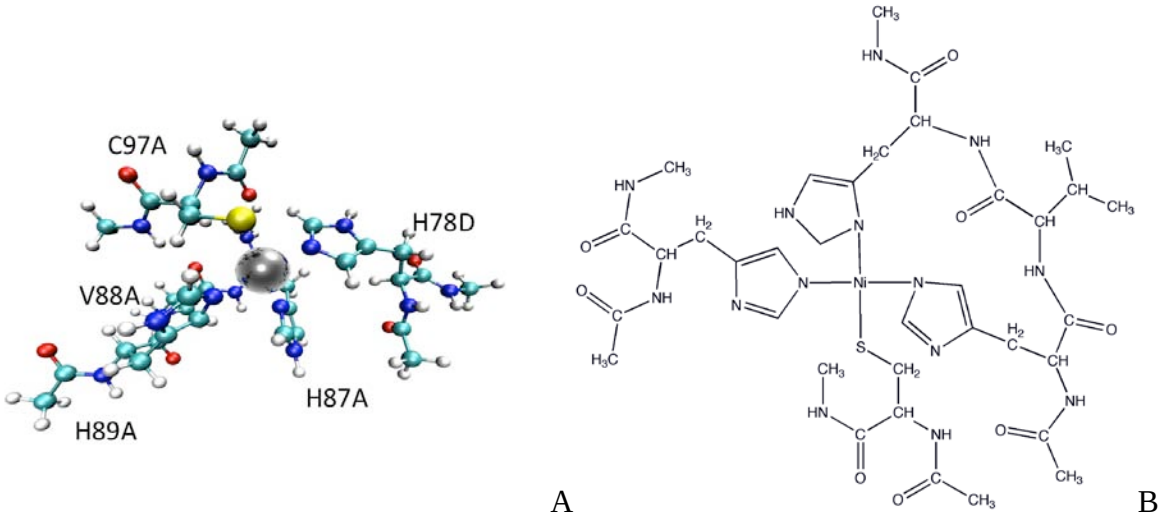


Figure 8-2. Sketch of various possible forms of NikR. Rectangles represent metal binding domains (each color represents a different chain, purple and lime represent intertwined chains). Trapezoids represent the wing-like DNA binding domains (each has contributions from two chains). Grey circles represent primary nickel binding sites. Only the three with names written in blue were found experimentally in the absence of DNA. In the cis form, NikRs DBDs (wings) snugly fit into the major groove of DNA.



Figures 8-3A,B. Representations of the "small" metal cluster. Atoms included in this model were used to obtain force constants via DFT calculations. A) A ball-and-stick representation. Labels indicate residue identity and chain ID B) A 2D representation.



Figures 8-4A,B. Representations of the “large” metal cluster. Atoms included in this model were used to obtain partial atomic charges via DFT. A) A ball-and-stick representation. B) A 2D representation.

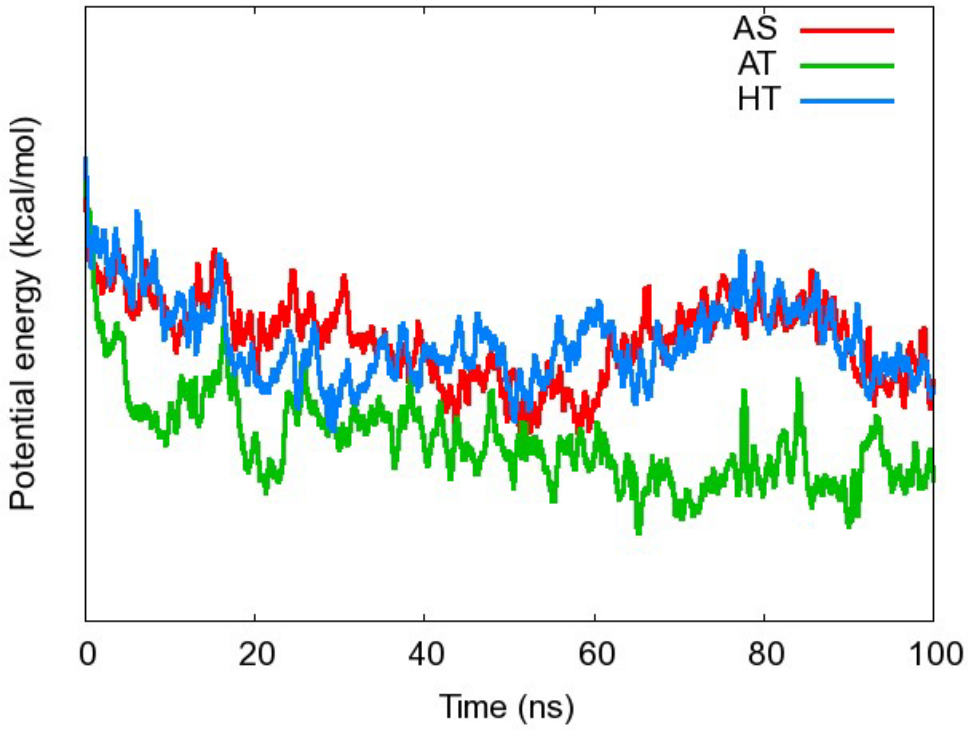
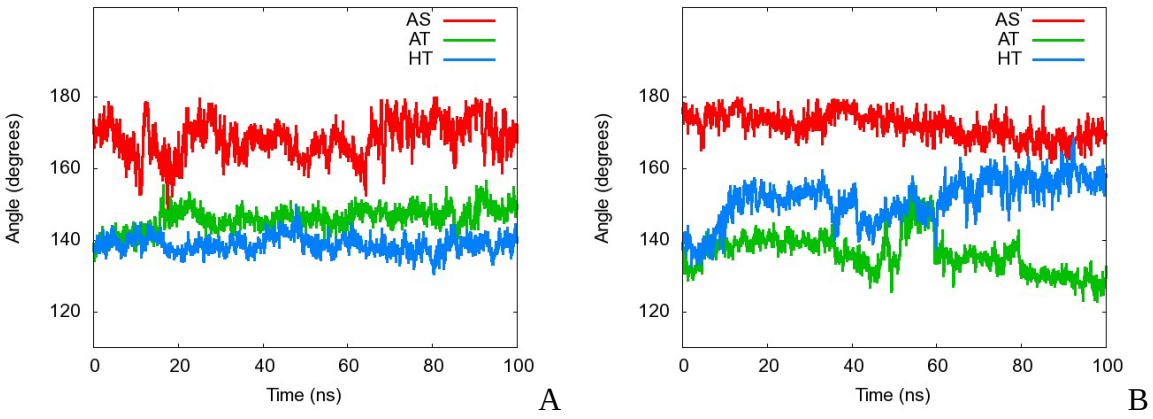


Figure 8-5. Averaged potential energy time series for three simulations of NikR. The smoothing window size is 50 ps.



Figures 8-6A,B. Angles between centers of mass of adjacent domains. A) Angle made by DBD-AB, MBD-AB, and MBD-CD. B) Angle made by MBD-AB, MBD-CD, DBD-CD.

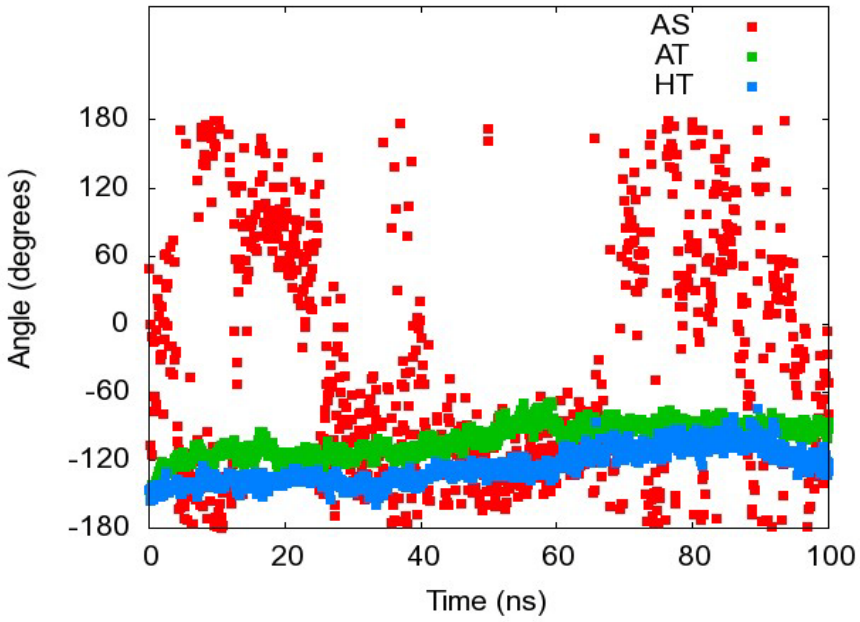


Figure 8-7. Dihedral angle between the centers of mass of the DBD-AB, MBD-AB, MBD-CD, and DBD-CD domains vs time.

