Improved Conformational Sampling Methods for Molecular Dynamics Simulations

A Dissertation Presented

by

Asım Okur

to

The Graduate School

in Partial fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Chemistry

Stony Brook University

May 2007

Stony Brook University

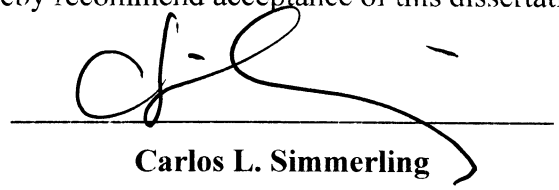
The Graduate School

Asım Okur

We, the dissertation committee for the above candidate for the

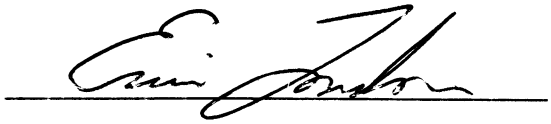
Doctor of Philosophy degree,

hereby recommend acceptance of this dissertation.



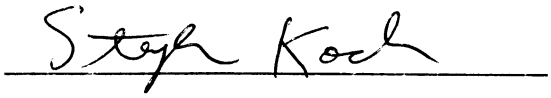
Carlos L. Simmerling

Associated Professor, Department of Chemistry



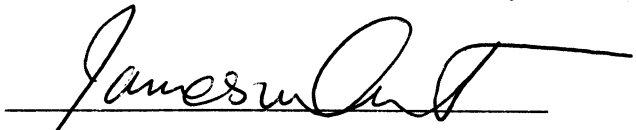
Erwin London

Professor, Department of Chemistry



Stephen A. Koch

Professor, Department of Chemistry



James W. Davenport

Senior Physicist, Director, Computational Science Center

Head, Theory and Computation, Center for Functional NanoMaterials

Brookhaven National Laboratory

This dissertation is accepted by the Graduate School

Dean of the Graduate School

Abstract of the Dissertation

Improved Conformational Sampling Methods for Molecular Dynamics Simulations

by

Asım Okur

Doctor of Philosophy

in

Chemistry

Stony Brook University

2007

Understanding conformational dynamics of biomolecules such as proteins is a fundamental challenge in structural biology. Native conformations of proteins can be determined experimentally via X-Ray Crystallography and NMR spectroscopy but usually such methods provide time averaged data. All-atom simulations are commonly used to supplement experimental observations where time dependent trajectories for complex systems can be obtained. However there are major challenges in computer simulations.

For successful simulations the potential function has to be accurate enough to correctly rank the local and global energy minima and the barriers in between for the simulated system. We developed an efficient method to test the accuracy of force field parameters where the energies of pre-generated conformations (decoys) were calculated for each parameter set in question and the identified energy minima were compared to

experimental measurements. After generating decoy sets the evaluation of force field or other simulation parameters can be done quickly and efficiently. We used this decoy screening procedure to identify α-helical bias in existing force fields in AMBER and to develop improved force field parameters.

Another major challenge in simulations is sampling because the time scales reached with standard simulations are 3-6 orders of magnitude shorter than actual comformational transitions observed in proteins. There are several new sampling methods available where transitions between energy minima are enhanced through the use of high temperatures. Such methods are still very computationally demanding and can only be applied to small systems. We have developed two new methods to further enhance the conformational sampling to reduce computational demands and increase the convergence speed of simulations.

To my beloved father…

Table of Contents

List of Figures

List of Tables

Acknowledgements

First of all, thanks to my advisor Dr. Carlos Simmerling. You have been a great mentor over the years. Thanks for providing the opportunities, the resources necessary, patiently answering questions, guiding me in the right direction, always having an open mind to new ideas, and knowing when to be a friend and when to push me to do better. I have learned a lot from you and I hope I can perform to your standards in the future.

Thanks to my committee members, Dr. Erwin London and Dr. Stephen Koch, for always being there when I needed help and always showing me ways for improvements. I would also like to thank my outside member Dr. James Davenport who agreed to be in the committee in short notice and gave me lots of useful positive criticism and advice for my thesis work.

Thanks to Dr. Daniel Raleigh, Dr. Robert Rizzo, Dr. David Green and Dr. Adrian Roitberg for many useful conversations about science and otherwise.

Thanks to the Simmerling lab, past and present members. Over the years we've become one big family, being there for each other, working together and having lots of fun. I cannot imagine going through graduate school without such friends.

Special thanks to Dr Guanglei Cui, for helping me immensely in the beginning, teaching me everything about simulations, patiently answering my silly questions, and always pointing me towards the right direction. Also many thanks to Dr. Bentley Strockbine, for being there when I needed a friend, keeping me on the right track and becoming one of my dearest friends.

I would like to thank all my friends at Stony Brook for sharing so many great things together. I will never forget F1 weekends, weekly international dinners, happy

hour, parties, karaoke, volleyball, movie nights, 24, dancing, bowling, darts tournaments and many other events. I know these friendships will last a lifetime.

Finally, I would like to thanks my family, especially my parents and my lovely sister, for your love and support all these years. I could not have done this without you. I love you all.

Chapter 1

Introduction

1.1 Structural Biology

Proteins are essential molecules for life where they play important roles in cellular processes such as enzymatic catalysis, transport, immune recognition, cellular control, mechanical structure, growth, replication, communication and differentiation. Because of their importance many diseases are caused by function or misfunction of certain proteins [1].

Proteins are polymers formed by 20 amino acids where the sequence is coded through DNA for each organism. Each protein has to adopt a specific three dimensional structure to perform its task. Determining this structure and understanding how the proteins fold into this structure are important problems in biology since they provide information about diseases and may help identifying potential drug targets.

Protein structures are usually determined experimentally through X-Ray Crystallography and NMR Spectroscopy (40354 structures available in Protein Data Bank as of November 28, 2006). These methods provide structural information for proteins in their native state and the resolution of the data obtained has been increased significantly over the years (See the review by Campbell for the evolution of structural biology [2]). Although these methods are commonly used they usually provide only snapshots of native state or time averaged data. Studying dynamics of protein folding is difficult through these experiments. Therefore computer simulations are an attractive

approach for such studies and are commonly used to supplement experimental observations with dynamic information in atomic detail.

Anfinsen has showed that the information for a protein to find its native structure is present in its amino acid sequence (also known as Anfinsen's Hypothesis) [3]. However determining the native structure of a protein using its sequence alone remains unsolved because of the number of available conformations gets very large for an average sized protein. All atom structure prediction using only sequence data has been successful only for small proteins [4] [5]. Simulations of larger systems usually focus on conformational changes on a portion of the protein such as loop modeling. The studies by Hornak et al. [6, 7] are good examples showing conformational transitions of the loop region between bound and unbound forms of HIV-1 Protease.

As mentioned before the number of accessible conformations for proteins gets very large even with small proteins. Levinthal suggested that there are many more possible states than a protein can visit in the time it has to fold, therefore it has to go through a sequence of events or pathways that lead to its native state [8, 9]. This makes the use of molecular dynamics an attractive approach to study dynamics of protein motions because if the force field is accurate the protein should follow same pathways as in real life. However inaccuracies in force fields and insufficient computational resources prevent us from simulating folding process for an average protein in full detail.

1.2 Force fields

Force fields provide the driving force on molecular simulations. Common simulation methods like molecular dynamics (MD) and Monte Carlo (MC) methods rely on force fields to calculate forces on each particle during simulations. The typical force field equation looks like Equation 1-1 where each term represents a different type of interaction such as bond lengths, bond angles, dihedral angles and non-bonded interactions.

$$E_{Pair} = \sum_{bonds} K_r (r - r_{eq})^2 + \sum_{angles} K_\theta (\theta - \theta_{eq})^2 + \sum_{torsions} \frac{V_n}{2}\left[1 + \cos(n\omega - \gamma)\right] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right]$$

Equation 1-1 Molecular mechanics force field equation.

The parameters used in the molecular mechanics force field equation are called parameter sets or force fields. Commonly used parameter sets for biological simulations are CHARMM [10], AMBER [11], GROMOS [12] and OPLS [13] force fields. Several review articles discuss the current status and future directions of biological force fields [14-16].

Among AMBER force fields ff94 [11] and ff99 [17] implementations are most commonly used. However several problems have been reported about both force fields suggesting they have strong bias for α-helical conformations. Several groups suggested that the backbone dihedral parameters were the source of the problem and many groups empirically modified them to obtain better simulation data for their systems [18-20]. Duan and coworkers adopted a different approach where they parameterized the entire force field using a different partial charge derivation scheme [21]. Chapter 2 describes our efforts of confirming the helical bias and modifying dihedral parameters to obtain a better parameter set for the test peptides (ffGA parameter set) [22]. Later we discovered

that the ffGA force field had strong preferences for β-strands and was only used in investigations of the Trpzip2 β-hairpin (Chapter 3) and dPdP three stranded β-sheet [23]. Force field development continued in our lab where the backbone parameters of ff99 were completely refit to quantum mechanics energies for Alanine and Glycine tetrapeptides [24]. Decoy screening procedure was used to test the accuracy of the available AMBER force fields on various peptides and small proteins having different secondary structures.

1.3 Solvent Models

Interactions with solvent play a central role in the thermodynamics and structure of macromolecules. In particular, the stability and functionality of proteins and nucleic acids are dictated by both specific and bulk solvent effects. The effects and importance of solvents for proteins and nucleic acids have been summarized by Makarov et al. [25]. Solvent properties in the proximity of protein surfaces can differ significantly from bulk solvent (e.g. see review by Bagchi [26]). Therefore, it is important to include solvent effects as accurately as possible or successful simulation studies.

Another important consideration when choosing a method to treat solvation is the impact it may have on the system size and thus the computational requirement of the simulations. Explicit representation of solvent molecules significantly increases the number of atoms in the simulated system. Periodic Boundary Conditions (PBC) are usually applied with an Ewald method [27] such as Particle Mesh Ewald (PME) [28] which take advantage of system periodicity to efficiently calculate long-range electrostatic interactions. While this may be reasonable for simulations of compact states,

it can become prohibitive when the solvent box is made large enough to enclose unfolded conformations of peptides and proteins. The growth in system size results in increased computational cost to calculate forces and integrate equations of motion for the solvent molecules. As a result, large explicitly solvated systems typically cannot be simulated for biologically relevant timescales.

Continuum solvent models, like those based on Poisson Boltzmann (PB) formalism or the semi-analytical Generalized Born (GB) model [29], estimate the free energy of solvation based solely on coordinates of solute atoms. The neglect of explicit solvent molecules can significantly reduce the computational cost of evaluating energies and forces for the system. Continuum solvent models are thus and attractive approach to enabling the study of larger systems with MD. Among various models that have been developed, the GB approach is commonly used with MD due to its computational efficiency, permitting use at each time step. However, continuum models can also have significant limitations. Since the atomic detail of the solvent is not considered, modeling specific effects of structured water molecules with any implicit model can be challenging [30, 31].

In the case of protein and peptide folding, it appears likely that the current generation of GB models do not have as good a balance between protein-protein and protein solvent interactions as do more widely tested explicit solvent models [32, 33]. More particularly, it has been reported [32-36] that ion pairs were frequently too stable in the GB implicit water model, causing salt-bridged conformations to be oversampled in MD simulations, thus altering the thermodynamics and kinetics of folding for small peptides.

1.4 Conformational Sampling

Sampling is probably the biggest challenge in computational biology today since the energy landscapes of real proteins are rugged preventing transitions between various local minima. Even with sufficiently accurate force fields it is still very difficult to simulate folding process for proteins. With the computational resources available, it is not possible to simulate folding for an average sized protein in full atomic detail. Difficulties encountered on sampling have been reviewed on several articles [37, 38].

One major problem encountered in simulations is quasi-ergodicity where simulations may appear converged when observing certain simulation parameters but in reality they may be trapped in local minima. Different simulations starting from different initial conditions may give different results (See example by Smith et al. [39]). This makes identifying the problem for an inaccurate simulation or testing the performance for force fields or other parameters difficult.

The method chosen to treat solvent effects can have a direct impact on system size and complexity since the number of degrees of freedom increases significantly with explicit representation of solvent. Implicit methods such as Generalized Born approach [29] are often used to reduce complexity and enhance conformational transitions through the lack of viscosity. However they usually have limitations which are discussed in Chapter 4 in detail. Several hybrid implicit/explicit solvent treatments are proposed to reduce system sizes while keeping solvent molecules close to the solute for increased accuracy. The strengths and weaknesses of such implementations are discussed in the recent review by Okur and Simmerling [40].

To overcome quasi-ergodicity several enhanced sampling methods have been developed. Currently the Parallel Tempering [41] or Replica Exchange Method (REMD) [42] are commonly used to increase sampling efficiency through the use of combining high temperature simulations with low temperature ones where conformational transitions between local minima is enhanced through higher temperatures.

Replica Exchange Method is successful exploring energy landscapes of peptides and small proteins. However obtaining converged results with larger systems or with explicit treatment of solvent molecules gets computationally very expensive preventing REMD to be used for large proteins. We have developed two methods to improve the sampling efficiency of REMD simulations to make them more applicable to larger systems. These methods are discussed in Chapter 4 and Chapter 5.

1.5 Outlines of Research Projects

This dissertation contains three projects that describe efficient methods for conformational sampling in molecular dynamics simulations. Chapter 2 describes decoy screening method where by generating a set of conformations only once, many tests on force fields or other simulation parameters can be performed very quickly. Such decoy structures can even be used to improve such parameters for improved simulations. This work is published in Journal of Computational Chemistry in 2003 [22]. Through this method we identified a helical bias in the commonly used Cornell et al. [11] force field (denoted ff94 in AMBER). This procedure was also used to test the accuracy of the parameter set developed by our group (ff99SB in AMBER 9) [24]. Chapter 3 describes the folding and unfolding study of the Tryptophan Zipper 2 β-hairpin using the force field

parameters obtained in Chapter 2. Chapter 4 describes a hybrid solvent approach to be used with Replica Exchange method to reduce the number of replicas required for explicit solvent simulations. This work is published in Journal of Chemical Theory and Computation in 2006 [43]. Chapter 5 describes another improvement in Replica Exchange method where when the replicas are coupled to a pre-generated high temperature reservoir, the convergence speed of the simulations is increased. This work is currently in press in Journal of Chemical Theory and Computation [44].

1.5.1 Decoy Screening

The transferability of molecular mechanics parameters derived for small model systems to larger biopolymers such as proteins can be difficult to assess. Even for small peptides, molecular dynamics simulations are typically too short to sample structures significantly different than initial conformations, making comparison to experimental data questionable. We employed a PC cluster to generate large numbers of native and non-native conformations for peptides with experimentally measured structural data, one predominantly helical and the other forming a β-hairpin. These atomic-detail sets do not suffer from slow convergence and can be used to rapidly evaluate important force field properties. In this case a suspected bias toward α-helical conformations in the ff94 and ff99 force fields distributed with the AMBER package was verified. The sets provide critical feedback not only on force field transferability, but may also predict modifications for improvement. Such predictions were used to modify the ff99 parameter set, and the resulting force field was used to test stability and folding of model peptides. Structural behavior during molecular dynamics with the modified force field is found to

be very similar to expectations, suggesting that these basis sets of conformations may themselves have significant transferability among force fields. We continue to improve and expand this data set and plan to make it publicly accessible. The calculations involved in this process are trivially parallel and can be performed using inexpensive personal computers with commodity components.

1.5.2 Folding and Unfolding Simulations of a β-hairpin

Understanding how proteins fold to a well defined structure is a complex problem of great interest. Using computational methods, we studied the folding and unfolding behavior of a small model peptide via nearly 4.5 μs of molecular dynamics simulation. We studied folding and unfolding pathways using non-equilibrium temperature jump simulations and validated these results against free energy data obtained from replica exchange molecular dynamics simulations. The unfolded state is observed to have a high tendency to sample a β-turn, along with non-specific hydrophobic contacts. Folding involves an increased specificity of these contacts and formation of native backbone hydrogen bonds, with both events occurring at the folding free energy barrier. Under simulation conditions, folding behavior does not appear to be a two-state process. We demonstrate that each one of our observed exponential processes is itself composed of multiple pathways with similar relaxation times. While the overall folding and unfolding behavior for this β-hairpin are highly related, some interesting deviations appear to indicate that in order to understand this complex process, a more thorough approach may be required than is typically performed.

1.5.3 Hybrid Solvent Replica Exchange Method

The use of parallel tempering or replica exchange molecular dynamics (REMD) simulations has facilitated the exploration of free energy landscapes for complex molecular systems, but application to large systems is hampered by the scaling of number of required replicas with increasing system size. Use of continuum solvent models reduces system size and replica requirements, but these have been shown to provide poor results in many cases, including overstabilization of ion pairs and secondary structure bias. Hybrid explicit/continuum solvent models can overcome some of these problems through an explicit representation of water molecules in the first solvation shells, but these methods typically require restraints on the solvent molecules and show artifacts in water properties due to the solvation interface. We propose an REMD variant in which the simulations are performed with fully explicit solvent, but the calculation of exchange probability is carried out using a hybrid model, with the solvation shells calculated on the fly during the fully solvated simulation. The resulting reduction in the perceived system size in the REMD exchange calculation provides a dramatic decrease in computational cost of REMD, while maintaining very good agreement with results obtained from standard explicit solvent REMD. We applied several standard and hybrid REMD methods with different solvent models to alanine polymers of 1, 3 and 10 residues, obtaining ensembles that were essentially independent of initial conformation, even with explicit solvation. Use of only a continuum model without a shell of explicit water provided poor results for Ala_3 and Ala_{10}, with significant bias in favor of α-helix. Likewise, using only the solvation shells and no continuum model resulted in ensembles that differed significantly from the standard explicit solvent data. Ensembles obtained

from hybrid REMD are in very close agreement with explicit solvent data, predominantly populating polyproline II conformations. Inclusion of a second shell of explicit solvent was found to be unnecessary for these peptides.

1.5.4 Reservoir Replica Exchange Method

Parallel tempering or replica exchange molecular dynamics (REMD) significantly increases the efficiency of conformational sampling for complex molecular systems. However, obtaining converged data with REMD remains challenging, especially for large systems with complex topologies. We propose a new variant to REMD where the replicas are also permitted to exchange with an ensemble of structures that have been generated in advance using high-temperature MD simulations, similar in spirit to J-walking methods. We tested this approach on two model systems, a β-hairpin and a 3-stranded β-sheet and compared the results to those obtained from very long (>100ns) standard REMD simulations. The resulting ensembles were indistinguishable, including relative populations of different conformations on the unfolded state. Use of the reservoir is shown to significantly reduce the time required for convergence.

Chapter 2

Using PC Clusters to Evaluate the Transferability of Molecular Mechanics Force Fields for Proteins

2.1 Introduction

Computer simulations are an attractive approach to supplementing experimental data for complex systems. They have the potential to provide thermodynamic information comparing relative stability of alternate conformations, as well as kinetic information describing interconversion between these structures. Simulations have the additional advantage that the resulting data need not be time- or ensemble-averaged, a limitation found in most experimental methods. Such calculations are computationally demanding, particularly when long-range interactions are significant and complex energy functions are employed. However, the timescale of many processes of interest, such as large conformational transitions, currently necessitates the use of a relatively simple form for the energy function. The simplification can typically involve many types of approximations, including simplified solvation models and neglect of environment-dependent charge redistribution. Many of the important parameters for these approximate molecular mechanics functions are developed to reproduce relative quantum mechanical energies for a handful of conformations for small model systems, such as individual

nucleic acid bases for DNA and RNA or very short peptides for protein parameters [11, 17, 45-47].

Two key issues are involved in successful use of these force fields: accuracy and transferability. If one is unable to achieve an accurate fit to the training data, it is unlikely that the force field will perform acceptably in the larger systems. Even if a parameter set is successful in accurately reproducing the behavior of its model systems, testing transferability to the larger systems and properties of interest is of critical importance. This may fail if the training systems were not representative models of the behavior one wishes to study. However, testing transferability is far from trivial. When the goal is to reproduce sequence-dependent structure for biopolymers, an equilibrium ensemble of structures must be sampled to be confident that the simulated properties are representative of the underlying force field. For all but the smallest systems, however, the computational cost of obtaining such ensembles is prohibitive. Thus, the behavior observed in a typical simulation is likely a result of barriers to local conformational change, rather than the ability of the force field to reproduce correct equilibrium properties. A molecule at room temperature, even if simulated for tens of nanoseconds, is not likely to travel far in conformational space. Individual simulations on the nanosecond timescale therefore cannot be used to reliably evaluate the transferability of a force field.

Recently, free energy landscapes for the C-terminal fragment of protein G were reported by two groups [36, 48]. Both studies used a replica-exchange approach [42] to sample conformational preferences in atomic detail with explicit solvation. The results were in disagreement concerning the amount of helical structure present in the ensemble of structures. While this could be an artifact of incomplete sampling, Zhou et al.

suggested that the origin of the problem might lie in the use of different force fields in the two studies.

To investigate this issue, we take an approach analogous to that currently used in the design and evaluation of scoring functions for protein structure prediction, the creation of sets of misfolded structures [49-51]. Such "decoy" databases have become quite valuable in the design and testing of potential functions for protein structure prediction [52-57]. Recently, an atomic-detail force field was shown to recognize misfolded structures in such databases with an impressive 90% accuracy [58]. Here we investigate not only whether a given force field can select the native conformation among several decoys, but also whether the native conformation is the most favorable that could be sampled with that force field. The latter requires significant local, as well as global structural variation, but can potentially provide greater insight into the effect of the parameters.

We generate large numbers of independent simulations on a cluster of personal computers to create sets of reasonable, but non-native decoy structures. We employ two model peptides with experimentally determined structural features, including helix, strand, turn and unstructured regions. The decoy sets contain much greater conformational variability than is likely to be seen in individual simulations, even those covering several μs. We used the decoy sets to evaluate the ff94 parameters [11] (denoted "parm94" in AMBER versions before 7.0) and confirm that a bias toward helical conformations is present even in these solvated peptides. The same bias is present in the related but more recent ff99 force field [17] (previously denoted "parm99"). In both cases, helical decoys are significantly more stable than any other structure, in disagreement with experimental data.

We also investigated the possibility of using our decoy sets to carry out optimization of the force field parameters. This process has two goals: an improved parameter set may be obtained, but perhaps more important is evaluation of the transferability of the atomic-detail decoy sets to force fields that were not used in their creation. This provides feedback on the degree of local and global sampling represented in the decoy sets. Simulations using the empirical decoy-based parameters have properties that are very similar to expectations based on decoy analysis, demonstrating that decoy transferability may be acceptable. Compared to the original ff99 parameter set, the test parameters result in simulations that are in much better agreement with experimental data. However, further investigation of the use of decoys to optimize the atomic-detail force field parameters is required and will be reported elsewhere.

2.2 Methodology and Model Systems

2.2.1 Model Peptides

The tryptophan zipper is a structural motif that greatly stabilizes ß-hairpin conformations through tryptophan – tryptophan crosstrand pairs [59]. Folding information for this peptide was determined by NMR and CD spectroscopy, and a family of structures (pdb code 1HRX) was refined using restraints from NMR experiments [59]. Among the trpzips, trpzip2 has the most cooperative melting curve and highest stability (~90%) at 300K, therefore it was selected for use in this study. The sequence is SWTWENGKWTWK, with a type 1' β-turn at the NG portion of the sequence.

The second peptide was chosen because of similar length to trpzip2, but different secondary structure propensity, allowing testing of structure based primarily on sequence rather than length. α-lactalbumin (αlac) 101-111 corresponds to residues 101 to 111 in the protein α-lactalbumin. The sequence is IDYWLAHKALA (the native Cys111 in α-lactalbumin was replaced by Ala for the NMR experiment). NMR experiments in aqueous solution have shown that residues 101 to 107 are highly ordered, with residues 103 to 107 predominantly adopting a helical conformation terminated at His107 [60, 61]. Residues 108 to 111 are not well defined by the family of structures (pdb code 1CB3).

2.2.2 Simulation Details

All simulations were carried out using the AMBER molecular modelling program suite (version 6) [62]. The NMR structures were taken from PDB and LEaP was used to prepare the systems for simulation. C-terminal groups were neutral in both cases, but N-terminal residues were acetylated for αlac 101-111 and positively charged for trpzip2, in accord with the respective experiments. All MD simulations used a temperature of 300K and 2fs time step unless otherwise noted. The SANDER module in AMBER6 was modified to include removal of rigid-body motion during GB dynamics, targeted MD simulation, improved scaling on the PC cluster and use of SHAKE [63] on all bonds during MPI simulation.

For explicit solvent simulations, peptides were solvated with TIP3P [64] water molecules in a rectangular periodic box, with a 5Å buffer between solute and box boundary. Long range electrostatic interactions were calculated by the PME method [28] with an 8Å cutoff for the real-space nonbonded interactions. Simulations were carried out

in the NPT ensemble at 1atm and 300K. Time constants for the temperature and pressure coupling were 0.2ps and 0.02ps, respectively. Systems were equilibrated with 50ps simulation with harmonic positional restraints on solute atoms, followed by minimizations with gradually reduced positional restraints and three 5ps MD simulations with gradually reduced restraints. Production simulations were carried out with weaker temperature and pressure coupling constants of 1ps and 0.2ps, respectively. SHAKE was applied to constrain the length of all bonds involving hydrogen. The high temperature simulations were run in the NVT ensemble with a 1fs time step.

Implicit solvent simulations used the Generalized Born (GB) implicit solvent model [29] as implemented in AMBER6. Translational and rotational motion was removed every 10000 steps. No cutoffs were used in energy calculations, all nonbonded interactions were evaluated at each MD timestep and SHAKE was used to constrain all bond lengths. Other parameters were the same as for explicit solvent calculations. Under these conditions we obtain 4ns per day on a single 1.4ghz AMD Athlon CPU.

2.2.3 Targeted Molecular Dynamics

During targeted molecular dynamics [65] an additional term was added to the energy function (Equation 2-1). A reference structure and a target RMSD value were given as additional input. When the calculated best-fit RMSD value differed from the target value, the atomic derivatives of this term forced the system toward or away from the target (depending on the sign of K).

$$E_{RMSD} = KN(RMSD_{current} - RMSD_{target})^2$$

Equation 2-1. Targeted MD energy. K is the force constant and N is the number of atoms.

2.2.4 Decoy Generation

The structures for the decoy sets were generated through the following sets of simulations with GB solvation, all at 300K unless noted. The following force fields were employed: ff94, ff99 and ff94 without ϕ/ψ dihedral terms. MD simulations were performed with backbone atoms restrained to the native conformations. Unrestrained MD simulations starting from the native conformation provided additional local fluctuations in structure. Targeted MD simulations were employed to unfold and refold the structure 45 times, with different force constants, by linearly scaling the backbone target RMSD each 2ns between 8.0Å and 1.0Å. A cluster analysis of a trajectory at 800K was performed, and representative structures from the clusters were saved, quenched and simulated at 300K to locally explore these basins of attraction. A similar approach was used for each model peptide. A total of nearly 500,000 decoys were generated for trpzip2 and 250,000 for αlac 101-111.

2.2.5 Genetic Algorithm

A program was written to carry out the genetic algorithm (GA) procedure for modification of the force field parameters. Input to the program included energy, RMSD and backbone ϕ/ψ dihedral values for each structure in the decoy sets. The energy values were evaluated without the ϕ/ψ dihedral terms. The gene consisted of phase (γ) and amplitude (V_n) values for each of 4 cosine terms in the Fourier series (Equation 2-2) for ϕ

and ψ, resulting in a gene length of 16. The series was limited to n=4 in analogy with ff94, which used terms up to n=4. One gene was initialized to zero for all variables, all other population members were assigned random initial values.

Amplitude values were allowed to take any value; this could result in an asymmetric Ramachandran plot that may not be desirable for glycine. However, the point of the fitting in this case is to obtain a force field that favors native conformations rather than one that is ideally transferable. A separate parameter set could be fit to glycine, but with more variables an increase in the number of decoy sets would be desirable to avoid overfitting.

$$E_{dihedral} = \sum_{n=1,4} \frac{V_n}{2}\left(1 + \cos\left(n\phi - \gamma\right)\right)$$

Equation 2-2. The energy for dihedral angle (ϕ) calculated in AMBER or the GA program (Dihedral term of the molecular mechanics equation shown in Equation 1-1).

One parent was chosen biased by fitness using a roulette-wheel scheme, the other was chosen randomly. Crossover points were selected randomly, with two points employed so that swap of internal gene segments was permitted (but not required). A mutation rate of 0.1 was used for each gene element, and mutation consisted of replacement of the value with a new random value. Amplitudes were limited to 0.5 kcal/mol. Offspring that were duplicates of parents were discarded. Elitism was employed, with the highest ranking half of the parents and offspring carried over to the next generation.

The fitness function had 2 components for each sequence: the energy gap between native and non-native structures, and the energy vs. RMSD slope. Native conformations

were defined as those with RMSD values under 1.0, while RMSD values above 1.7 were considered non-native. This difference reduced the arbitrary nature of using a cutoff value for native conformations. Average energies were calculated for the 1000 lowest energy native structures and the 1000 lowest energy non-native structures. The energy gap was defined as the difference between these two values, with a positive value indicating that the global minimum is native-like. The fitness value was calculated as the geometric mean of the energy gap and slope for each of the two sequences. A large gap has been suggested to be critical to the existence of a folded state .The combination of energy gap and increase in energy with decreasing similarity to native conformations has been used in the past to optimize non-atomic detail energy functions [54, 55]. A detailed discussion of the selection of the fitness function and its properties are not essential to the focus of this work will and will be presented in a separate publication [24].

The program was written to employ the MPI parallel library. During every generation, 50 MPI processes each evaluated the fitness of one gene (parameter set). This included calculation of the ϕ/ψ energies for all structures for that gene, and adding these values to the input energies for all other terms in the force field. The population was allowed to evolve for 200 generations, resulting in over 7x106 energy evaluations (roughly equivalent to 15μs MD simulation with a 2fs time step). Since only the parameters for ϕ and ψ were optimized, it was not necessary to evaluate the other terms in the energy function, resulting in a dramatic speedup compared to re-calculation of these terms for each new parameter set. Inter-process communication involved only collection of fitness values for each parameter set, thus nearly perfect scaling was achieved even with commodity 100 mbit/sec network interfaces.

2.2.6 Data Analysis

Cluster analysis, dihedral angle evaluation and RMSD calculations were carried out using MOIL-View [66]. Unless stated otherwise, RMSD values for trpzip2 were calculated using the backbone atoms in residues 2-11 since those are well defined in the family of NMR structures. For αlac 101-111, backbone atoms in residues 2-8 (corresponding to 102-108 in the intact protein) were used.

2.2.7 Cluster Configuration

All calculations were carried out on our Beowulf-type Linux cluster consisting of 20 dual 800MHz PentiumIII nodes and 50 1.4GHz AMD Athlon nodes. All nodes have 256MB RAM and are configured as diskless machines writing data to a SCSI RAID array on a central file server. The entire cluster is on a private (i.e. non-routable IP) network with nodes interconnected by an HP ProCurve 4000M Ethernet switch at 100Mb/s. The programs were compiled with GNU compilers and the publicly available MPICH library. The cost of the cluster was ~$70,000. We have provided further information such as cluster details, benchmarks and AMBER MACHINE files on the main AMBER web site at http://amber.scripps.edu/cluster_info/index.html .

2.3 Results and Discussion

2.3.1 MD simulation with ff94 and ff99

We performed simulations of trpzip2 using ff99 in explicit solvent at 300K and monitored the RMSD of backbone atoms as a function of time (Figure 2-1). Trpzip2 remained stable with an average RMSD value from the initial conformation of ~ 1.0Å for approximately 15ns. After this time an increase in RMSD was observed, with the deviation remaining above 2Å for the remainder of the ~135ns simulation. Analysis of the backbone structure revealed that residues 1-3 and 10-12 underwent a transition from extended to α-helical conformation (Figure 2-2). This was accompanied by a decrease in average energy of the solvated system of ~7 kcal/mol (Figure 2-3). Simulations with ff94 behaved similarly, with transition to helical conformation in the terminal residues observed within 6ns.

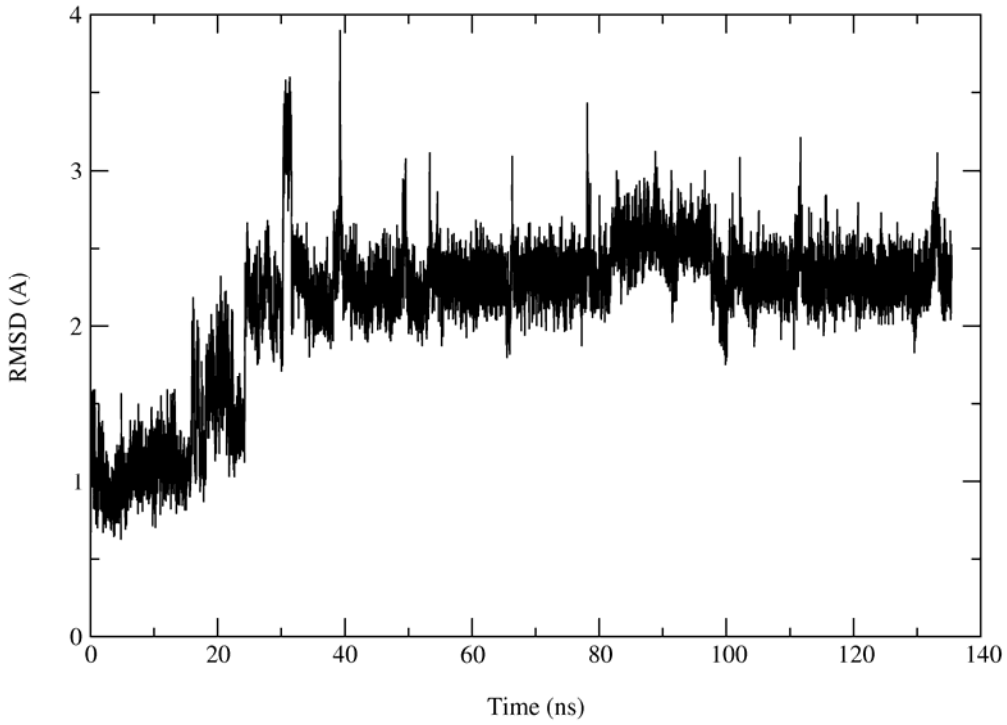


Figure 2-1. Backbone RMSD (residues 2-11) vs. simulation time for trpzip2 in explicit solvent at 300K using ff99. After staying native-like for ~15 ns an increase in the RMSD observed.

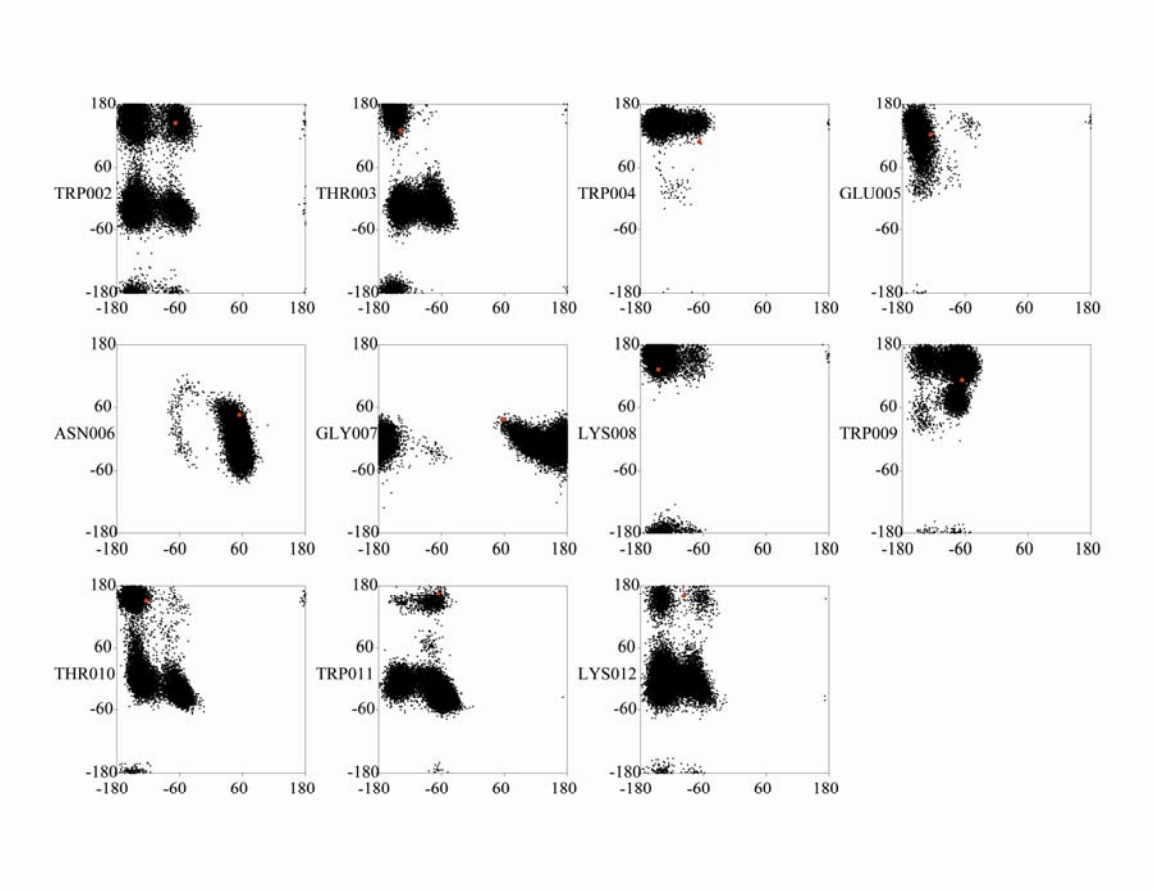


Figure 2-2 Ramachandran plots for each trpzip2 residue during explicit solvent simulation using ff99 at 300K. Residues 2-3 and 10-12 sample non-native α-helical conformations.

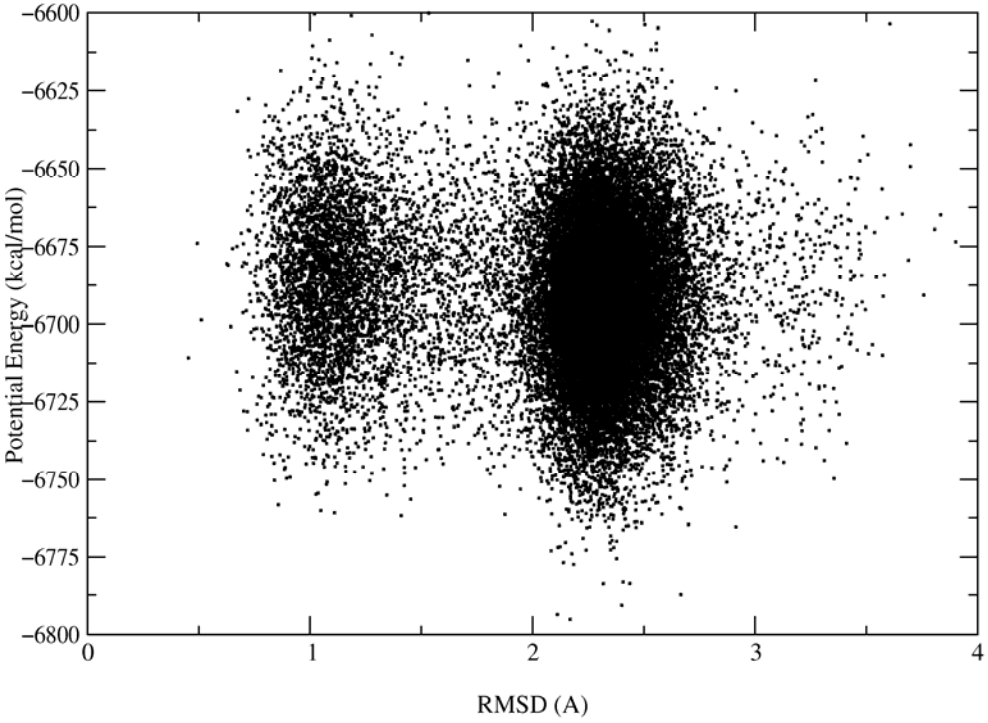


Figure 2-3. Energy vs. RMSD graph for the explicit solvent simulation of trpzip2 using ff99. The first cluster represents the native conformation and the second and more populated cluster (at RMSD values of ~2.5Å) represents the structures where the residues at the ends sample helical conformations. Some structures in the second cluster have lower energies than the native one.

Even though the simulation was relatively long (135ns), only two structural families were observed and it is likely that only local equilibration has been achieved. Several possibilities for increasing the transition rate were explored. First, simulation in explicit solvent at 550K was performed. In this case, the peptide structure converted within 3ns to an α-helix which remained stable for the remainder of the ~12ns simulation. This should be surprising, since this sequence is one of the most stable short β-hairpins that has been studied [59]. In addition, 550K is well above the 345K trpzip2 melting temperature [59]

and no significant structure should be observed, suggesting an incorrect stabilization of helical conformations.

The second method that we tested to increase transition rates was to employ the GB solvation model during MD simulation, obtaining increased rates due to the lack of solvent friction. In this case, the same α-helical conformation was found at 550K, but the timescale was ~300ps, roughly 10x faster. The simulation was repeated at 300K and the helical conformation was located at 9ns. This correspondence of the converged structures obtained from the continuum and explicit solvent models suggests that using GB to explore structural properties may be more efficient and provide similar results.

For the αlac 101-111 GB simulations with ff94 and ff99 at 300K showed a strong tendency for the helix to extend beyond His107 to the full length of the sequence (Figure 2-4). This is in contrast to the family of structures refined using experimental NMR data, which indicate that the helix is terminated after His107, and the final 4 residues may be disordered. This helical conformation persisted throughout 35ns simulation in both cases. A 50ns simulation initiated with a fully extended conformation with ff99 converged rapidly to the same helix, which remained stable. This again suggests an overstabilization of the helical conformation. The average structures obtained with ff99 for the trpzip2 and αlac 101-111 sequences differed by less than 1Å, demonstrating no sequence dependence in stark contrast to experimental observations.

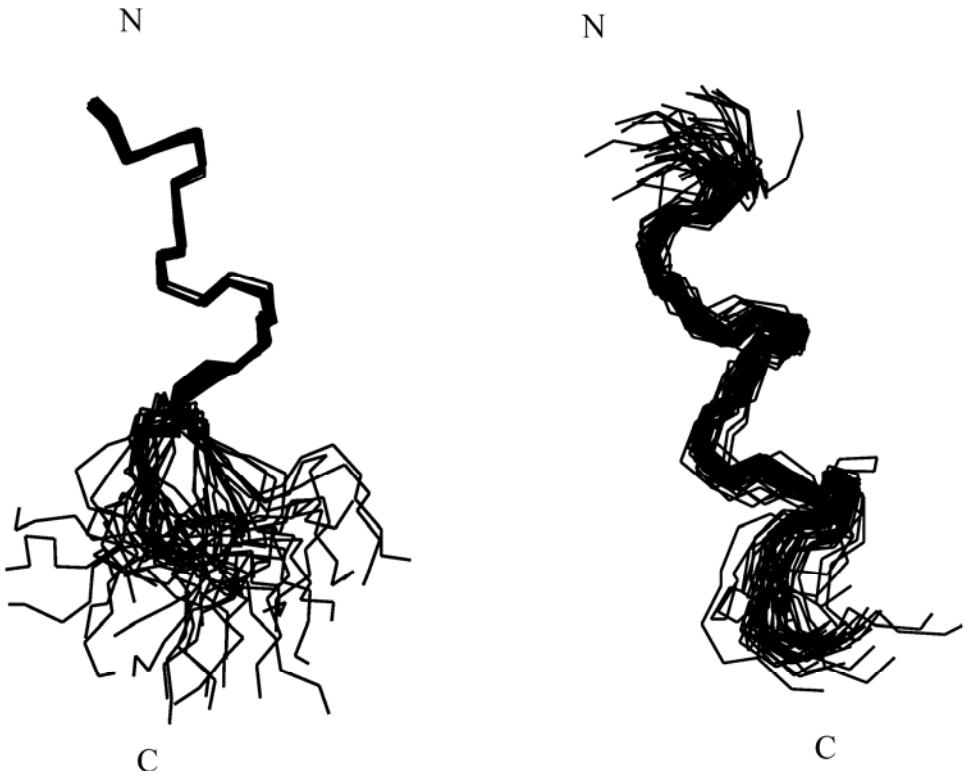


Figure 2-4. Backbone structures of αlac 101-111, with the family of NMR structures shown on the left and snapshots from 35ns MD with ff99 on the right. Residues 108-111 are not well defined in the NMR family but are always helical in the simulation.

2.3.2 Generating and analyzing decoy structures

While multi-ns length simulations are becoming routine, they still require significant resources and it is difficult to obtain information about conformations far from initial structures. It is also impractical to repeat these simulations for each new variation of a force field in order to test the behavior. Perhaps the most important drawback is that unless full exploration of phase space can be achieved, the behavior observed in such simulations may be determined to a great extent by the barriers to conformational transitions, rather than the relative energies of different conformations. Therefore, this is

not an optimal approach for evaluation of the thermodynamic properties of a force field. It is clear that simulations need to be extended to much longer timescales, even on small systems such as those discussed in this article. Length of the simulation is not the only important factor; great care must be taken with approximations that increase timescale at the expense of accuracy.

The approach that we investigated involved generating large sets of "decoy" structures that were reasonable conformations for the peptide at 300K, including local and global structural variation. These sets of structures can provide a more complete view of the conformational preferences of the sequence with a given force field than is possible from a single long simulation. The structures were not generated randomly, but rather with MD simulations at 300K to ensure that all were reasonable structures for these sequences (see Methodology section for details), however, they do not reflect a distribution in any particular ensemble. Altogether nearly 750,000 structures were obtained from ~1.5μs of simulation data. Although these decoy sets represent microsecond-length MD at 300K, the diversity of structures is likely to be much greater than observed in a single 1 μs simulation due to the many initial structures and forced conformational sampling.

The diversity of the structures can be evaluated in many ways. Two examples are presented; the coverage of Ramachandran space for the trpzip2 decoy structures is demonstrated in Figure 2-5, and a histogram of RMSD values is shown in Figure 2-6. The Ramachandran plots demonstrate significant sampling of fluctuations about the traditional secondary structure basins of attraction and look similar to those extracted from large sets of known protein structures. The Gly7 distribution is much broader and

nearly symmetric, consistent with expectations due to the lack of a side chain. Larger structural variations are also represented, with significant numbers of structures with RMSD values ranging up to 8Å. The properties of the decoy set for αlac 101-111 are similar, with RMSD values ranging up to 5Å.

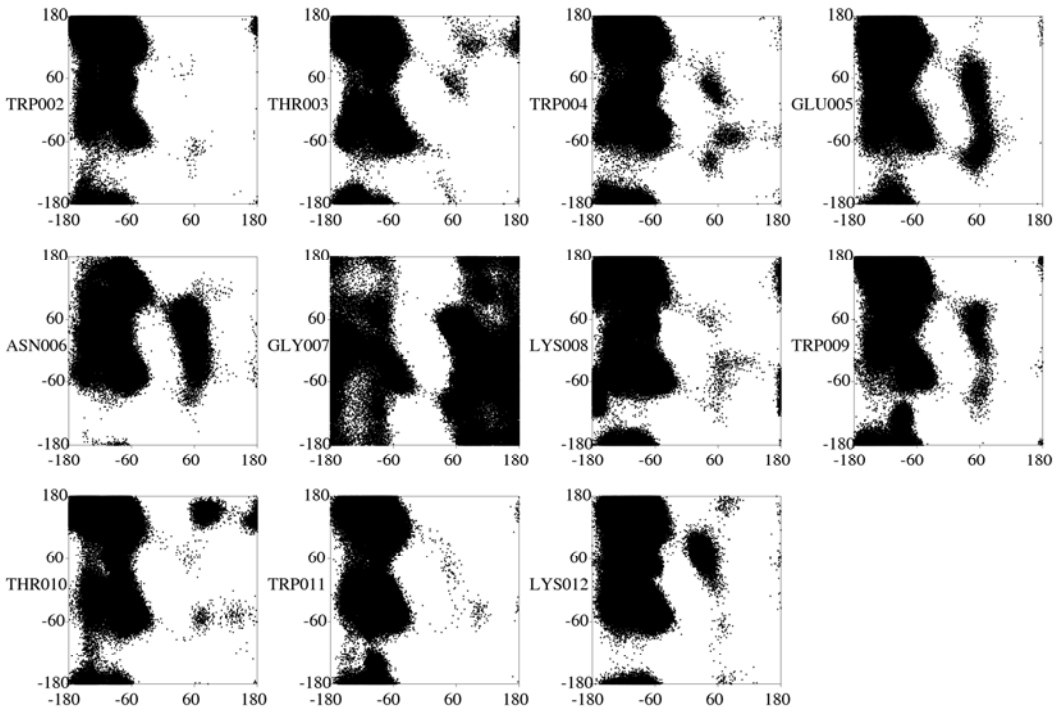


Figure 2-5. Ramachandran plots for each trpzip2 residue for the structures in the decoy set. All secondary structure areas are sampled extensively, with Gly7 showing the expected broad and nearly symmetric distribution.

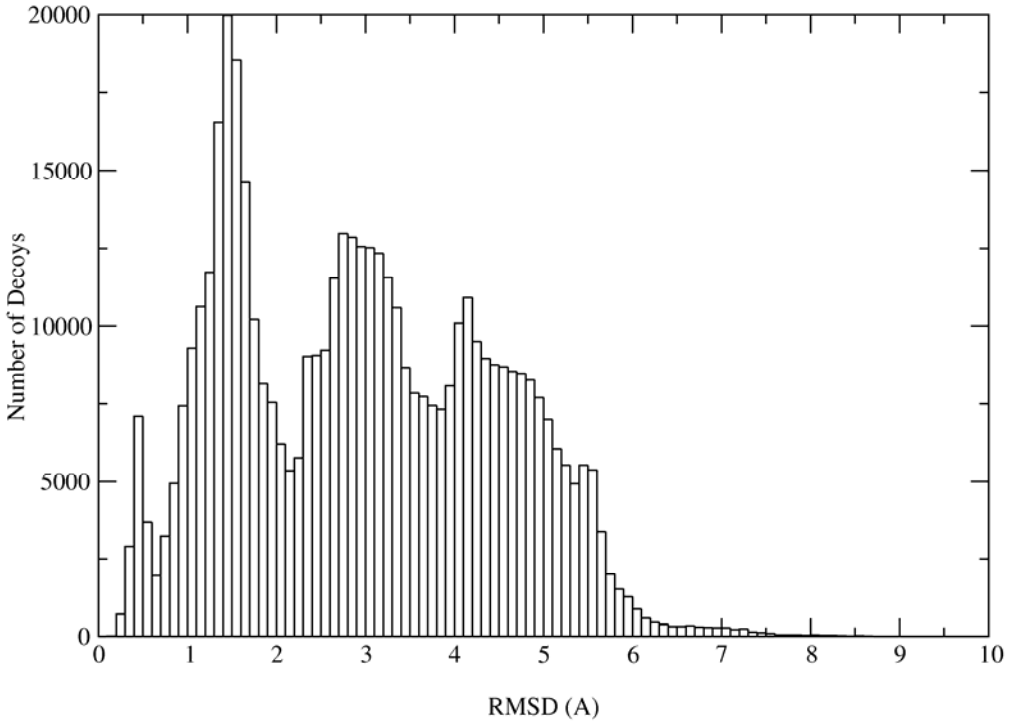


Figure 2-6. Histogram of RMSD values for the trpzip2 decoys, showing significant numbers of structures with RMSD values up to 6Å.

The conformational preferences of each sequence in a particular force field can be estimated by recalculating energies for this set of reference structures. Efficiency is increased 1000-fold because the number of energy evaluations is far less than was required to produce all of these structures, since snapshots were saved every 1000 steps of MD. The analysis requires the number of energy evaluations required for ~1.5ns of MD simulation and can currently be performed in a matter of hours. Even further speedup can be obtained since this evaluation is trivial to parallelize on a PC cluster by splitting the decoy set into subsets for each processor, requiring no inter-process communication.

In Figure 2-7 we show the potential energy for the decoy structures as a function of RMSD to the native conformation for ff94. This is simply a 1-dimensional projection of the energy landscape of the peptide. For trpzip2, it is clear that the lowest potential energy values are not at low RMSD values, in fact the global minimum is near 4Å. When these low energy structures are analyzed it is found that they are helical, consistent with the results of the long MD simulations. Although only residues 101-107 of αlac 101-111 are well-defined in the NMR family of structures, the RMSD is shown for all backbone atoms compared to the fully helical structure. This allows us to observe that the longer helix is ~11 kcal/mol more stable than alternate conformations, again consistent with the results of the long single MD simulations.

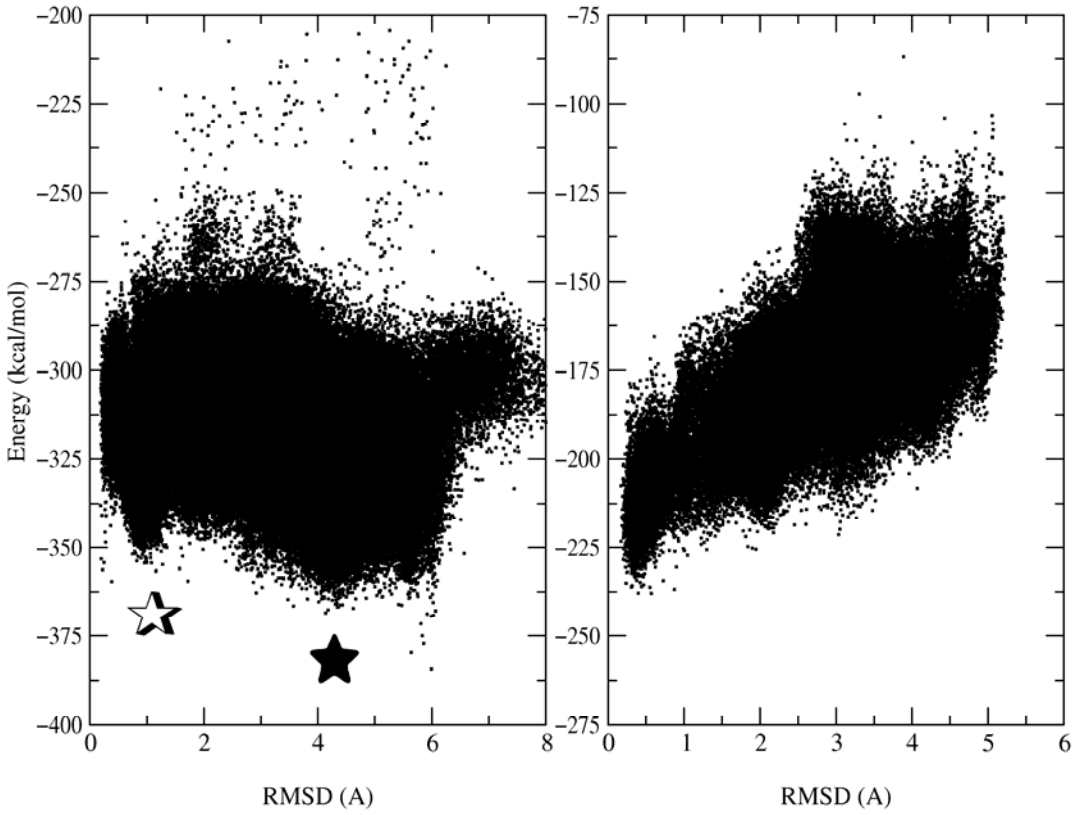


Figure 2-7. Energies of the decoy structures calculated with ff94. For trpzip2 (left) the global energy minimum (filled star) is helical and differs significantly from the native (open star) conformations. Energies of the αlac 101-111 decoy structures (right) calculated with ff94. RMSD values are calculated with a completely helical reference structure, demonstrating that this full helix is ~11 kcal/mol lower in energy than other structures.

In Figure 2-8 we show the decoy results using the more recent ff99 parameters, in which the ϕ/ψ dihedral parameters were re-fit in order to improve the relative energies of alternate peptide conformations. It is perhaps surprising to note that the decoy energy profiles are very similar to those obtained with ff94, despite the difference in the ϕ/ψ parameters in these force fields (Figure 2-9). These plots demonstrate the important result that ff94 and ff99 do not reproduce the experimentally determined sequence dependence

of peptide structure; instead a similar helix (1Å RMSD) is the global energy minimum for both sequences in both parameter sets.

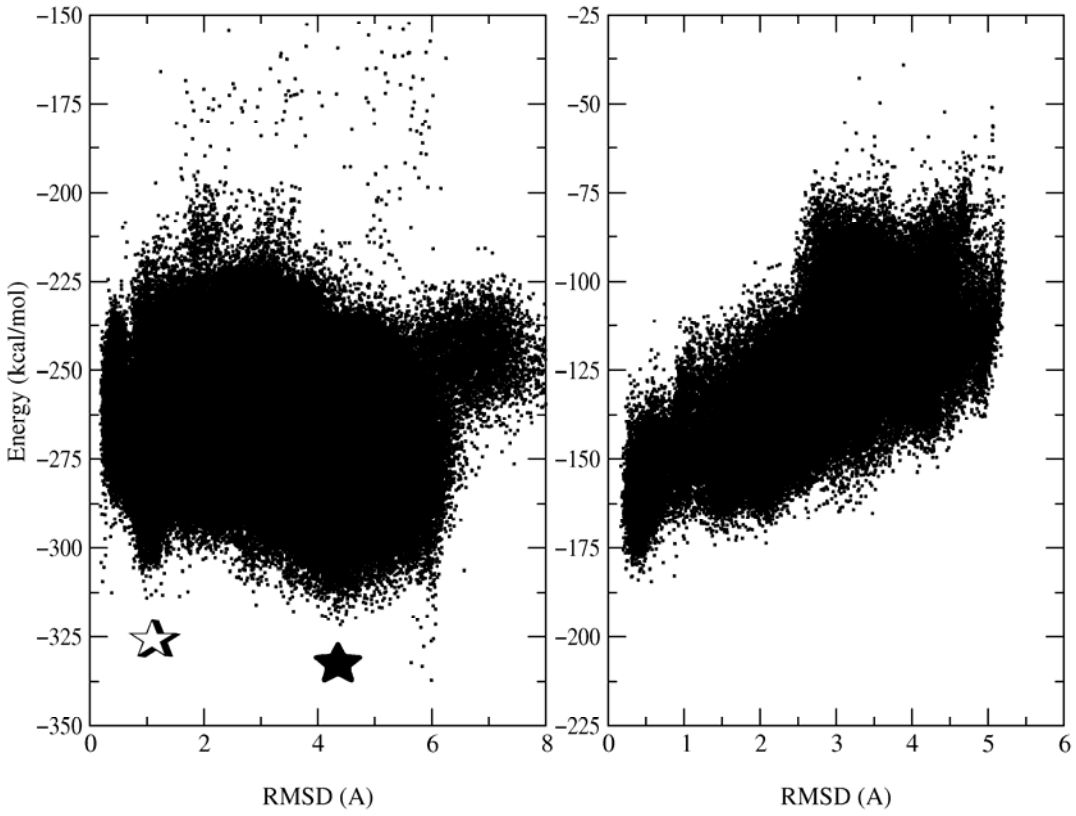


Figure 2-8. Energies of the decoy structures calculated with ff99. The profiles are very similar to that for ff94 (Figure 2-7). For trpzip2 (left plot), the global energy minimum (filled star) still differs from native conformations (open star). The helical conformation of αlac 101-111 is lowest in energy.

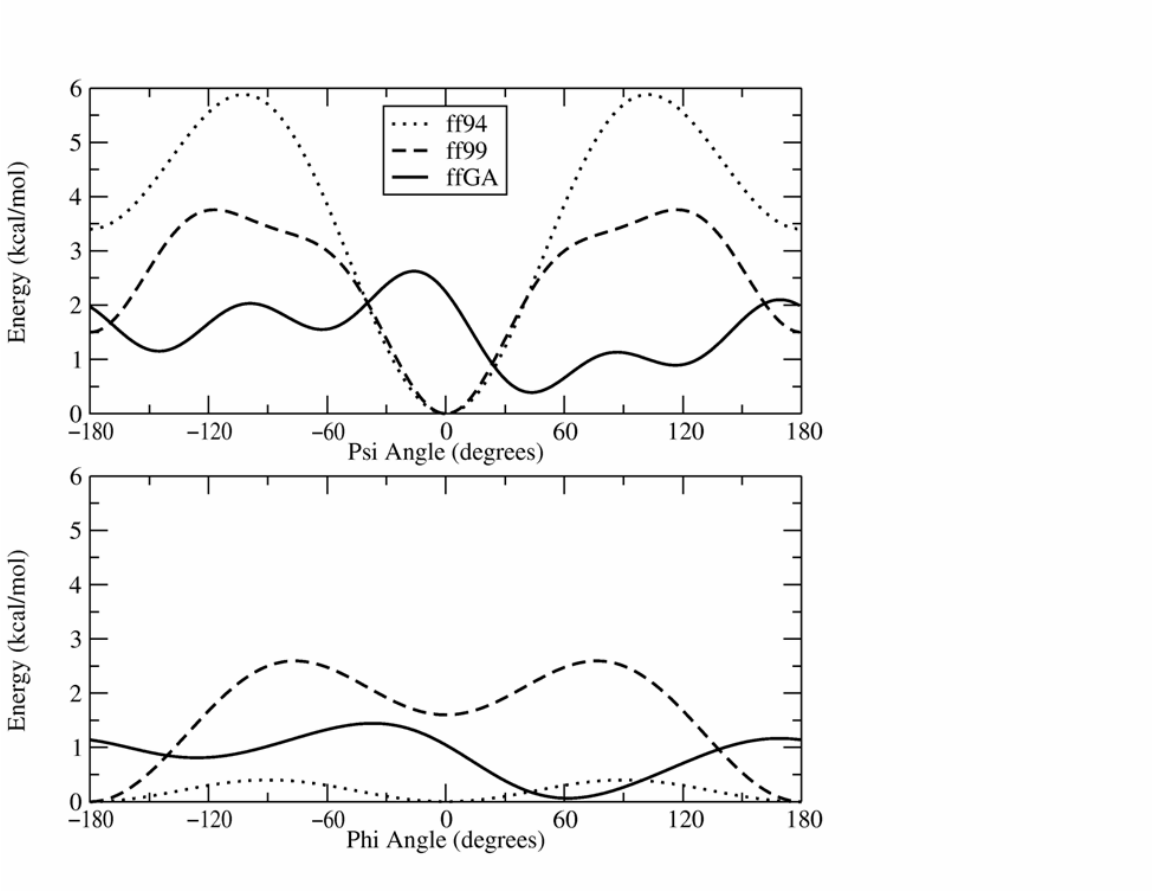


Figure 2-9. Energy profiles for rotation about ϕ (lower) or ψ (upper) backbone dihedrals. Profiles are shown for the three parameter sets discussed.

2.3.3 Interpretation of decoy results

For both model sequences, very long MD simulations with a given force field converged to structures that were similar to the lowest energy decoy structures. This indicates that decoy evaluation is a useful way to test the preferences of a given parameter set. For ff94 and ff99, a suspected bias toward helical conformations was confirmed and both fail the evaluation in that native conformations are not energetically preferred. However, passing the decoy test is not a sufficient condition to conclude that a force field is "correct". The relative energies for members of the non-native ensemble are

not considered in this procedure, yet these energy differences directly influence important properties such as entropy contributions, equilibrium constants and melting temperatures. Further analysis is therefore needed to determine if a parameter set will reproduce the actual conformational preferences of a particular sequence beyond identifying or predicting the native conformation.

Since ff94 and ff99 failed the simpler decoy test, it is not appropriate to use them to test these advanced properties of force fields. Additional existing parameter sets could be tested, however more insight into the properties of decoy sets may be gained by modifying the ff94/ff99 parameter sets based on the decoy results. Each set is necessarily a limited subset of all possible conformations for the sequence, and the completeness of this conformational basis set influences the transferability of the results and the validity of the insight that can be extracted.

The conformational preferences of ff94 and ff99 should be well represented by the set of decoys since each was used to generate many of the structures. If using the same set to evaluate a different force field is acceptable, many thousands of CPU hours could be saved for each parameter set. However, if the decoy set does not represent an adequate range of local fluctuations within a given basin of attraction, subtle shifts in the positions of minima might preclude calculation of accurate relative energies for alternate force fields. If the decoy sets do not sample enough low-energy basins, a poor parameter set could pass the test by having the absent non-native basin as the global minimum. We address these questions by using the existing decoy data to modify the ff99 parameter set and comparing properties predicted using the decoys to those actually obtained with the resulting parameters. The goal of this fitting is not necessarily to obtain an ideal set of

parameters, but rather to demonstrate how decoy sets could be used for empirical fitting, test the completeness of the decoy sets and to obtain a parameter set that has native-like conformations as lowest in energy to permit further analysis of entropic effects. For maximum transferability, inclusion of more sequences with greater diversity in secondary structure would be highly desirable.

2.3.4 Using decoy results to guide modification of force-field parameters

Since the evaluations described above indicated that the largest problem with the ff94/ff99 parameters is a sequence-independent secondary structure bias, we focused on the Fourier series parameters for the ϕ and ψ rotational energy profiles.

We directly used the decoy set energy vs. RMSD profiles to re-optimize these parameters by using a genetic algorithm (GA) program written for this purpose, with further detail provided in the Methodology section. After 200 generations, the parameter set (denoted ffGA, Table 2-1) with the largest fitness value was selected for further analysis. Even though periodicities up to n=4 were allowed, the n=3,4 terms for ϕ have amplitudes of nearly zero and therefore have little contribution. The dihedral energy profiles for these parameters are shown in Figure 2-9.

	ϕ		ψ	
N	amplitude (kcal/mol)	phase (radians)	amplitude (kcal/mol)	phase (radians)
1	0.40	4.58	0.48	4.78
2	0.41	5.29	0.45	5.39
3	0.02	5.02	0.12	5.76
4	0.02	5.81	0.45	5.52

Table 2-1. Optimized parameters for ffGA. Other force field parameters were the same as ff99.

Before this optimization was performed, it was also unclear if refitting only ϕ and ψ dihedral parameters could remedy the problems described above. The energy vs. rmsd profiles for decoy structures with this parameter set (Figure 2-10) differ significantly from ff94/ff99, indicating the sensitivity of the decoy relative energies to these backbone dihedral parameters. The native conformations of both sequences are now the global energy minima, demonstrating that a single parameter set is able to correctly and simultaneously identify native conformations for both secondary structure types. For trpzip2, the energy difference between the lowest energy native and non-native conformations increased from the ff99 value of –16 (native was less stable) to 8.1 kcal/mol. It is interesting to note that the αlac 101-111 energy gap was decreased from the ff99 value of 10.7 to 5.5 kcal/mol. Thus requiring a single parameter set to simultaneously prefer native conformations for both secondary structures reduced the maximum attainable αlac 101-111 stability.

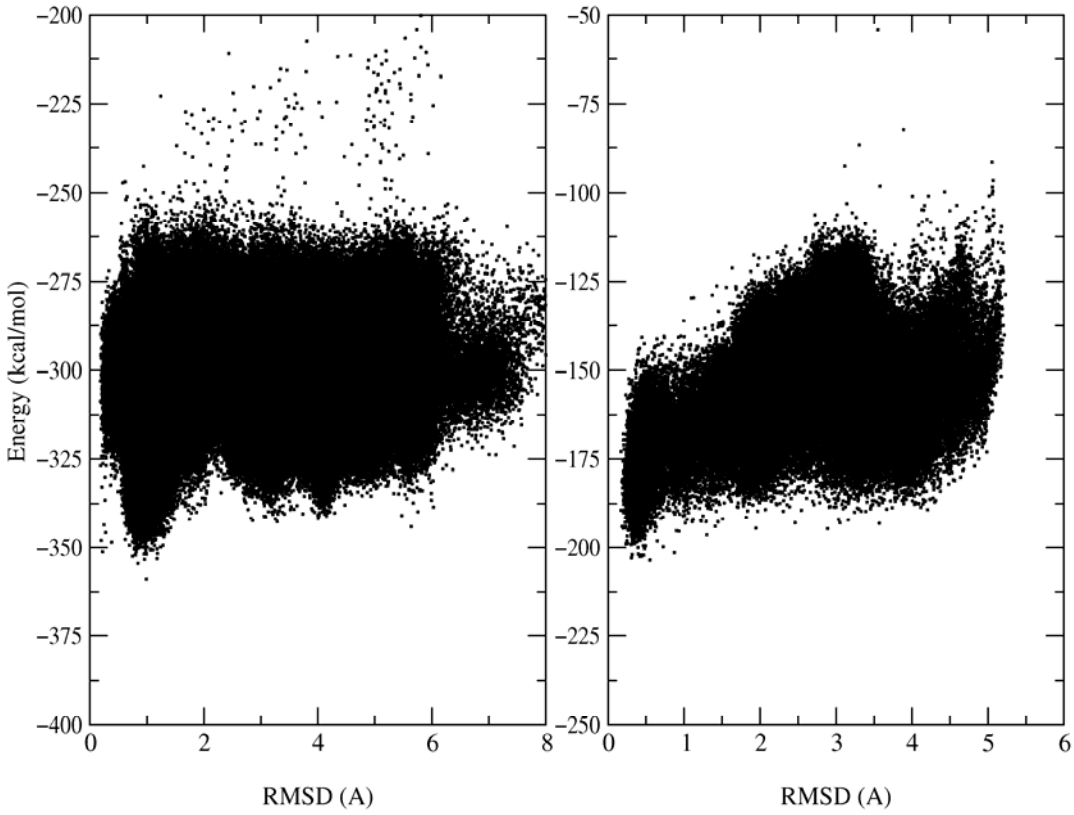


Figure 2-10. Energies of the decoy structures calculated with ffGA. In both sequences, the global energy minimum is now native-like, but the stability of the αlac 101-111 helix is reduced compared to ff94 and ff99.

2.3.5 Testing the transferability of decoy set: MD simulation with ffGA

2.3.5.1 αlac 101-111

Three simulations were carried out for αlac 101-111 with ffGA starting from conformations obtained from ff99. In each case, the experimentally determined helical conformation near 101-107 was reproduced, but the region from 108-111 was substantially more flexible than was observed with ff94 or ff99 (Figure 2-11). This is

consistent with the family of NMR structures (Figure 2-4) and the analysis by Demarest et al. indicating that the structure for this region was not well defined [61]. This demonstrates that the simulations with parameters based on decoy set analysis are able to reproduce not only sequence-dependent structure, but also sequence dependent stability at a given temperature. In two of three simulations, transient loss of helical content is observed, consistent with the reduced energy gap in the decoy set when using ffGA as compared to ff94 and ff99. It is clear, however, that the reduction in stability of the long helix is not distributed evenly in the sequence, but predominantly in residues 108-111, consistent with experimental observations.

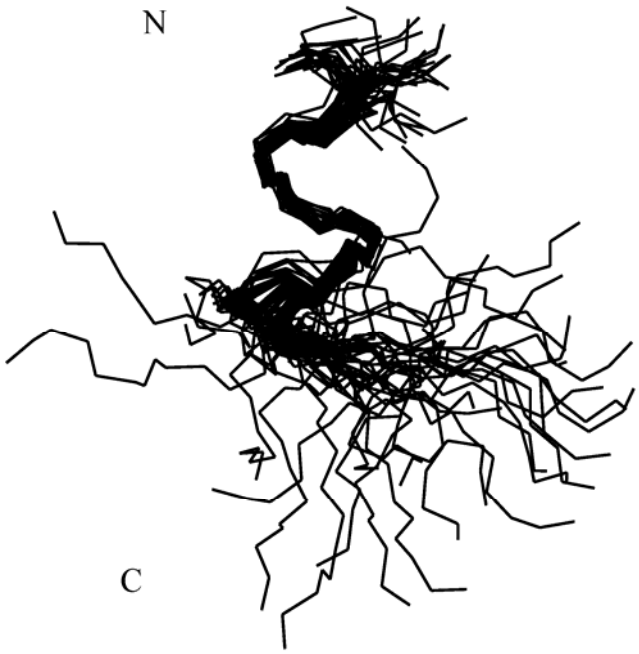


Figure 2-11. Snapshots of αlac 101-111 from MD simulation with ffGA. The helical conformation in residues 101-107 is stable, but 108-111 show significant flexibility, similar to the family of NMR-derived structures (Figure 2-3).

2.3.5.2 Trpzip2

MD simulations of trpzip2 using ffGA were carried out at 300K starting from the native conformations of each sequence. In contrast to ff99, the hairpin conformation was stable using ffGA, with an average RMSD of ~1.0Å during the entire 30ns simulation and no unfolding events. The large twist in the native conformation was maintained, as well as the close interaction between the Trp side chains. While this is an encouraging result, it merely indicates that the native conformation is a stable, though perhaps local, minimum in this force field. Such stability can be the result of kinetic trapping at this temperature rather than favorable thermodynamics. It does not demonstrate that the decoy set was complete enough to represent a sufficient number of thermally accessible basins of attraction.

Direct comparison of fractional population can provide additional validation. Since reaching a true equilibrium state at 300K is likely beyond our present capability, an additional 60ns native simulation was performed at 350K, slightly above the experimental melting point of 345K. Multiple unfolding/refolding events were observed. A 100ns simulation starting from a distorted hairpin structure was able to convert to the observed native conformation within ~3ns, and thereafter resulted in similar behavior. The time course and histogram of RMSD values for both simulations are shown in Figure 2-12. The large peaks near RMSD values of 0.8 demonstrates that significant native population is present at this temperature. Using a somewhat arbitrary cutoff value of 1.5Å for native conformations (the end of the first peak), the calculated native fraction of 0.3 to 0.35 for each simulation is in excellent agreement with the value of 0.41 determined by experiment [67]. However, folding/unfolding transitions are separated by

~20ns even at this elevated temperature, and the distributions are not identical. Increased statistics are required to ensure that sampling of phase space was adequate both in the decoy set and in the ffGA test simulations.

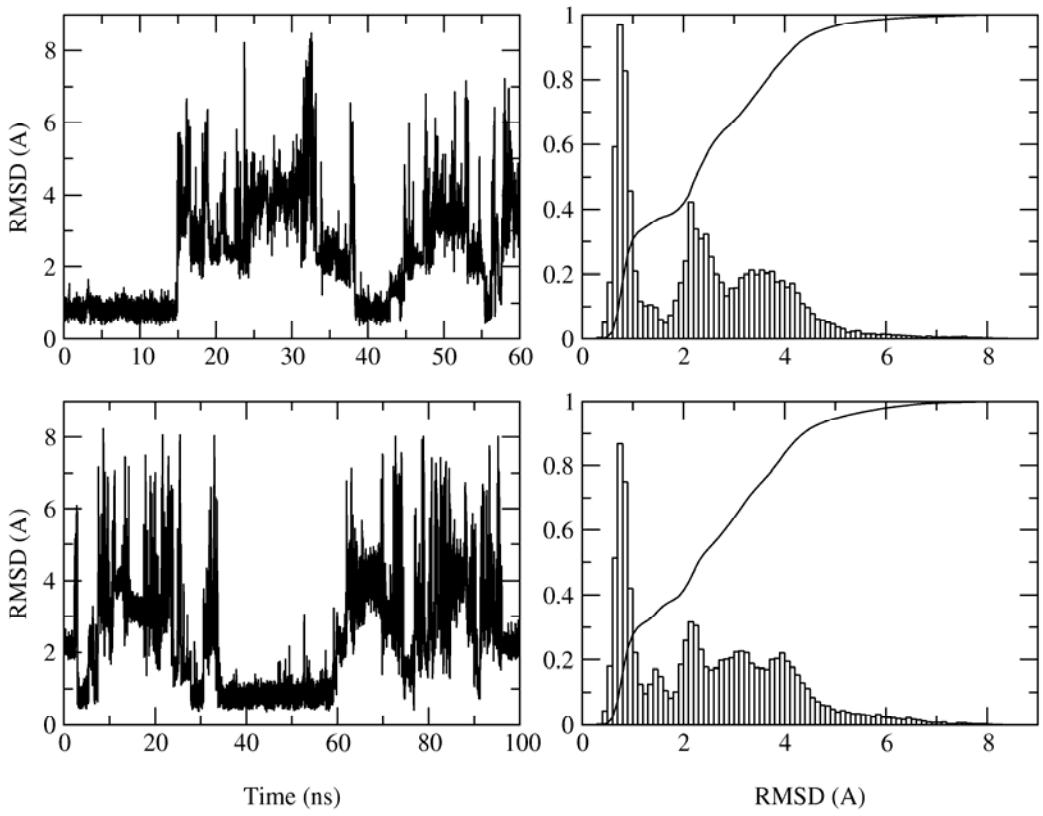


Figure 2-12. RMSD values (left) during two trpzip2 simulations using ffGA at 350K. Each shows multiple folding/unfolding events. Histograms of the RMSD values with integration curves are shown on the right, and are similar for the two independent simulations. A native-like fraction of 0.3 to 0.35 is calculated in each case, in excellent agreement with the experimental observation of 0.4 at this temperature.