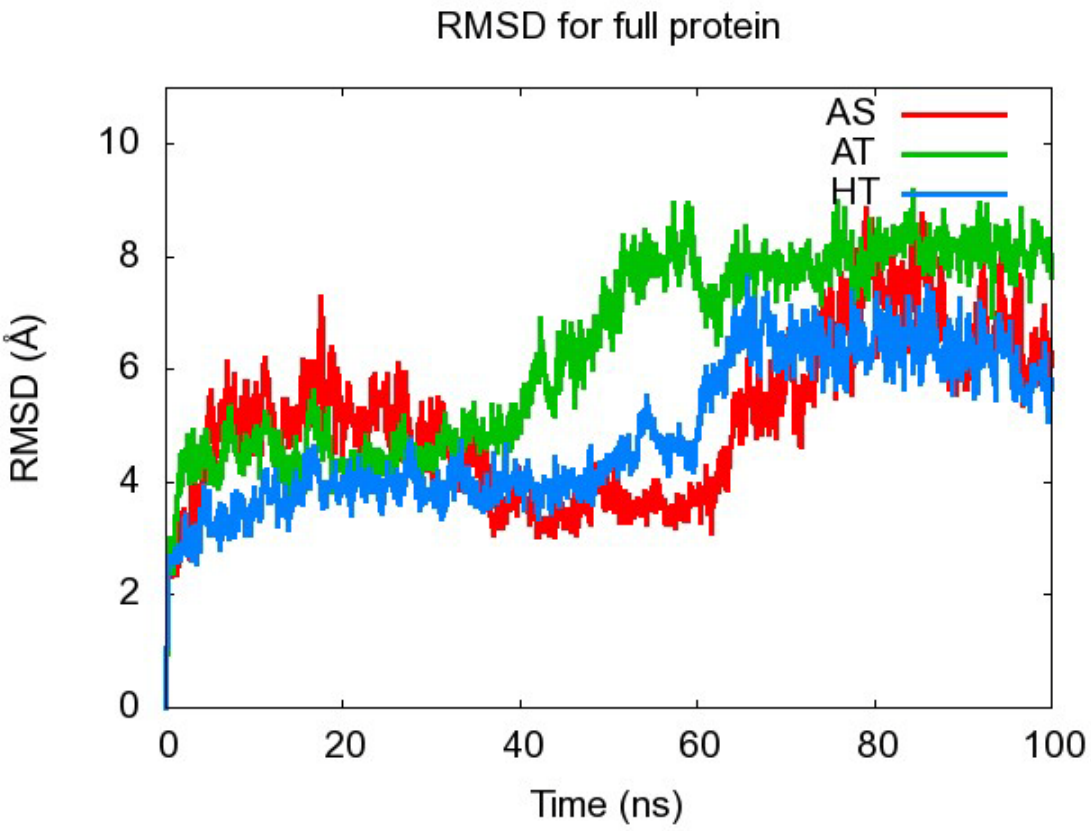


Figure 8-8. C_α RMSD for entire protein in the AS, AT and HS simulations.

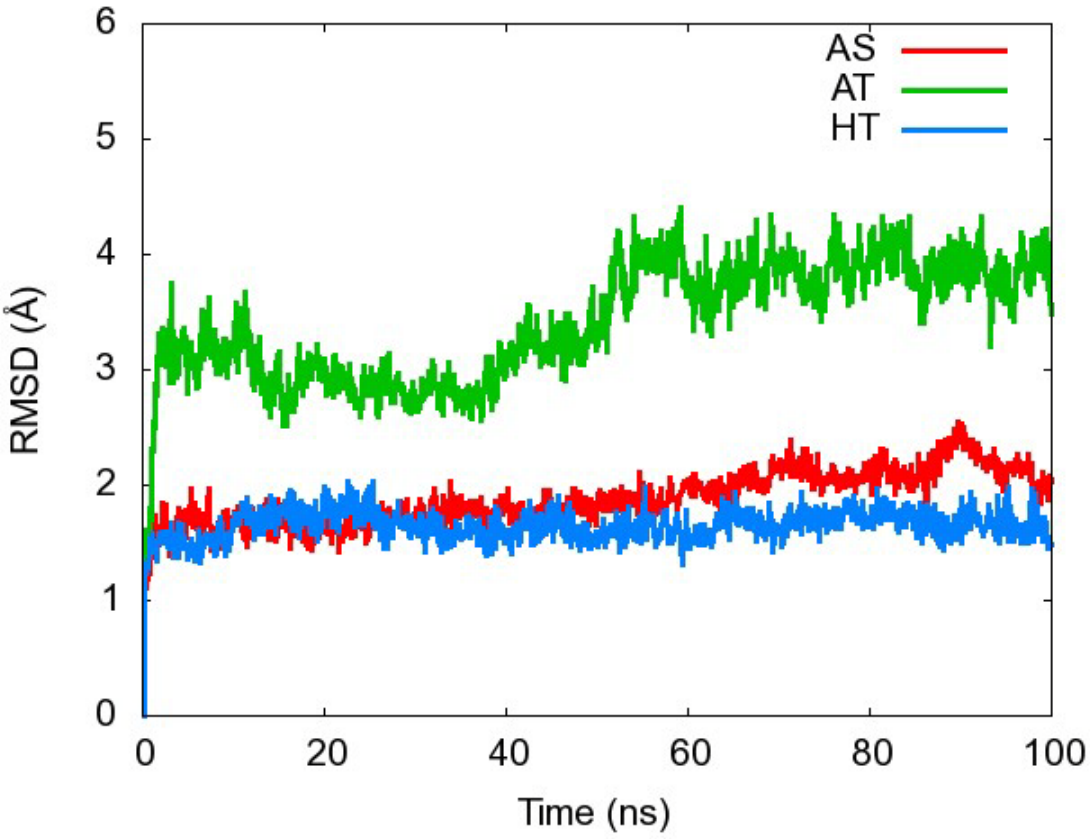


Figure 8-9. C_α RMSD for the metal binding domain for all three simulations.

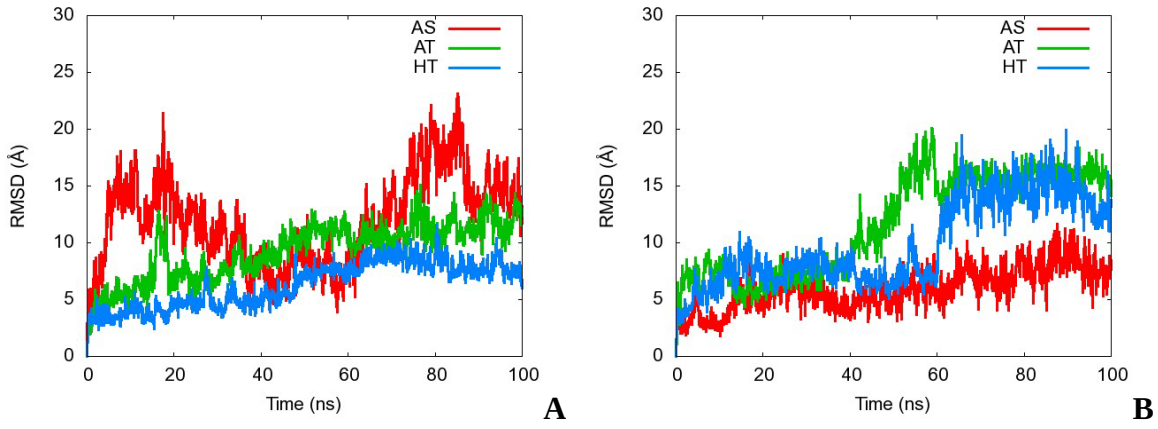


Figure 8-10A,B. C_α RMSD for DBD "wings" from each simulation with a fit to the MBD rather than the DBDs themselves.