A simulation at 300K starting from the helical structure obtained from ff99 was performed, and the peptide underwent a series of changes resulting in formation of the native hairpin at ~40ns, which persisted for the remainder of the ~80ns simulation. The folded structure is very similar to that determined by NMR (Figure 2-13), with the

exception that the stacking of the outer Trp side chains differs somewhat from the published structures.

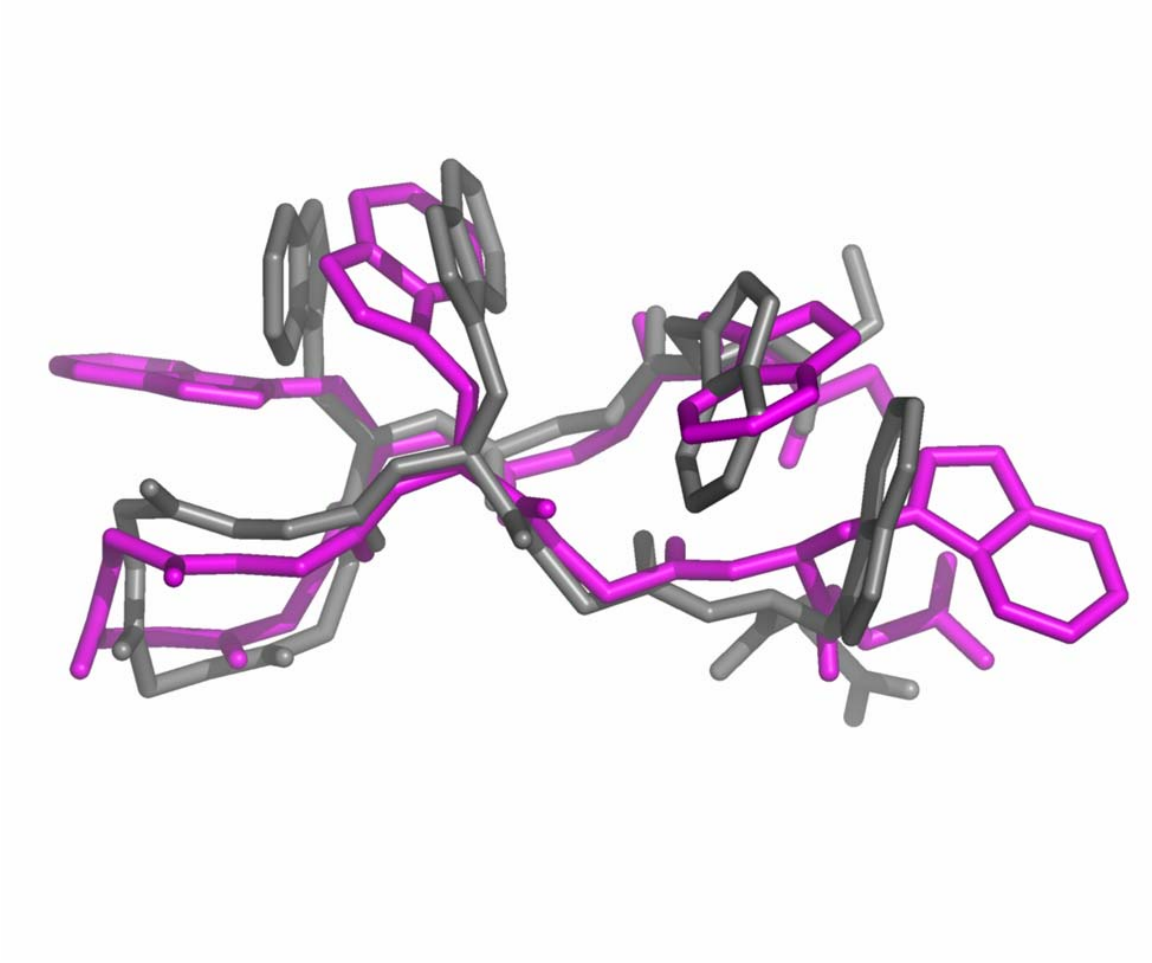


Figure 2-13. Overlap of the representative folded conformation of trpzipo2 using ffGA (purple) to the average NMR conformation (gray). For clarity, only the backbone and Trp side chains are shown. The backbone structures are very similar, including large twist in the β-sheet. The stacking of the outer Trp side chains differs from the published structures.

Since this still represents only a single folding event and has little statistical validity, 16 additional folding simulations were performed, initiated from representative structures from an ensemble generated at 800K with ff99. Each employed 4 CPUs and all were carried out simultaneously on the cluster, for a total of 72 CPUs and a speedup of ~50x.

A total of 650ns of data was collected at 350K. The evolution of RMSD with time is shown for the simulations in Figure 2-14. Fourteen of the 16 simulations (88%) folded to the native hairpin, with several simulations showing multiple unfolding/refolding events. This large fraction of folding is required to have confidence that the structure is the true native conformation of the force field, since it is possible to introduce bias when simply observing that a small fraction of simulations converts to what is known to be the actual folded form. Analysis of first crossing times gives an approximately exponential folding curve (Figure 2-15). These observations provide additional confidence that sampling of phase space in these simulations was adequate and that it is unlikely that the structures are kinetically trapped on this timescale. However, detailed analysis of the folding landscape is beyond the scope of the present article.

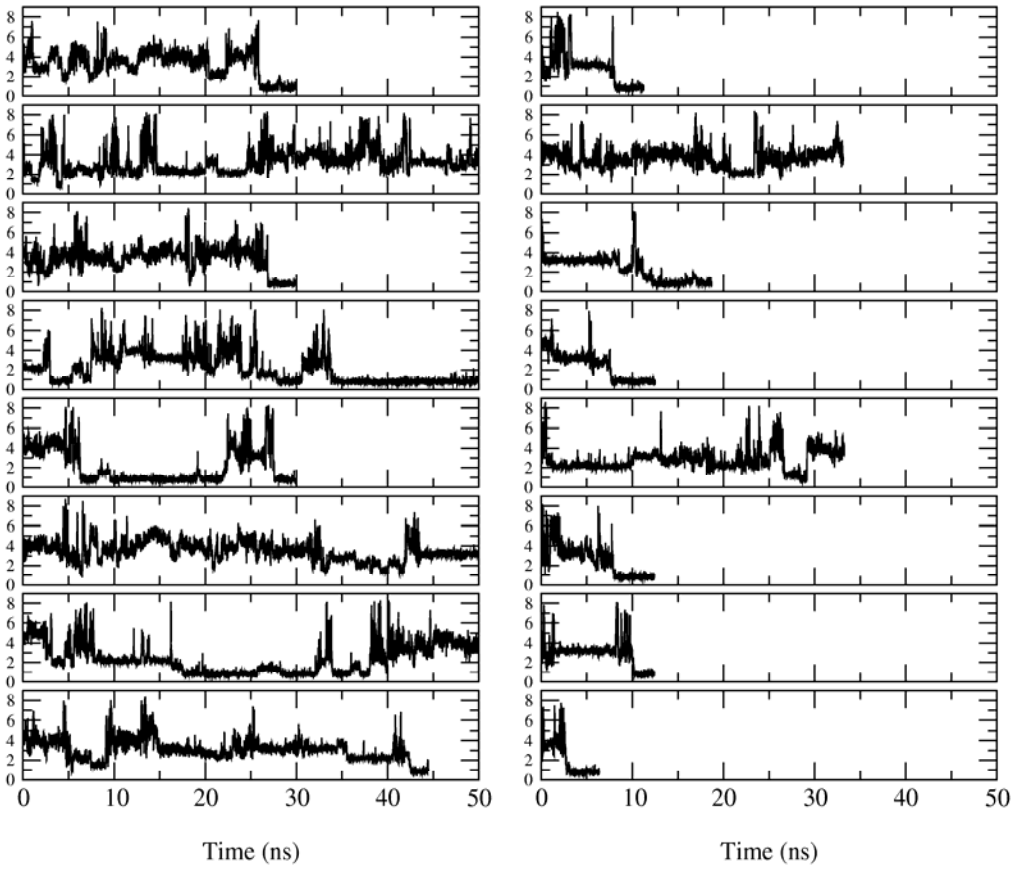


Figure 2-14. RMSD vs. time for 16 folding simulations of trpzip2 at 350K. Fourteen of the 16 simulations locate the native conformation, and several show unfolding/refolding events.

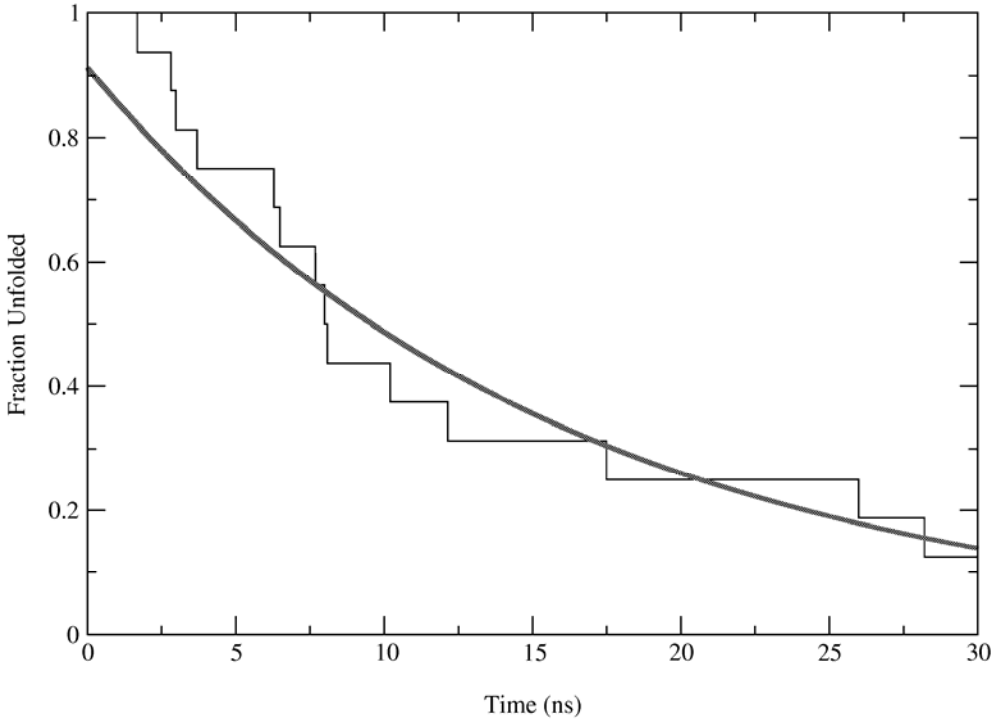


Figure 2-15. Trpzip2 folding curve (jagged line) calculated from the first crossing times observed in the folding simulations depicted in Figure 2-16. 88% of the simulations fold. The curve is approximately exponential (smooth line), suggesting reasonable statistics.

These simulations now permit evaluation of the transferability of the decoy set to a different set of backbone dihedral parameters. Structures from the 350K folding simulations were combined, and the ffGA energy vs. RMSD profile for these 325,000 snapshots is found to be very similar to that observed for the decoy set (Figure 2-16) despite the fact that ffGA was not used to create any of the decoys. The average energies are higher by the amount expected due to the larger thermal fluctuations, and the range of energies is reduced since all structures were generated with the ffGA parameters. Even with this extensive additional sampling, the native conformation remained the global

energy minimum. This demonstrates that the decoy set was complete enough to avoid creating a non-native global minimum during parameter fitting. If such structures were found, the decoy set could be expanded through an iterative process [54].

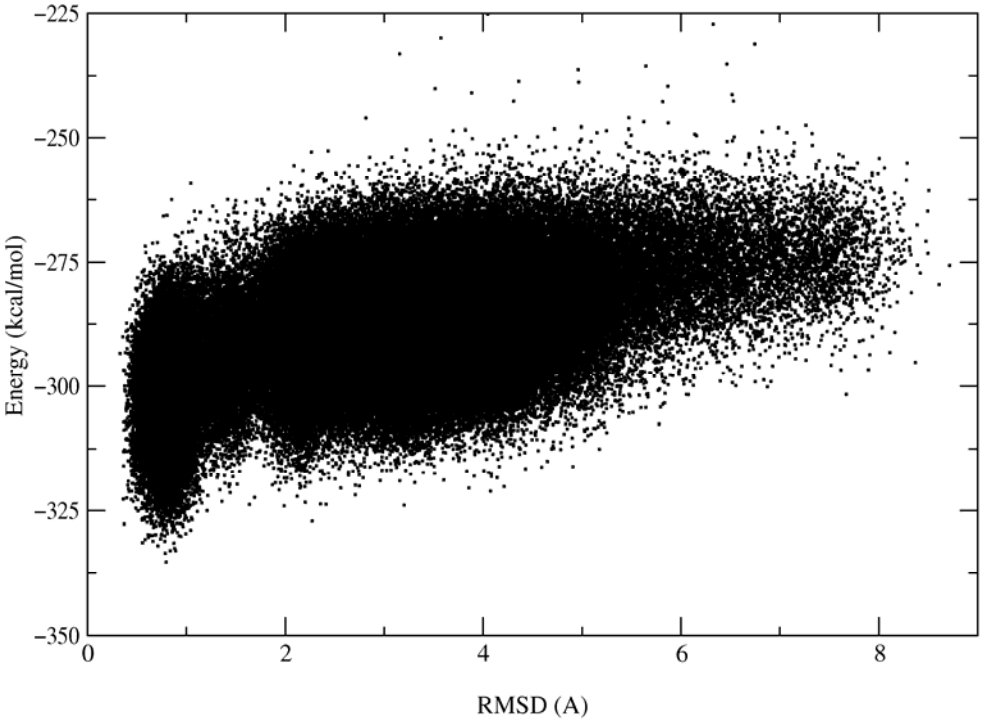


Figure 2-16. Energy profile for trpzip2 structures calculated from 650ns of ffGA simulation at 350K. The energy profile is very similar to that obtained with this force field for the decoy set, suggesting good transferability of the decoys.

The energy gap between native and non-native structures was predicted to be 8.1 kcal/mol from the decoy set, while the actual ffGA structures have a gap of 8.6 kcal/mol. This represents an error of only 2% of the predicted 24 kcal/mol change in relative energies from modification of the ff99 parameter set. The agreement is excellent

considering the differences between the conditions used to generate the two sets of structures, and indicates that such sets of decoy structures may have useful transferability.

2.4 Conclusions

Simulations with explicit solvent tend to be trapped in local minima and undergo significant conformational transitions only after tens of nanoseconds. Even when extended to over 100ns, these expensive simulations are unconverged and sample only a few backbone conformations. When a continuum solvent model is employed, ensembles sampled during 100ns still depend on initial conditions. The timescale of obtaining converged simulations even for short peptides remains daunting, and further efforts to improve conformational sampling are needed.

This indicates that even very long simulations starting from native conformations may not be reliable indicators of the quality of a biomolecular force field. Instead, long simulations from many different starting points provides a reference set of "decoy" structures, with extensive sampling of many basins of attraction. The explicit inclusion of many basins makes the decoy set significantly more valuable than single simulations of similar timescale, and permit evaluation of the transferability of parameter sets to systems of interest. Multiple folding simulations are also needed to ensure that the native conformation is the same as that identified through energy analysis. In this case a large fraction of the ensemble should fold; observation of small numbers of folding events cannot evaluate the force field in this manner. The multiple simulations also permit comparison of ensemble properties and fractional population to values determined experimentally. Temperature dependence of these properties, though not presented in this

article, can also be generated in an analogous manner and compared directly to experimental data. PC clusters are well suited to this task since very large numbers of processors can be used but high-speed communication is not required.

We generated a total of ~1.5μs of simulation to create sets of decoy structures, and demonstrated that a helical bias exists in the ff94 and ff99 force fields. While the bias had been suspected, this thermodynamic analysis demonstrated that it was clearly present even in solvated simulations of biomolecules. In order to determine whether the decoy sets could also provide critical feedback about improving, rather than just evaluating force fields, we used the data to modify the ff99 parameter set. The changes in properties predicted by the decoy analysis were very similar to those actually obtained from the simulations using the parameters, suggesting transferability of the decoy set.

The correspondence between simulation and experiments was dramatically improved with these modified parameters as compared to ff99, and it appears that the removal of helical bias allowed important sequence-dependent structural details to emerge. However, the current parameters should not be viewed as being generally applicable unless a more diverse set of sequences is employed during the fitting and testing procedures. The present results suggest that such an approach may be worth exploring. The near-linear scaling would permit many more sequences to be used for decoy generation and genetic algorithm optimization while still employing an inexpensive PC cluster with commodity components.

Chapter 3

Multiple pathways in β-hairpin folding and unfolding simulations

3.1 Introduction

An important aspect of the protein folding problem lies in understanding the process by which proteins locate their native conformations from the vast available phase space. Computational methods are an attractive way to tackle this problem since they can provide non-averaged data, in contrast to many experiments that supply only averages over time and/or macroscopic sample sizes. This is particularly important when multiple folding pathways may be involved, as we demonstrate below.

However, serious limitations also apply to computational approaches. The computational cost can make simulating folding for even a very small protein unfeasible due to the level of detail and relatively long timescale involved. Thus it is desirable to validate methods on small model peptides (stable secondary structure units) prior to future application on larger, potentially more interesting systems.

In contrast to experiments, a major drawback to simulation is the difficulty in obtaining well-converged ensemble-averaged data. This may contribute to the lack of consensus arising from simulation studies. Direct observation of folding events in unrestrained simulations is extremely challenging. Generating a sufficient number of these folding trajectories to obtain reliable insights into "typical" behavior of the

ensemble is rarely practical. An alternate approach is to generate thermodynamic properties with enhanced sampling techniques (such as Parallel Tempering [41]or Replica Exchange [42]), usually losing explicit time-dependent behavior (and "observation of folding") in the process. Using simulations in which the process was not observed to describe folding usually relies on interpretation of (free) energy barriers observed in a reduced dimensionality and/or along pre-determined order parameters. These may not accurately reflect the actual barriers or even the minima encountered during folding of individual members of the ensemble.

While both approaches have drawbacks, we combine them in this study and draw on the strengths of each. Comparing the two types of data for otherwise identical simulated systems allows us to identify which features of a particular simulation reliably represent characteristic behavior of the system as compared to individual observations. This provides additional validation of the methods, as any inconsistencies may indicate non-converged data or unreliable interpretations.

In the present case we focus on β-hairpin secondary structure, characterization of the native and unfolded ensemble and the changes that occur through the folding transition. β-hairpins have been studied extensively experimentally and computationally [36, 68-79] and several folding mechanisms were proposed where the difference in these is generally the balance between hydrophobic collapse and hydrogen bond formation, and the order in which the hydrogen bonds form.

Recently the designed Tryptophan Zippers (trpzips) [59] are becoming popular systems for experimental and computational studies because of their unusual thermodynamic stability and small size. Trpzips were designed by Cochran and

coworkers by mutating the hydrophobic residues of the well known and studied GB1 hairpin to tryptophans for increased stability [59]. Our group did the first simulation study of the trpzip2 [22] and we could not reproduce the proposed face-to-face tryptophan stacking as proposed in the original experimental paper [59] (PDB code: 1HRX) (Figure 2-13). The structure was then refined and the tryptophan orientations were changed to edge-to-face packing as seen in our results [80](Refined PDB code: 1LE1). Snow and coworkers have studied the folding kinetics of trpzip peptides via T-jump spectroscopy and molecular dynamics [81]. In their study the molecular dynamics simulations were carried out using GB/SA solvent model with an extra viscosity term. Simulations were performed using distributed computing and the folding rates were estimated assuming irreversible and two-state folding and they were in close agreement with experiments. Yang and coworkers studied the thermal and chemical unfolding of trpzip2 using CD spectroscopy and replica exchange molecular dynamics [82]. At different denaturant concentrations they measured change in the melting temperature and in their simulations they calculated it using different order parameters and observed some differences as well. They concluded that under optimal folding conditions the hairpin has unusual folding kinetics. Later when they do kinetic measurements using T-jump spectroscopy, they see different pathways and their folding kinetics looks like a double exponential decay [83]. They concluded that there are kinetic traps on the energy landscape and there are some misfolded structures having incorrect tryptophan pairings and it may be possible that folding from the unfolded state is much faster than trapped structures. Wang and coworkers studied Trpzip2 further and they too observed differences in melting temperatures between CD and IR spectra [84]. They saw double

exponential of the folding behavior as well however they conclude that the first fast phase may be an artifact of the experimental method used. Latest studies on Trpzip peptides via T-Jump experiments for various Trpzips with different turn sequences show that the folding rate is strongly dependent on turn formation [85].

In our simulations of the trpzip2 we find that the unfolded ensemble has a significant tendency to form the β-turn, along with non-specific hydrophobic contacts. Folding involves an increase in contact specificity coincident with formation of native backbone, and unfolding generally reflects the reverse process with differences in the sequence of events. While a single exponential describes the unfolding process, the presence of a slow phase in folding results in double exponential behavior only apparent when a large fraction of the ensemble undergoes folding. Our ability to separate the ensemble, based on structural properties, into two sets that show single-exponential decay thus gives a physical justification for the overall double-exponential fit as arising from independent folding pathways. Finally, we demonstrate that each exponential decay in folding and unfolding actually arises from multiple distinct pathways that have similar intrinsic relaxation times.

3.2 Methods

3.2.1 Model System

The model system chosen was the tryptophan zipper (trpzip) developed by Cochran and coworkers[59]. This β-hairpin structural motif is stabilized through cross-strand tryptophan pairs. Trpzip2 (SWTWENGKWTWK, with a type I' β-turn at NG) has

the most cooperative melting curve and highest stability (~90% at 300K) among the trpzips; therefore, it was selected for use in this study. Thermodynamic properties for this peptide have been determined by NMR and CD spectroscopy, and a family of structures was refined using restraints from NMR experiments[59] [80] (PDB code 1LE1). The N-terminal of the peptide was acetylated and the C-terminal was amidated, in accord with the experiments[59].

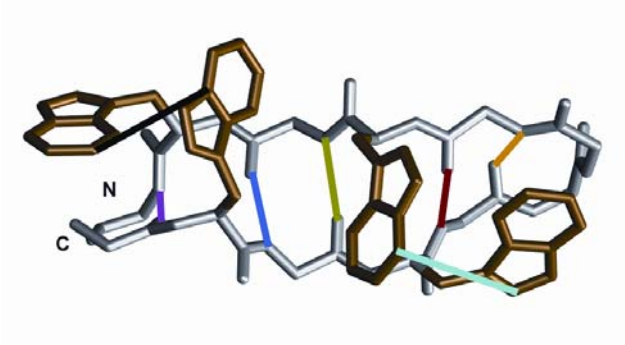


Figure 3-1. NMR-based conformation of trpzip2 (pdb code 1LE1). Side-chains are shown only for Trp residues. Native contacts defined in the text are shown as color-coded lines, with the colors matching data curves for these contacts as shown in subsequent figures. The number of native backbone hydrogen bonds that are not present defines the "HBlost" order parameter.

A set of native contacts was chosen as an alternative to RMSD to quantify the progression of folding (Figure 3-1). The contacts were defined using specific atom pair distances and corresponded to the 5 backbone hydrogen bonds [E5O-K8H, K8O-E5H, T3O-T10H, T10O-T3H, and S1O-K12H, 4 Å contact cutoff] and two native tryptophan packing contacts [W2-W11, W4-W9, 6Å cutoff]. Specific atom pairs were used for the Trp contacts to require native orientation as well as contact distance.

3.2.2 Replica Exchange Simulations

Replica Exchange Molecular Dynamics (REMD) were performed to generate converged equilibrium data at various temperatures to evaluate the thermodynamic data [42]. 14 replicas with GB solvent model were prepared at temperatures ranging from 251.7K to 554.7K. The temperature range was selected to obtain an exchange ratio about 15%; extra replicas were added around experimental melting temperature to ensure better sampling of the melting curve. Exchanges between replicas were attempted every picosecond and coordinates of each replica were saved every picosecond. Each replica is run under the same conditions as the individual simulations except translational and rotational motion is removed every 250 steps. Two sets of simulations were run, one starting all replicas from experimental native conformation and the other one starting from an unfolded conformation. Each REMD simulation was run for about 85000 exchanges (85ns per replica) and first 10000 exchanges were discarded to remove the bias introduced by starting conformations. Detailed description of REMD methodology and exchange probabilities can be found in Chapter 4.

3.2.3 Thermodynamic Analysis

Replica exchange molecular dynamics simulations were performed to obtain a melting curve. Structures were classified as native when RMSD from the experimentally determined structure was under 1.7Å. Fractions of folded and unfolded structures were calculated at each temperature. These data were fit to the Gibbs-Helmholtz equation to obtain values for melting temperature and enthalpy of melting:

$$\Delta G = [\Delta H_m (1 - {}^{T}/_{Tm})] - \Delta Cp [(T_m - T) + T \ln({}^{T}/_{Tm})]$$

Equation 3-1. Gibbs-Helmholz Equation

The 350K temperature trajectories from both REMD runs were combined and treated as single ensemble resulting in ~150000 structures. Several types of analysis of this ensemble were performed. The number of native backbone hydrogen bonds was calculated for each structure, using a distance cutoff of 2.9Å between the hydrogen and carbonyl oxygen. We also calculated the number of contacts between hydrophobic Trp side chains. In this case, all 6 Trp pairs were considered in order to obtain a measure of non-specific hydrophobic clustering of these residues. A hydrophobic contact was considered present if the distance between any of the heavy atoms on two given Trp residues was less than 4.8 Å. This smaller cutoff was used since any atoms pairs in the Trp pair could identify these non-specific contacts.

Potential Energy (PE) as a function of number of lost native backbone hydrogen bonds (HBlost) was calculated by averaging the energy of all structures with a given HBlost value. Free energies as a function of order parameters were obtained from multi-dimensional population histograms; free energy values shown are relative to the most populated histogram bin.

Lower limits for uncertainties in thermodynamic values and contact fractions were estimated as follows: each value was calculated using both REMD runs and the uncertainty was reported as the average error for calculated properties.

3.2.4 Temperature Jump Simulations

Non-native structures were generated through simulation at 800K. Forty-nine snapshots with proper stereochemistry and *trans* peptide bonds were chosen randomly. Backbone RMSD values ranged from 2Å to 8Å from native conformation. The ensemble of structures was subjected to a temperature jump by instantaneously changing the bath temperature to 350K. First passage times were calculated as the time at which the instantaneous backbone RMSD for residues 2-11 fell below 0.6Å to ensure that the native basin was reached**.** After folding, the simulations were terminated. The fraction of structures that had not yet folded was then calculated as a function of time. This procedure eliminates a contribution from the unfolding rate in this simulated relaxation experiment.

Unfolding was studied using an analogous procedure. Initial structures for 53 unfolding trajectories were obtained by assigning different velocities to a native conformation that was equilibrated at 300K. First passage times were identified when the backbone RMSD rose above 3.0Å where native and near native conformations are no longer present. The fraction of structures that had not yet unfolded was calculated as a function of time.

3.2.5 Simulation Details

All simulations were carried out using a locally modified version of AMBER (versions 6 and 8) [62]. The systems were coupled to a bath to maintain constant temperature (350K) with a temperature coupling constant of 1.0 ps unless otherwise noted [86]. Overall translational and rotational motion was removed every 10000 steps.

All non-bonded interactions (with no cutoffs) were evaluated at each MD time step (2 fs) and SHAKE [63] was used to constrain all bond lengths. All simulations used the Generalized Born (GB) implicit solvent model[29] with GB^{HCT} implementation in AMBER[87], without additional friction terms. This lack of viscosity prevents direct comparison of simulated and experimental rate constants. The force field was ff94[11], with modifications made to reduce over stabilization of α-helical conformations[22]. Data was analyzed using the programs MOIL-View [66]and ptraj.

3.3 Results and Discussion

3.3.1 Hairpin Structure and Stability: Equilibrium Simulations

Before carrying out a detailed analysis of the simulated folding for any system, it is important to validate the approach by ensuring that the simulations reproduce the experimentally determined structure and stability. It is insufficient to demonstrate only that the native conformation can be transiently sampled or that an initial native conformation was not lost on a particular timescale. We therefore calculated free energy profiles through REMD simulations and analyzed the 350K temperature trajectories in which multiple folding and unfolding events were observed. In Figure 3-2 we show free energy as a function of two order parameters: backbone RMSD and number of native backbone hydrogen bonds lost (HBlost). Using this convention, the NMR conformation has a value of zero for both parameters and increasing values correspond roughly to decreasing similarity to the NMR conformations.

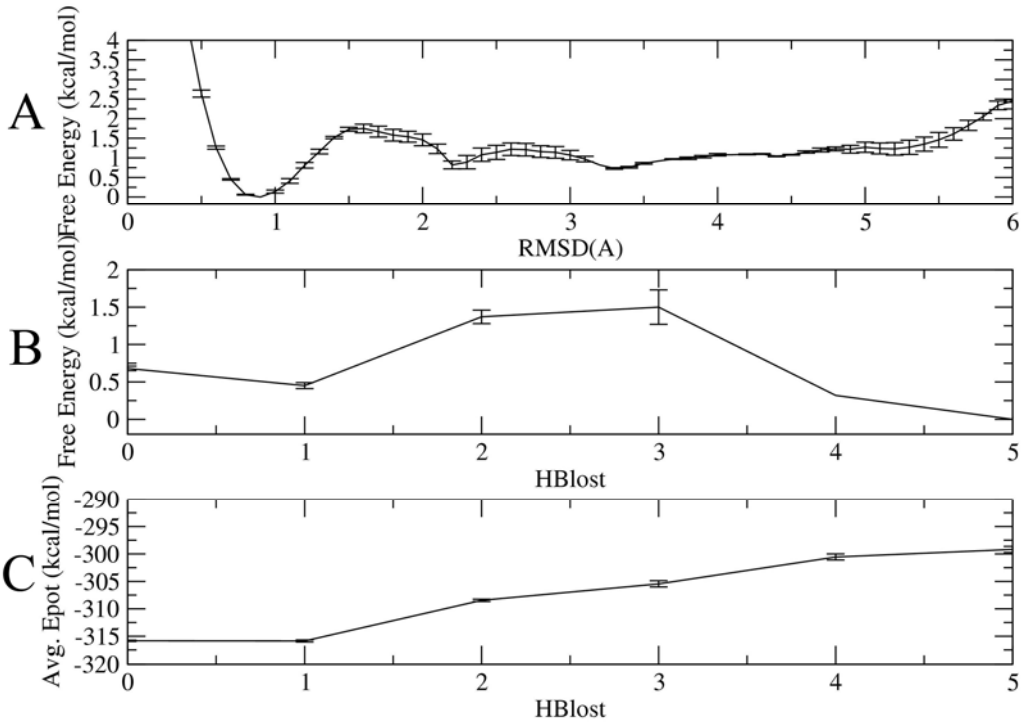


Figure 3-2. Free energy (A,B) and average potential energy (C) as a function of folding order parameters RMSD and HBlost (# native backbone hydrogen bonds not present) at 350K. The native state is thus on the left in all plots. Free energies show a barrier for folding while average potential energy does not. Error bars reflect statistical uncertainties.

The free energy *vs.* RMSD curve shows multiple minima, including the narrow global minimum located at RMSD ~0.8Å compared to the experimental structure. The broad minimum at large RMSD values corresponds to the unfolded state, and local minima near 2.2Å and 3.2Å represent specific misfolded structures that are described in detail below. The free energy *vs.* HBlost shows only 2 minima, corresponding to the native and non-native states. It is interesting to note that the stability of the native state appears to differ for alternate order parameters where the native conformation is about

1.0 kcal/mol more stable using RMSD as order parameter and the unfolded state is about 0.4 kcal more favorable using HBlost parameter. This observation is consistent with different melting curves reported by Yang et al. [82] using different variables.

Both of the profiles show an apparent free energy barrier for folding and unfolding (with the transition state located near 1.7Å RMSD or 3 HBlost). The minimum located at RMSD ~2.2Å is not encountered during folding and reflects an *off-pathway* intermediate (discussed below). This artifact of projection of the multidimensional surface onto a single dimension obscures the location of the transition state as encountered during folding events.

In contrast to the free energy data, when we calculate the average potential energy (PE) along the HBlost coordinate, the PE decreases steadily during folding, with no apparent barrier. The magnitude of the potential energy change is also much larger (~18 kcal/mol) than the corresponding free energy change, consistent with the expected large entropy contribution in the unfolded state.

In order to estimate the thermal stability of the hairpin, the fraction of native conformation was calculated for every temperature trajectory of the REMD simulations where conformations having a backbone RMSD less than 1.7Å were classified as native. The resulting melting curve is shown in Figure 3-3, along with an analogous curve generated from experimental data [59]. At low temperatures REMD seem to overestimate the native population and at elevated temperatures the calculated values seem to be in good agreement with experiments since the experimental curve lies within the error bars at calculated temperatures.

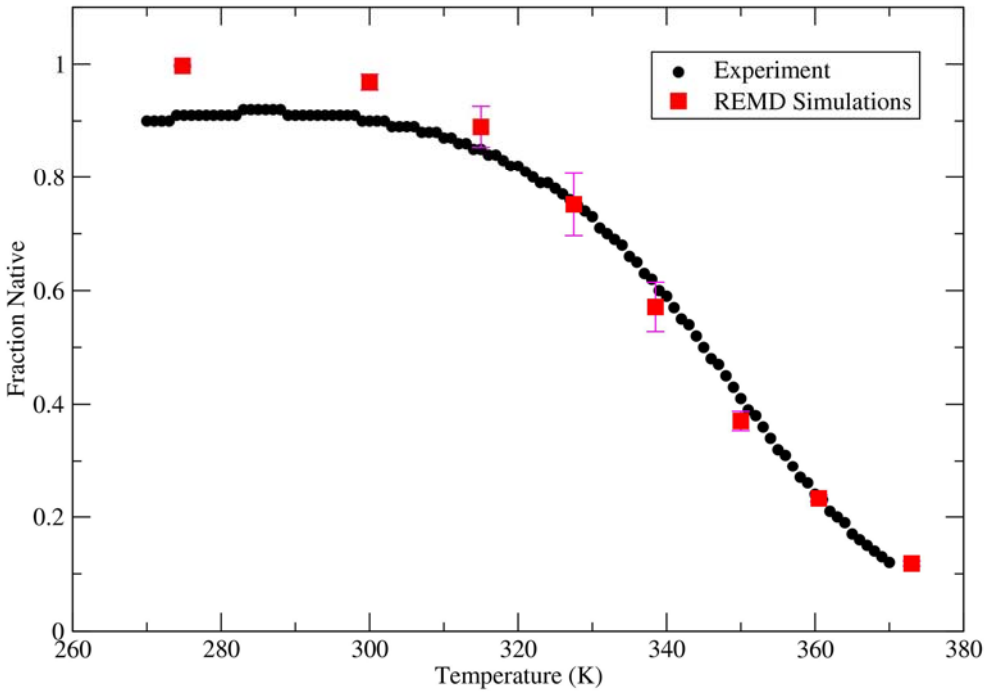


Figure 3-3. Fraction native structure vs. temperature obtained from REMD simulations (red) and experimental data (reproduced from thermodynamic data reported by Cochran et al. [59]) (black)

We used this data to determine the melting temperature (T_m) and changes in enthalpy and heat capacity for folding (Table 3-1, experimental and simulated values). The ΔH values are in very good agreement (within ~10% of the experiment). As expected, the simulated ΔH value is also very similar to the average ΔE shown in Figure 3-2. The differences in heat capacity change are larger, with simulations underestimating the experimentally determined value. This underestimation is consistent with incomplete sampling in the unfolded ensemble, but is also likely to arise from the use of a continuum

solvent model that does not properly account for the temperature-dependent properties of aqueous solvation.

Parameter	Experiment [59]	REMD Simulations
ΔH_m, cal·mol^{-1}	16770 ± 60	16100 ± 1200
ΔC_p, cal·mol^{-1}·K^{-1}	281 ± 2	8 ± 2
T_m, K	345.0 ± 0.1	342 ± 3

Table 3-1. Changes in enthalpy and heat capacity along with melting temperature calculated from simulations compared to those obtained from experimental measurements.

The ability to closely reproduce the native hairpin structure and several key thermodynamic parameters suggests that the simulations provide a useful model for folding of this peptide. An additional validation of our approach was provided by a more detailed analysis of the packing of the Trp side chains that stabilize the native hairpin. In particular, the face-to-face stacking of the indole rings originally observed in the NMR-based conformations (pdb code 1HRX, now withdrawn) differs from that obtained after a further refinement stage that included chemical shift data[80] (pdb code 1LE1). In our simulations, the native conformations adopted the edge-to-face packing seen in the more accurate NMR-based conformations, even though all simulations that began with "native" conformations used 1HRX (See Figure 2-13).

3.3.2 Characterization of the Non-native Ensemble

The interaction of the indole rings of the Trp side chains was suggested to be a dominant stabilizing factor for the trpzip hairpin[59]. We investigated the amount and type of these interactions in our equilibrium ensemble and the fraction of each possible Trp-Trp contact pair is shown as a function of the HBlost folding order parameter in Figure 3-4. The plot shows that the packing in the native state (HBlost=0) is highly specific and nearly complete for native Trp pairs 2:11 and 4:9. The middle Trp pair (2:9) does not have significant contact in the native structure, which is reflected in a contact fraction of only ~0.5. The average number of Trp pairs in the native state is thus ~2.5 (the sum of these individual values). These observations are all consistent with the NMR-derived conformation (Figure 3-1). As HBlost increases during unfolding, Trp clustering becomes non-specific in nature, with the most common pairs in the unfolded state being non-native interactions between Trp pairs close in sequence (2:4, 9:11). In addition, there is a slight decrease in the average total number of Trp pair contacts (1.7 contacts with HBlost=5). This change in the contact specificity occurs during the region of 2-3 HBlost, the same region previously identified as the location of the free energy barrier to folding (Figure 3-2). This increase in Trp packing specificity, along with the configurational entropy loss resulting from formation of backbone hydrogen bonds, is likely a significant entropic contribution to the free energy barrier for folding.

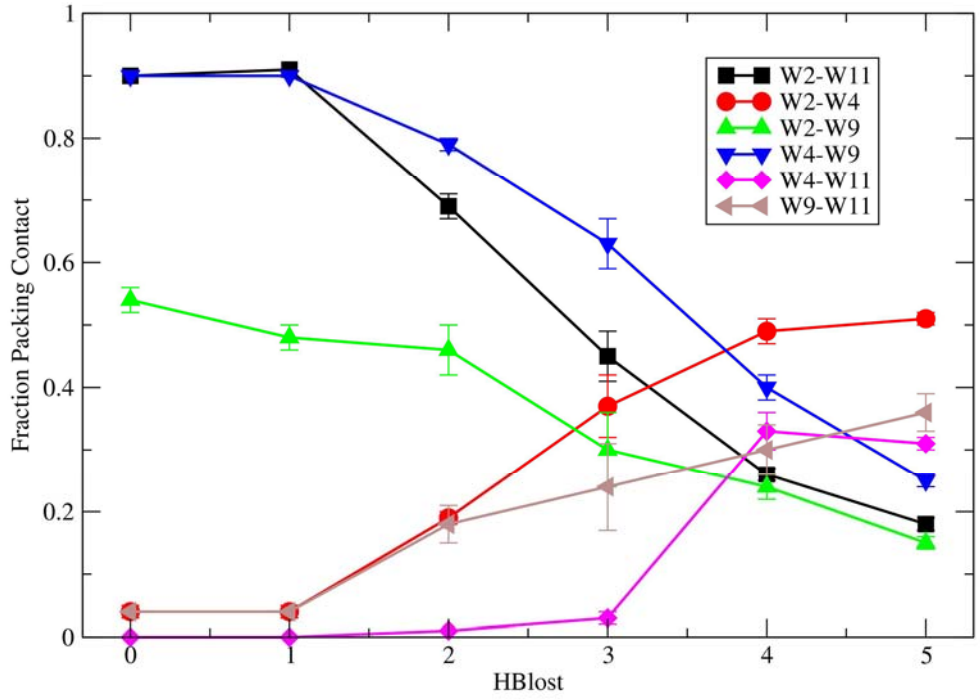


Figure 3-4. Average population of each Trp pair contact for structures with different HBlost order parameter, calculated from equilibrium MD data. Native conformations are on the left. Unfolding is accompanied by loss in specificity of Trp contacts.

3.3.3 Temperature-jump Simulations

Ensembles of folded and unfolded structures were subjected to a temperature-jump as described above and simulations were continued until first passage was recorded for 85% (folding) or 100% (unfolding) of the ensembles. This large fraction is critical to ensuring that all relevant pathways have been sampled since the earliest observed events may not be relevant to the behavior of the majority of the ensemble [88]. The total simulation time for folding and unfolding simulations was 2.1μs.

We initially attempted to fit the data with a single exponential**,** assuming simple 2-state kinetics [89]**.** The decay (folding) or rise (unfolding) in the fraction of non-native conformations is shown in Figure 3-5. Unfolding can be fit by a single exponential with a relaxation time of 13ns (Figure 3-5B). In contrast, folding data requires at least two exponentials with approximately equal weights (Figure 3-5A), with relaxation times differing by nearly an order of magnitude (4.5ns and 38.5ns). The double exponential fit suggests the presence of kinetic partitioning [90], in this case through at least two single exponential processes. Similar partitioning has been encountered in the folding kinetics of proteins [91] and recent experiments suggest this for trpzip2 as well [83]. The single exponential fit for unfolding does not imply a single unfolding pathway (or even a single rate constant); parallel reactions initiated from the same basin will always give rise to single-exponential behavior. This point will be revisited below.

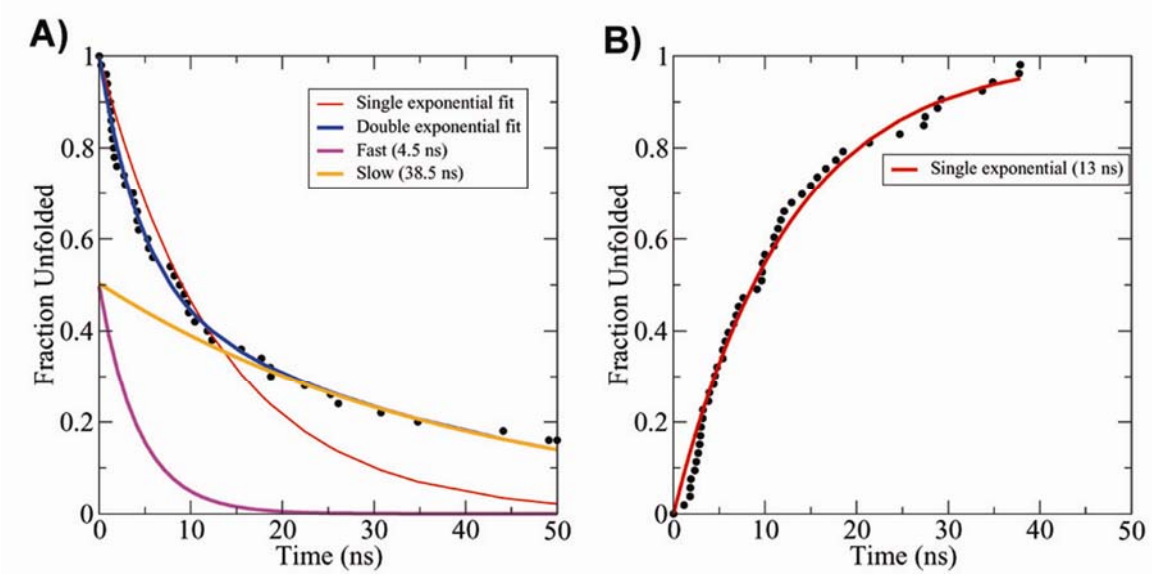


Figure 3-5. Fraction non-native structure as a function of time during folding (left, 800K→350K) and unfolding (right, 300K→350K) following the simulated temperature jump. Black circles represent times at which a member of the ensemble underwent a folding or unfolding transition and the simulation was terminated. Folding data was

poorly fit by single exponential (thin line) and at least two exponentials are required. Unfolding data can be represented with a single exponential.

These curves reflect behavior of the ensemble as a whole, and our goal is to justify the multi-exponential fit and elucidate the properties of the simulated folding process that result in this observation. An advantage to simulations as compared to analogous experiments is that the behavior of each member of the ensemble is available in atomic detail with a time resolution limited only by the frequency of saving coordinate snapshots (each picosecond in this case). The difficulty arises when attempting to assign a particular trajectory (ensemble member) to one of the exponential decay processes.

In the present case, the large difference in relaxation times for folding permitted analysis of events that occurred at timescales for which the faster process is nearly complete (>20ns, Figure 3-5A). This revealed sampling of two different non-native metastable hairpin conformations. One of these (Figure 3-6A) has a γ-turn at Gly7, resulting in a switch in up/down pattern of side chains on the C-terminal strand. Cross-strand Trp pairs are thus positioned on opposite faces of the hairpin. The other structure type (Figure 3-6B) has the Trp indole rings on the same side of the hairpin but the location of the type I' β-turn is shifted by one residue (G7-K8). In both structure families, the entire backbone hydrogen bond pattern is non-native, the N-terminal strand is extended past the C-terminal and inter-strand stacking of Trp indole rings is not present. Neither type was present in the set of initial structures representing our unfolded ensemble (before the T-jump).

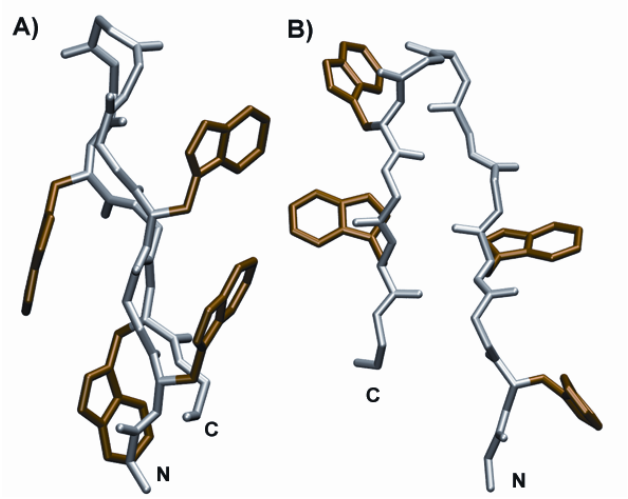


Figure 3-6. Non-native hairpins that give rise to kinetic partitioning. Structure A has a γ-turn at Gly7 while structure B has the β-turn at G7-K8 instead of the native N6-G7. Neither have significant Trp pair packing nor any native backbone hydrogen bonds.

In order to confirm that these non-native hairpin structures were responsible for the slow folding phase, simulations sampling either type were separated from the rest of the ensemble of T-jump refolding trajectories. Simulations that sampled misfolded hairpins showed single-exponential folding to the true native state with a relaxation time very similar to the slower phase (31 *vs*. 38ns) of the double exponential behavior seen for the entire ensemble.

We find that the misfolded hairpin structures represent *off-pathway* intermediates (Figure 3-7). The transition from unfolded to misfolded state occurs on a timescale similar to the transition from unfolded to native state (~4ns in each case). In addition, none of the unfolding trajectories sampled the misfolded structures prior to their transition into the unfolded state. These results demonstrate that the transition *into* the misfolded structures is not responsible for the slow folding behavior. Examination of refolding trajectories revealed that the transition between misfolded and native

conformation is not direct; misfolded structures always show significant unfolding prior to reaching the native state (Figure 3-8 shows one of such trajectories). The misfolded structures have RMSD values near 2.2Å and 3.2Å (same values for the local minima observed in REMD simulations, see Figure 3-2A), yet RMSD values rise to ~5Å before successful folding occurs. RMSD is thus a poor reaction coordinate in this case since the free energy barrier occurs at a larger value than misfolded (reactant) or native (product) conformations.

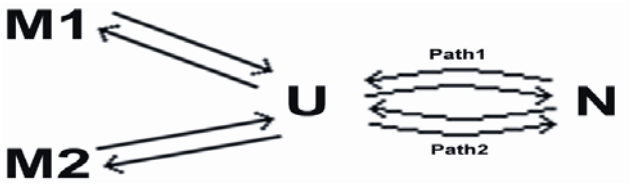


Figure 3-7. Observed folding mechanism of trpzip2 in simulations. From the unfolded state (U) trpzip can either misfold (M1 and M2) or fold to the native state (N) by one of two pathways. Passing through the unfolded state is necessary to access the native state.

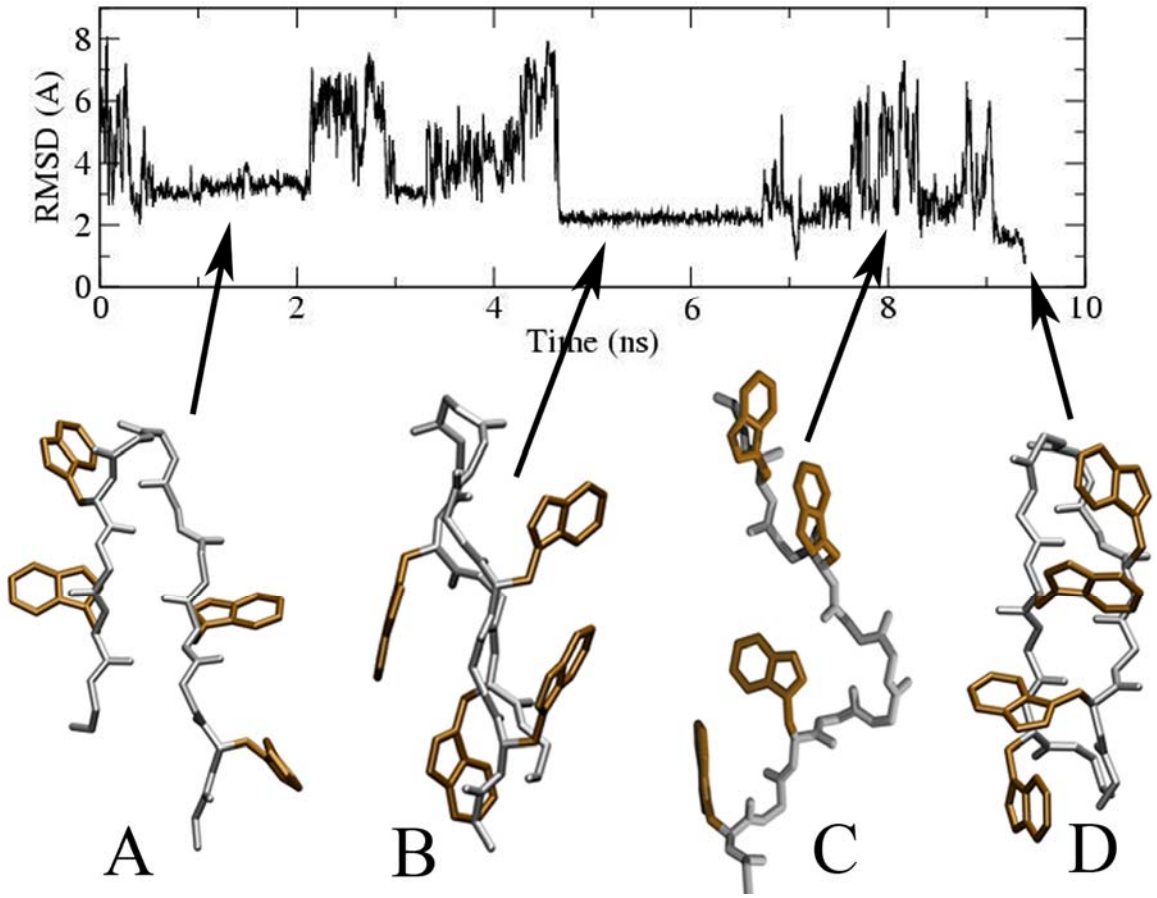


Figure 3-8. Backbone RMSD vs. time during a refolding simulation that samples both misfolded structure types (A,B). These always unfold (C) before reaching the native hairpin (D).

We next re-calculated first passage times to the folded state starting from the first snapshot after leaving the misfolded basin, and obtained single-exponential decay with a timescale nearly identical to the fast folding phase, consistent with our observation that a transition back to the unfolded state occurred (data not shown). Thus the slow folding is not due to entering the misfolded basins, nor to folding after leaving the misfolded basins; rather the rate limiting step is the transition from trapped structure back to the unfolded state.

We described above that the combined misfolding trajectories give rise to a single-exponential folding process. If we separate these slow-phase trajectories into two sets based on the type of misfolded hairpin sampled, we obtain single-exponential decay in both cases with slightly different timescales for the two sets (30ns *vs.* 45ns). This is direct atomic-level evidence that an observed single-exponential folding process can arise from physically distinct folding pathways that have similar intrinsic folding rates.

Finally, the ensemble of trajectories that never sampled these misfolded hairpins (~50%) show single exponential folding with nearly identical relaxation time as the faster phase of the double-exponential fit to all simulations. This indicates that the simulations sampling the two incorrect hairpins give rise to the entire slow phase of folding. Our ability to separate the ensemble, based on structural properties, into two sets that each show single-exponential decay thus gives a physical justification for the overall double-exponential fit as arising from independent folding pathways.

3.3.4 Analysis and Comparison of Folding and Unfolding Pathways

As we noted above, folding events were not observed to originate directly from the misfolded hairpins. We thus examined the folding pathway using the ensemble of folding trajectories that did not become kinetically trapped in these basins. For unfolding, the entire ensemble of unfolding trajectories was used. For each ensemble, we calculated (Figure 3-9) the time-dependent fraction of 7 native contacts (as defined in Figure 3-1). Comparison of relative rates of forming each contact provides insight into the sequence of contact formation (in an ensemble-averaged manner). Comparison of folding and unfolding also highlights any differences in the two processes.

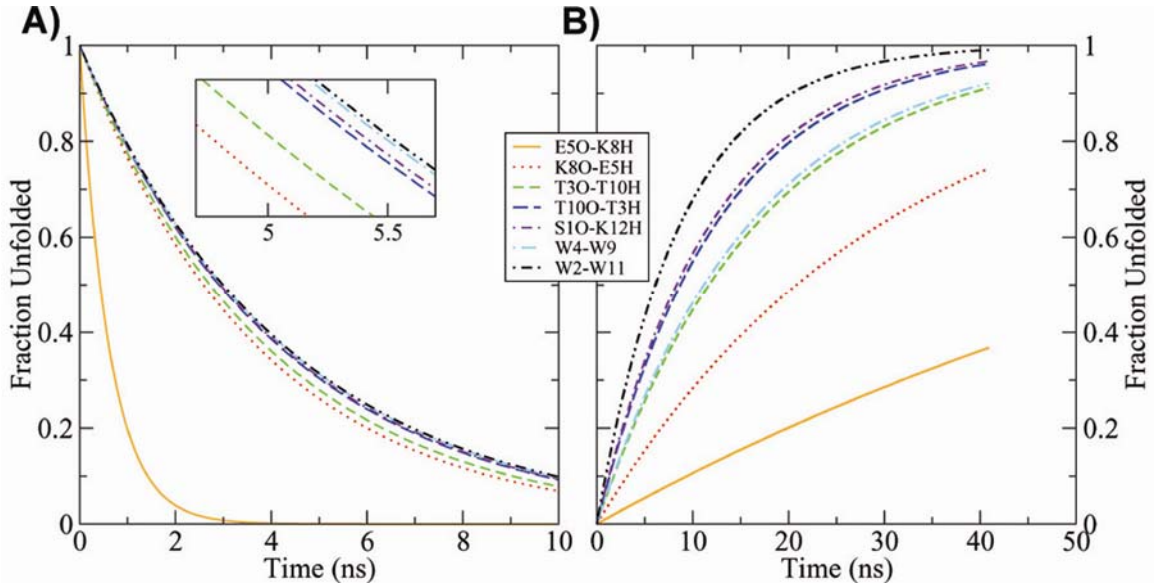


Figure 3-9. Average contact loss as a function of time for folding (left) and unfolding (right) ensembles. Colors correspond to contact definitions in Figure 3-1. Contact loss is shown for consistency with Figure 3-5. The inset on the left shows detail for the contacts that have similar timescales.

The most apparent feature of this data is that the contact corresponding to the β-turn forms on a much more rapid timescale than any of the other contacts. During unfolding, this contact was lost most slowly and was retained by a large fraction of the ensemble even after the remaining contacts were nearly completely lost. Both sets of observations imply a high tendency to form the turn in the unfolded state, consistent with our analysis of the equilibrium data.

For the remainder of the contacts, it is interesting to note that the rates of contact formation during folding vary by less than 15%, while nearly 300% variation is seen during unfolding under the same conditions. In both cases, however, the ordering of backbone hydrogen bond formation or loss is consistent, with zipping occurring from the turn out, and unzipping from the termini toward the turn.

Although unfolding is the reverse of folding with respect to the backbone, Trp packing shows an important difference. During folding (Figure 3-9A), native Trp-Trp contacts formed only after the hairpin was complete**,** with Trps 4:9 forming before 2:11. In contrast, unfolding (Figure 3-9B) occurs by initial loss of a single Trp pair contact (usually Trps 2:11, see below), followed by loss of the adjacent backbone hydrogen bonds, then the second Trp pair contact, and finally the last set of hydrogen bonds. Thus formation of the two native Trp pairs is the last step during folding, but loss of both pairs is <u>not</u> the first step during unfolding.

These trends in the ensemble data were confirmed by visual inspection of multiple individual trajectories. We discovered that unfolding actually occurs simultaneously by two very different pathways. In the predominant pathway (90% of the unfolding simulations) unzipping proceeds by successive loss of inter-strand hydrogen bonds from the termini towards the turn, consistent with the order of contact loss seen in Figure 3-9B**.** However, a second minor unfolding pathway also exists (10%) in which a hydrogen bond near the turn (E5H:K8O) is initially lost and unzipping proceeds away from the turn. This pathway is not apparent from the contact loss curves obtained these simulations, presumably due to the lower weight of this unfolding pathway in the ensemble data. It is also of interest to note that no reverse of the minor unfolding pathway was seen for any member of the folding ensemble (hydrogen bonds for the open end of the hairpin never formed before those near the turn).

A free energy landscape was constructed using the two distances (as sampled during REMD) corresponding to the initial hydrogen bond lost in each of the unfolding pathways (Figure 3-10). Two alternate low-energy pathways are apparent and correspond

to unzipping of the hairpin from either end. The broad unfolded state shows a slight depression along a nearly constant value for the E5H:K8O distance; this arises from the tendency to form the turn in the unfolded ensemble.

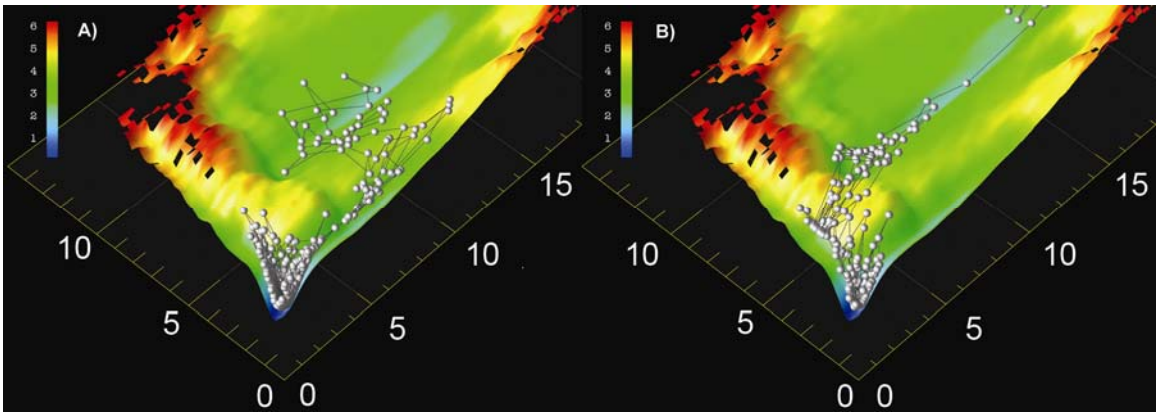


Figure 3-10. Both figures show the same free energy surface, calculated from populations obtained from REMD simulations. The Y axis corresponds to hydrogen bond E5H:K8O and the X axis is S1O:K12H. The native conformation is in front and the broad unfolded basin toward the rear. Height and color of the surface correspond to free energy relative to the global minimum. White spheres are positioned at values sampled by snapshots during major (left figure) and minor (right figure) pathways as explored during temperature-jump unfolding simulations. Backbone conformations for representative structures are shown along each pathway.

Snapshots from simulations sampling each type of pathway were projected onto the landscape. Each process begins in the narrow native basin and ends in the broad unfolded region, but both pathways are explored. The higher incidence of unzipping toward the turn is consistent with the lower free energy barrier for the exit from the native basin, relative to that traversed by unzipping away from the turn (~1.2 kcal/mol higher for the minor pathway). In each pathway, this initial step is followed by crossing of a second free energy barrier along the alternate coordinate. The landscape thus suggests an explanation for the more cooperative folding process as compared to unfolding (Figure

3-9). Multiple barriers are encountered during both of the unfolding pathways, yet no significant free energy barrier to folding is present once either native hydrogen bond has formed. We note however, since this landscape employs backbone hydrogen bond order parameters, it does not provide insight into the different coupling of these parameters to Trp pair contact formation observed between folding and unfolding simulations.

3.4 Conclusions

Multiple microseconds of equilibrium and non-equilibrium simulations were combined to provide a thorough analysis of the native state, unfolded ensemble and folding/unfolding transitions for trpzip2 in the simulation environment. While the alternate approaches yielded a consistent and often complementary view, interesting differences between the order of events occurring during folding and unfolding were noted. In both cases, we demonstrated that the single-exponential behavior clearly arises from multiple pathways with similar intrinsic rates. In addition, we justified the slow phase of a double-exponential fit of folding data as arising from two different off-pathway intermediates that give rise to kinetic partitioning during folding.

The β-hairpin folding pathway we describe here differs in detail from those previously reported, but shares general features with most. Folding proceeds by rapid formation of the β-turn (and a corresponding slow loss during unfolding). The unfolded ensemble samples a high level of non-specific hydrophobic clustering, but side chain rearrangement and a rise in the specificity of these contacts accompany formation of the native backbone conformation while traversing the free energy barrier. Folding and unfolding differed in a statistically meaningful manner in the details of this

rearrangement. In addition, an alternate unfolding pathway was found that was never sampled during folding simulations but was distinctly present in the free energy landscape.

While these simulations were carried out using an approximate solvent model, we feel that such approaches are currently required in order to obtain sufficient sampling of conformations and folding transitions. Obtaining converged populations and following significant fractions of ensembles through folding and unfolding for larger and more complex systems remains an immense challenge. With increased computational resources, we feel that approaches such as we have taken here will not only provide useful insights but also supply an essential test of the internal self-consistency and reliability of any conclusions drawn from simulations.

Chapter 4

Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model

4.1 Introduction

The potential energy surfaces of biological systems have long been recognized to be rugged**,** hindering conformational transitions between various local minima. This sampling problem can preclude success even when a sufficiently accurate Hamiltonian of the system is used in the simulations. Thus, significant effort has been put into devising efficient simulation strategies that locate low-energy minima for these complex systems. Conformational sampling was recently reviewed[38] and is also the subject of a recent special journal issue[37].

One approach that has seen a recent increase in use for biomolecular simulation is the replica exchange method[41, 92, 93]. In replica exchange molecular dynamics (REMD)[42] (also called parallel tempering[41]), a series of molecular dynamics simulations (replicas) are performed for the system of interest. In the original form of REMD, each replica is an independent realization of the system, coupled to a heat bath at a different temperature. The temperatures of the replicas span a range from low values of interest (such as 280K or 300K) up to high values (such as 600K) at which the system can rapidly overcome potential energy barriers that would otherwise impede conformational transitions on the timescale simulated.

At intervals during the otherwise standard simulations, conformations of the system being sampled at different temperatures are exchanged based on a Metropolis-type criterion[94] that considers the probability of sampling each conformation at the alternate temperature (described in more detail in Methods). In this manner, REMD is hampered to a lesser degree by the local minima problem, since simulations at low temperature can escape kinetic traps by "jumping" directly to alternate minima being sampled at higher temperatures. Likewise, the structures sampled at high temperatures can anneal by being transferred to successively lower temperatures. Moreover, the transition probability is constructed such that the canonical ensemble properties are maintained during each simulation, thus providing potentially useful information about conformational probabilities as a function of temperature. Due to these advantages, REMD has been applied to studies of peptide and small protein folding[32, 34, 36, 41, 42, 48, 95-99].

For large systems, however, REMD becomes intractable since the number of replicas needed to span a given temperature range increases with the square root of the number of degrees of freedom in the system[100-103]. Several promising techniques have been proposed[102, 104-106] to deal with this apparent disadvantage to REM.

The method chosen to treat solvent effects can have a direct impact on the system size and thus the computational requirement of employing REMD. Explicit representation of solvent molecules significantly increases the number of atoms in the simulated system, particularly when the solvent box is made large enough to enclose unfolded conformations of peptides and proteins. The growth in system size results in a need for many more replicas to span the same temperature range. This increase in

computational cost is in addition to that added by the need to calculate forces and integrate equations of motion for the explicit solvent molecules.

Continuum solvent models like the semi-analytical Generalized Born (GB) model[29] estimate the free energy of solvation of the solute based on coordinates of the solute atoms. The neglect of explicit solvent molecules can significantly reduce the computational cost of evaluating energies and forces for the system, but a larger effect with REMD can arise from the reduction in number of replicas due to fewer degrees of freedom. This factor can determine whether REMD is a practical approach to model the system. For example, in the 10-residue peptide model presented below, 40 replicas are needed when solvent is included explicitly while only 8 are sufficient for the same peptide with a continuum solvent model. Larger systems would be expected to show even greater differences; the number of peptide atoms increases approximately linearly with sequence length, while the volume of a sphere (and thus the number of solvent atoms) needed to enclose extended conformations increases with the peptide length to the third power. Thus one can roughly estimate that the difference in number of replicas required for explicit vs. continuum solvation of a system will increase with the number of solute degrees of freedom to the $^3/_2$ power.

Continuum solvent models are thus an attractive approach to enabling the study of larger systems with REMD. Among the various models that have been developed, the GB approach is commonly used with molecular dynamics due to its computational efficiency, permitting use at each time step. However, these models can also have significant limitations. Since the atomic detail of the solvent is not considered, modeling specific effects of structured water molecules can be challenging. In the case of protein and

peptide folding, it appears likely that the current generation of GB models do not have as good a balance between protein-protein and protein-solvent interactions as do the more widely tested explicit solvent models[32, 33]. More particularly, is has been reported[4, 33-35] that ion pairs were frequently too stable in the GB implicit water model, causing salt bridged conformations to be oversampled in MD simulations, thus altering the thermodynamics and kinetics of folding for small peptides. A clear illustration was given by Zhou and Berne[33] who sampled the C-terminal ß-hairpin of protein G (GB1) with both a surface-GB (SGB)[107] continuum model and explicit solvent. The lowest free energy state with SGB was significantly different from the lowest free energy state in explicit solvent, with incorrect salt bridges formed at the core of the peptide, in place of hydrophobic contacts. Zhou extended this study on GB1 by examining several force field–GB model combinations, with all GB models tested showing erroneous salt-bridges[35].

The more rigorous models based on Poisson-Boltzmann (PB) equations are generally considered to be more accurate. Historically, the increased cost of evaluating solvation free energy with these methods results in their use primarily to post-process a small number of conformations, or snapshots sampled during an MD simulation in explicit solvent[108]. However, some researchers have reported using PB as a solvent model for molecular dynamics simulation[109, 110]. PB approaches do not necessarily overcome the difficulty of modeling non-bulk effects in the first solvation shells.

In order to benefit from the efficiency of implicit solvents while incorporating these first shell effects, several hybrid explicit/implicit models have been proposed. These typically employ explicit solvent only for the first 1-2 solvation shells of the solute, often

surrounded by a continuum representation of various types[111-123]. However, these methods have drawbacks in that the explicit water typically must be restrained to remain close to the solute to avoid diffusion into the "bulk" continuum. These restraints, as well as boundary effects at the explicit/implicit interface, can have a dramatic effect on solute behavior. In a recent implementation, Lee et al. employed a hybrid TIP3P/GB solvation model with excellent results[119], but they pointed out drawbacks typical for these models, such as the need for a fixed solute volume and shape for the solvation cavity, preventing large-scale conformational changes of the type that is necessary for detailed analysis of conformational ensembles using enhanced sampling techniques like REMD. In addition, they demonstrated that solvent properties such as radial density and dipole distributions showed significant artifacts due to boundary effects.

Recognizing that the main difficulty in applying REMD with explicit solvent lies in the number of simulations required, rather than just the complexity of each simulation, we propose a new approach in which each replica is simulated in explicit solvent using standard methods such as periodic boundary conditions and inclusion of long-range electrostatic interactions. However, the calculation of exchange probabilities (which determines the temperature spacing and thus the number of replicas) is handled differently. Only a subset of closest water molecules is retained, with the remainder temporarily replaced by a continuum representation. The energy is calculated using the hybrid model, and the exchange probability is determined. The original solvent coordinates are then restored and the simulation proceeds as a continuous trajectory with fully explicit solvation. This way the perceived system size for evaluation of exchange probability is dramatically reduced and fewer replicas are needed.

An important difference from existing hybrid models is that our system is fully solvated throughout the entire simulation, and thus the distribution functions and solvent properties should not be affected by the use of the hybrid model in the exchange calculation. In addition, no restraints of any type are needed for the solvent, and the solute shape and volume may change since the solvation shells are generated for each replica on the fly at every exchange calculation. Nearly no computational overhead is involved since the calculation is performed infrequently as compared to the normal force evaluations. Thus the hybrid REMD approach can employ more accurate continuum models that are too computationally demanding for use in each time step of a standard molecular dynamics simulation.

In this study we have tested the hybrid REMD method on varying lengths of polyalanine peptides (dipeptide, tetrapeptide and Ala_{10}). Many helical design studies have used polyalanines with charged residues[124-127], N-capping[128] and C-capping interactions[129] to solubilize the peptides and stabilize helical structure. Recently, experimental studies with CD, NMR, and UV resonance Raman have been able to characterize primarily polyproline type II (P_{II}) structure in short polyalanines [130-132] and in the denatured state of longer alanine peptides[133]. MD simulations of polyalanines have further substantiated these experimental observations[134]. The quality of solvent model is expected to be critically important since it has been proposed that specific solvation of backbone amide groups plays a key role in the stabilization of P_{II} conformations[135, 136].

For each peptide we first obtained conformation ensembles using standard REMD in explicit solvent. We used this data as a reference in order to remove the influence of the

protein force field parameters from this study of solvation models. For each sequence, two sets of REMD simulations in explicit solvent were run with different initial conformations until convergence was indicated by reasonable agreement between the data sets. For example, the populations of conformation clusters in the two Ala_{10} runs in TIP3P solvent were highly correlated (R^2=0.974), demonstrating high similarity not only in the types of structures sampled in these two simulations, but also in their probability in these independently generated ensembles. This level of convergence gives us confidence that the differences we observe between the various solvent models are predominantly due to solvation effects and not poorly converged ensembles with large uncertainties in the resulting data.

We then employed pure GB REMD simulation using both models available in Amber (GB^{HCT} [87] and GB^{OBC} [137, 138]) as well as the hybrid REMD approach using the same GB models. We also performed REMD where only the first 1 or 2 solvation shells were retained for the exchange calculations (without a continuum model). Comparison of these results to each other and to the standard explicit solvent REMD results provides insight into the performance of the GB models, the improvement obtained by retaining the first solvation shell in the calculation of exchange probability (the hybrid model), and the need for the reaction field surrounding the solvation shells.

We compared ensemble distributions of properties such as chain end-to-end distance, backbone ϕ/ψ free energy maps, and cluster populations among the methods. While all of the solvation models provided similar results for alanine dipeptide, the GB models failed to reproduce the TIP3P ensemble data for Ala_3 and Ala_{10} even at a qualitative level, providing ensembles that were dominated by α-helical conformations. Simulations using

hybrid REMD using GB^{OBC} and only a single shell of explicit water were in good accord with the reference simulations, with a high degree of similarity between structure populations (R^2=0.93), with lack of significant α-helix and a strong preference for P_{II} conformation. This agreement was obtained despite a significant reduction in computational cost; for Ala_{10}, 40 replicas were used for standard REMD in TIP3P, while only 8 were needed for pure GB or hybrid GB/TIP3P REMD.

4.2 Methods

4.2.1 Replica Exchange Molecular Dynamics (REMD)

We briefly summarize the key aspects of REMD as they relate to the present study. In standard Parallel Tempering or Replica Exchange Molecular Dynamics[41, 42], the simulated system consists of M non-interacting copies (replicas) at M different temperatures. The positions, momenta and temperature for each replica are denoted by $[q^{[i]}, p^{[i]}, T_m]$, i = 1,…,M ; m = 1,…, M. The equilibrium probability for this generalized ensemble is

$$W(p^{[i]},q^{[i]},T_m) = \exp\{-\sum_{i=1}^{M}\frac{1}{k_B T_m}H(p^{[i]},q^{[i]})\}$$

Equation 4-1

where the Hamiltonian $H(p^{[i]},q^{[i]})$ is the sum of kinetic energy $K(p^{[i]})$ and potential energy $E(q^{[i]})$. For convenience we denote $\{p^{[i]},q^{[i]}\}$ at temperature T_m by $x_m^{[i]}$ and further define X = $\{ x_1^{[i(1)]}, \ldots, x_M^{[i(M)]} \}$ as one state of the generalized ensemble. We

now consider exchanging a pair of replicas. Suppose we exchange replicas *i* and *j*, which are at temperatures T_m and T_n respectively,

$$X = \{...; x_m^{[i]};...; x_n^{[j]};...\} \rightarrow X' = \{...; x_m^{[j]};...; x_n^{[i]};...\}$$

Equation 4-2

In order to maintain detailed balance of the generalized system, microscopic reversibility has to be satisfied, thus giving

$$W(X)\,\rho(X \rightarrow X') = W(X')\,\rho(X' \rightarrow X)$$

Equation 4-3

where ρ(X→X') is the exchange probability between two states X and X'. With the canonical ensemble, the potential energy *E* rather than total Hamiltonian *H* will be used simply because the momentum can be integrated out. Inserting Equation 4-1 into Equation 4-4, the following equation for the Metropolis exchange probability is obtained:

$$\rho = \min\left(1, \exp\{(\frac{1}{k_B T_m} - \frac{1}{k_B T_n})(E(q^{[i]}) - E(q^{[j]}))\}\right)$$

Equation 4-4

In practice, several replicas at different temperatures are simulated simultaneously and independently for a chosen number of MD steps. Exchange between a pair of replicas is then attempted with a probability of success calculated from equation 4. If the exchange is accepted, the bath temperatures of these replicas will be swapped, and the

velocities will be scaled accordingly. Otherwise, if the exchange is rejected, each replica will continue on its current trajectory with the same thermostat temperature.