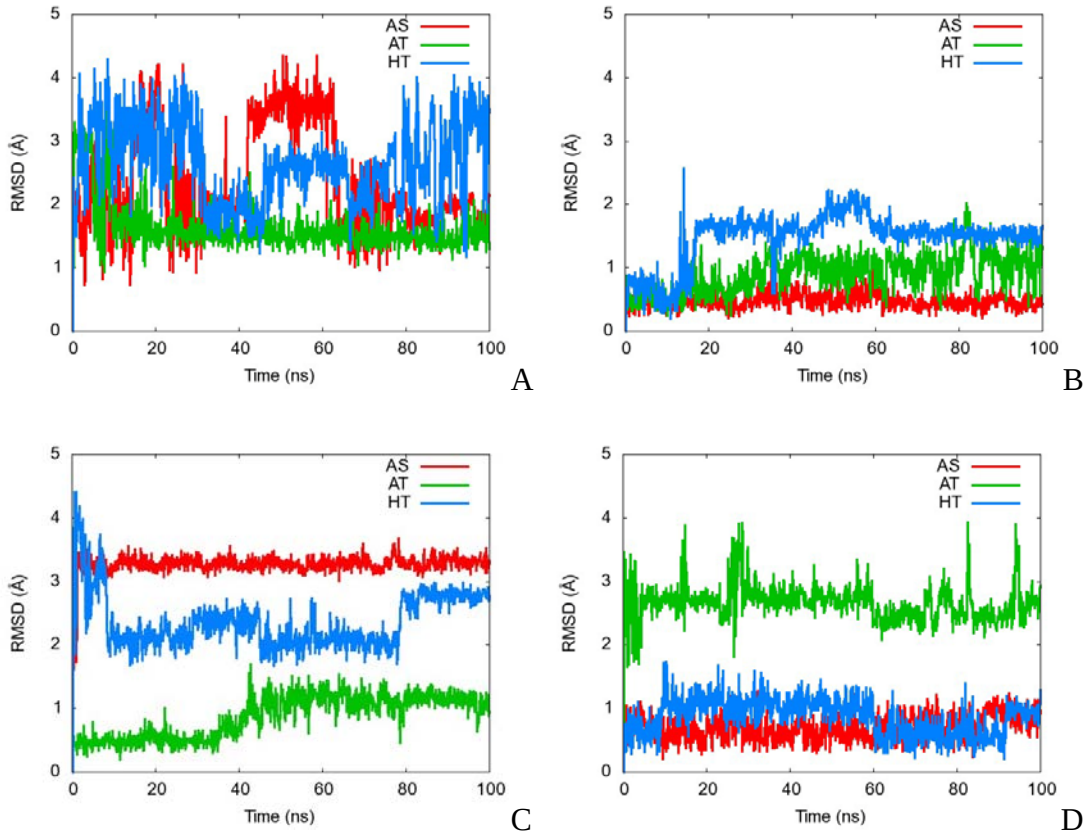


Figure 8-11A-D. C_α RMSD for the C-terminal loops for all simulations. Figures A-D are for chains A,B,C and D respectively.

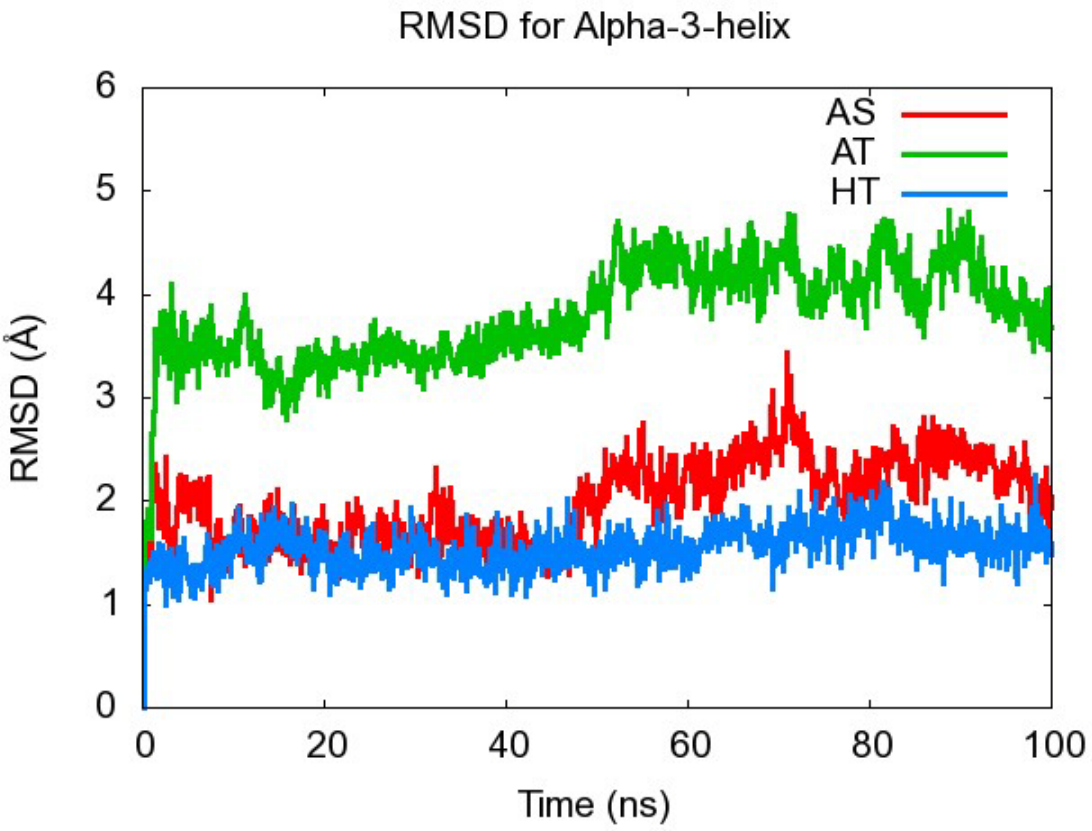
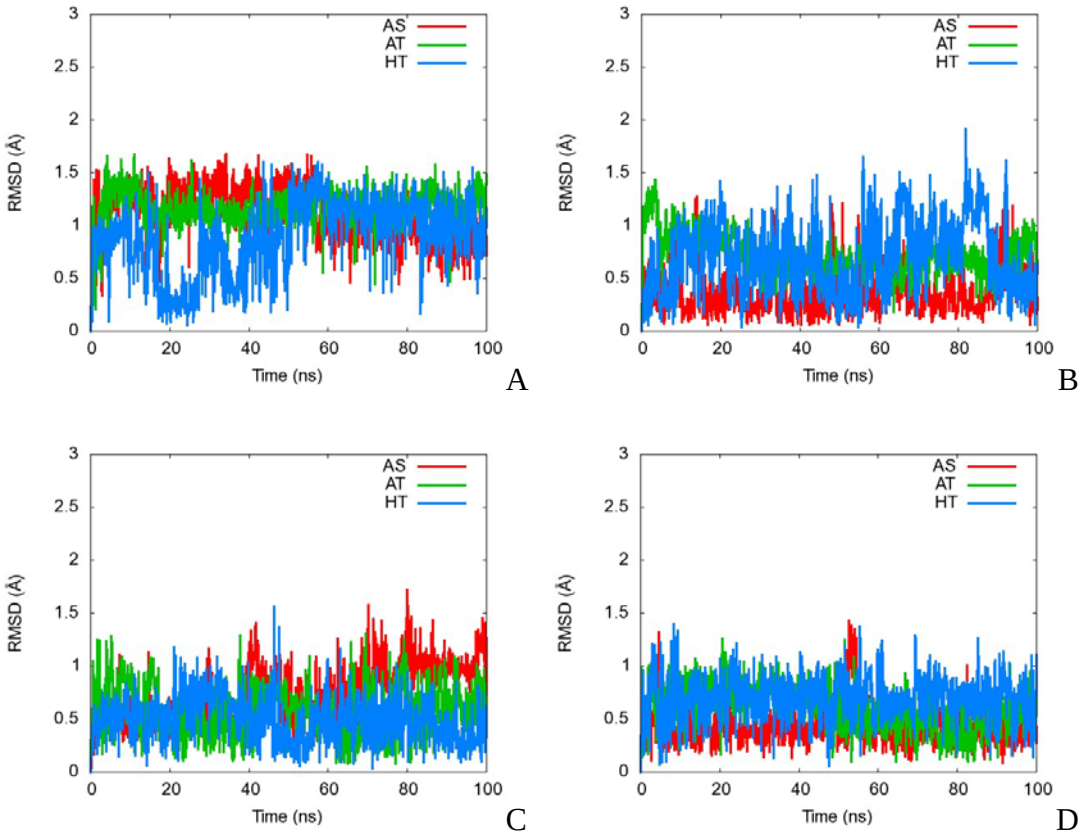


Figure 8-12. C_α RMSD for the α_3-helix.