As we described above, one of the major limitations of REM is that the number of replicas needed to span a temperature range grows proportionally to the square root of number of degrees of freedom in the simulated system. While a more rigorous analysis of the acceptance probability in REM trials has been given recently using a Gaussian energy distribution model[103, 139], one can also approximate from equation 4 that the overall exchange probability P_{acc} is proportional to $exp(-\Delta T^2/T^2)$, which implies that a greater acceptance ratio requires a smaller temperature gap ΔT or a more dense temperature distribution to reach. On the other hand, ΔT should be as large as possible so as to span a wide temperature range with a small number of replicas. The relationship can be estimated through consideration of potential energy fluctuations of two replicas sampling at the target temperature T_n and T_{n-1} (Figure 4-1). The instantaneous energy fluctuation δE in a given simulation at temperature T scales as $\sqrt{f}$ T, and the average energy gap ΔE between two neighboring replicas is proportional to $f\Delta T$, where f is the number of degrees of freedom and $\Delta T = T_n - T_{n-1}$. Obtaining a reasonable acceptance ratio relies on keeping the replica energy gap comparable to the energy fluctuations, thus ΔE/δE should be near unity. Since ΔE/δE is proportional to $\Delta T\sqrt{f}/T$, the acceptable temperature gap between neighboring replicas therefore decreases with larger systems as $\Delta T \sim 1/\sqrt{f}$, and more simultaneous simulations are needed to cover the desired temperature range.

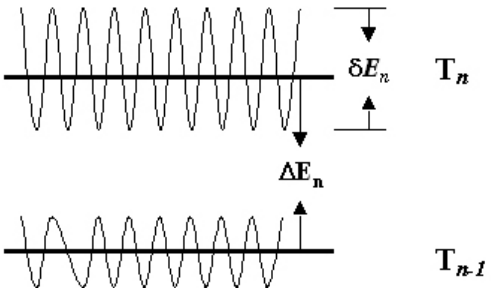


Figure 4-1. Schematic diagram illustrating the energy fluctuations for simulations at two temperatures for neighboring replicas. In order to obtain high exchange probabilities, the energy fluctuations δE in each simulation should be of comparable magnitude to the mean energy difference ΔE.

4.2.2 Model Systems and Simulation Details

We simulated three polyalanine sequences: alanine dipeptide (Ala_1), alanine tetrapeptide (Ala_3) and polyalanine (Ala_{10}), all with acetylated and amidated N- and C-termini, respectively. All simulations employed the Amber ff99 force field[11, 17], with modifications[24] to reduce α-helical bias. Explicit solvent and hybrid REMD used the TIP3P water model[64]. The standard REMD simulations in explicit solvent and in pure GB were run using our REMD implementation as distributed in Amber (version 8) [62]. The hybrid solvent REMD calculations were performed with a locally modified version of Amber 8. All bonds involving hydrogen were constrained in length using SHAKE[63]. The time step was 2 fs. Temperatures were maintained using weak coupling[86] to a bath with a time constant of 0.5 ps^{-1}.

Secondary structure basin populations for central residues were calculated based on ϕ/ψ dihedral angle pairs. The dihedral angle ranges defining for those regions are provided in Table 4-1. The solvent accessible surface areas (SASA) for simulated

peptides were calculated using the gbsa=2 option in AMBER**.** The end-to-end distances for Ala_{10} were calculated between Cα atoms of Ala2 and Ala9 (omitting terminal residues) using the ptraj module of Amber. Cluster analysis for Ala_{10} was performed using Moil-View[66], using backbone RMSD for Ala2-9 and a similarity cutoff of 2.5Å.

Secondary Structure	**ϕ**	**φ**
α	-160 to -50	-60 to +30
β	-180 to -110	+110 to +180
$\mathbf{P^{II}}$	-110 to -40	+110 to +180
α^{L}	+20 to +70	-30 to +70

Table 4-1. The ranges used to determine residue based secondary structure populations.

4.2.3 Explicit Solvent REMD

The Ala_{10} peptide in α-helical conformation was solvated in a truncated octahedral box using 983 TIP3P water molecules for a total of 3058 atoms. The system was equilibrated at 300K for 50ps with harmonic positional restraints on solute atoms, followed by minimizations with gradually reduced solute positional restraints and three 5ps MD simulations with gradually reduced restraints at 300K. Long range electrostatic interactions were calculated using PME[140]. Simulations were run in the NVT ensemble.

40 replicas were used at temperatures ranging from 267K to 571K, which were optimized to give a uniform exchange acceptance ratio of ~30%. Exchange between neighboring temperatures was attempted every 1 ps and each REMD simulation was run

for 50,000 exchange attempts (50 ns). The first 5ns of each simulation was discarded to remove initial structure bias.

In order to provide a stringent test of data convergence for greater conformational diversity expected for Ala_{10}, two sets of REMD simulations were performed, starting from different initial conformations. In one set, all replicas were started from a fully α-helical conformation; in the other an extended conformation was employed. In the case of Ala_1 and Ala_3, lower bounds for uncertainty were estimated by separating the full simulation data into halves and reporting the difference between values calculated for each half.

A similar procedure was used for Ala_1 and Ala_3. Ala_1 was solvated in a truncated octahedral box using 341 TIP3P water molecules. Ala_3 required 595 water molecules. For both systems the same equilibration procedure as used for Ala_{10} was employed. To cover the same temperature range 20 replicas for Ala_1 and 26 replicas for Ala_3 were needed. Both systems were simulated for ~ 40000 exchanges, and the first 5000 exchange attempts were discarded as equilibration.

4.2.4 Implicit Solvent REMD

Solvent effects were calculated through the use of two Generalized Born implementations in Amber (GB^{HCT} and GB^{OBC} (note that GB^{OBC} is model 2 in reference 59)). Two sets of intrinsic Born radii were used, both adopted from Bondi[141] with modification of hydrogen [142]. Unless otherwise noted, the GB^{HCT} model was used with the mbondi radii, and the GB^{OBC} model was employed with mbondi2 radii (as recommended in Amber). Scaling factors were taken from the TINKER modeling

package[143]. No cutoff on non-bonded interactions was used. All other simulation parameters were the same as used in explicit solvent.

For $Ala_{10,}$ the use of the continuum solvent model resulted in a total of 109 atoms considered explicitly in the simulations (~28 times fewer than in the explicitly solvated system). The much smaller system size permitted the use of 8 replicas to cover the same temperature range that required 40 replicas in explicit solvent, while obtaining the same 30% exchange acceptance probability. Exchanges were attempted every 1 ps and the REMD simulation was run 50,000 exchange attempts (50 ns). Simulations were initiated with the same two initial conformation ensembles as were used for the explicit solvent REMD calculations, with comparison of the two runs providing a lower bound for the uncertainty in resulting data. For Ala_1 and Ala_3 the same approach was used, with 4 replicas used to cover the temperature space for each system. Simulations were run for 50000 exchange attempts, and the first 5000 exchanges were discarded.

4.2.5 Hybrid Solvent REMD

All simulation parameters in the hybrid solvent REMD simulations were the same as those employed for standard REMD in explicit solvent, with the exception that the number of replicas (8 for Ala_1, Ala_3 and Ala_{10}) and the target temperatures were the same as those used for the pure GB REMD simulations for Ala_{10}. It is important to note that the hybrid solvent model was used only for calculation of exchange probability; the simulations themselves were performed on fully solvated systems with truncated octahedral periodic boundary conditions and PME for calculation of long-range electrostatic interactions.

We determined the number of water molecules to retain in the hybrid model based on analysis of the number of waters in the first solvation shell of Ala_{10} in the ensemble of structures sampled in the standard REMD explicit solvent simulations. We found that 100 water molecules were sufficient even for the most extended conformations (data not shown). Thus this number was used for all replicas and all exchanges. For $Ala_{1,}$ 30 water molecules were enough to incorporate the 1st solvation shell and 60 water molecules for the 1st and 2nd solvation shells. These numbers increase to 50 waters and 100 waters for 1st solvation shell and 1st and 2nd solvation shells of Ala_3 respectively. Ala_1 and Ala_3 hybrid simulations were run for ~ 30000 exchanges and the first 5000 were discarded.

At each exchange step, the distance between the oxygen atom of each water molecule and all solute atoms was calculated. Water molecules were then sorted by their closest solute distance, and all water molecules except the X with shortest solvent-solute distances were temporarily discarded (where X is the numbers of waters retained in each system, as described above). The energy of this smaller system was then recalculated using only these close waters and the GB solvent model. This energy was used to calculate the exchange probability, and then all waters were restored to their original positions and the simulations were continued (Figure 4-2). In this manner the simulations using the hybrid solvent model were continuous simulations with fully solvated PBC/PME and the hybrid model was used only for calculation of exchange probabilities.

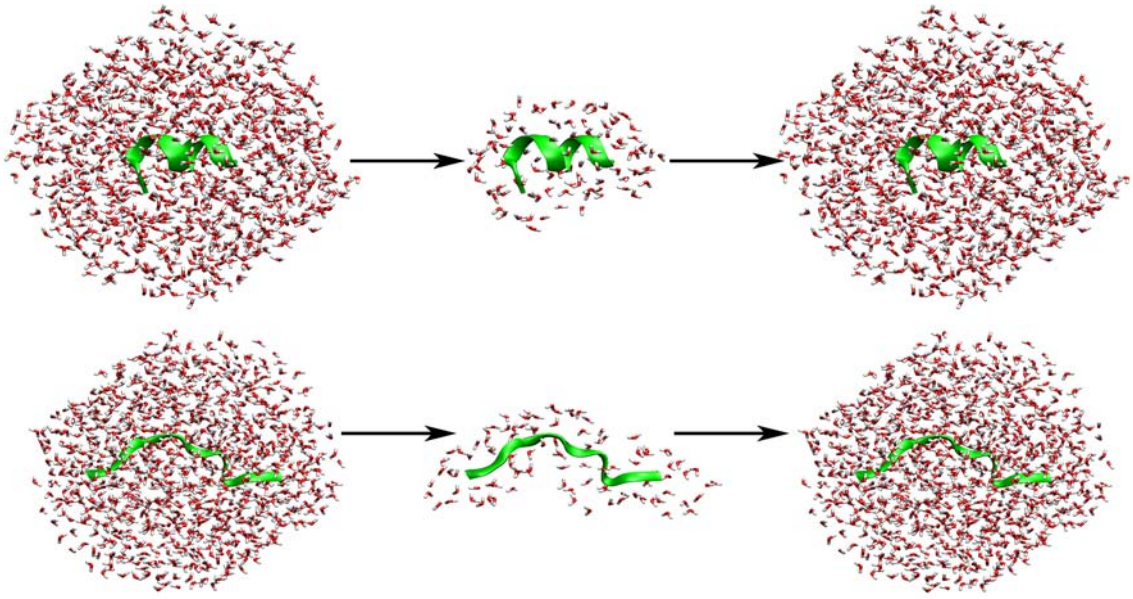


Figure 4-2. Schematic description of hybrid solvent REMD. The fully solvated Ala10 (with truncated octahedral boundary conditions) is simulated between exchanges (left). The exchange energy is calculated by retaining only the closest 100 waters (center), with bulk solvent properties calculated using the GB solvation model. After the exchange calculation the explicit solvent is restored and the dynamics continues under periodic boundary conditions. This approach allows on the fly calculation of the solvation shell, whose shape adjusts automatically to the solute conformation (top: α-helical structure, bottom: extended structure). As a result, many fewer replica simulations are required.

4.3 Results and Discussion

4.3.1 Comparison of exchange efficiency for hybrid and standard REMD in Ala_{10}

Even though REMD has become a useful tool to improve conformational sampling, REMD simulations are highly computationally expensive, particularly when solvent is treated explicitly. The increase in cost arises not only from the additional effort involved

in calculating forces in a given simulation, but from the increase in the number of simulations (replicas) needed to span a particular temperature range. This increase is due to the much larger number of degrees of freedom present in the explicitly solvated system as compared to that in continuum solvent models. In the case of Ala_{10}, our largest model system, the number of replicas needed to span the range of 267K to 517K increases from 8 to 40 when switching from implicit to explicit solvation.

We evaluated the utility of the hybrid solvent model during the calculation of exchange probability on several levels, using Ala_{10} as its size is most relevant to the larger systems that would benefit most from this method. First, we validated that fewer replicas were needed to obtain efficient exchange with the hybrid model as compared to the number required when retaining the full periodic box of explicit water molecules during the exchange probability calculation (Equation 4-4). Efficient exchanges were obtained with the hybrid model even when using the same number of replicas as was needed for the pure continuum solvent REMD simulations. Next, we evaluated whether the use of the hybrid model affected the data obtained from the simulations, with particular emphasis on the conformational distributions sampled by the model peptides. These distributions were also compared to those obtained for REMD with only the continuum solvent model.

An important benefit of REMD is the ability to obtain improved sampling at low temperatures of interest by exchanging conformations with higher temperature simulations that have less likelihood to become kinetically trapped. As described in Methods, the probability of successful exchange of conformations between two temperatures depends on the overlap in potential energy distributions at those

temperatures. Figure 4-3 shows the potential energy distributions vs. temperature for sets of simulations with explicit solvent (A) and those with GB (B) between 267K and 571K. The graph illustrates why fewer replicas are required for the GB model; the energy range spanned is smaller for the smaller system and fewer replicas are still able to achieve the required overlap. In contrast, when the explicit solvent model is used with only the 8 replica temperatures that are successful with GB, no significant overlap in the distributions is observed (Figure 4-3C).

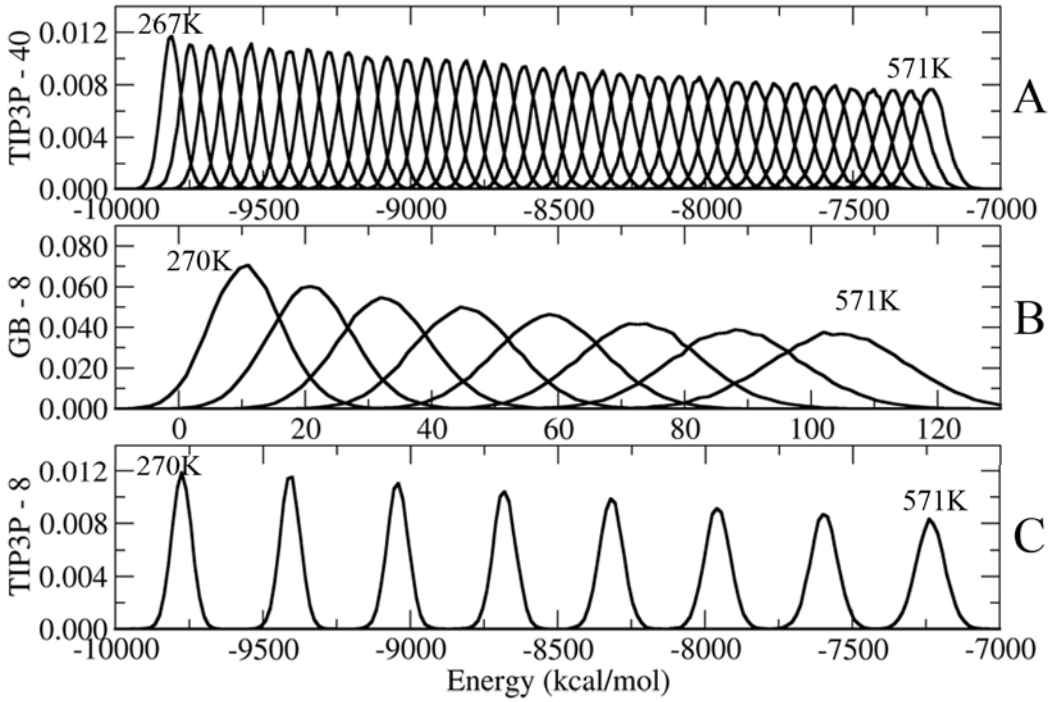


Figure 4-3. Potential Energy distributions for Ala10 simulations over a range of temperatures using (A) explicit solvent REMD with 40 replicas, (B) GB REMD with 8 replicas and (C) explicit solvent REMD with 8 replicas using the same temperature distribution as GB REMD. GB simulations involve fewer degrees of freedom and are able to span the energy range with fewer replicas. In contrast, no overlap is obtained

when using explicit solvent with the same replica and temperature selection as GB. This implies that no exchanges would be permitted and the benefits of REMD would be lost.

Based on Figure 4-3, exchanges between replicas at neighboring temperatures are expected to occur with high probability when using 40 replicas in explicit solvent or 8 replicas with GB. No exchanges are expected for explicit solvent with only 8 replicas. Figure 4-4 shows the temperature histories of the first 2 replicas in the same explicit solvent and GB REMD simulations as were shown in Figure 4-3. As expected, the replicas visited all available temperatures during the run (the other replicas showed similar behavior and are not shown for clarity). However, the explicit solvent REMD with only 8 replicas showed no exchanges even after 25,000 attempts (25ns simulation), and all replicas remained at their initial temperatures. This REMD simulation is identical to 8 standard MD simulations at different temperatures, and therefore no sampling improvement is obtained. Thus, in order for replicas to sample a range of temperatures, more replicas (and thus significantly more computational resources) are required for simulations in explicit solvent. Reducing this requirement while maintaining fully explicitly solvated simulations is the goal of our hybrid model.

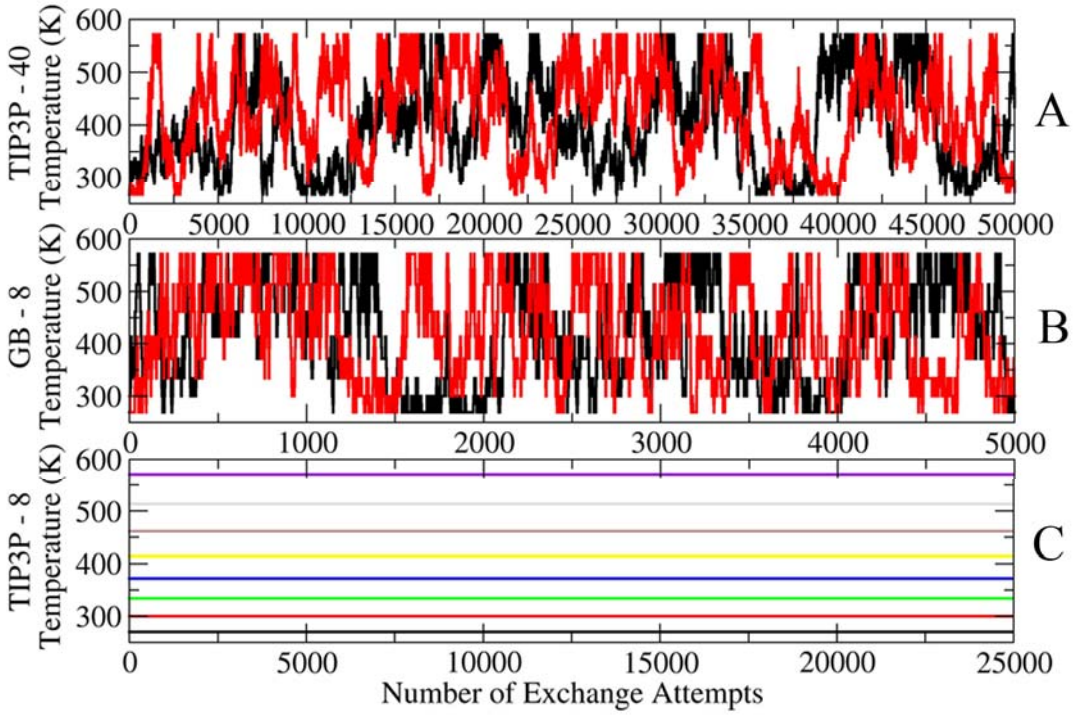


Figure 4-4. Temperature histories for Ala10 replicas using (A) explicit solvent with 40 replicas, (B) GB with 8 replicas and (C) explicit solvent with 8 replicas. For clarity only the first two replicas for A and B and only the first 5000 exchanges of B are shown. Consistent with the potential energy distributions shown in Figure 4-3, exchanges are only obtained when sufficient overlap in potential energy distributions is present. If too few replicas are used (C), the result is a series of standard MD simulations.

These exchange efficiencies are all consistent with previously reported REMD simulations and the known scaling with system size of the number of replicas required for efficient exchange. In our case this data provides an important context for evaluation of the use of hybrid solvation models during the calculation of exchange probability. We performed REMD simulations using the same explicitly solvated system as shown above, but with only the 8 replicas/temperatures that gave efficient exchange with pure GB solvation. With standard REMD, this system showed no overlap in potential energy

distributions and was unable to generate any successful exchanges (Figure 4-4C). We employed the hybrid solvent model only for calculation of the exchange probability (Equation 4-4) for this fully explicit solvent system. The distributions of the potential energies for the different temperatures during 10,000 exchange attempts (10 ns) are shown in Figure 4-5. Use of the hybrid solvent model permits the simulations to achieve nearly the same level of energy distribution overlap as we obtained for the pure GB model. Consistent with this observation, multiple exchanges are observed despite the relatively small number of replicas employed. The replicas are able to traverse the entire temperature range on the nanosecond timescale. It is interesting to note that this is more rapid than seen for the standard REMD explicit solvent run, most likely due to the larger temperature step taken with each successful exchange with the hybrid solvent model (due to larger ΔT between neighboring replicas). The standard REMD run requires more exchanges to traverse the same total temperature range. This suggests that the hybrid calculation may have additional advantages beyond simply reducing the number of replicas as compared to standard REMD; however such analysis is outside the scope of the present article.

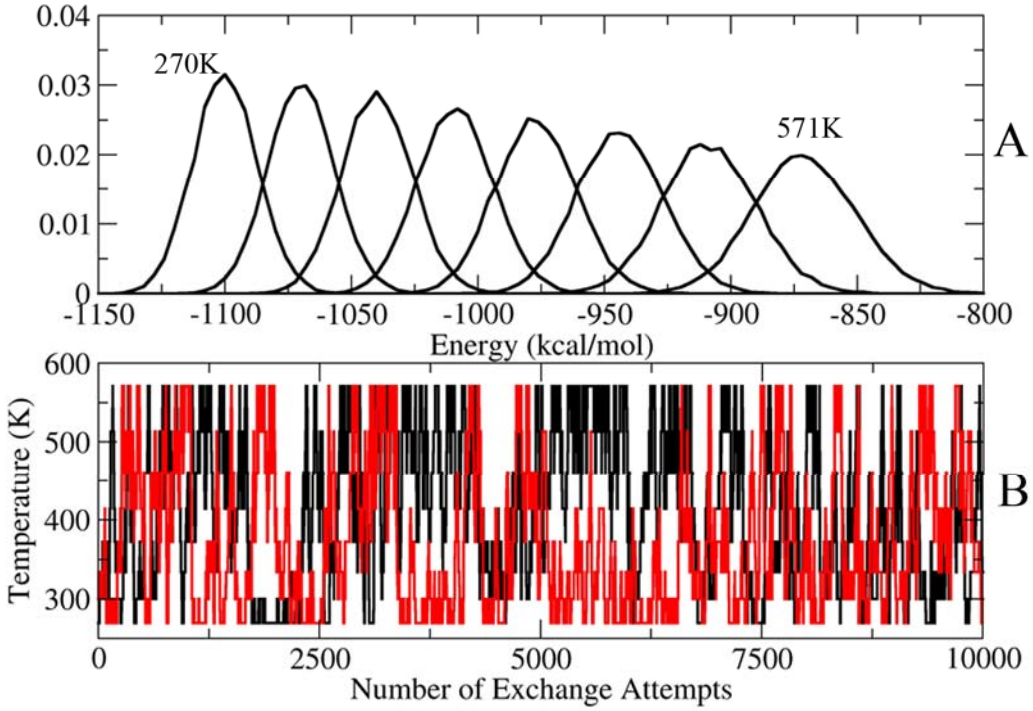


Figure 4-5. Potential energy distributions (A) and temperature histories of 2 Ala10 replicas (B) using 8 replicas in periodic boxes with fully explicit solvent, but with the hybrid solvent model for calculation of exchange probability. Use of the hybrid model gives overlap between neighboring temperatures and allows replicas to span a range of temperatures, in sharp contrast to the total lack of exchanges for the same simulated system with standard REMD Figure 4-3C and Figure 4-4C). For clarity only the first 10000 exchanges are plotted and only 2 replicas are shown in the lower figure.

4.3.2 Analysis of conformational sampling in hybrid and standard REMD

After establishing the ability of the hybrid REMD model to reduce the number of replicas required to obtain efficient exchanges, we examine the ability of the hybrid approach to reproduce ensemble data obtained with standard REMD in explicit solvent.

We also investigate whether the reaction field beyond the solvation shells is required, and the dependence of the results on the number of solvation shells included in the exchange calculation. For the larger Ala_{10}, the computational demands of obtaining high-precision data for various hybrid models (which require fully solvated simulations) prevented exhaustive testing. Thus, these more detailed tests were performed on the smaller models alanine dipeptide (blocked Ala_1) and alanine tetrapeptide (blocked Ala_3).

4.3.2.1 Alanine Dipeptide

We first compared results obtained for standard REMD with TIP3P to those from 2 different GB models, as well as to TIP3P but using the hybrid solvent model for calculation of exchange probability. The hybrid model employed either a first solvent shell (30 TIP3P waters) or first and second shells (60 waters). The population of minima corresponding to alternate secondary structure types (see Methods for details) are shown in Table 4-2. The largest population is found for the polyproline II basin (~35%), followed by α-helix and β-sheet (each ~25%), and a much lower population of left handed α-helix or turn conformation (1-3%). We make the observation that all of these solvent models provide essentially the same results. Use of either GB^{OBC} or GB^{HCT} with no explicit solvent either in MD or in the exchange calculation provides populations for each of the basins with an error of ~2% population as compared to the standard REMD in explicit solvent. Similarly, the average SASA is nearly identical for all models. These data indicate that the hybrid model is at least performing adequately and does not have any obvious and serious problems, and that similar results are obtained for either first and second solvation shells or only the first shell. This insensitivity is expected since the GB

simulations adequately reproduced the explicit solvent data with no explicit solvent shell. The insensitivity of the results to solvent model strongly indicates that alanine dipeptide is not a good test case for evaluation of the effects of inclusion of explicit solvent.

Alanine dipeptide	α	β	$\mathbf{P^{II}}$	α^{L}	**SASA**
Explicit solvent	28.1 ± 1.0	25.1 ± 0.1	36.2 ± 0.5	2.6 ± 0.1	355.8 ± 0.0
GB^{OBC}	29.3 ± 0.8	26.5 ± 0.5	35.1 ± 0.2	0.7 ± 0.1	356.5 ± 0.0
GB^{HCT}	28.5 ± 0.2	27.6 ± 0.1	34.0 ± 0.2	0.8 ± 0.2	356.5 ± 0.1
Hybrid 1^{st} shell + GB^{OBC}	29.7 ± 1.8	24.7 ± 0.4	35.0 ± 1.5	2.5 ± 0.1	355.8 ± 0.1
Hybrid 1^{st} and 2^{nd} shells + GB^{OBC}	30.3 ± 1.5	24.7 ± 0.3	36.0 ± 0.2	1.3 ± 0.8	355.9 ± 0.1

Table 4-2. Populations of basins on the alanine dipeptide ϕ/ψ energy landscape corresponding to alternate secondary structures, along with average solvent accessible surface areas. The results for the pure GB and hybrid REMD models are all similar to those obtained using standard REMD with full explicit solvent.

4.3.2.2 Alanine Tetrapeptide

We next turn to results from alanine tetrapeptide to evaluate whether the agreement between all solvent models tested for alanine dipeptide is maintained in larger systems. In

Table 4-3 we show populations for secondary structure basins for the central alanine residue using standard REMD with explicit solvent, GB^{OBC} or GB^{HCT}. Data is also shown for several hybrid models, as discussed below.

Alanine tetrapeptide	α	β	$\mathbf{P^{II}}$	α^{L}	**SASA**
Explicit Solvent	23.6 ± 0.1	23.4 ± 1.3	40.2 ± 1.4	5.1 ± 0.1	565.3 ± 0.1
GB^{OBC}	50.5 ± 2.4	17.5 ± 0.9	22.9 ± 0.6	1.1 ± 0.4	557.4 ± 1.0
GB^{HCT}	57.8 ± 1.0	15.2 ± 0.2	18.2 ± 0.4	1.2 ± 0.1	552.4 ± 0.4
Hybrid 1^{st} shell noGB	41.4 ± 0.8	13.5 ± 0.9	23.4 ± 1.0	13.1 ± 0.8	552.7 ± 0.1
Hybrid 1^{st} and 2^{nd} shells noGB	29.5 ± 0.2	14.1 ± 0.2	24.1 ± 0.5	23.4 ± 0.3	550.8 ± 0.2
Hybrid 1^{st} Shell GB^{OBC}	21.6 ± 0.9	21.2 ± 0.2	41.1 ± 0.3	7.6 ± 1.0	563.2 ± 0.1
Hybrid 1^{st} and 2^{nd} Shells GB^{OBC}	28.3 ± 1.7	22.2 ± 0.9	37.7 ± 0.2	3.8 ± 0.1	563.8 ± 0.2
Hybrid 1^{st} Shell + GB^{HCT}	23.5 ± 1.1	22.1 ± 0.8	42.8 ± 1.0	2.3 ± 0.0	566.4 ± 0.2
Hybrid 1^{st} and 2^{nd} Shells + GB^{HCT}	14.9 ± 0.2	25.6 ± 0.1	49.4 ± 0.4	1.9 ± 0.4	569.6 ± 0.1

Table 4-3. Data for the central alanine in alanine tetrapeptide (blocked Ala3). Populations of basins on the ϕ/ψ energy landscape corresponding to alternate secondary structures are shown, along with average solvent accessible surface areas. Data is discussed in the text.

For standard REMD in explicit solvent, we observe that the populations have not changed significantly from those obtained for alanine dipeptide, with a slight increase in population of the polyproline II conformation that dominates the ensemble. In this case, however, we observe that both of the pure GB models are in significant disagreement, with α-helical conformations dominating the ensemble (over 50% for each GB model). The two GB models are similar to each other. Overstabilization of salt bridges in GB has been reported[33-35], but no salt bridges are present in this system.

Next, we performed REMD simulations in explicit solvent, but retain only the first (50) or the first and second (100) solvation shells in the exchange calculation. Importantly, no GB model was included in these simulations. Using only a single solvation shell results in a significant bias in favor of α-helical conformations (41% vs. ~24% for standard REMD), much too little polyproline II conformation and nearly three times the α^L / turn conformation than was sampled in standard REMD. Inclusion of a second shell (without GB) resulted in an even greater shift of the ensemble toward turn structures. Notably, both of these shell models show significantly smaller average SASA than obtained with standard REMD in explicit solvent, consistent with a drive toward compact conformations that reduce the water/vacuum interface that is present without a reaction field to surround the solvent shells.

We next examine the data obtained from the hybrid model in which GB solvation was employed in addition to shells of explicit solvation. We note that all of these models are in significantly better agreement with the standard TIP3P REMD data, regardless of

the GB method or number of shells. The more recent GB^{OBC} model performed best, with errors in population of only ~3% for all basins with the exception of the α-helix conformation with the first and second shell model, which had an error that was less than 5%. The average SASA was also in excellent agreement with standard REMD. We conclude that this hybrid model is significantly better than the pure GB REMD or inclusion of only the solvation shells with no reaction field. The addition of a second shell in the exchange calculation appears to make no significant difference as compared to a single shell.

As described above, the MD simulations between exchanges in the hybrid model are performed with full explicit solvation. We thus do not need to restrain the explicit water and since the solvation shells are surrounded by bulk explicit solvent, we expect no effect on the water geometries as have been reported when using a hybrid GB+explicit water model for dynamics[119]. To test this hypothesis, we calculated the radial distribution function for water oxygens around the carbonyl oxygen in the central Ala2, and found that the function obtained in the hybrid model was indistinguishable from that in the standard REMD in explicit solvent (Figure 4-6). Since this data is obtained from the entire set of structures, this close agreement is also a further indicator of the similarity of the ensembles obtained using hybrid or standard REMD.

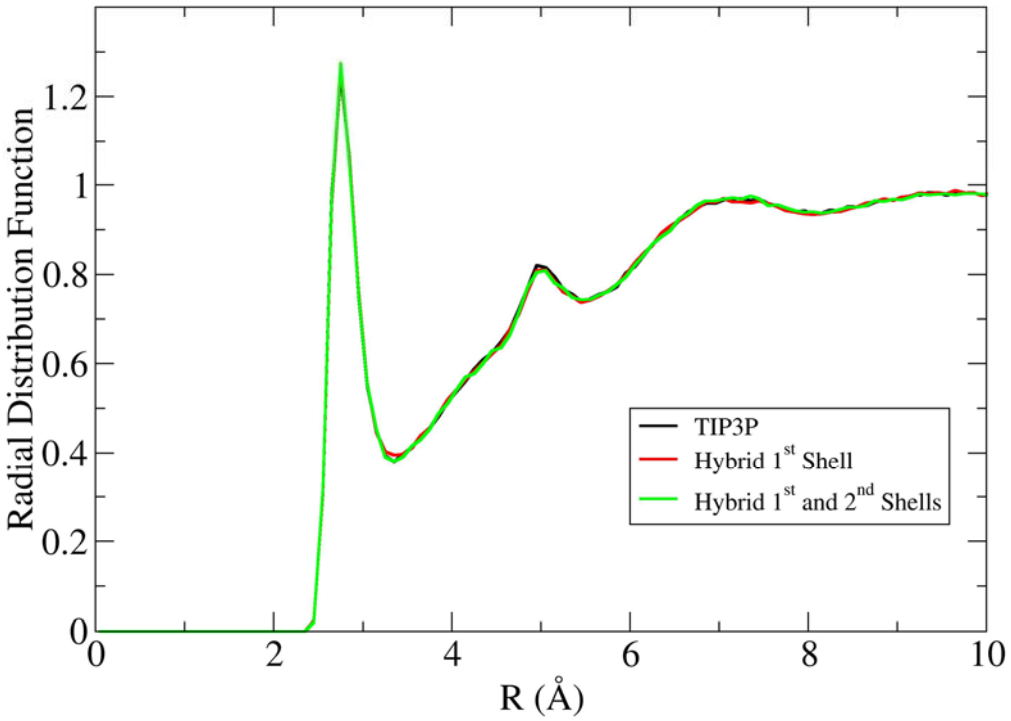


Figure 4-6. Radial distribution functions for water oxygen atoms around the carbonyl of Ala2 in alanine tetrapeptide, calculated using ptraj. The distributions for the hybrid models using either 1st or 1st and 2nd shells are nearly indistinguishable from those obtained using the reference standard REMD in explicit solvent.

The hybrid model using GB^{HCT} performed comparably to GB^{OBC} when only a single shell was used, but the first+second shell model showed a marked reduction in α-helix conformation (from 23.5% to 14.9%). This was accompanied by an increase in average SASA. These effects with GB^{HCT} are even more apparent in Ala_{10} and will be discussed in more detail below.

4.3.2.3 Polyalanine (Ala_{10})

The conformational variability available to Ala_{10} is significantly greater than for alanine dipeptide or tetrapeptide. We thus performed a more stringent evaluation of data convergence in this case to ensure that the differences we observe between the different solvent models are statistically significant. We performed two completely independent REMD simulations for each of the solvent models, in each case starting from 2 different initial ensembles (fully extended or fully helical). This allows us to evaluate the influence of the solvent model within the context of intrinsic uncertainties in each data set.

We also consider separately the local ϕ/ψ conformations and more global properties of this larger peptide, such as end to end distance distributions and conformation cluster analysis.

4.3.2.3.1 Comparison of local conformational preferences

In Table 4-4 we show secondary structure basin populations for the central Ala5 residue. Free energy surfaces for these simulations are provided in Figure 4-7. For the reference standard REMD simulations in explicit solvent, the polyproline II conformation is again favored with the same ~40% population as we obtained for alanine dipeptide and tetrapeptide. In comparison, both GB models show very large bias in favor of α-helix conformations (~70-80%).

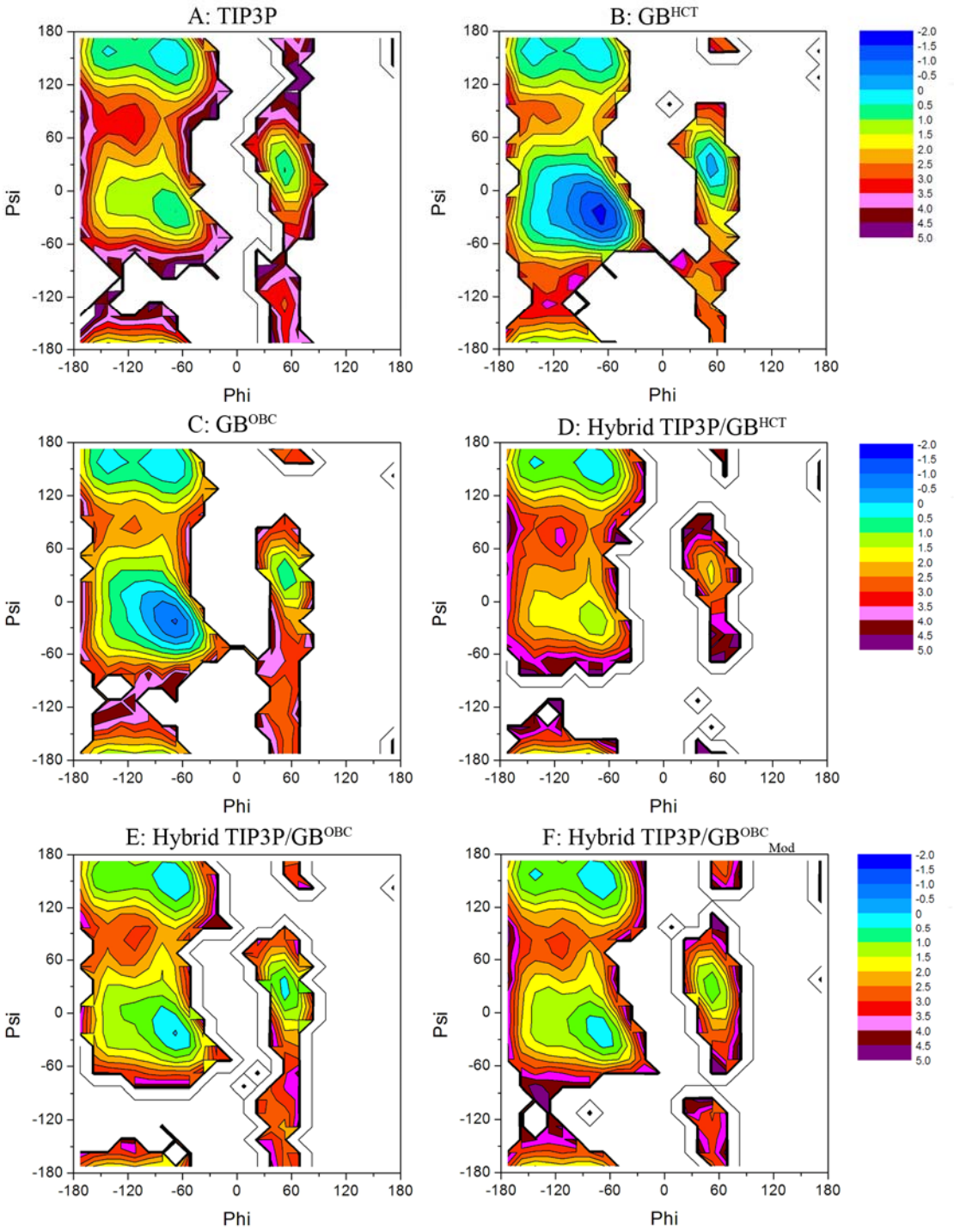


Figure 4-7. Free energy profiles at 300K for the central Ala5 residue from REMD in multiple solvent models. Contour levels are spaced 0.5 kcal/mol apart. Solvent models are (A) TIP3P explicit solvent, (B) GB^{HCT}, (C) GB^{OBC}, (D) GB^{HCT}/TIP3P hybrid, (E) GB^{OBC}/TIP3P hybrid and (F) GB^{OBC}/TIP3P hybrid with intrinsic Born radius on hydrogen bonded to oxygen reduced by 0.05Å. (D), (E) and (F) correspond to fully solvated REMD simulations with the hybrid model used only for calculation of exchange

probability. Basins corresponding to the major secondary structure types are all similar in free energy for models using explicit solvent; however both pure GB models show strong bias (2-3 kcal/mol) favoring α-helical conformations. Free energy landscapes were calculated using two dimensional histogram analyses of the dihedral angles of Ala5. For easier comparison between models, free energy values were normalized using the TIP3P REMD global minimum (the bin corresponding to $-75° < \varphi < -60°$, $150° < \psi < 165°$) as a free energy of zero.

Consistent with the results obtained for alanine tetrapeptide, the GB^{HCT} hybrid model favors extended conformations with large SASA too strongly (β and P_{II}), despite the bias in favor of α-helix for the pure GB^{HCT} simulations. This suggests that the explicit water shell is solvated too strongly by this GB model. The GB^{OBC} hybrid model shows a more balanced profile in good agreement with the full TIP3P data. The strong bias favoring α-helix in the pure GB^{OBC} model is nearly completely eliminated when a single solvent shell is retained, although some remains with approximately 10% too much α-helix present in the GB^{OBC} hybrid.

Ala_{10}	**α**	**β**	**P^{II}**	**α^{L}**	**SASA**
Explicit Solvent	24.9 ± 0.8	19.5 ± 0.6	39.5 ± 0.4	8.4 ± 2.0	1195.4 ± 5.6
GB^{OBC}	67.8 ± 1.8	8.3 ± 0.7	12.5 ± 0.8	4.2 ± 0.1	1098.6 ± 0.4
GB^{HCT}	83.1 ± 0.1	3.2 ± 0.1	5.0 ± 0.0	2.3 ± 0.1	1038.3 ± 1.6
Hybrid GB^{OBC} +1^{st} shell	35.7 ± 6.2	17.3 ± 0.2	29.0 ± 5.3	6.6 ± 0.7	1140.8 ± 4.4
Hybrid GB^{HCT} + 1^{st} shell	12.3 ± 0.2	28.3 ± 0.3	50.5 ± 1.2	2.1 ± 1.1	1275.4 ± 2.5
Hybrid $GB^{OBC'}$ +	29.8 ± 1.6	18.5 ± 1.6	34.3 ± 0.5	8.9 ± 0.3	1167.8 ± 2.5

1st shell					

Table 4-4. Data for the central Ala5 in blocked Ala10. Populations of basins on the ϕ/ψ energy landscape corresponding to alternate secondary structures are shown, along with average solvent accessible surface areas. GBOBC' refers to the hybrid model using GBOBC with slight adjustment of the Born radius on H bonded to O. Uncertainties reflect differences between independent simulations from different initial structures. Data is discussed in the text.

In addition to differences in the method for calculating GB effective Born radii, the GB^{HCT} and GB^{OBC} simulations employed different intrinsic Born radii (denoted in Amber as mbondi and mbondi2 sets, respectively), consistent with recommendations for these models. In order to determine the relative influence of these two differences, we repeated the calculations, swapping the GB models and radii (GB^{HCT} with mbondi2, GB^{OBC} with mbondi). We found that the results depended nearly exclusively on the set of radii and were less sensitive to the GB models themselves (data not shown). This is consistent with the aim of the GB^{OBC} model, which was designed to provide improved properties for larger systems than our current model[137]. We note that the strong bias toward extended structures seen in the hybrid models using mbondi radii likely arises from the use of 0.8 Å for hydrogen atoms bonded to oxygen. In the more recent mbondi2 set, this value was restored to the default Bondi value of 1.2 Å. This larger value appears to have an improved balance of hydrogen bonding of the explicit solvent to the solute or to the bulk (continuum) solvent.

4.3.2.3.2 Comparison of global structural properties

Our analysis of alanine dipeptide and tetrapeptide focused on local backbone conformation; in the larger Ala_{10} we supplement this analysis with more global properties of the chain. We calculated the end-to-end distance distributions for Ala10 in the 300K ensembles obtained from each of the different REMD simulations. In Figure 4-8 we show the results of the 2 explicit solvent REMD simulations that were initiated from fully α-helical or extended conformations, respectively. A broad distribution of distances is observed, suggesting that no particular conformation is preferred, consistent with the local backbone preferences for the central Ala5. Consistent with the small uncertainties in the ϕ/ψ basin populations, we observe that the initial conformation has essentially no effect on the distribution, indicating that the REMD simulations are well-converged on this timescale. Similar behavior is observed for other temperatures. As expected, standard MD simulations at 300K were trapped near the initial conformation on this timescale (data not shown).

In Figure 4-8, we show the distance distributions at 300K obtained from GB REMD using the two GB models (HCT and OBC). In contrast to the relatively flat profiles seen in the explicit solvent REMD data, a sharp peak near 11Å is obtained using either GB model, with essentially no sampling of extended conformations with end to end distances greater than ~15-20Å, unlike the explicit solvent REMD that shows a nearly flat distribution out to ~22Å. This is consistent with the strong bias toward α-helix in the pure GB models as shown in Table 4-4. The bias is somewhat less pronounced with the GB^{OBC} model than with GB^{HCT}. We note that these differences between the various solvent models are much larger than the differences obtained from alternate initial conformations using the same solvent model.

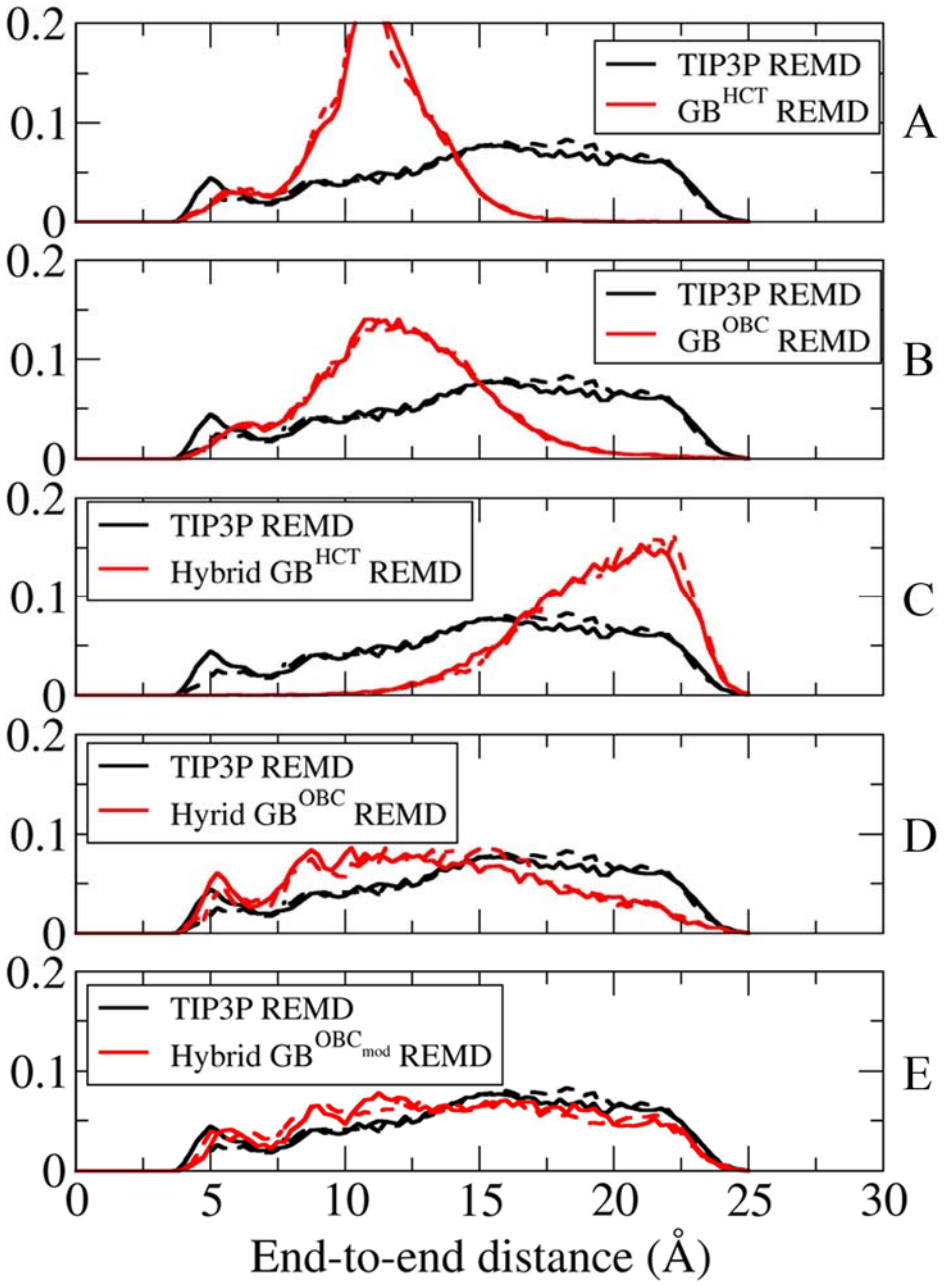


Figure 4-8. Ala_{10} end-to-end distance distributions at 300K obtained in REMD using alternate solvent models (red): (A) pure GB^{HCT}, (B) pure GB^{OBC}, (C) hybrid REMD with GB^{HCT} and mbondi radii, (D) hybrid REMD with GB^{OBC} and mbondi2 radii (HO=1.2 Å) and (E) hybrid REMD with $GB^{OBC'}$ (mbondi2 radii with H^{O}= 1.15 Å). In each case the

results are independent of initial conformation (solid/dashed lines). Data from standard REMD with explicit solvent is shown in each graph for comparison (black).

In Figure 4-8 we also show end to end distance distributions at 300K obtained from REMD with the same hybrid variations shown in Table 4-4, each of which retained only the first shell (100 closest) water molecules combined with different GB models in the exchange calculation. When GB^{HCT} was used in the hybrid model (Figure 4-8C), the distributions differ significantly from the reference explicit solvent REMD data, consistent with the large increase in polyproline II backbone conformations and average SASA for this model shown in Table 4-4. This bias toward more extended conformations in the hybrid using GB^{HCT} is also consistent with what we observed for alanine tetrapeptide (Table 4-3).

We next analyzed the distributions obtained from the GB^{OBC} hybrid model (Figure 4-8D). In this case, much better agreement with the reference data is seen than with either GB^{OBC} alone or the explicit/GB^{HCT} hybrid. However, the sampling of the most extended conformations (longest end to end distances) is slightly reduced in the hybrid REMD simulations.

The good convergence of our data suggested the possibility of using it for minor empirical adjustment of the mbondi2 values for use with the GB^{OBC} hybrid model. We adjusted the radii of hydrogen bonded to either N or O by 0.05 Å. Modification of H on N had little effect on the resulting distributions (data not shown), but reduction of the radius of H on O from 1.2 Å to 1.15 Å ($GB^{OBC'}$) resulted in an end to end distance distribution in improved agreement with standard explicit solvent REMD data (Figure 4-8E and Table 4-4). This slight reduction in the hydrogen radius is consistent with the increased

electronegativity of oxygen [142]. This change does not affect the pure GB calculations since Ala_{10} has no H bonded to O.

The $GB^{OBC'}$ hybrid model showed improved agreement with the pure TIP3P data, with all basin populations within 5% of the standard explicit solvent REMD. Some slight bias favoring α-helix at the expense of some polyproline II conformation remains in this model and will be the subject of future investigation. We repeated the simulations of alanine dipeptide and tetrapeptide using this modified radius and found that the populations (Table 4-5) remained in good agreement with standard REMD with explicit solvent.

ALA1	**α**	**β**	**P^{II}**	**α^{L}**	**SASA**
Hybrid Mod 1st shell	28.0 ± 0.5	24.5 ± 0.1	35.7 ± 0.6	3.5 ± 0.3	355.8 ± 0.0
Hybrid Mod 1st and 2nd shells	29.3 ± 1.0	23.8 ± 0.1	36.0 ± 0.1	2.8 ± 0.7	355.8 ± 0.1
ALA3	**α**	**β**	**P^{II}**	**α^{L}**	**SASA**
Hybrid Mod 1st shell	26.4 ± 0.5	22.3 ± 0.3	36.7 ± 2.1	6.9 ± 1.6	561.9 ± 0.3
Hybrid Mod 1st and 2nd shells	21.1 ± 1.6	22.6 ± 0.4	43.5 ± 2.2	4.4 ± 1.2	556.4 ± 1.1

Table 4-5. Populations of basins on the ϕ/ψ energy landscape corresponding to alternate secondary structures, along with average solvent accessible surface areas. These simulations employed the modified intrinsic Born radius for hydrogen bonded to oxygen, as described in the text.

Since the backbone conformation populations suggest that the P_{II} basin is the global free energy minimum in both the standard explicit solvent and the hybrid solvent models (Table 4-4 and Figure 4-7), we performed cluster analysis to determine the extent to which this local preference was reflected in the conformation of the entire polymer chain. Once again we compare results from independent ensembles generated by REMD with different initial conformations to ensure the convergence of our data.

The most populated cluster for Ala_{10} at 300K in both standard explicit solvent REMD runs was an extended P_{II} conformation (over 98% of the local backbone conformations in this cluster are P_{II}, data not shown). This fully P_{II} cluster comprised ~20% of the overall ensemble in both explicit solvent simulations (19.5% vs. 21.2%). Representative structures for the clusters obtained from the independent simulations differed by only 1.3Å in backbone RMSD (Figure 4-9A). Once again, the high level of consistency between the data sets and independence of not only the conformation but the absolute population of the clusters give us confidence in the converged nature of our data. The relatively low population of this cluster in both simulations is also consistent with the broad distribution of end to end distances (Figure 4-8). A more detailed analysis of the ensemble of structures sampled by Ala_{10} will be presented elsewhere, but this preference for P_{II} conformations is consistent with the experimental and simulation reports described previously.

As was demonstrated with the analyses presented above, the pure GB^{HCT} and GB^{OBC} REMD simulations do not reproduce the data obtained in explicit solvent, nor are they consistent with experimental data. The most populated cluster in both cases is fully α-helical (Figure 4-9B shows the GB^{OBC} structure), comprising ~48% of the overall

ensemble for GB^{HCT}, and 25.4% for GB^{OBC}. This analysis is consistent with the α-helical bias apparent in the Ramachandran free energy surfaces shown in Figure 4-7.

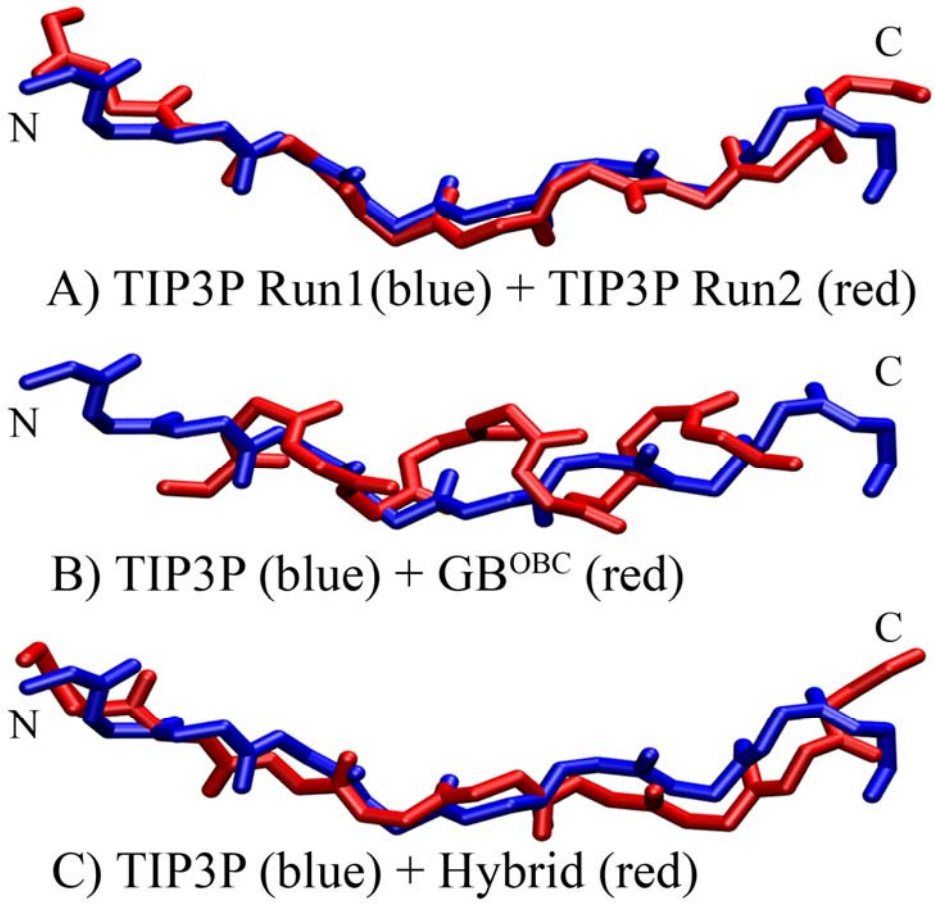


Figure 4-9. Representative structures for the most populated clusters in 300K ensembles obtained using various solvent models. (A) Very similar PII structures are obtained from 2 independent standard REMD simulations with explicit solvent, initiated in extended and fully helical conformations. (B) Comparison of structures from GBOBC and TIP3P. GBOBC prefers α-helical conformations, in disagreement with explicit solvent simulations. (C) Using GBOBC' with the hybrid model provides structures in close agreement with standard REMD in TIP3P. Terminal residues were not included in the cluster analysis.

We next performed cluster analysis on the ensembles obtained with the $GB^{OBC'}$ hybrid model with modified mbondi2 radii. Consistent with the standard explicit solvent REMD runs, the most populated cluster at 300K was also an extended P_{II} conformation. Representative structures were within 1.5 Å backbone RMSD from those obtained in explicit solvent (Figure 4-9C), again suggesting that the hybrid model is able to capture the dominant effects of explicit solvent in the exchange calculation despite the need for many fewer replicas.

Since the most populated clusters were in close agreement between both TIP3P REMD simulations and $GB^{OBC'}$ hybrid model, we compared the populations of all clusters observed. Smith et al. showed[39] that cluster analysis of simulations was a much more stringent test of convergence than other measures that they tested, including energy, RMSD or diversity of hydrogen bonds sampled. This was particularly useful when analyzing coordinate sets obtained by merging two independent trajectories. They examined the 5 ns dynamics of an 11-residue peptide and showed that the two trajectories sampled essentially none of the same clusters.

We adapted this approach to our analysis, but we emphasize not just the existence of conformation families in two data sets, but also the fractional population of each cluster in 300K ensembles sampled in independent simulations. All trajectories from TIP3P REMD, GB^{OBC} REMD and hybrid $GB^{OBC'}$ simulations were combined and the resulting data set was clustered. A total of 44 clusters contained 99% of the structures; the fraction of the ensemble corresponding to that cluster was calculated for each REMD simulation. We compared the population of each cluster in the different ensembles, including those generated with the same or different solvent models.

First we evaluated the convergence of our standard REMD simulations with TIP3P by comparing cluster sizes between the independent runs with different initial conformations (extended and fully α-helical). Not only were the same conformations sampled in each run (20.3±0.9%), but the populations of clusters in each ensemble were highly correlated (Figure 4-10A, R^2=0.974 and a slope of 1.02). This indicates that the relative population of each structure type is highly converged in these data sets.

In stark contrast, when the TIP3P and GB^{OBC} ensembles are compared, no correlation between cluster populations is observed (Figure 4-10B, R^2=0.075), and the largest cluster in each (~20%) has less than 2% population in the other model. Much better results are obtained from the $GB^{OBC'}$ hybrid data, with a correlation coefficient of 0.935 with the standard TIP3P REMD data (Figure 4-10C). All clusters larger than 5% have the same rank order in the two models. There is a relatively small difference in the size of the single cluster that is the largest for both models (15.9±0.6% and 20.3±0.9% for hybrid and standard TIP3P REMD, respectively). This corresponds to an error of only 0.15 kcal/mol for the free energy of this cluster between the two models, compared to 0.05 kcal/mol difference obtained between data sets from the same model. For comparison, the error in the free energy of this conformation using GB was more than ten times larger (1.6 kcal/mol).

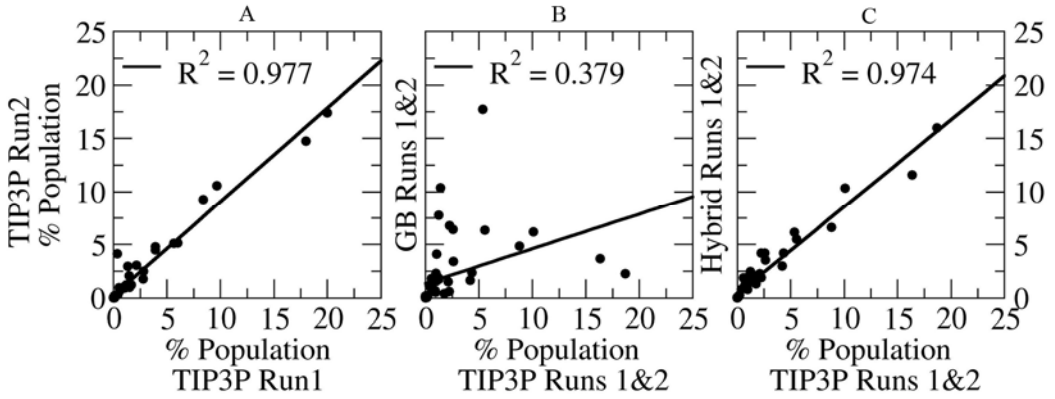


Figure 4-10. Cluster populations at 300K from REMD for TIP3P Run1 vs. Run2 (A), TIP3P Runs 1&2 vs. GB^{OBC} Runs 1&2 (B) and TIP3P Runs 1&2 vs. hybrid $GB^{OBC'}$ Runs 1&2. High correlations between individual TIP3P simulations and between TIP3P and hybrid simulations are observed, with the difference in the largest cluster in (C) corresponding to an error in free energy of only 0.18 kcal/mol. No correlation between TIP3P and GB^{OBC} is observed; note also in plot (B) that the largest cluster in each solvent model has very low population in the other model (indicated by arrows).

Since the standard explicit solvent REMD and hybrid solvent using $GB^{OBC'}$ have the same most populated cluster, we investigated the timescale required for each model to adopt this conformation as the dominant member of their ensemble. This is important since the standard REMD simulation employed many more replicas, possibly facilitating earlier location of the P_{II} conformation that would then be adopted in the lowest temperature ensembles. In Figure 4-11 we show the fractional size of this cluster in the structures sampled as a function of time for standard REMD and hybrid REMD, including data from both initial conformations in each model. Data is shown at 300K, and the first 5 ns were discarded in each case to remove biasing of the populations by the initial conformations that were not sampled at later points. The level of agreement is impressive; the long-time averages for both simulations of the 2 models are all ~20%,

with convergence to this value occurring at approximately 5ns in all cases (in addition to the 5 ns that was discarded).

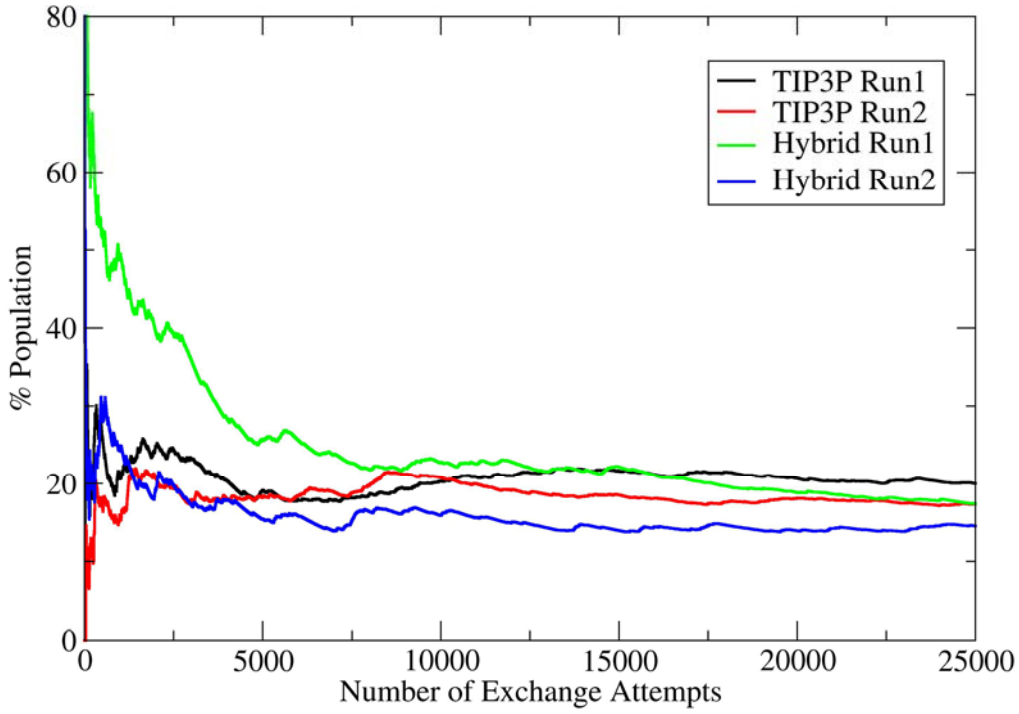


Figure 4-11. Population of the cluster corresponding to polyproline II helix (Figure 4-10) as a function of time for REMD simulations in explicit solvent, with the 2 independent simulations using the full system energy in the exchange calculation shown in black/red and the $GB^{OBC'}$ hybrid shown in green/blue. At ~5ns, all four simulations converge to a population of 16-20% (the largest cluster in each of the ensembles), with a slightly lower population in the hybrid models that is consistent with Figure 4-10C.

4.4 Conclusions

We introduced a new variant of replica exchange molecular dynamics in which simulations are performed with a fully explicit representation of solvent, but those

solvent molecules beyond the first solvation shell are replaced with a continuum description only for the purpose of calculating the exchange probability. This reduces the effective system size governing the number of replicas required to span a given temperature range, and therefore significantly reduces the computational cost of REMD simulations. This approach is similar in spirit to hybrid explicit/continuum models that have been proposed for use during each step of MD simulation; in the present case, however, the solvent is fully explicit during the dynamics and no restraints are needed to maintain a solvation shell. However, since the Hamiltonian used for the exchange differs from that employed during dynamics, these simulations are approximate and are not guaranteed to provide correct canonical ensembles. It is important to determine the extent to which this approximation affects the resulting ensembles; in this article we introduce the method and investigate some of these effects on several short alanine-based peptides.

Recently, another approach to reducing the number of replicas required for explicit solvent REMD simulations was proposed[144] in which the water-water interaction energy was temperature-dependent. That study employed alanine dipeptide as a model to show that their less computationally demanding method provided a similar ensemble to that obtained with standard REMD. In the present work we show that alanine dipeptide conformations are nearly insensitive to the solvent models that we tested, with results from full explicit solvent, two different GB models and several hybrid models all providing similar ensembles. In contrast, several of these models provided ensembles for the longer peptides that were in significant disagreement with standard REMD in explicit solvent, indicating that larger model systems should be included in evaluation of solvent models.

We further tested the method by calculation of conformational ensembles of Ala_{10} using the TIP3P explicit solvent model, two GB models available in Amber, and hybrid variants using TIP3P and each GB model, all using the same underlying protein force field parameters. Ensembles from standard REMD in explicit solvent were considered the standard, and convergence of this data set was validated by a high correlation (R^2=0.974) between the fractional populations of conformation families in simulations initiated with completely different initial structure ensembles. While a broad distribution of conformations was sampled, the predominant cluster for Ala_{10} adopted a P_{II} structure. This preference is consistent with reported experimental and computational results for short polyalanine peptides [145].

Simulations using the hybrid model with GB^{OBC} were in excellent agreement with the reference data for local backbone conformations, end to end distance, SASA and populations of each conformation family in the ensemble. The difference in population in the largest cluster indicates that the hybrid model introduced and error of less than 0.2 kcal/mol in free energy while reducing the computational expense by a factor of five.

In contrast, REMD using only the GB models provided ensembles that bore no resemblance to the reference data, with the GB ensembles incorrectly dominated by α-helical conformations. This may be indicative of general errors in these GB models, or they may arise from neglect of structure in the first solvation shells of the peptide. Mezei et al. recently reported[136] free energy calculations in explicit solvent, showing that solvation strongly favors the P_{II} conformation over α-helix. Solvation free energy was shown to be highly correlated with the energy of interaction between the peptide and its first solvation shell.

It is important to note that several challenges remain for more general use of the proposed hybrid approach. In particular, the present work studied the effects on alanine-based peptides. Future studies should be performed on other sequences with a more diverse representation of functional groups in the side chains. In particular, it will be important to determine whether the hybrid model is able to overcome known issues with GB models and ions pair interactions. The inclusion of explicit counterions in the exchange calculation may also be problematic. Additionally, we demonstrated that inclusion of a single shell of explicit water was sufficient for alanine dipeptide and analine tetrapeptide. In both cases similar results were obtained using one or two shells, but we were unable to perform these comparisons for Ala_{10}. Although our approach reduces the number of replicas required for REMD, the simulations are still fully solvated during each step of MD and obtaining well converged data requires a significant investment of computational resources.

The results obtained from these model systems provide additional evidence that explicit representation of water in the first solvation shell can significantly improve the performance of the GB continuum models, providing data similar to standard REMD with fully explicit solvent but at a greatly reduced cost. This reduction in computational requirements can enable simulations on longer time scales for the same system size, or permit application of REMD to the study of much larger systems. We also showed that use of one or two explicit solvent shells alone was inadequate and that adding a reaction field was essential for obtaining reasonable results. Adaptation of this method to other continuum models (such as the more rigorous PB) should be straightforward. Since the

continuum solvent is only used for the infrequent exchange calculations, models that are too complex for use at each step of dynamics can be readily employed.

Chapter 5

Improving Convergence of Replica Exchange Simulations through Coupling to a High Temperature Structure Reservoir

5.1 Introduction

Conformational sampling remains one of the largest challenges in simulating biologically relevant events in atomic detail. Even when a sufficiently accurate Hamiltonian of the system is used, the rugged and complex potential energy surfaces usually result in simulations being trapped, prohibiting complete exploration of conformational space. Thus, significant effort has been put into devising efficient simulation strategies that locate low-energy minima for these complex systems. The challenges in conformational sampling has been discussed in several reviews [37, 38].

One major problem for molecular simulations is quasi-ergodicity where simulations may appear converged when observing some simulation parameters, but in reality large energy barriers may prevent them from sampling important regions of the energy landscape. Another simulation initiated in a different conformation may look converged as well, but comparison may show that only partial equilibration was achieved. An example of this behavior has been demonstrated by Smith et al. who reported that MD simulations of short peptides starting from different initial

conformations were in poor agreement despite apparent convergence in some measured properties [39].

One popular approach to overcoming quasi-ergodicity in biomolecular simulation is the replica exchange method [41, 92, 93]. In replica exchange molecular dynamics (REMD) [42] (also known as parallel tempering[41]), a series of molecular dynamics simulations (replicas) are performed for the system of interest. In the original form of REMD, each replica is an independent realization of the system, coupled to a thermostat at a different temperature. The temperatures of the replicas span a range from low values of interest (experimentally accessible temperatures such as 280 or 300K) up to high values (such as 600K) at which the system is expected to rapidly overcome potential energy barriers that would otherwise impede conformational transitions on a computationally affordable timescale.

At intervals during the otherwise standard simulations, conformations of the system being sampled a different temperatures are exchanged based on a Metropolis-type criterion[94] that considers the probability of sampling each conformation at the alternate temperature (further details are discussed in Methods). In this manner, REMD is hampered to a lesser degree by the local minima problem, since simulations at low temperatures can escape kinetic traps by "jumping" directly to alternate minima being sampled at higher temperatures. Moreover, the transition probability is constructed such that the canonical ensemble properties are maintained during each simulation, thus providing potentially useful information about conformational probabilities as a function of temperature. Due to these advantages, REMD has been widely applied to studies of peptide and small protein folding [18, 23, 34, 36, 41, 42, 48, 95, 97-99].

For large systems, REMD can become intractable since the number of replicas needed to span a given temperature range increases with the square root of the number of degrees of freedom in the system[100-103]. Since the number of accessible conformations also typically increases with system size, the current computational cost for REMD simulations of large systems limits the simulation lengths to tens of nanoseconds per replica, which limits the ability to obtain converged ensembles for large systems. Several promising techniques have been proposed[43, 102, 104, 106, 146] to deal with this apparent disadvantage of REMD. To our knowledge converged REMD simulations in explicit solvent from independent starting conformations have been reported only for short helical or unstructured peptides. [32, 43, 147]

Several studies have compared the sampling efficiencies of standard MD and REMD. Sanbonmatsu and Garcia reported a fivefold increase in sampled conformations using REMD over MD in the 5 residue Met-enkephalin peptide in explicit solvent [96]. Zhang et al. showed that REMD enhances sampling over conventional MD by 15 – 70 times at different temperatures for the 21 residue Fs peptide in continuum solvent [148]. A recent study by Zuckerman and Lyman investigated the sampling efficiency of REMD through consideration of the rate acceleration afforded by increased temperature [149]. For slower converging systems (such as β-hairpins or more complex topologies where folding time is in the order of microseconds) REMD simulations typically initiated from the native conformation (see recent example by Zhang et. al. [150]) where unfolding through high temperature replicas is obtained and temperature dependent properties are calculated from the resulting structures.

REMD simulations increase conformational sampling over standard MD simulations, but obtaining reliable results for non-trivial systems remains challenging. It is possible that REMD does not provide even greater efficiency gains for peptides and proteins because the temperature dependence of the folding rate tends to be more weakly temperature dependent than the unfolding rate, as has been shown experimentally [74, 151-154] and computationally[155, 156]. When starting from non-native conformations, high temperature replicas give limited advantage for finding native states since more minima on the free energy landscape become accessible at higher temperatures, further complicating the search. Furthermore, when a high temperature REMD replica locates a favorable low-energy basin (such as the native structure), this conformation is exchanged to lower temperature and the high temperature replica needs to repeat the search process. Importantly, during the search by the high-temperature replicas, all replicas continue to be simulated. Thus a very large set of simulations, all of which are long enough for the high-temperature replicas to sample multiple folding events, can be required to achieve correct Boltzmann-weighted ensembles across the range of replicas. From another perspective, REMD drives the generation of correct equilibrium ensembles of structures by employing an exchange criterion that explicitly assumes that structures being considered for exchange have Boltzmann-weighted probability of being sampled (see Methods for details). However, this assumption is only true after the generalized ensemble has already reached convergence and is typically incorrect at the start of the REMD simulation. Thus until all temperatures sample an equilibrium ensemble, none of the temperatures would be expected to have correct distributions due to coupling of replicas through an incorrect exchange probability.

An approach to reducing quasi-ergodicity that is conceptually similar to REMD was reported by Frantz et al. for Monte Carlo (MC) simulations of atomic clusters [157]. In their approach, called jump-walking (or J-walking) they coupled one MC simulation to another at higher temperature. Somewhat analogously to REMD, the low temperature simulation was used to sample local minima and provide thermodynamic ensemble data at the temperature of interest, while the high temperature simulation was used to facilitate barrier crossing. Periodically the low temperature structures escape local minima by "jumping" to basins sampled at high temperatures. The Boltzmann distribution generated by the high temperature walker becomes the sampling distribution for attempted jumps by the low temperature walkers. One drawback is that too large a temperature difference results in poor acceptance probabilities for the jump, comparable to the need to optimize the spacing between REMD temperatures. Variations of the J-walking scheme were tested by employing high temperature simulations on a different time scale than the low temperature simulation or using multiple high temperature simulations. They determined that the most efficient method is running the high temperature walker to obtain an adequate distribution, and using the stored conformations for jumps in a MC run at slightly lower temperature. The results of this lower temperature run were then used as the seed set for a new J-walking run at even lower temperature. They validated this approach using simple double well potentials where comparison to analytical results was possible, and in simulations of Argon clusters of various sizes. Similar approaches to J-walking have been developed, such as Smart Walking (S-Walking) [158] , Smart Darting [159] and Cool Walking [160]. The J-walking scheme has been adapted to REMD simulations that employ a resolution exchange scheme, where replicas were run using a

coarse grained model to obtain conformations to be subsequently sampled by an all-atom model[161, 162].