Figures 8-13A-D. C_α RMSD for "DNA touching loop", residues 64-67 of each chain. Figures A-D show chains A,B,C and D respectively.

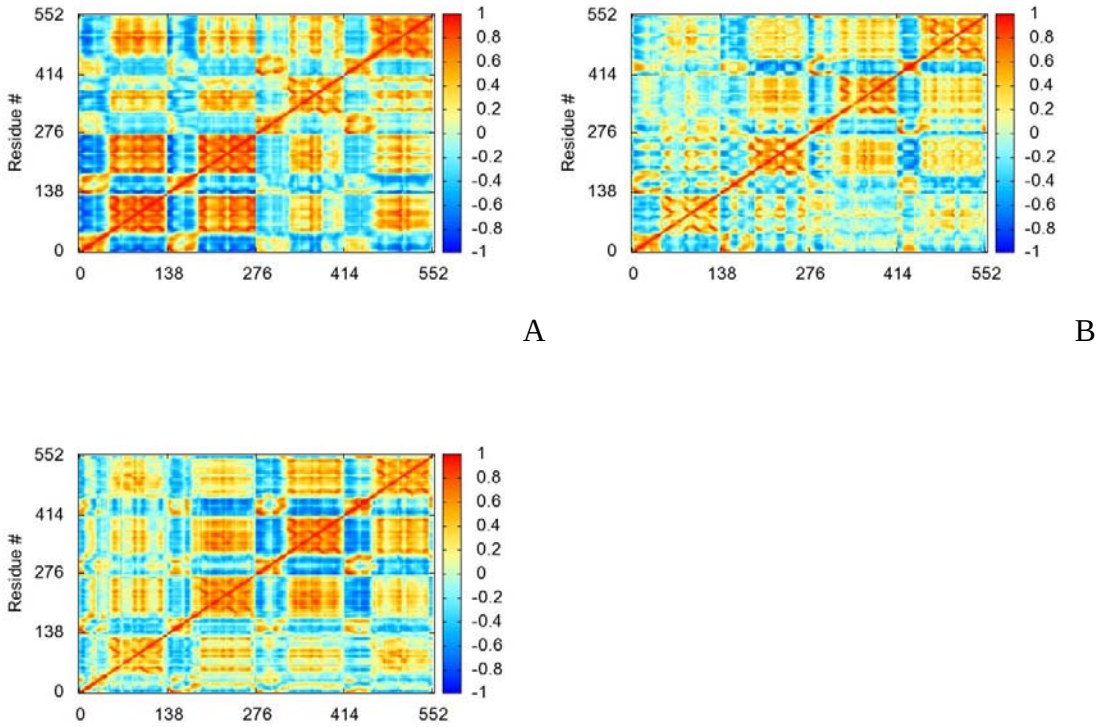


Figure 8-14A-C. Correlation matrices for all three simulations. A, B, and C are the AS, AT, and HT simulations, respectively.

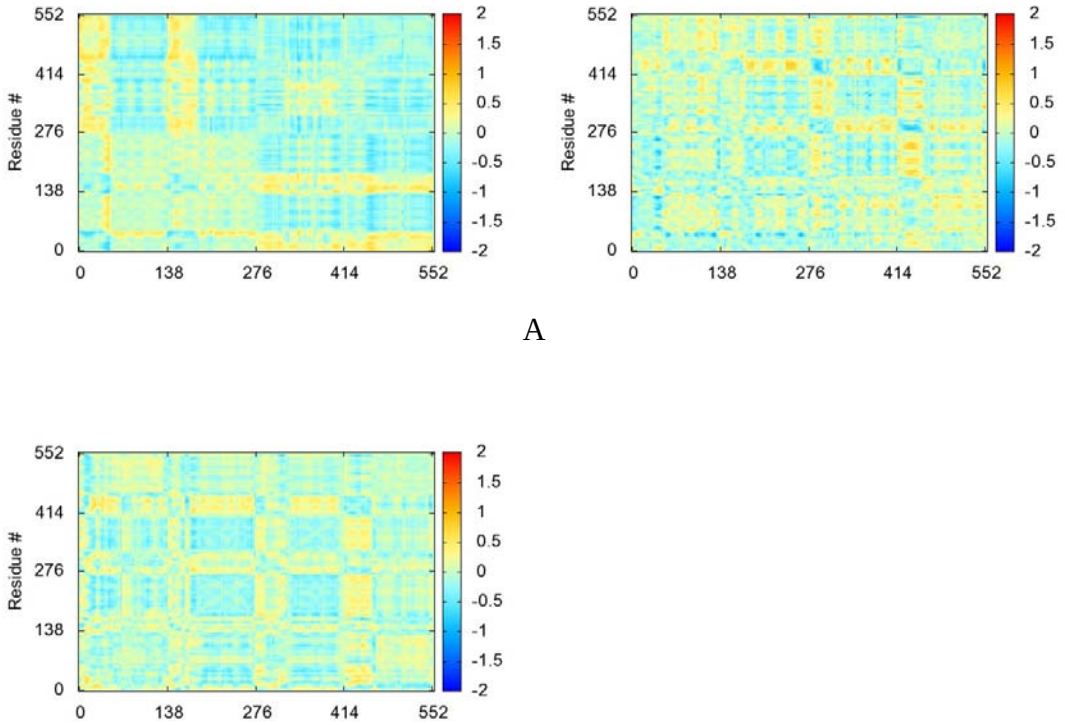
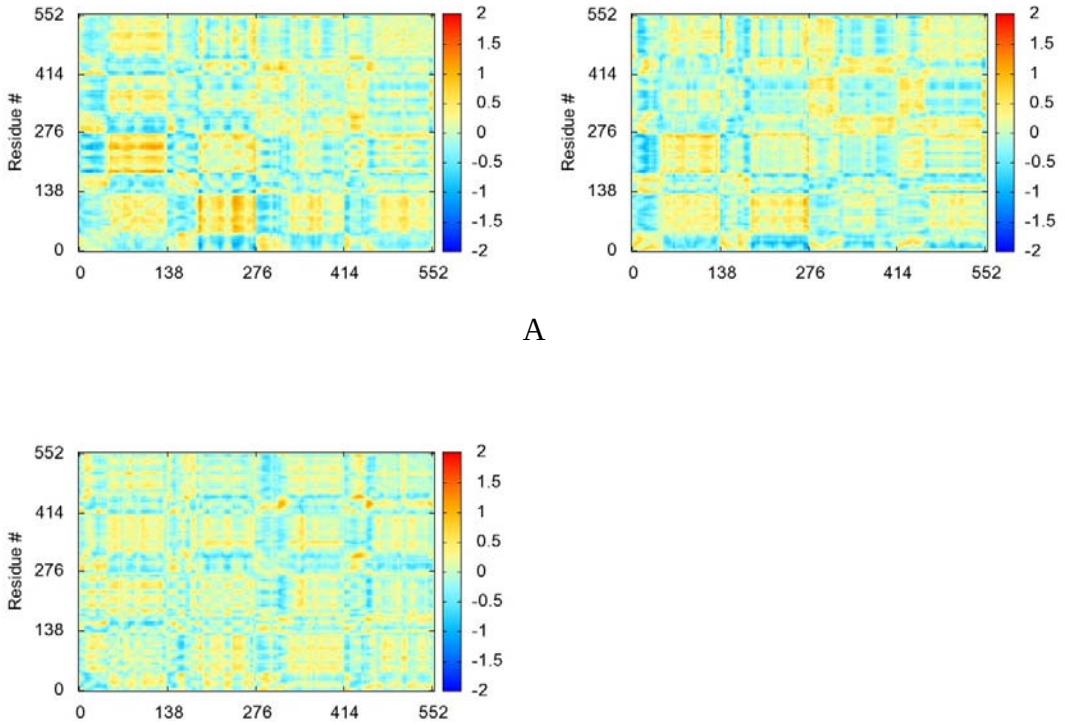