Here we introduce a variant to REMD where we draw upon the strengths of the J-walking approach to overcome the slow convergence and high computational expense of REMD. Similar to J-walking, an ensemble of structures is generated using standard MD simulation at high temperature. Instead of reducing the temperature stepwise and re-equilibrating the ensemble in stages, an REMD run is used to link in a single step the high-temperature ensemble to the low temperature of interest. Periodic exchanges are made between randomly chosen conformations from the reservoir set and the highest temperature replica. This process formally provides correct ensembles at lower temperature with free energies that reflect the proper relative populations of minima. Importantly, the convergence speed of the REMD run is greatly enhanced since exchanges are attempted from an already converged Boltzmann ensemble and thus the exchange probabilities are correct at the start of the REMD run. We call this method Reservoir REMD (R-REMD) since REMD is coupled to a high temperature reservoir.

One major advantage of the reservoir approach with REMD is that a converged ensemble of conformations has to be generated only once and only for one temperature. After extensive conformational search at one temperature, the remaining temperatures can sample from and anneal these structures to rapidly construct equilibrium distributions consistent with their thermostat temperature. This is in contrast to the typical REMD approach where all replicas are run simultaneously, and the computational expense for running long simulations must be paid for each of the replicas even though only a few high-temperature ones may be contributing to the sampling of new basins. Another

advantage is that the exchanges with the reservoir need not be time correlated with the replica simulations. Folding events sampled during reservoir generation can provide multiple native structures for the other replicas, in contrast to standard REMD where an independent folding event is required for each temperature that will have substantial native population. Overall, the convergence rate for the set of replicas is greatly enhanced by exchanging with a previously converged ensemble.

We have implemented the reservoir REMD approach in the Amber [163] simulation package and have tested it on two models peptides, the trpzip2 β-hairpin [59] and the dPdP [164] three-stranded antiparallel β-sheet. These systems were selected due to the complexity and slow folding of β-sheets and hairpins as compared to α-helices, which fold rapidly enough that the performance advantage of R-REMD may not be apparent. For both systems, reservoir ensembles were generated at 400K using Generalized Born [29] (GB) implicit solvent model using multiple simulations with different initial conditions. Subsequent R-REMD simulations we compared to standard REMD calculations with the same temperature ranges. In all cases, simulations were extended until close agreement was obtained between results obtained from independent runs with different initial structure ensembles (folded and unfolded). For both peptides the use of reservoir structures is shown to provide the same structure ensembles and thermal melting profiles as standard REMD, with a reduction in overall computational cost of 5 to 20 times, including the generation of the reservoir.

5.2 Methods

5.2.1 Replica Exchange Molecular Dynamics (REMD)

We briefly summarize the key aspects of REMD as they relate to the present study. In standard Parallel Tempering or Replica Exchange Molecular Dynamics [41, 42], the simulated system consists of M non-interacting copies (replicas) at M different temperatures. The positions, momenta and temperature for each replica are denoted by ($q^{[i]}$, $p^{[i]}$, T_m), i = 1,…,M ; m = 1,…, M. The equilibrium probability for this generalized ensemble is

$$W(p^{[i]}, q^{[i]}, T_m) = \exp\left\{-\sum_{i=1}^{M} \frac{1}{k_B T_m} H(p^{[i]}, q^{[i]})\right\}$$

Equation 5-1

where the Hamiltonian $H(p^{[i]}, q^{[i]})$ is the sum of kinetic energy $K(p^{[i]})$ and potential energy $E(q^{[i]})$. For convenience we denote $\{p^{[i]}, q^{[i]}\}$ at temperature T_m by $x_m^{[i]}$ and further define X = $\{ x_1^{[i(1)]}, \ldots, x_M^{[i(M)]} \}$ as one state of the generalized ensemble. We now consider exchanging a pair of replicas. Suppose we exchange replicas i and j, which are at temperatures T_m and T_n respectively,

$$X = \{\ldots; x_m^{[i]}; \ldots; x_n^{[j]}; \ldots\} \rightarrow X' = \{\ldots; x_m^{[j]}; \ldots; x_n^{[i]}; \ldots\}$$

Equation 5-2

In order to maintain detailed balance of the generalized system, microscopic reversibility has to be satisfied, thus giving

$$W(X)\,\rho(X \rightarrow X') = W(X')\,\rho(X' \rightarrow X)$$

Equation 5-3

where ρ(X→X') is the exchange probability between two states X and X'.

A key step in the derivation of the exchange criterion [42] is the substitution of the Boltzmann factor for the weight of each conformation into Equation 5-3, yielding Equation 5-4. We note that this is not strictly correct until equilibrium has been reached, at which point the structures are actually considered for exchange with this probability.

$$\exp\left\{-\frac{1}{k_B T_m} H\left(p^{[i]}, q^{[i]}\right) - \frac{1}{k_B T_n} H\left(p^{[j]}, q^{[j]}\right)\right\} \cdot \rho(X \rightarrow X') = \exp\left\{-\frac{1}{k_B T_m} H\left(p^{[j]}, q^{[j]}\right) - \frac{1}{k_B T_n} H\left(p^{[i]}, q^{[i]}\right)\right\} \cdot \rho(X' \rightarrow X)$$

Equation 5-4

In the canonical ensemble, the potential energy E rather than total Hamiltonian H can be used because the momentum can be integrated out [42]. By rearranging Equation 5-4 the following Metropolis exchange probability is obtained (Equation 5-5).

$$\rho = \min\left(1, \exp\left\{\left(\frac{1}{k_B T_m} - \frac{1}{k_B T_n}\right)\left(E(q^{[i]}) - E(q^{[j]})\right)\right\}\right)$$

Equation 5-5

It is important to reiterate that Equation 5-4 is valid only for equilibrated ensembles that follow Boltzmann distributions. This assumption is true at the end of the

simulation, and use of this exchange probability drives each replica towards adoption of the correct ensemble.

In standard REMD, several replicas at different temperatures are simulated simultaneously and independently for a chosen number of MD steps. Exchange between a pair of replicas is then attempted with a probability of success calculated from Equation 5-5. If the exchange is accepted, the bath temperatures of these replicas will be swapped, and the velocities will be scaled accordingly. Otherwise, if the exchange is rejected, each replica will continue on its current trajectory with the same thermostat temperature.

5.2.2 Reservoir REMD (R-REMD)

Reservoir REMD simulations (R-REMD) were run using same simulation parameters as standard REMD simulations. The only difference is that the highest temperature replica is replaced with a previously generated structure reservoir (replica R^N). Standard replicas (MD simulations) were used for each of the lower temperatures (replicas R^1 to R^{N-1}). Exchanges are attempted based on the same criterion as used for standard REMD (Equation 5-4). During exchange attempts for replicas between R^1 and R^{N-1} the exchange calculation is performed using current simulation coordinates. The only difference between R-REMD and REMD is when an exchange is attempted between replica R^{N-1} and the reservoir set R^N. The exchange attempt is made between the current structure of R^{N-1} and a randomly selected structure from the reservoir. If the exchange is accepted, the coordinates and velocities from R^N are sent to replica R^{N-1}. Formally the coordinates from replica R^{N-1} would be placed into the reservoir, however for simplicity it is discarded since we assume that the reservoir constitutes a complete representation of

the ensemble and that the inclusion of the new coordinates will have a negligible effect on the reservoir.

5.2.3 Model Systems

The first model system chosen was the tryptophan zipper (trpzip) developed by Starovasnik and coworkers [59]. This β-hairpin structural motif is stabilized through cross-strand tryptophan pairs. Trpzip2 (SWTWENGKWTWK, with a type I' β-turn at NG) has the most cooperative melting curve and highest stability (~90% at 300K) among the trpzips, and was selected for use in this study. Thermodynamic properties for this peptide have been determined by NMR and CD spectroscopy, and a family of structures was refined using restraints from NMR experiments [59] (PDB code 1LE1). The N-terminal of the peptide was acetylated and the C-terminal was amidated, in accord with the experiments.

The second model system was created from the sequence of DPDP (VFITSdPGKTYTEVdPGOKILQ, dP=D-proline, O=ornithine) except that lysine was substituted for the ornithine. Replacing ornithine with lysine in a related peptide analogous to the C-terminal hairpin of DPDP caused no detectable effect on the structure [165]. The termini were amidated and acetylated in accordance with experiments. DPDP was designed with a net charge of +2 to prevent aggregation, and our model retains this net charge[164].

5.2.4 REMD Simulations

For both systems standard REMD simulations were carried out with Amber version 8 [163]. For trpzip2 all covalent bonds were constrained using SHAKE [63]. For dPdP only the bonds involving hydrogen atoms were constrained. A 2fs time step was used and temperatures were maintained using weak coupling [86] to a bath with a time constant of 0.5 ps^{-1}. All non-bonded interactions were calculated at each time step (i.e. no cutoff was used). In order to permit comparison to our previously published data, both peptides were simulated with the Amber ff99 force field with modified backbone parameters [22]. Steepest descent energy minimization was performed for both systems for 500 steps prior to REMD simulations. Both systems were simulated with Generalized Born solvation model [29] with GB^{HCT} [87] implementation in Amber. Scaling factors were taken from the TINKER modeling package [143].

Standard REMD simulations were performed for both systems using 14 replicas for trpzip2 and 12 replicas for dPdP, covering a temperature range of ~260 – 570K with an expected exchange probability of 15%. For trpzip2 additional replicas were manually placed between 300 K and 370 K to increase statistics around the experimentally observed melting transition. Exchanges between neighboring replicas were attempted at 1ps intervals.

For both systems two independent replica exchange simulations were run. For trpzip2 one simulation initiated all replicas in the published native conformation. The other simulation started from a compact non-native conformation where no hairpin backbone hydrogen bonds were present. Both REMD simulations were run to 155000 exchange attempts (155 ns per replica). dPdP simulations were run as explained in our

previous work [23], with a simulation starting with all replicas in fully extended and another with all replicas in a compact non-native structure. dPdP simulations were carried out for 170000 exchange attempts (170ns per replica).

5.2.5 Generation of Reservoir Structures

The reservoir structures were generated through molecular dynamics at 400K with the same simulation parameters as used for REMD simulations. For trpzip2 four independent MD simulations of ~38 ns in length were run starting from an extended conformation. Multiple folding and unfolding transitions were observed for each trajectory. For dPdP a single long trajectory of 260ns was generated. Multiple folding and unfolding transitions were observed. For both systems velocities and coordinates were saved each 1 ps. In the present implementation of R-REMD in Amber, coordinates and velocities for the reservoir were loaded into memory at the start of the R-REMD run. To minimize memory requirements, the reservoir ensembles were reduced to 10000 structures by selecting equidistant snapshots from the trajectories.

5.2.6 Reservoir REMD Simulations

For trpzip2 four replicas were used below the 400K reservoir with temperatures of 300 K, 323 K, 350 K and 373 K. No additional replicas were used since these four replicas were sufficient to provide a 25-30% exchange ratio. Two sets of R-REMD simulations were each run for 50000 exchange attempts, starting from the same native or unfolded initial conformations as used for the standard REMD calculations.

Since dPdP is a larger system, R-REMD simulations used 6 replicas below 400K with the same temperature distribution as the standard REMD reported by Roe et al. [23]. One R-REMD simulation starting from extended conformations was run for 50000 exchange attempts.

5.2.7 Analysis

The trajectories obtained from standard and reservoir REMD simulations were analyzed using the Amber ptraj module. Trpzip2 simulations were compared to the experimentally determined native structure [59] (Model 1 of PDB code 1LE1) where backbone RMSD's were calculated for residues 2 to 11. Terminal residues were omitted to remove the effects of fluctuations. An RMSD cutoff of 1.7Å was used to determine native structures based on free energy profile along RMSD where the native minimum reached up to 1.7 Å (data not shown). For dPdP the fraction of native contacts were calculated and the native population was calculated using a cutoff of 0.50 for both hairpin1 and hairpin2 contacts (as described in Roe et al. [23]).

Melting curves were generated by calculating the average population of native structures at each temperature. For trpzip2 simulations, data from the first 55000 exchange attempts were discarded for each standard REMD simulation to remove initial structure bias. For dPdP REMD simulations data from the first 20000 exchange attempts were discarded.

Native fractions as a function of time were calculated by averaging the native population up to that time point for both systems using their respective criteria. For all systems the rate of convergence was observed by comparing populations starting from

different initial conformations. When both simulations show similar observables and a flat profile is obtained for all temperatures, the simulations are classified as converged.

Cluster analysis was performed as described previously [43] using the Moil-View program [66]. The trajectories from standard and reservoir REMD simulations were combined. Cluster analysis was performed on the combined set, and then normalized populations for each cluster type were calculated for each of the original simulations. This process permits direct comparison of the populations since the structure families are defined using the combined trajectories.

5.3 Results and Discussion

We apply the R-REMD method to two model systems (trpzip2 and dPdP) that we have studied previously using standard REMD. In order to validate the R-REMD approach, we first compare the resulting structure ensembles to those obtained with standard REMD to validate that R-REMD provides accurate results. Next, we examine whether R-REMD provides these results more efficiently than standard REMD.

5.3.1 Trpzip2 REMD Simulations

We performed 2 independent REMD simulations of the trpzip2 peptide, one starting with all replicas in the published NMR structure (native) and one from a compact non-native structure. Both simulations were run ~155000 exchange attempts (equivalent to 155ns per replica) where 14 replicas were used to cover a temperature range of 260K – 570K.

Even though trpzip2 is a small system, long simulations were required to obtain good agreement between simulations with different initial conformations. Throughout the simulations the melting profiles were monitored and compared. After ~150 ns both REMD simulations showed similar melting profiles which no longer changed with increasing simulation times. The convergence rates of each simulation will be discussed later in this section. Since significant time was required to overcome the bias from initial conformations, data from the first 55000 exchange attempts (55 ns) were discarded for constructing the melting curves (Figure 5-1). It should be noted that the amount discarded is larger than the total simulation time of most published REMD studies. Simulations starting from unfolded conformations show slightly higher stability than those initiated with the native state, suggesting that these differences involve fluctuations in the data and do not reflect initial structure bias. As determined by fitting of the native fractions to the Gibbs – Helmholz equation, both simulations show comparable thermodynamic properties, with melting temperatures of 342.4K and 352.4K and ΔH_m of -15.90 kcal/mol and -16.46 kcal/mol. These values are in excellent agreement with the experimental melting temperature of 345 K and ΔH_m of -16.8 kcal/mol[59]. While the accuracy of the force field is not the subject of this study, it indicates that we are evaluating the performance of the R-REMD method under conditions that are relevant to experimental observations.

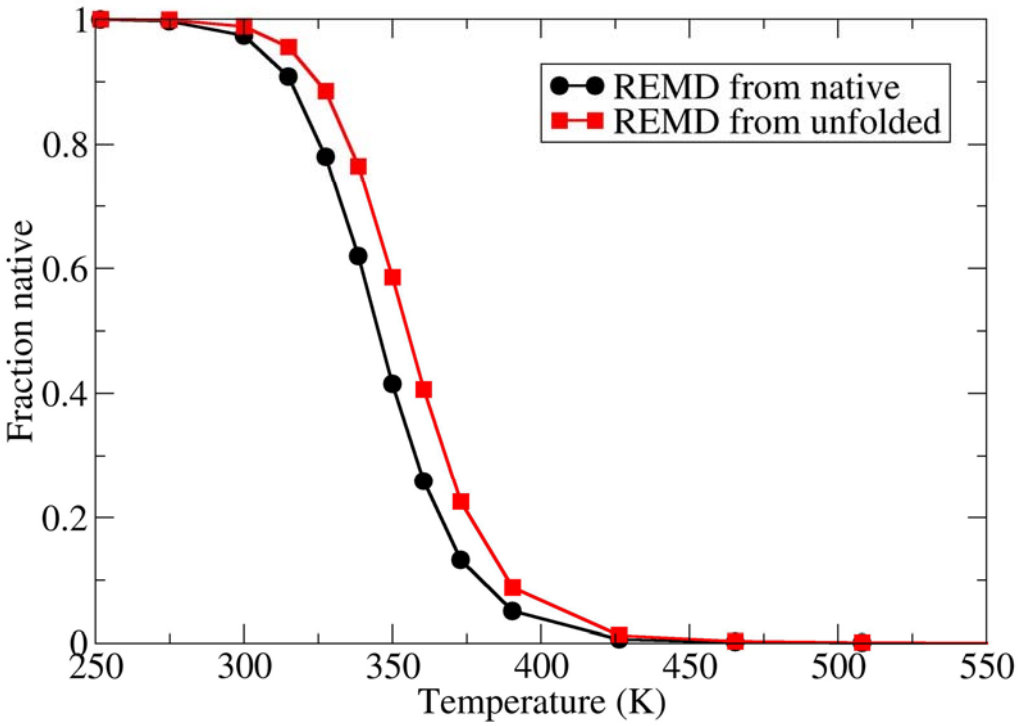


Figure 5-1. Melting curves for trpzip2 REMD simulations starting from native and unfolded conformations. Symbols represent temperatures at which simulation data is obtained. The similar profiles suggests that the data is reasonably well converged. Simulations show melting temperatures of 342.3K and 352.4K, in excellent agreement with the experimentally measured value of 345K.

5.3.2 Testing the accuracy of R-REMD

After establishing benchmark results obtained using converged standard REMD simulations, we generated the high temperature reservoir ensemble at 400K. We chose 400K because it is high enough to allow rapid conformational transitions and it is well above the Tm, thus requiring R-REMD to significantly transform the reservoir ensemble to obtain accurate ensembles at lower temperatures.

4 standard molecular dynamics simulations were performed at 400K using identical conditions as standard REMD simulations. Each simulation was run for ~38 ns with a cumulative simulation time of ~152ns, where multiple folding and unfolding transitions were observed for each trajectory (Figure 5-2). The presence of reversible folding transitions during standard MD is a reasonable indicator that the ensemble is fairly well converged (discussed in more detail below). Due to the elevated temperature, rapid unfolding takes place after each folding event and the native population for each simulation is between 1 and 5 %, in good agreement with the melting curves shown in Figure 5-1 (3.3% and 6.1% native populations at 400K, calculated using the Gibbs – Helmholz equation and native fractions at the other temperatures).

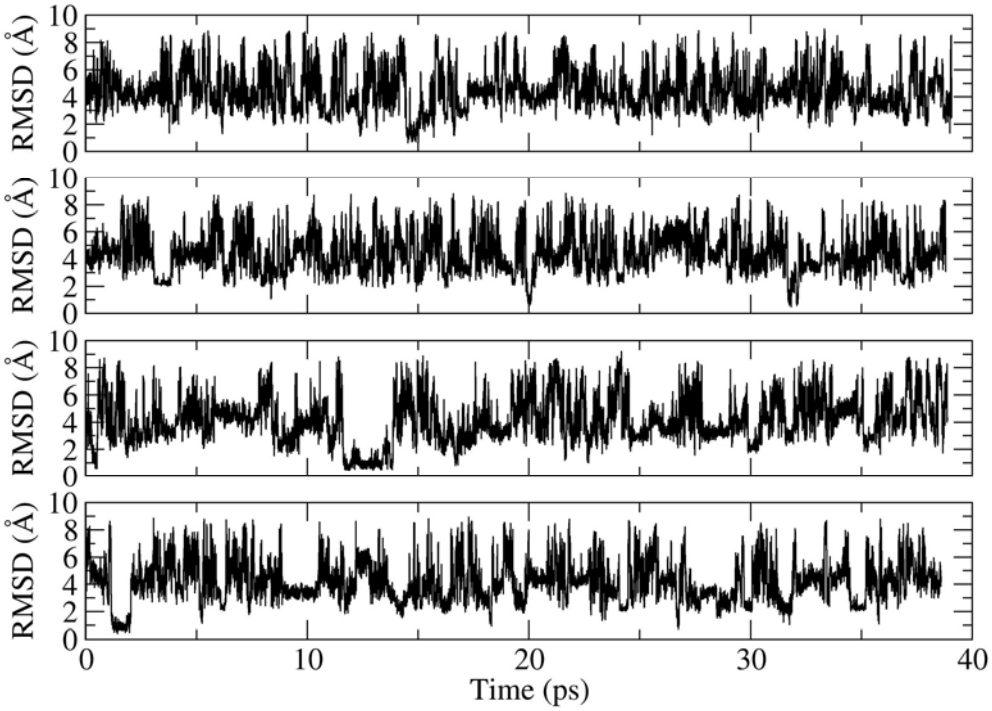


Figure 5-2. Trpzip2 backbone RMSD vs. time during the four simulations at 400K used to generate the R-REMD reservoir. All simulations show reversible folding with a low population of the native β-hairpin.

Monitoring RMSD values with respect to native structure throughout each trajectory can be a good indicator for whether the native conformation is accessible during the simulation, but this provides little information on the convergence of sampling for the unfolded ensemble. Monitoring the number of clusters sampled during a simulation has been suggested for evaluation of simulation convergence [39]. Following our previously published work[43], we extend that approach to evaluate the population of each cluster to determine whether independent simulations provide the same ensembles. Cluster analysis was performed on the combined set of structures to assign structure

families (clusters), and the fraction of the ensemble populating each cluster was then calculated for each simulation. It is important to note assignment of clusters using the combined trajectories permits a direct comparison of the populations sampled by the different simulations. Cluster analysis resulted in 136 structure families. In Figure 5-3, we compare the populations of each family sampled in the first two trajectories to the populations from the other two trajectories. A good correlation is observed, suggesting that the simulations not only sample the same types of structures, but that the relative population of each structure family is similar. While the composition of the unfolded ensemble will be discussed elsewhere, it is important to note that the most populated clusters (10-15% of the ensemble) are non-native at this elevated temperature, with a native population of only ~3%.

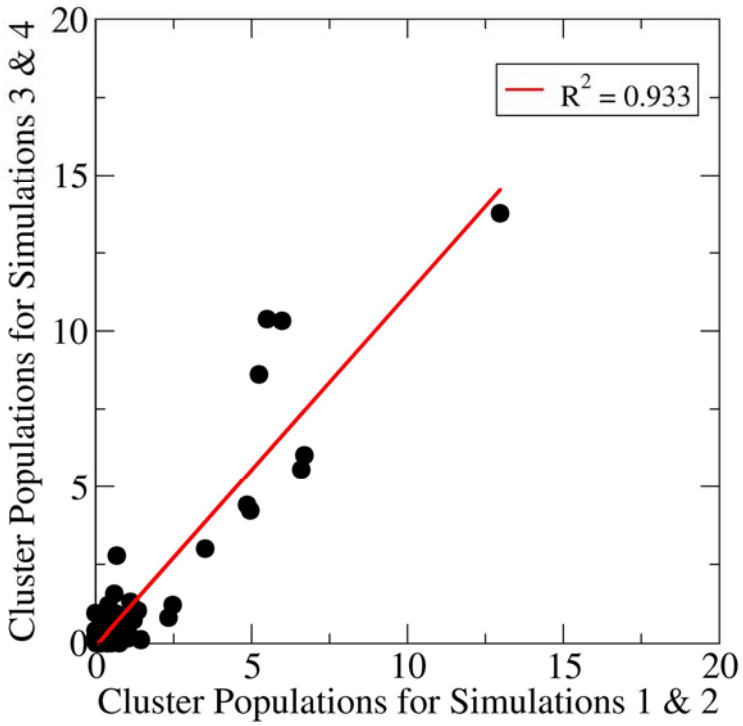


Figure 5-3. Populations of different trpzip2 structure clusters sampled by standard MD simulations. Populations of the first two trajectories are compared to populations of the same clusters in the remaining two trajectories. All clusters with large populations in runs 1&2 are also present with similar populations in runs 3&4, suggesting good convergence.

This pool of 10000 structures (coordinates and velocities) was used as the reservoir set for the R-REMD simulations. Four replicas were used with temperatures 300K, 323K, 350K and 373K, where the 373K replica periodically attempted to exchange with the 400K reservoir as described in Methods. Two sets of R–REMD simulations with different initial structures were run for 50000 exchange attempts.

Figure 5-4 shows the potential energy distributions of each replica, including the reservoir, and the expected Gaussian distributions for the energies are obtained. It should also be pointed out that the lower temperatures show narrower energy distributions where

mostly native and native-like conformations are sampled while the higher temperatures show broader distributions where most of the conformation space becomes thermally accessible. As expected the broadest distribution for both methods is near the transition midpoint of 350K, where both native and non-native conformations are present. During the simulations 25-30% exchange ratios are observed between replicas and a 30% ratio between the 373K replica and the 400K reservoir.

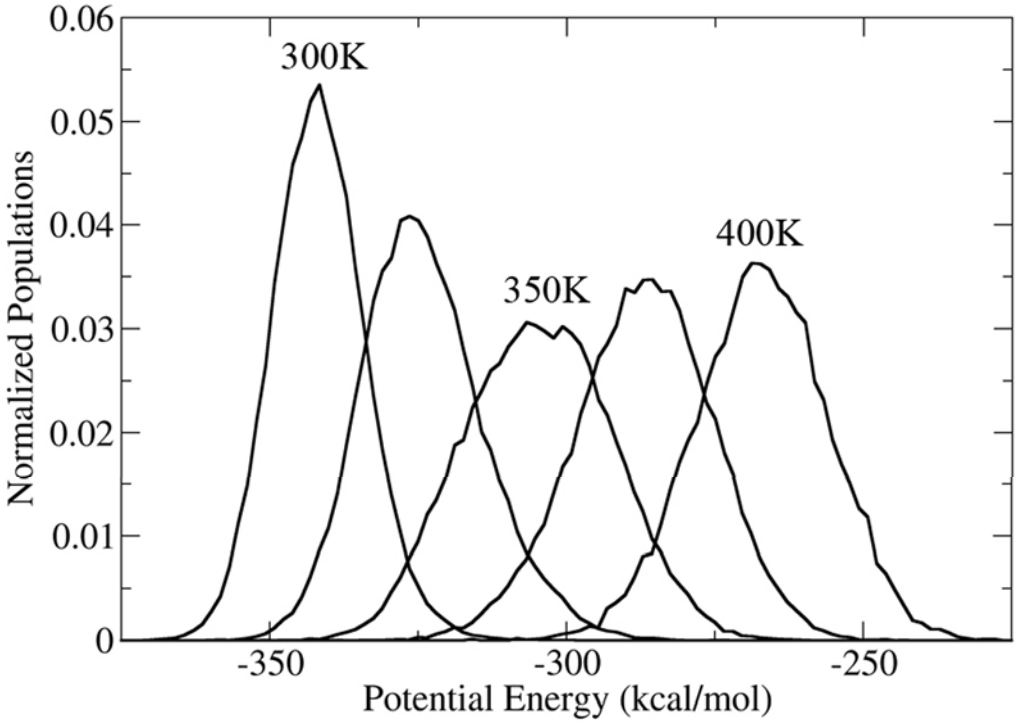


Figure 5-4. Potential energy distributions for the trpzip2 ensembles sampled in R-REMD simulation. As expected good overlaps are observed between neighboring replicas and between the highest temperature replica and the reservoir.

We next evaluated whether the use of the reservoir had any negative impact on the accuracy of the simulations. We calculated the thermal melting profiles for ensembles from the R-REMD simulations using the same procedure as was used for the standard REMD data. In Figure 5-5 we show the comparison of these melting curves to those from standard REMD. Excellent agreement is observed; the melting curves from the two R-REMD simulations lay within the bounds defined by the curves obtained from the two standard REMD simulations. Importantly, the R-REMD ensembles at low temperature are nearly fully native despite the low (3%) native population in the reservoir; thus the REMD replicas are capable of accurately transforming the ensemble in the reservoir to what should be sampled at alternate temperatures. This result also suggests that it is possible to use this method for structure prediction, since the native conformation at low temperature is correctly identified despite the fact that it is not the most populated structure type in the high temperature reservoir (Figure 5-3).

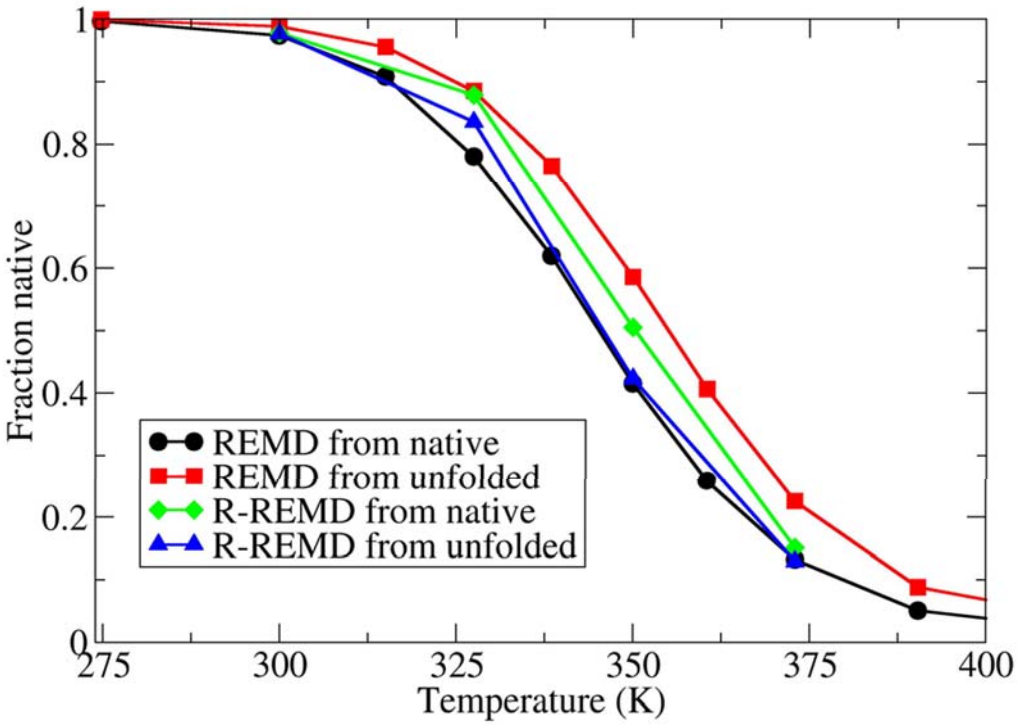


Figure 5-5. Thermal melting profiles for trpzip2 obtained from standard REMD (black and red) and R-REMD simulations (blue and green). Symbols represent temperatures at which simulation data is obtained. Standard REMD simulations are shown in black and red and R-REMD results are shown in green and blue. For easier comparison only temperatures below 400K are shown. Both R-REMD simulations are in good agreement with each other and lie fully within the precision range defined by the standard REMD results.

Figure 5-5 shows a striking agreement between the melting profiles obtained using standard REMD and reservoir REMD simulations. As we noted above, however, analysis of only native populations gives an incomplete view of the composition of an ensemble of structures. To be able to more fully evaluate the ensembles provided by R-REMD one must compare populations not only of the native conformation but for all accessible states. We selected the ensemble at 350K for this analysis; the proximity to the

Tm makes this an excellent temperature to characterize the ensemble under conditions where native and non-native conformations are well populated. We performed cluster analysis on the combined set of structures sampled at 350K in all REMD and R-REMD simulations and then calculated the population for each cluster in the ensemble from each simulation run (two from standard REMD and two from R-REMD with different initial conformations). This analysis resulted in 63 clusters with the native conformation being the highest populated cluster in each simulation (Figure 5-6). We note that using the same clustering method and cutoff fewer clusters were obtained at 350K than the 400K reservoir described above (63 clusters vs. 136 clusters) and the most populated cluster is different at these temperatures (native at 350K and non-native at 400K).

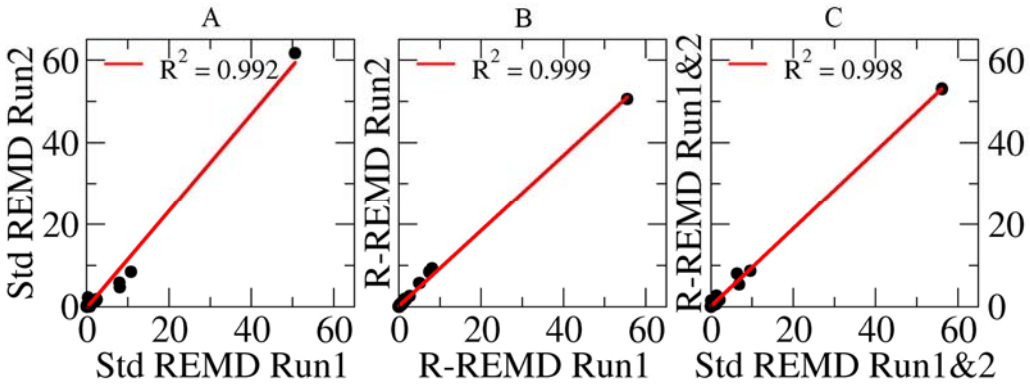


Figure 5-6. Comparison of the populations of a set of trpzip2 structure types sampled in different simulations. Structure families are defined using the combined set of structures, permitting direct comparison of populations between trajectories. (A) comparison of standard REMD from native vs standard REMD from unfolded, (B) comparison of R-REMD from native vs. R-REMD from unfolded (C) comparison of the combined data from standard REMD and the combined data from R-REMD. High correlations were observed in each case (R^2~0.99), and the most populated cluster is the same in all runs. Regression analysis after discarding the most populated cluster results in a similar level of agreement.

Standard REMD simulations starting from different initial conformations show high correlation between cluster populations (R^2>0.99), suggesting that the ensembles are well converged and the data are suitable as a reference to evaluate R-REMD results (Figure 5-6A). Similarly both R-REMD simulations starting from different conformations are in excellent agreement with $R^2 > 0.99$ (Figure 5-6B). Having thus validated the precision of the results from each method, we compare the populations of different structures in the ensemble obtained from standard REMD to that from R-REMD (Figure 5-6C). The agreement between the two data sets is impressive, with $R^2 = 0.998$ and a slope of 0.932.

The regression data obtained using all clusters may be biased by a single cluster with large population (native). We repeated the regression analysis for the data shown in Figure 5-6 after removal of this data point, thus comparing the preference to sample the various weakly populated structures in the unfolded state. For all cases the resulting fit is similar to the original, with correlation coefficients of 0.974 (A), 0.997 (B) and 0.966 (C) between the unfolded ensembles sampled in the REMD and R-REMD simulations. Thus we conclude that the ensemble obtained from R-REMD is essentially indistinguishable from that obtained using standard REMD, including the relative populations of the various conformations that make up the unfolded state.

5.3.3 Testing the efficiency of R-REMD

We have demonstrated that R-REMD can produce the same ensembles of structures as standard REMD, validating the accuracy of the approach. We next investigate whether R-REMD offers any advantage over standard REMD in terms of

computational cost. To analyze the rate of convergence for simulations using each of the two methods, the population of native conformation with respect to simulation time was calculated for each simulation and temperature. By comparing the results from simulations initiated with different structure sets we can observe how long it takes to obtain a particular level of precision in the native populations. We expect that after sufficient time the independent simulations will show similar behavior with population sizes that fluctuate around the same average values.

Figure 5-7A shows the native populations vs. time for several temperatures in the two independent standard REMD simulations. As expected the values undergo very large fluctuations at the beginning of the REMD run and slowly approach their equilibrium values (obtained by combining the two data sets and discarding a significant amount of data to remove bias from initial conditions as described for Figure 5-5). After 155000 exchange attempts (155ns per replica), populations near the melting temperatures still fluctuate and do not show a flat profile with increasing simulation time. It is interesting to note that the simulation initiated with all replicas in the native conformation still underestimates the equilibrium native population. Data near the thermal melting transition (where native and non-native conformations are both sampled) is critically important for characterizing the folding landscape. Even at 100ns per replica, the population values differ significantly from the final values. Importantly, the populations from the two independent simulations provide similar values (i.e. good precision) at times where the population value is dramatically different from the final value, indicating that precise results for the native population are not a reliable indicator of the overall convergence of the data. As an example, if we perform cluster analysis on the ensembles

sampled up to 50000 exchange attempts, the native population in both simulations is similar (55% and 60%). However, the correlation coefficient for the populations of unfolded conformations is only 0.796, showing that even though the largest cluster populations agree with each other the overall sampling is not complete and the unfolded state and folding landscape may be poorly converged.

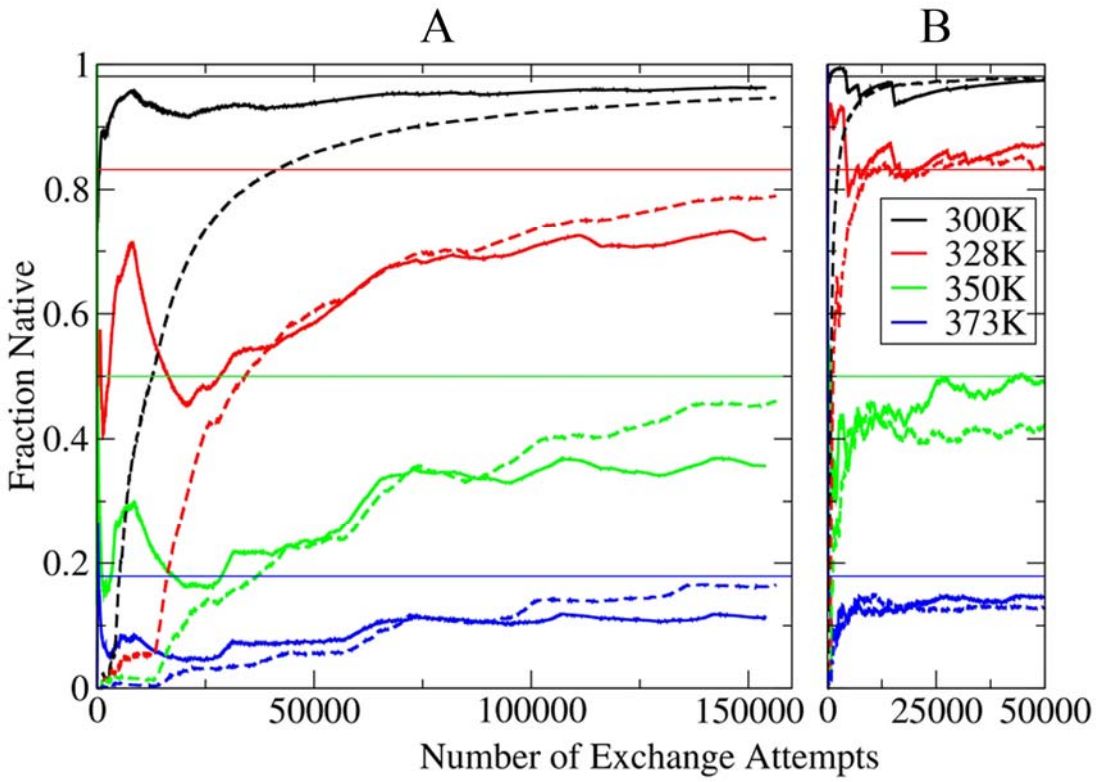


Figure 5-7. Convergence of native population in standard REMD runs (left) and R-REMD runs (right) vs. number of exchange attempts. Solid lines represent simulations starting from native conformation and dashed lines represent simulations starting from unfolded conformations. Thin lines on both graphs represent the average equilibrium values obtained from the standard melting curves (Figure 5-5). For both graphs, the X-axis is on the same scale. For standard REMD (left) the results fluctuate at the beginning of the simulations and slowly converge to their equilibrium values. Even though the simulations were extended to 155000 exchange attempts the average native populations show about 10% deviation between the two runs at multiple temperatures and plateau

values have not been reached. R-REMD simulations (right) converge much faster (~5000 to 10000 exchange attempts).