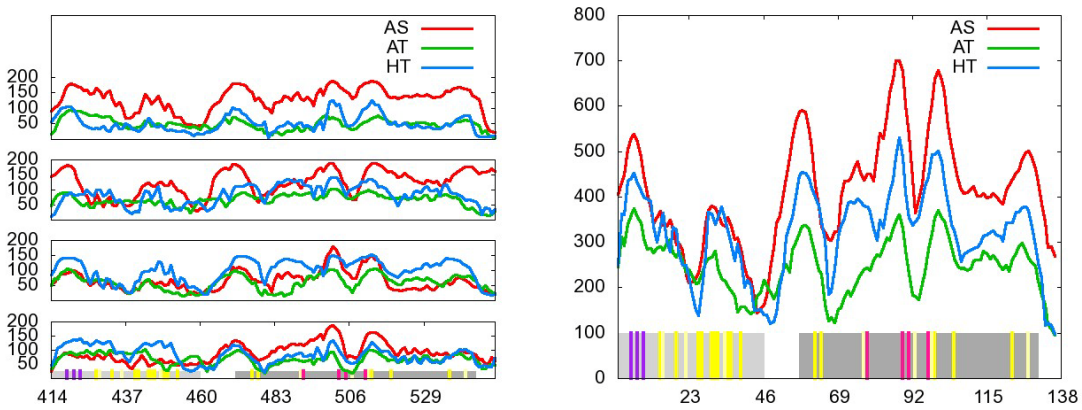
Figure 8-15A-C. Correlation change between first and second halves of each simulation. A, B and C show correlation difference for AS, AT, and HT simulations respectively.



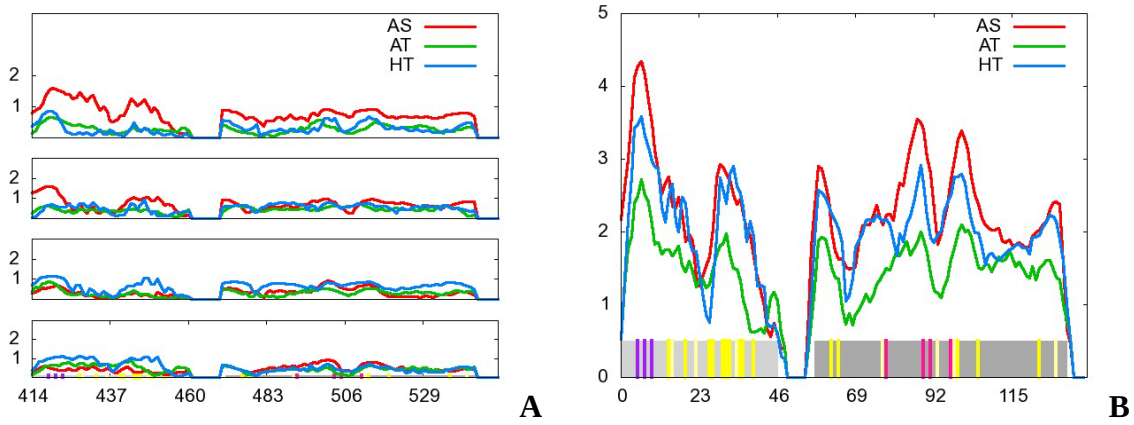
A B

C

Figures 8-16A-C. Correlation difference between simulations. Figures 16A-C represent $C_{ij}^{AS} - C_{ij}^{AT}$, $C_{ij}^{AS} - C_{ij}^{HT}$, and $C_{ij}^{HT} - C_{ij}^{AT}$ respectively.



Figures 8-17A,B. Correlation relevance graphs for each simulation. A) Correlation relevance by residue (REL_i). B) Correlation relevance summed over chains ($cREL_i$)



Figures 8-18A,B. Cross-correlation relevance graphs for each simulation. A) Cross-correlation relevance by residue (xDREL_i). B) Cross-correlation relevance summed over chains (cx$DREL_i$)

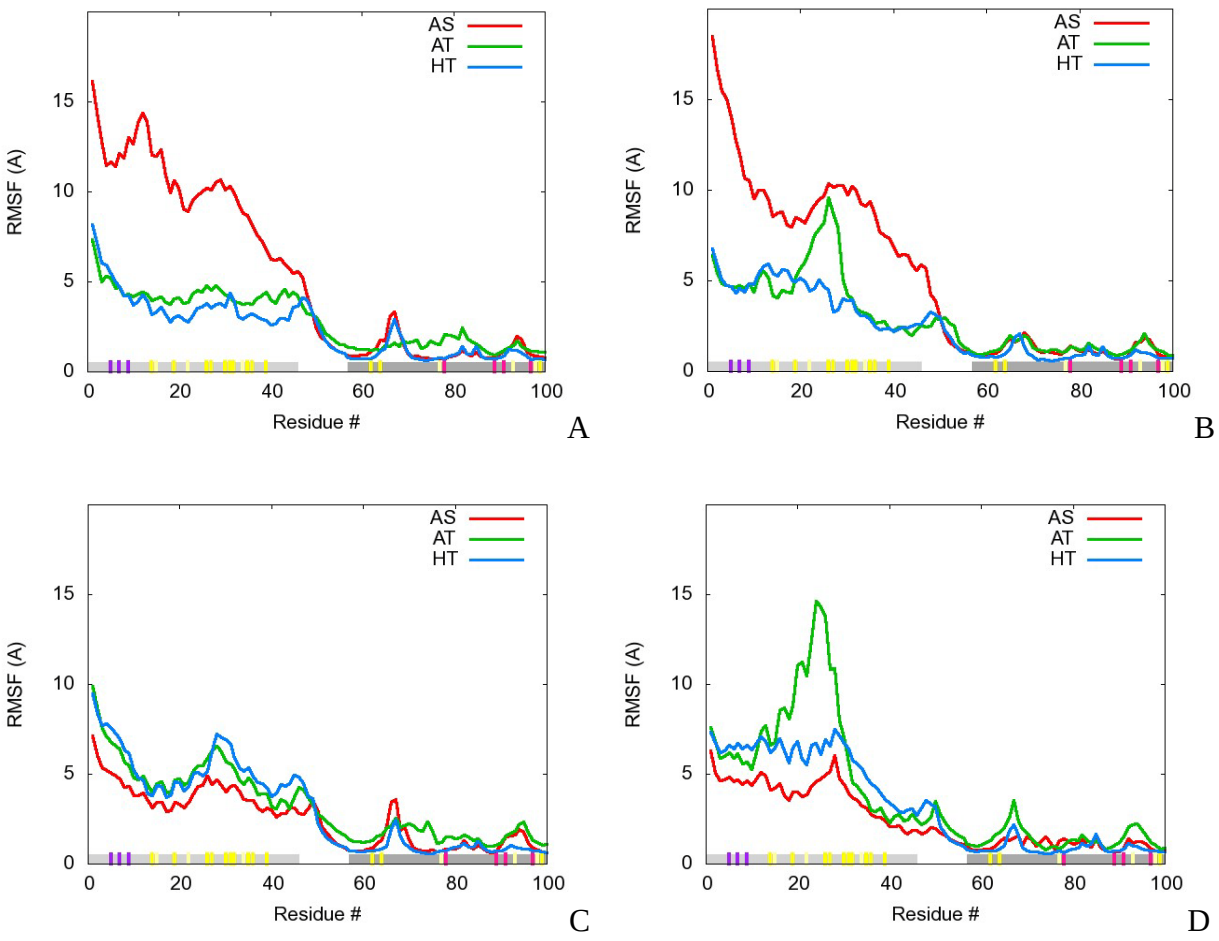


Figure 8-19A-D. RMS fluctuations of the C_α atom for each chain of each simulation. Figures A-D represent chains A, B, C, and D respectively.

APPENDIX
LIST OF FORCE FIELD PARAMETERS OBTAINED FOR NICKEL BOUND FORM OF NIKR

A.1 Partial Charges

Table A-1. Partial charges for Holo NikR using AMBER terminology.

C97A		H87A		H89A		H78D	
Amber Atom ID/Type	Charge	AMBER Atom ID/Type	Charge	AMBER Atom ID	Charge	AMBER Atom ID	Charge
N/N	-0.463	N/N	-0.4157	N/N	-0.4157	N/N	-0.4157
HN/H	0.252	H/H	0.2719	H/H	0.2719	H/H	0.2719
CA/CT	0.035	CA/CT	0.0188	CA/CT	-0.0581	CA/CT	0.0188
HA/H1	0.048	HA/H1	0.0881	HA/H1	0.136	HA/H1	0.0881
CB/CT	0.21679	CB/CT	-0.32182	CB/CT	0.0838	CB/CT	-0.04799
HB3/H1	0.01975	HB2/H1	0.18394	HB2/HC	0.06662	HB2/HC	0.07776
HB2/H1	0.01975	HB3/H1	0.18394	HB3/HC	0.06662	HB3/HC	0.07776
SG/S1	-0.52432	CG/CC	-0.1281	CG/CC	0.04537	CG/CC	0.10735
C/C	0.616	ND1/NA	-0.04524	ND1/NY	0.0069	ND1/NA	-0.3257
O/O	-0.504	HD1/H	0.29485	CE1/CR	-0.1906	HD1/NA	0.37104
		CE1/CR	0.04268	HE1/H5	0.20125	CE1/CR	0.04183
		HE1/H5	0.14435	NE2/NA	-0.08532	HE1/H5	0.20837
		NE2/NZ	-0.66101	HE2/H	0.31809	NE2/NX	-0.16857
		CD2/CV	0.21993	CD2/CW	-0.22892	CD2/CV	-0.24011
		HD2/H4	0.03232	HD2/H4	0.19764	HD2/H4	0.14371
		C/C	0.5973	C/C	0.5973	C/C	0.5973
		O/O	-0.5679	OO	-0.5679	O/O	-0.5679

A.2 Bond Stretching Parameters

Table A-2. Bond stretching parameters for Holo NikR using AMBER terminology

Bonds	Force Constant (Kcal/mol/A^2)	Distance (A)
CT-S1	237	1.81
HS-S1	274	1.336
LP-S1	600	0.7
CB-NZ	414	1.391
CK-NZ	529	1.304
CC-NZ	410	1.394
CR-NZ	488	1.335
CV-NZ	410	1.394
LP-NZ	600	0.2
CB-NY	414	1.391
CK-NY	529	1.304
CC-NY	410	1.394
CR-NY	488	1.335
CV-NY	410	1.394
LP-NY	600	0.2
CB-NX	414	1.391
CK-NX	529	1.304
CC-NX	410	1.394
CR-NX	488	1.335
CV-NX	410	1.394
LP-NX	600	0.2
NI-NZ	97.34	2.038
NI-NY	66.24	2.062
NI-S1	92.81	2.206
NI-NX	89.89	2.017

A.3 Angle Bending Parameters

Table A-3. Angle bending parameters for Holo NikR using AMBER terminology

ANGL	Kcal/mol/rad^2	Angle (degrees)
H1-CT-S1	50	109.5
CT-CT-S1	50	108.6
CT-S1-HS	43	96
HS-S1-HS	35	92.07
CT-S1-LP	50	90
LP-S1-LP	50	180
HS-S1-LP	50	90
C-CB-NZ	70	130
CA-CB-NZ	70	132.4
CB-CB-NZ	70	110.4
H5-CK-NZ	50	123.05
N*-CK-NZ	70	113.9
CT-CC-NZ	70	120
CW-CC-NZ	70	120
H5-CR-NZ	50	120
NA-CR-NZ	70	120
CC-CV-NZ	70	120
H4-CV-NZ	50	120
CB-NZ-CK	70	103.8
CC-NZ-CR	70	117
CR-NZ-CV	70	117
CB-NZ-LP	50	126
CC-NZ-LP	50	126
CK-NZ-LP	50	126
CR-NZ-LP	50	126
CV-NZ-LP	50	126
C-CB-NY	70	130
CA-CB-NY	70	132.4
CB-CB-NY	70	110.4
H5-CK-NY	50	123.05
N*-CK-NY	70	113.9
CT-CC-NY	70	120
CW-CC-NY	70	120
H5-CR-NY	50	120
NA-CR-NY	70	120
CC-CV-NY	70	120
H4-CV-NY	50	120
CB-NY-CK	70	103.8
CC-NY-CR	70	117

Table A-3. Continued.

ANGL	Kcal/mol/rad^2	Angle (degrees)	
CB-NY-LP		50	126
CC-NY-LP		50	126
CK-NY-LP		50	126
CR-NY-LP		50	126
CV-NY-LP		50	126
C-CB-NX		70	130
CA-CB-NX		70	132.4
CB-CB-NX		70	110.4
H5-CK-NX		50	123.05
N*-CK-NX		70	113.9
CT-CC-NX		70	120
CW-CC-NX		70	120
H5-CR-NX		50	120
NA-CR-NX		70	120
CC-CV-NX		70	120
H4-CV-NX		50	120
CB-NX-CK		70	103.8
CC-NX-CR		70	117
CR-NX-CV		70	117
CB-NX-LP		50	126
CC-NX-LP		50	126
CK-NX-LP		50	126
CR-NX-LP		50	126
CV-NX-LP		50	126
CV-NX-NI		119.045	120.926
S1-NI-NX		135.003	85.768
NZ-NI-NY		98.051	88.68
NZ-NI-S1		125.6	95.101
NZ-NI-NX		86.915	176.994
CR-NX-NI		114.211	126.433
CR-NY-NI		100.446	115.058
CC-NY-NI		109.757	135.465
CV-NZ-NI		110.423	131.46
CT-S1-NI		128.037	107.681
NY-NI-S1		96.174	173.922
NY-NI-NX		103.035	90.697
CR-NZ-NI		110.488	118.865

LIST OF REFERENCES

[1] G. D. Purvis and R. J. Bartlett, J. Chem. Phys. **76** (4), 1910 (1982).

[2] C. Levinthal, J. Chim. Phys. Phys.-Chim. Biol. **65** (1), 44 (1968).

[3] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **268** (5213), 960 (1995).

[4] S. G. Itoh, H. Okumura, and Y. Okamoto, Mol. Simul. **33** (1-2), 47 (2007).

[5] B. a. Berg and T. Neuhaus, Phys. Rev. Lett. **68** (1), 9 (1992).

[6] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsovvelyaminov, J. Chem. Phys. **96** (3), 1776 (1992).

[7] A. Mitsutake, Y. Sugita, and Y. Okamoto, Biopolymers **60** (2), 96 (2001).

[8] R. H. Swendsen and J. S. Wang, Phys. Rev. Lett. **57** (21), 2607 (1986).

[9] M. C. Tesi, E. J. J. vanRensburg, E. Orlandini, and S. G. Whittington, Journal of Statistical Physics **82** (1-2), 155 (1996).

[10] U. H. E. Hansmann, Chem. Phys. Lett. **281** (1-3), 140 (1997).

[11] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65** (6), 1604 (1996).

[12] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314** (1-2), 141 (1999).

[13] G. E. Moore, Proceedings of the Ieee **86** (1), 82 (1998).

[14] L. Verlet, Physical Review **159** (1), 98 (1967).

[15] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).

[16] R. Zwanzig, J. Stat. Phys. **9** (3), 215 (1973).

[17] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79** (2), 926 (1983).

[18] R. Constanciel and R. Contreras, Theor. Chim. Acta **65** (1), 1 (1984).

[19] G. M. Torrie and J. P. Valleau, J. Chem. Phys. **66** (4), 1402 (1977).

[20] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63** (12), 1195 (1989).

[21] B. Roux, Comput. Phys. Commun. **91** (1-3), 275 (1995).

[22] F. G. Wang and D. P. Landau, Physical Review E **64** (5) (2001).

[23] F. G. Wang and D. P. Landau, Phys. Rev. Lett. **86** (10), 2050 (2001).

[24] M. R. Shirts and J. D. Chodera, J. Chem. Phys. **129** (12) (2008).

[25] Y. Okamoto, J. Mol. Graphics Modell. **22** (5), 425 (2004).

[26] R. H. Zhou, B. J. Berne, and R. Germain, Proc. Natl. Acad. Sci. U. S. A. **98** (26), 14931 (2001).

[27] A. E. Garcia and K. Y. Sanbonmatsu, Proteins-Structure Function and Genetics **42** (3), 345 (2001).

[28] K. Y. Sanbonmatsu and A. E. Garcia, Proteins-Structure Function and Genetics **46** (2), 225 (2002).

[29] J. W. Pitera and W. Swope, Proc. Natl. Acad. Sci. U. S. A. **100** (13), 7587 (2003).

[30] F. Rao and A. Caflisch, J. Chem. Phys. **119** (7), 4035 (2003).

[31] M. Cecchini, F. Rao, M. Seeber, and A. Caflisch, J. Chem. Phys. **121** (21), 10748 (2004).

[32] R. H. Zhou, Proc. Natl. Acad. Sci. U. S. A. **100** (23), 13280 (2003).

[33] S. Trebst, M. Troyer, and U. H. E. Hansmann, J. Chem. Phys. **124** (17) (2006).

[34] F. Calvo, J. Chem. Phys. **123** (12) (2005).

[35] D. A. Kofke, J. Chem. Phys. **117** (15), 6911 (2002).

[36] A. Kone and D. A. Kofke, J. Chem. Phys. **122** (20) (2005).

[37] N. Rathore, M. Chopra, and J. J. de Pablo, J. Chem. Phys. **122** (2) (2005).

[38] C. Predescu, M. Predescu, and C. V. Ciobanu, J. Phys. Chem. B **109** (9), 4189 (2005).

[39] H. Fukunishi, O. Watanabe, and S. Takada, J. Chem. Phys. **116** (20), 9058 (2002).

[40] E. Lyman, F. M. Ytreberg, and D. M. Zuckerman, Phys. Rev. Lett. **96** (2) (2006).

[41] S. W. Rick, J. Chem. Phys. **126** (5) (2007).

[42] X. L. Cheng, G. L. Cui, V. Hornak, and C. Simmerling, J. Phys. Chem. B **109** (16), 8220 (2005).

[43] A. Okur, L. Wickstrom, M. M. Layten, K. Song, V. Hornak, and C. Simmerling, Abstracts of Papers of the American Chemical Society **231** (2006).

[44] A. Okur, D. R. Roe, G. L. Cui, V. Hornak, and C. Simmerling, J. Chem. Theory Comput. **3** (2), 557 (2007).

[45] A. E. Roitberg, A. Okur, and C. Simmerling, J. Phys. Chem. B **111** (10), 2415 (2007).

[46] H. Z. Li, G. H. Li, B. A. Berg, and W. Yang, J. Chem. Phys. **125** (14) (2006).

[47] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **329** (3-4), 261 (2000).

[48] N. Metropolis, Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E., J. Chem. Phys. **21**, 1087 (1953).

[49] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, and K. A. Dill, J. Chem. Theory Comput. **3** (1), 26 (2007).

[50] L. Wickstrom, A. Okur, K. Song, V. Hornak, D. P. Raleigh, and C. L. Simmerling, J. Mol. Biol. **360** (5), 1094 (2006).

[51] W. Zhang, C. Wu, and Y. Duan, J. Chem. Phys. **123** (15) (2005).

[52] S. B. Opps and J. Schofield, Physical Review E **6305** (5) (2001).

[53] Y. M. Rhee and V. S. Pande, Biophys. J. **84** (2), 775 (2003).

[54] D. Sindhikara, Y. L. Meng, and A. E. Roitberg, J. Chem. Phys. **128** (2) (2008).

[55] H. X. Lei, C. Wu, H. G. Liu, and Y. Duan, Proc. Natl. Acad. Sci. U. S. A. **104** (12), 4925 (2007).

[56] Y. Z. Ohkubo and C. L. Brooks, Proc. Natl. Acad. Sci. U. S. A. **100** (24), 13916 (2003).

[57] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, J. Comput. Chem. **26** (16), 1668 (2005).

[58] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, Proteins: Struct., Funct., Bioinf. **65** (3), 712 (2006).

[59] A. Onufriev, D. Bashford, and D. A. Case, Proteins: Struct., Funct., Bioinf. **55** (2), 383 (2004).

[60] D. J. Lockhart and P. S. Kim, Science **260** (5105), 198 (1993).

[61] P. A. Thompson, W. A. Eaton, and J. Hofrichter, Biochemistry **36** (30), 9200 (1997).

[62] I. K. Lednev, A. S. Karnoup, M. C. Sparrow, and S. A. Asher, J. Am. Chem. Soc. **123** (10), 2388 (2001).

[63] A. E. Garcia and K. Y. Sanbonmatsu, Proc. Natl. Acad. Sci. U. S. A. **99** (5), 2782 (2002).

[64] H. Nymeyer and A. E. Garcia, Proc. Natl. Acad. Sci. U. S. A. **100** (24), 13934 (2003).

[65] W. Kabsch and C. Sander, Biopolymers **22** (12), 2577 (1983).

[66] W. Yang, R. Bitetti-Putzer, and M. Karplus, J. Chem. Phys. **120** (20), 9450 (2004).

[67] M. J. Abraham and J. E. Gready, J. Chem. Theory Comput. **4** (7), 1119 (2008).

[68] Y. Sugita, A. Kitao, and Y. Okamoto, J. Chem. Phys. **113** (15), 6042 (2000).

[69] A. Mitsutake and Y. Okamoto, Chem. Phys. Lett. **332** (1-2), 131 (2000).

[70] R. D. Mountain and D. Thirumalai, J. Phys. Chem. **93** (19), 6975 (1989).

[71] D. Thirumalai, R. D. Mountain, and T. R. Kirkpatrick, Phys. Rev. A **39** (7), 3563 (1989).

[72] J. S. Hu and A. Bax, J. Am. Chem. Soc. **119** (27), 6360 (1997).

[73] M. E. Tuckerman and G. J. Martyna, J. Phys. Chem. B **104** (2), 159 (2000).

[74] H. J. C. Berendsen, J. P. M. Potsma, W. F. van Gunsteren, A. D. DiNola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).

[75] S. Nose, Mol. Phys. **52** (2), 255 (1984).

[76] W. G. Hoover, Phys. Rev. A **31** (3), 1695 (1985).

[77] D. J. Evans and B. L. Holian, J. Chem. Phys. **83** (8), 4069 (1985).

[78] H. C. Andersen, J. Chem. Phys. **72** (4), 2384 (1980).

[79] H. C. Andersen, J. Chem. Phys. **72**, 2384 (1980).

[80] P. Hunenberger, in *Advanced Computer Simulation Approaches for Soft Matter Sciences I* (Springer Berlin / Heidelberg, 2005), Vol. 173, pp. 105.

[81] M. Braxenthaler, R. Unger, D. Auerbach, J. A. Given, and J. Moult, Proteins: Struct., Funct., Genet. **29** (4), 417 (1997).

[82] B. P. Uberuaga, M. Anghel, and A. F. Voter, J. Chem. Phys. **120** (14), 6363 (2004).

[83] Y. Le Jan, Annales de l'I.H.P. Probabilités et statistiques **23** (1), 111 (1987).

[84] S. Fahy and D. R. Hamann, Phys. Rev. Lett. **69** (5), 761 (1992).

[85] a. Maritan and J. R. Banavar, Phys. Rev. Lett. **72** (10), 1451 (1994).

[86] S. Lise, A. Maritan, and M. R. Swift, Journal of Physics a-Mathematical and General **32** (28), 5251 (1999).

[87] M. Ciesla, S. P. Dias, L. Longa, and F. A. Oliveira, Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys. **6306** (6) (2001).

[88] L. Longa, E. M. F. Curado, and F. A. Oliveira, Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys. **54** (3), R2201 (1996).

[89] D. S. Cerutti, R. Duke, P. L. Freddolino, H. Fan, and T. P. Lybrand, J. Chem. Theory Comput. (2008).

[90] M. Hagen, B. Kim, P. Liu, R. A. Friesner, and B. J. Berne, J. Phys. Chem. B **111** (6), 1416 (2007).

[91] H. J. Shen, C. Czaplewski, A. Liwo, and H. A. Scheraga, J. Chem. Theory Comput. **4** (8), 1386 (2008).

[92] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics Modell. **14** (1), 33 (1996).

[93] T. Wang, D. G. Du, and F. Gai, Chem. Phys. Lett. **370** (5-6), 842 (2003).

[94] S. Kim and A. E. Roitberg, J. Phys. Chem. B **112** (5), 1525 (2008).

[95] B. Nolting, R. Golbik, and A. R. Fersht, Proc. Natl. Acad. Sci. U. S. A. **92** (23), 10668 (1995).

[96] C. D. Snow, L. L. Qiu, D. G. Du, F. Gai, S. J. Hagen, and V. S. Pande, Proc. Natl. Acad. Sci. U. S. A. **101** (12), 4077 (2004).

[97] N. Sreerama and R. W. Woody, in *Numerical Computer Methods, Pt D* (Academic Press Inc, 2004), Vol. 383, pp. 318.

[98] G. Marsaglia, B. Narasimhan, and A. Zaman, Comput. Phys. Commun. **60** (3), 345 (1990).

[99] V. C. Culotta, M. Yang, and T. V. O'Halloran, Biochimica Et Biophysica Acta-Molecular Cell Research **1763** (7), 747 (2006).

[100] L. A. Finney and T. V. O'Halloran, Science **300** (5621), 931 (2003).

[101] N. S. Dosanjh and S. L. J. Michel, Curr. Opin. Chem. Biol. **10** (2), 123 (2006).

[102] J. S. Valentine, P. A. Doucette, and S. Z. Potter, Annu. Rev. Biochem. **74**, 563 (2005).

[103] A. C. Rosenzweig, Acc. Chem. Res. **34** (2), 119 (2001).

[104] E. Gaggelli, H. Kozlowski, D. Valensin, and G. Valensin, Chem Rev **106** (6), 1995 (2006).

[105] J. S. Beckman, A. G. Estevez, J. P. Crow, and L. Barbeito, Trends Neurosci. **24** (11 Suppl), S15 (2001).

[106] D. L. Huffman and T. V. O'Halloran, Annu. Rev. Biochem. **70**, 677 (2001).

[107] M. J. Maroney, Curr. Opin. Chem. Biol. **3** (2), 188 (1999).

[108] S. B. Mulrooney and R. P. Hausinger, FEMS Microbiol. Rev. **27** (2-3), 239 (2003).

[109] K. De Pina, V. Desjardin, M. A. Mandrand-Berthelot, G. Giordano, and L. F. Wu, J. Bacteriol. **181** (2), 670 (1999).

[110] P. T. Chivers and R. T. Sauer, Protein Sci. **8** (11), 2494 (1999).

[111] P. T. Chivers and R. T. Sauer, J. Biol. Chem. **275** (26), 19735 (2000).

[112] M. Contreras, J. M. Thiberge, M. A. Mandrand-Berthelot, and A. Labigne, Mol. Microbiol. **49** (4), 947 (2003).

[113] P. T. Chivers and R. T. Sauer, Chem. Biol. **9** (10), 1141 (2002).

[114] P. E. Carrington, P. T. Chivers, F. Al-Mjeni, R. T. Sauer, and M. J. Maroney, Nature Structural Biology **10** (2), 126 (2003).

[115] E. R. Schreiter, M. D. Sintchak, Y. Y. Guo, P. T. Chivers, R. T. Sauer, and C. L. Drennan, Nature Structural Biology **10** (10), 794 (2003).

[116] E. R. Schreiter, S. C. Wang, D. B. Zamble, and C. L. Drennan, Proc. Natl. Acad. Sci. U. S. A. **103** (37), 13676 (2006).

[117] C. Dian, K. Schauer, U. Kapp, S. M. McSweeney, A. Labigne, and L. Terradot, J. Mol. Biol. **361** (4), 715 (2006).

[118] P. T. Chivers and T. H. Tahirov, J. Mol. Biol. **348** (3), 597 (2005).

[119] E. R. Schreiter, P. T. Chivers, and C. L. Drennan, FASEB J. **18** (8), C188 (2004).

[120] C. M. Phillips, E. R. Schreiter, Y. Guo, S. C. Wang, D. B. Zamble, and C. L. Drennan, Biochemistry **47** (7), 1938 (2008).

[121] G. Cui and K. M. Merz, Biophys. J. **94** (10), 3769 (2008).

[122] M. J. Bradley, P. T. Chivers, and N. A. Baker, J. Mol. Biol. **378** (5), 1155 (2008).

[123] B. Zagrovic and V. Pande, J. Comput. Chem. **24** (12), 1432 (2003).

[124] V. Hornak, A. Okur, R. C. Rizzo, and C. Simmerling, J. Am. Chem. Soc. **128** (9), 2812 (2006).

[125] V. Hornak, A. Okur, R. C. Rizzo, and C. Simmerling, Proc. Natl. Acad. Sci. U. S. A. **103** (4), 915 (2006).

[126] R. H. Stote and M. Karplus, Proteins-Structure Function and Genetics **23** (1), 12 (1995).

[127] D. V. Sakharov and C. Lim, J. Am. Chem. Soc. **127** (13), 4921 (2005).

[128] J. Aqvist and a. Warshel, J. Mol. Biol. **224** (1), 7 (1992).

[129] Y. P. Pang, Proteins-Structure Function and Genetics **45** (3), 183 (2001).

[130] Y. P. Pang, K. Xu, J. El Yazal, and F. G. Prendergast, Protein Sci. **9** (10), 1857 (2000).

[131] M. Peters, edited by M. Peters (2008), Vol. 2008, pp. Utilities website.

[132] G. W. T. M. J. Frisch, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., 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, V. Bakken, 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 (Wallingford CT, 2004).

[133] K. E. Riley and K. M. Merz, J. Phys. Chem. A **111** (27), 6044 (2007).

[134] K. K. Kannan, M. Ramanadham, and T. A. Jones, Ann. N. Y. Acad. Sci. **429** (Jun), 49 (1984).

[135] B. H. Besler, K. M. J. Merz, and P. A. Kollman, J. Comput. Chem. **11** (4), 431 (1990).

[136] R. J. Woods, M. Khalil, W. Pell, S. H. Moffat, and J. V. H. Smith, J. Comp. Chem. **11** (3), 297 (1990).

[137] L. E. Chirlian and M. M. Francl, J. Comp. Chem. **8** (6), 894 (1987).

[138] U. C. Singh and P. A. Kollman, J. Comp. Chem. **5**, 129 (1984).

[139] D. E. Williams and J. M. Yan, Adv. Atomic Mol. Phys. **23**, 87 (1988).

[140] S. R. Cox and D. E. Williams, J. Comput. Chem. **2** (3), 304 (1981).

[141] C. A. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman, J. Phys. Chem. **97**, 10269 (1993).

[142] F. M. Floris, M. Persico, A. Tani, and J. Tomasi, Chem. Phys. Lett. **227** (1-2), 126 (1994).

[143] E. R. Schreiter and C. L. Drennan, FASEB J. **19** (4), A298 (2005).

BIOGRAPHICAL SKETCH

Daniel Jon Sindhikara was born in Point Pleasant, New Jersey in 1980 to Dr. Jean Sindhikara and Mr. Ian Sindhikara. He has an older sister, Ms. Rana Sindhikara. Daniel grew up in Brick Township, New Jersey and graduated from Brick Township Memorial High School. While in high school, he became interested in both computer programming (mostly in the interest of computer games) and biology.

Daniel went to undergraduate college at Richard Stockton College of New Jersey. There, inspired by Dr. Louis DeChiaro, he pursued computational physics. In 2004, he got his B.S. in applied physics with his senior research project involving computational thermal modeling of the then largest geothermal heat pump system in the world.

With an interest in computational science, biology, and physics, he naturally paired up with Professor Adrian Roitberg at University of Florida's Quantum Theory Project. There, he attained a variegated repertoire of projects all involving computational molecular biophysics. His graduate studies also allowed him to travel to do research in Japan. This life-changing trip inspired him to work on learning the language fluently. In addition to doing research, he currently studies both written and conversational Japanese. He graduated from University of Florida in May 2009 with a PhD in Physics.