Next we calculated the native population convergence behavior for the two independent R-REMD simulations. In marked contrast to the slow convergence obtained with standard REMD, both R-REMD simulations reach their equilibrium values after only 10000 exchange attempts and fluctuate around this value for each temperature. As observed in the melting curves, good agreement between the two methods over the temperature range is observed. The results seem to differ about 7-8% at 350K, which is reasonable since the melting transition is sharp around this temperature and this small difference corresponds to only ~0.16 kcal/mol difference in free energy (49.0% vs. 42.3%).

As seen from Figure 5-7, the R-REMD simulations converge to their equilibrium values much faster than standard REMD simulations. Standard REMD simulations have not reached their equilibrium values even at 150,000 exchange attempts (150ns per replica). In contrast, R-REMD simulations reach their equilibrium values in ~ 5000 to 10000 exchange attempts and remain near these values throughout the remainder of the simulations. This represents an improvement of over an order of magnitude in efficiency with R-REMD as compared to standard REMD.

Up to this point the R-REMD simulations were compared to standard REMD simulations that employed a much larger temperature range (up to 570K). As shown in Figure 5-5 and Figure 5-6, these differences had little effect on the converged ensembles. For examination of computational efficiency, however, a more direct comparison between standard and R-REMD would involve using the same number of replicas and

temperatures for each method. To test this, a new REMD simulation was prepared starting from the same unfolded conformation used for R-REMD but with 5 replicas: 4 matching the temperatures used in the R-REMD run and an additional replica at 400K. The only difference between this REMD run and the previous R-REMD run is that the 400K trajectory is a continuous simulation with exchanges that are synchronized with the other replicas instead of being chosen randomly from the pre-generated 400K structure reservoir used for R-REMD. The native populations vs. time for this REMD simulation are shown in Figure 5-8.

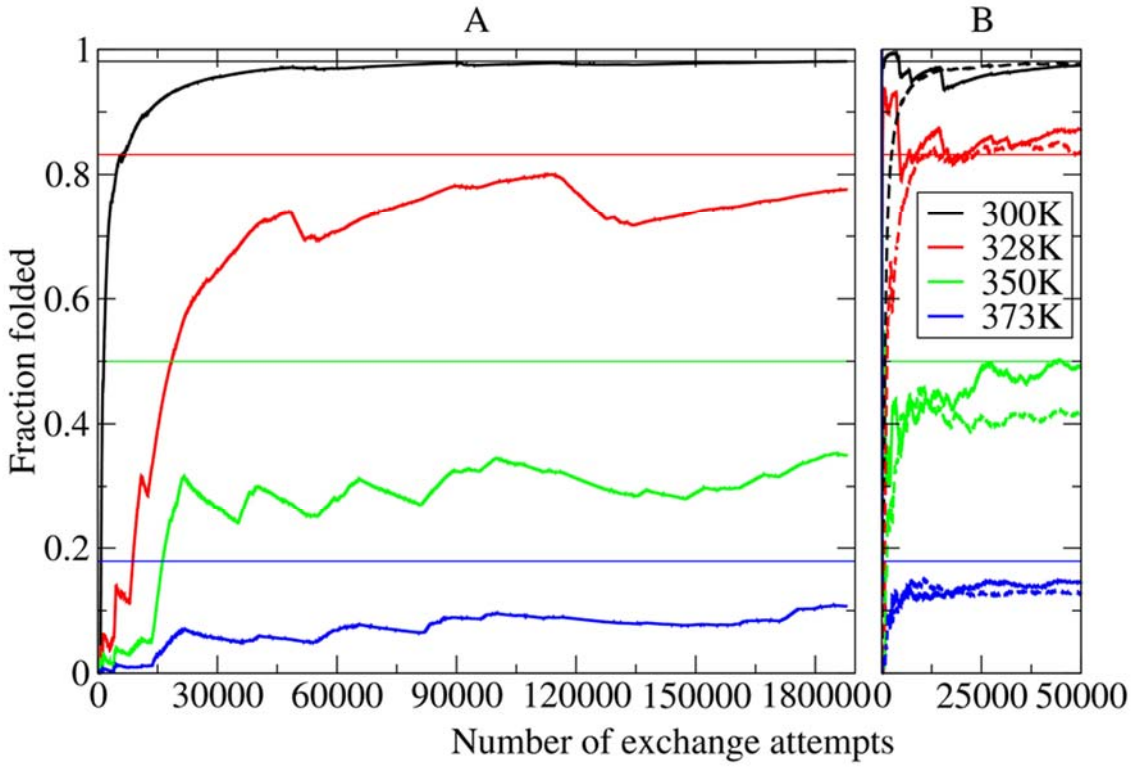


Figure 5-8. Native population at different temperatures vs. number of exchange attempts for (A) standard REMD using the same protocol as the R-REMD run (B) but using a 400K MD replica instead of a reservoir. Equilibrium populations from standard REMD with the higher temperature range are shown as solid lines. Very slow convergence is observed for standard REMD; even after 180,000 exchange attempts large fluctuations

are present at moderate temperatures. It should be pointed out that this convergence graphs has different scale and the x-axis covers much longer timescale than previous plots.

These standard REMD simulations with a highest temperature of 400K converge much more slowly than the original standard REMD which used more replicas covering a wider temperature range. After 180000 exchange attempts (180 ns per replica) the replicas still did not reach the equilibrium values determined from standard REMD runs, and they also show relatively little progress towards these values. This slow convergence is somewhat unexpected since this REMD simulation was run longer than the cumulative simulation time of our standard MD simulations at 400K (180ns per replica vs. 152ns of standard MD), and these standard simulations were shown to be reasonably well converged (Figure 5-2 and Figure 5-3). We believe that this difference in convergence between high temperature MD and REMD demonstrates the effect of "scavenging" of low-energy structures sampled at the highest T by the lower temperatures, slowing the convergence of the high T REMD ensemble. This interpretation is consistent with the observation that the lowest temperature converges within ~50ns to nearly fully native ensemble; once this structure is located at higher T and exchanged to the lowest T, it becomes trapped and no further exchanges take place (and will not until other low-energy basins are located at higher temperatures). Thus the rapid convergence of this low temperature is not an adequate measure for simulation convergence. As discussed above, temperatures such as 328K where the native state does not fully dominate the ensemble are likely to be much more useful in characterizing the folding landscape and composition of the unfolded state. The poor convergence of standard REMD at these temperatures and the rapid convergence of R-REMD under otherwise identical conditions

confirm that using an equilibrated structure reservoir instead of a synchronous high temperature replica significantly increases the rate of convergence of REMD simulations.

To summarize the efficiency comparisons described above, standard REMD simulations with 14 replicas were run for 155 ns per replica from two initial conformations resulting in a cumulative simulation time of ~ 4.3 μs simulation time and still did not fully converge. The R-REMD simulations were run using 4 replicas and two initial conformations and both runs reach their equilibrium values in under 10 ns per replica (40 ns total). Generation of the reservoir does require additional computational effort that must be included in the comparison. In the present case, four simulations of ~40 ns were employed (152 ns, almost as long as REMD simulations but only at 1 temperature). The ability to use multiple simulations provides the reservoir generation with parallel efficiency comparable to the REMD simulations. Thus the cumulative simulation time for R-REMD including the reservoir generation is about 232 ns, approximately 19 times more efficient than the less well converged 4.3 μsec standard REMD simulation. Comparison to the standard REMD that used 5 replicas is difficult since they remained poorly converged even when extended to 180ns per replica (0.9 μs total simulation time). Thus R-REMD is more than 4 times more efficient even when the same replicas are used.

One remaining question with the R-REMD simulations is how much the convergence rate and/or final results depend on the composition of the reservoir set. We tested this dependence by repeating the R-REMD run from an initial unfolded ensemble, but using only the first half of the original structure reservoir (corresponding to two of the four MD trajectories at 400K). The resulting pool of 5000 structures had a native

population of ~ 1.5%.The resulting melting curve is shown in Figure 5-9**,** along with those obtained from standard REMD and R-REMD with the larger reservoir. The thermal stability of trpzip2 in the R-REMD run with the smaller pool is somewhat lower, with ~15 K reduction in the midpoint of the melting transition. This likely arises from a lower population of native conformations in the smaller reservoir (~1.5 % vs. 3 %). Even with this much smaller native population in the reservoir, the R-REMD run shows good agreement at the lowest temperatures away from the reservoir and the native population at higher temperatures is reduced accordingly. The simulations still converge as fast as the R-REMD simulations using the full structure pool (data not shown), suggesting that repeating R-REMD simulations with independent reservoirs would be an excellent approach to validating data convergence.

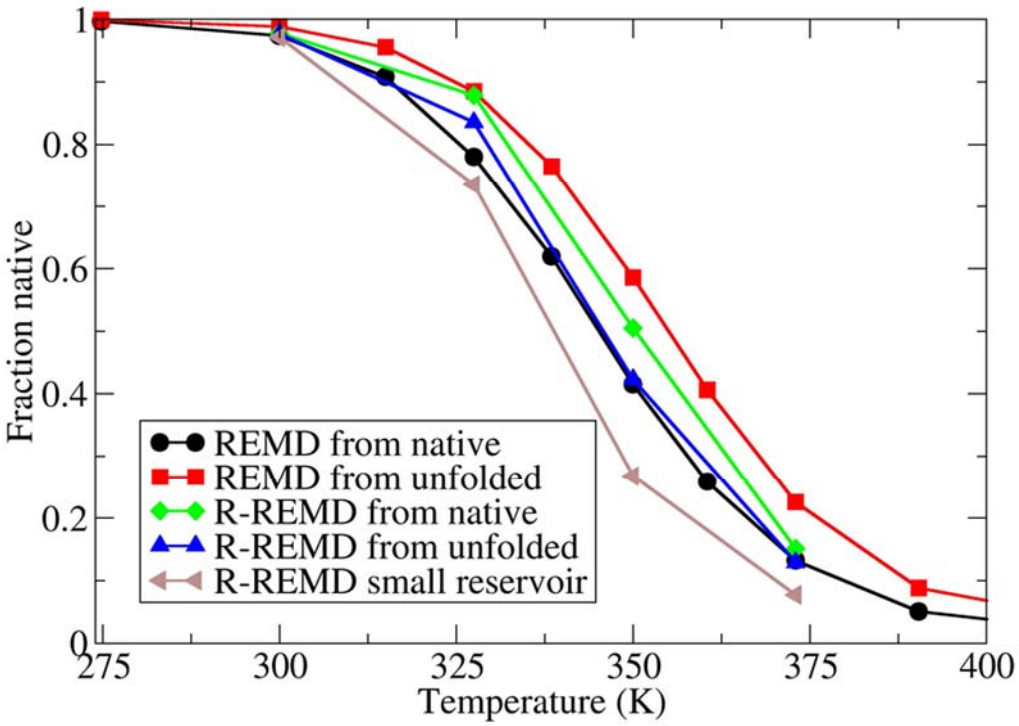


Figure 5-9. Melting curves of standard REMD, R-REMD and R-REMD with half of the reservoir simulations. Using only the first half of the reservoir, the peptide is less stable as indicated by ~15 K reduction in the melting temperature.

5.3.4 Testing R-REMD performance with and anti-parallel β-sheet

To test the efficiency of the R-REMD method on a different and more challenging system we simulated the peptide dPdP, which has been shown to adopt a 3-stranded anti-parallel β-sheet [165]. We previously reported results from independent standard REMD simulations starting from fully extended and from compact initial conformations [23]. Here we compare those results to data from new simulations performed using R-REMD, starting from a fully extended conformation.

We employed a single long MD simulation of dPdP to generate the structure reservoir (260ns, with 5 folding transitions observed). The reservoir was again generated at 400K, and the native content in the resulting ensemble was 7.7%, in reasonable agreement with data at 399K in our standard REMD simulations (4.5% and 4.7% in the independent REMD runs). Once again 10000 structures were selected at equal intervals for use as the structure reservoir. Since dPdP is a larger system than trpzip2, 6 replicas were used with the same temperature distribution as we employed in the standard REMD simulations resulting in 15-20% exchange ratios between replicas and 14% between highest replica and the 400K reservoir. No data were discarded, since within 28 exchange attempts every one of the initial fully extended conformations had been exchanged with the reservoir, as expected since the fully extended conformation is energetically less favorable than the MD-generated conformations in the reservoir.

We compare the dPdP melting curves from our standard REMD simulations with the R-REMD results in Figure 5-10. As we observed with trpzip2, good agreement is obtained between REMD and R-REMD, with the R-REMD melting profile falling within the precision bounds obtained from the two independent standard REMD runs. As we reported previously [23], these values are in good agreement with experimental observations.

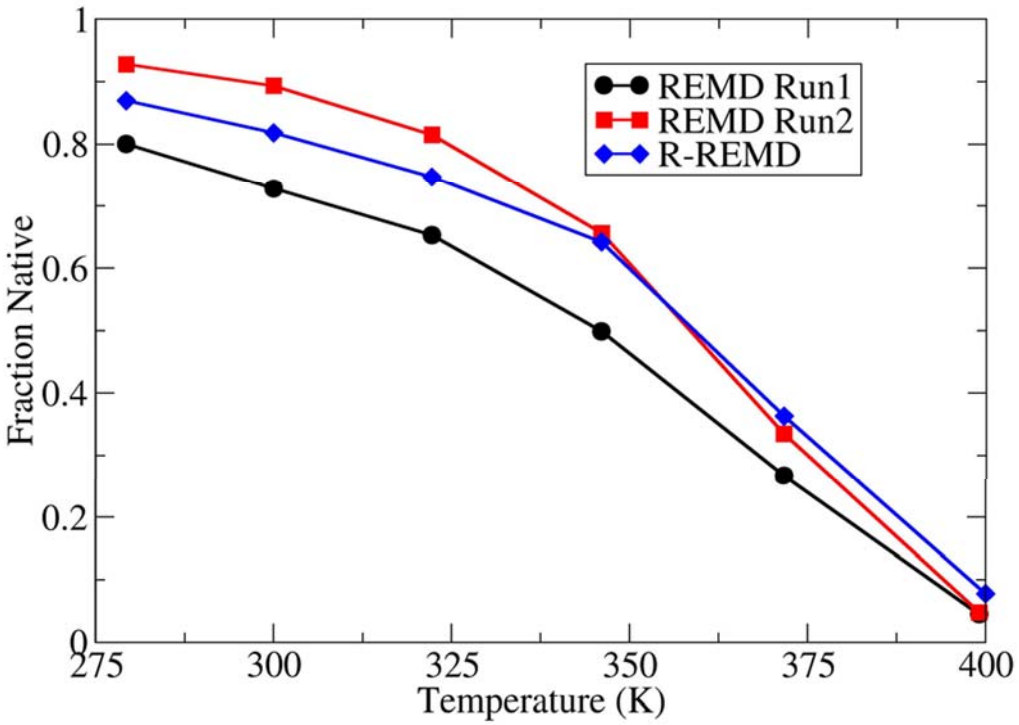


Figure 5-10. Comparison of dPdP melting curves from standard REMD simulations (black and red) and R-REMD simulation (blue). For standard REMD simulations, data from the first 20000 exchange attempts were discarded to remove bias introduced by initial conformations. For the R-REMD simulation the 400K population reflects the reservoir ensemble.

Having confirmed that R-REMD is once again able to accurately reproduce the thermal melting profiles obtained using standard REMD, we evaluated how long it took each simulation to reach these equilibrium values (Figure 5-11). Even after 170000 exchange attempts (~2μs per run) for standard REMD simulations it was not possible to conclude that the simulations were well converged since the populations at some temperatures varied more than 10% and in many cases a plateau had not yet been reached.

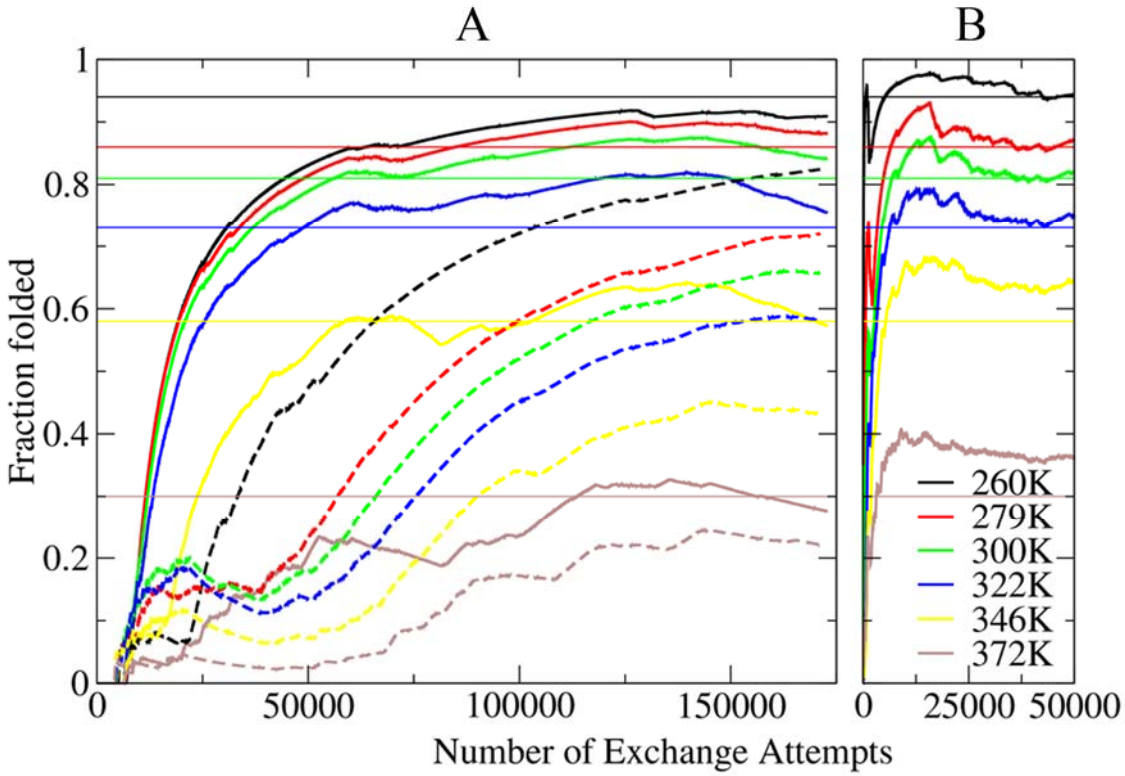


Figure 5-11. Native fraction vs. number of exchange attempts for standard REMD simulation (A) and R-REMD simulation (B). Solid lines in (A) represent simulations starting from compact non-native structure and dashed lines represent simulations starting from extended conformation. Even after 170,000 exchange attempts plateau values have not been reached. During R-REMD simulations (B) all replicas converge to their equilibrium values after ~10000 exchange attempt and show a flat profile thereafter.

In contrast with the standard REMD results, dPdP R-REMD simulations reach their equilibrium values within ~ 10000 exchange attempts and show an essentially flat profile after that point. Simulations were continued up to 50000 exchange attempts, with no significant changes for any of the simulated temperatures. Including reservoir generation (260 ns), the total simulation time used to obtain fully converged ensembles using R-REMD was ~320 ns, although we did not test whether a shorter reservoir

generation simulation would have been sufficient. Based on this time, we estimate that R-REMD is at least 6.4 times more efficient than standard REMD.

5.4 Conclusions

We introduced a new variant of the replica exchange method where slow convergence and high computational cost of REMD have been greatly improved by coupling of the REMD replicas to an ensemble of conformations that is generated in advance, similar in spirit to J-walking schemes. This approach builds on the hypothesis that the main contribution to sampling efficiency during REMD is obtained from the replicas exploring the free energy landscape at high temperatures. Rather than simulating all replicas during this search process, R-REMD performs the search for alternate local minima in advance and subsequently uses a relatively short REMD run to generate accurate Boltzmann-weighted ensembles at other temperatures. An important advantage is that exchanges with the reservoir need not be time-correlated with the replica simulations, permitting REMD replicas to obtain many low-energy (such as native) conformations from a smaller number of folding events; this is not possible with standard REMD, which may a contributing factor in slow convergence.

We tested R-REMD by comparing to standard REMD results for two systems, a β-hairpin and a three-stranded β-sheet, under conditions in which the standard REMD data were in good agreement with experimental observations. We find that the thermal melting profiles obtained from R-REMD simulations were highly accurate as compared to standard REMD, as expected due the lack of any approximations in development of the method. Furthermore, excellent agreement was noted between the compositions of the

structure ensembles obtained from standard REMD and R-REMD, including very high correlations between the two methods for the populations of native and non-native conformational families.

Analysis of convergence rates suggests that R-REMD 5 to 20 times more efficient than standard REMD and is limited mostly by the quality of the initial high-temperature ensemble used as the reservoir pool of structures during the R-REMD run. This reservoir can readily be generated through multiple independent MD simulations. We demonstrated that this high-temperature reservoir actually converged more rapidly than the corresponding temperature during REMD; the slow convergence of high temperature data during REMD likely arises from removal of low-energy conformations through exchange with lower temperatures. A key advantage is that replicas corresponding to the full temperature range needed for REMD do not need to be simulated during the reservoir generation; they are only simulated during the REMD phase which converged much more rapidly than analogous calculations performed without the reservoir.

Since the populations of alternate local minima in the reservoir formally influence the equilibrium properties at other temperatures, it is important to ensure that the reservoir is well converged. With the current implementation not only each minima have to be sampled, the relative populations for each conformation should be correct as well. However, it should be possible to modify the form of the exchange probability calculation to accommodate other well-defined probability distributions in the reservoir and still obtain correct canonical ensembles for the temperatures spanned by the REMD replicas. This may further reduce the computational requirements of the method; future studies will investigate this possibility.

Chapter 6

Future Plans

6.1 Decoy Analysis

Chapter 2 introduces the decoy screening technique for small peptides and its uses for identifying potential problems with parameter sets and improving available force fields. We have increased the number of decoys mentioned in Chapter 2. Currently we have decoy sets of four different peptides and proteins. As mentioned in Chapter 2 we still use the Trpzip2 decoy structures. In addition to Trpzip2 we have a helical decoy set generated through Replica Exchange simulations from a helical structure generated from Baldwin Helix [166]. We also use systems that show more than one secondary structure element as decoy structures such as Trp-cage miniprotein [167] and villin headpiece helical subdomain (HP36).

As mentioned in Chapter 2, there were problems with the force field parameters in AMBER. During that study we used the decoy structures to train the dihedral parameters of the force field to obtain a better force field. The resulting force field (ffGA) was very successful for Trpzip2 simulations discussed in Chapter 3. However later tests showed that the ffGA parameter set had strong bias towards β-turn conformation which turned other test peptides into left handed helices (see minimum around +60° for φ in Figure 2-9). It became apparent that more work was needed to obtain a better force field.

We took a different approach where we parameterized the entire backbone parameters by comparing energies to quantum mechanical energies for GLY and ALA

tetra-peptides. The resulting parameter set is called ff99SB [24]. While we worked in improving parameters other groups came with variations of existing force fields to reproduce experimental observations. We have used the decoy screening procedure to test the accuracy of the available and our new parameter sets (ff94 [11], ff99 [17], ff03 [21], ff94gs [18], ff99ϕ [19, 20]). Figure 6-1 shows how decoy analysis is performed now, where minimum energy profiles are compared for each parameter set tested on the same graph.

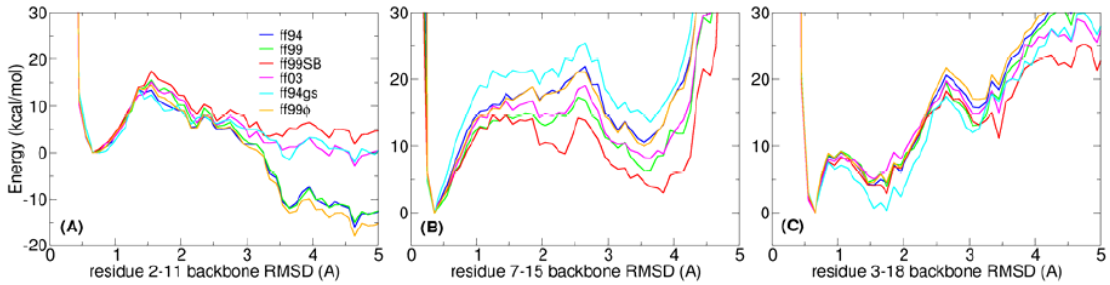


Figure 6-1 Lowest energy profiles for three decoy systems, each tested with six Amber force fields (ff94, ff99, ff99SB, ff03, ff94gs, ff99ϕ). (A) Trpzip2, (B) Baldwin Helix, (C) Trp-cage. RMSD values are calculated with respect to the experimentally determined structure. Ideally, a force field should show lowest energies for the lowest RMSD values.

Decoy testing is a quick way of identifying potential problems in force fields or other simulation parameters. It can be used to test other parameters such as different GB models when needed. However it has limitations. If a parameter set "passes" the decoy set it does not mean that it is a good set. There may still be problems with it and after successful decoy screening extensive simulations should be run to confirm the performance of the parameter set. If it fails the decoy screening it can be concluded that that parameter set has certain deficiencies and should be improved.

6.2 β-hairpin folding

We have investigated the folding and unfolding thermodynamics and kinetics of the Trpzip2 β-hairpin in Chapter 3. We used the force field parameters developed in Chapter 2 (ffGA) to perform this study. As mentioned before the ffGA force field was able to produce results in close agreement with experimental observations for Trpzip2. However it turned out that this force field is not transferable to other systems. It was only suitable for β-systems with similar turn types. The same force field was used successfully on the dPdP peptide which adopts a three stranded β-sheet [23].

We are in process of reproducing ffGA results with our new force field ff99SB. However discovering problems with the Generalized Born solvent model forced us to run our simulations in explicit solvation. This combined with slow transition kinetics of β-strand conformation results in unconverged simulations even with enhanced sampling methods such as Replica Exchange. REMD simulations started from unfolded conformations were able to identify native conformations and exchange them to lower temperatures but even after 350,000 exchange attempts (350ns/replics, 50 replicas) only a handful of folding events were observed and the calculated melting temperature is very low compared to experimental observation.

This low convergence suggests that even with REMD it is not possible to obtain reliable data for systems having slow transition kinetics and improvements to the existing methods are needed. We have two proposed methods to improve the sampling efficiency of standard REMD method (Chapter 4 and Chapter 5) where in Chapter 5 we used Trpzip2 as model system and compared the performance of the new Reservoir REMD method to the results discussed in Chapter 2.

6.3 Hybrid Solvent REMD

In Chapter 4 we discussed a new variant of REMD method where a hybrid solvation scheme was used to reduce the number of replicas for explicit solvent REMD. The method is currently tested on Alanine peptides of different sizes (1, 3 and 10 residues) and it is still in development stage. Alanines are simple residues and ployalanines lack complex sidechain – sidechain and sidechain – solvent interactions. The use of hybrid solvent exchange scheme eliminates the backbone effects of GB solvent model and produces similar results to standard explicit solvent REMD.

However other problems such as overstabilized salt bridges have been reported for GB models. To test the effect of the hybrid exchange potential on salt bridging residues we ran simulations on a test peptide of four residues where oppositely charged Arg and Glu residues were separated by two Alanines. Same procedure as polyalanine was used where REMD simulations were performed with TIP3P explicit solvent, Generalized Born and hybrid explicit/implicit solvent models. The salt bridge PMF was calculated using the distance between charged groups of Arg and Glu.

The helical backbone preference of GB was noticed on this small peptide as well. When similar backbone conformations were compared the salt bridge was significantly stronger in GB compared to TIP3P. Hybrid potential produced a similar free energy profile to fully solvated simulations suggesting that the inclusion of the first hydration shell is enough to overcome deficiencies of Generalized Born Solvent model and obtain fully solvated trajectories at reduced cost.

Even though hybrid explicit/implicit solvent REMD method is successful on test systems, more simulations with peptides and small proteins such as Trpzip or Trp-cage where experimental data is available should be performed for better accuracy tests. Also this method can be employed systems where we could not produce converged explicit solvent before such as Trpzip2.

6.4 Reservoir REMD

In Chapter 5 we have shown that the convergence speed of REMD simulations can be improved by coupling the high temperature replica with a pre-generated structure reservoir. In standard REMD, when a low energy structure is discovered by high temperature replicas it is exchanged to lower temperatures and high temperatures have to the search over again. For large systems with slow transition kinetics running converged REMD simulations may become computationally very expensive. Through Reservoir REMD (R-REMD) we showed that we can obtain converged results for slow converging β-hairpin and β-sheet using fraction of resources.

However the performance of R-REMD depends on the quality of reservoir. With the implementation described in Chapter 5 the reservoir has to have good sampling and the conformations have to have correct weights. Obtaining such a sampling even at elevated temperatures can be difficult for slow converging systems. We are testing a modified Reservoir REMD scheme where conformations between the reservoir and highest temperature replica are compared using a different non-Boltzmann exchange potential. Through this approach the reservoir structures do not need to have correct relative weights between conformations.

We have tested the reservoir REMD scheme on β-hairpin and β-sheet forming peptides using Generalized Born solvent model. However obtaining converged results using GB is possible using standard REMD method. Generating converged ensembles with explicit solvent REMD is difficult especially for systems with slow transition kinetics. The main reason for developing the Reservoir REMD scheme was to speed up the convergence speed of simulations employing explicit solvation. This method would enable us detailed thermodynamics analysis of current systems in explicit solvent where converged simulations were not possible through standard methods.

Bibliography

1. Dobson, C.M., *Protein-misfolding diseases: Getting out of shape.* Nature, 2002. **418**(6899): p. 729-730.

2. Campbell, I.D., *The march of structural biology.* Nature Reviews Molecular Cell Biology, 2002. **3**(5): p. 377-381.

3. Goldberger, R.F., C.J. Epstein, and C.B. Anfinsen, *Acceleration of Reactivation of Reduced Bovine Pancreatic Ribonuclease by a Microsomal System from Rat Liver.* Journal of Biological Chemistry, 1963. **238**(2): p. 628-&.

4. Simmerling, C., B. Strockbine, and A.E. Roitberg, *All-atom structure prediction and folding simulations of a stable protein.* Journal of the American Chemical Society, 2002. **124**(38): p. 11258-11259.

5. Moult, J., *A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction.* Current Opinion in Structural Biology, 2005. **15**(3): p. 285-289.

6. Hornak, V., A. Okur, R.C. Rizzo, and C. Simmerling, *HIV-1 protease flaps spontaneously close to the correct structure in simulations following manual placement of an inhibitor into the open state.* Journal of the American Chemical Society, 2006. **128**(9): p. 2812-2813.

7. Hornak, V., A. Okur, R.C. Rizzo, and C. Simmerling, *HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(4): p. 915-920.

8. Dill, K.A. and H.S. Chan, *From Levinthal to pathways to funnels.* Nature Structural Biology, 1997. **4**(1): p. 10-19.

9. Levinthal, C., *Are There Pathways for Protein Folding.* Journal De Chimie Physique Et De Physico-Chimie Biologique, 1968. **65**(1): p. 44.

10. Brooks, B.R., R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus, *Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations.* Journal of Computational Chemistry, 1983. **4**(2): p. 187-217.

11. Cornell, W.D., P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman, *A 2Nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules.* Journal of the American Chemical Society, 1995. **117**(19): p. 5179-5197.

12. Van Gunsteren, W.F., *GROMOS, Groningen Molecular Simulation Program Package, University of Groningen, Groningen*. 1987.

13. Jorgensen, W.L. and J. Tiradorives, *The Opls Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin.* Journal of the American Chemical Society, 1988. **110**(6): p. 1657-1666.

14. Mackerell, A.D., *Empirical force fields for biological macromolecules: Overview and issues.* Journal of Computational Chemistry, 2004. **25**(13): p. 1584-1604.

15. MacKerell, A.D., *Empirical Force Fields for Proteins: Current Status and Future Directions*. Annual Reports in Computational Chemistry, ed. D.C. Spellmeyer. Vol. 1. 2005: Elsavier. 91-102.

16. Ponder, J.W. and D.A. Case, *Force fields for protein simulations.* Protein Simulations, 2003. **66**: p. 27-+.

17. Wang, J.M., P. Cieplak, and P.A. Kollman, *How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?* Journal of Computational Chemistry, 2000. **21**(12): p. 1049-1074.

18. Garcia, A.E. and K.Y. Sanbonmatsu, *alpha-Helical stabilization by side chain shielding of backbone hydrogen bonds.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(5): p. 2782-2787.

19. Sorin, E.J. and V.S. Pande, *Exploring the helix-coil transition via all-atom equilibrium ensemble simulations.* Biophys J, 2005. **88**(4): p. 2472-93.

20. Sorin, E.J. and V.S. Pande, *Empirical force-field assessment: The interplay between backbone torsions and noncovalent term scaling.* Journal of Computational Chemistry, 2005. **26**(7): p. 682-690.

21. Duan, Y., C. Wu, S. Chowdhury, M.C. Lee, G.M. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J.M. Wang, and P. Kollman, *A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations.* Journal of Computational Chemistry, 2003. **24**(16): p. 1999-2012.

22. Okur, A., B. Strockbine, V. Hornak, and C. Simmerling, *Using PC clusters to evaluate the transferability of molecular mechanics force fields for proteins.* J Comput Chem, 2003. **24**(1): p. 21-31.

23. Roe, D.R., V. Hornak, and C. Simmerling, *Folding cooperativity in a three-stranded beta-sheet model.* Journal of Molecular Biology, 2005. **352**(2): p. 370-381.

24. Hornak, V., R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, *Comparison of multiple Amber force fields and development of improved protein backbone parameters.* Proteins-Structure Function and Bioinformatics, 2006. **65**(3): p. 712-725.

25. Makarov, V., B.M. Pettitt, and M. Feig, *Solvation and hydration of proteins and mucleic acids: A theoretical view of simulation and experiment.* Accounts of Chemical Research, 2002. **35**(6): p. 376-384.

26. Bagchi, B., *Water dynamics in the hydration layer around proteins and micelles.* Chemical Reviews, 2005. **105**(9): p. 3197-3219.

27. Ewald, P.P., *The calculation of optical and electrostatic grid potential.* Annalen Der Physik, 1921. **64**(3): p. 253-287.

28. Essmann, U., L. Perera, M.L. Berkowitz, T. Darden, H. Lee, and L.G. Pedersen, *A Smooth Particle Mesh Ewald Method.* Journal of Chemical Physics, 1995. **103**(19): p. 8577-8593.

29. Still, W.C., A. Tempczyk, R.C. Hawley, and T. Hendrickson, *Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics.* Journal of the American Chemical Society, 1990. **112**(16): p. 6127-6129.

30. Masunov, A. and T. Lazaridis, *Potentials of mean force between ionizable amino acid side chains in water.* Journal of the American Chemical Society, 2003. **125**(7): p. 1722-1730.

31. Yu, Z.Y., M.P. Jacobson, J. Josovitz, C.S. Rapp, and R.A. Friesner, *First-shell solvation of ion pairs: Correction of systematic errors in implicit solvent models.* Journal of Physical Chemistry B, 2004. **108**(21): p. 6643-6654.

32. Nymeyer, H. and A.E. Garcia, *Simulation of the folding equilibrium of alpha-helical peptides: A comparison of the generalized born approximation with explicit solvent.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(24): p. 13934-13939.

33. Zhou, R.H. and B.J. Berne, *Can a continuum solvent model reproduce the free energy landscape of a beta-hairpin folding in water?* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(20): p. 12777-12782.

34. Pitera, J.W. and W. Swope, *Understanding folding and design: Replica-exchange simulations of "Trp-cage" miniproteins.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(13): p. 7587-7592.

35. Zhou, R.H., *Free energy landscape of protein folding in water: Explicit vs. implicit solvent.* Proteins-Structure Function and Genetics, 2003. **53**(2): p. 148-161.

36. Zhou, R.H., B.J. Berne, and R. Germain, *The free energy landscape for beta hairpin folding in explicit water.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(26): p. 14931-14936.

37. Roitberg, A. and C. Simmerling, *Special issue: Conformational sampling.* Journal of Molecular Graphics & Modelling, 2004. **22**(5): p. 317-317.

38. Tai, K., *Conformational sampling for the impatient.* Biophysical Chemistry, 2004. **107**(3): p. 213-220.

39. Smith, L.J., X. Daura, and W.F. van Gunsteren, *Assessing equilibration and convergence in biomolecular simulations.* Proteins-Structure Function and Genetics, 2002. **48**(3): p. 487-496.

40. Okur, A. and C. Simmerling, *Hybrid Explicit/Implicit Solvation Methods*. Annual Reports in Computational Chemistry, ed. D.C. Spellmeyer. Vol. 2. 2006: Elsevier. 97-109.

41. Hansmann, U.H.E., *Parallel tempering algorithm for conformational studies of biological molecules.* Chemical Physics Letters, 1997. **281**(1-3): p. 140-150.

42. Sugita, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding.* Chemical Physics Letters, 1999. **314**(1-2): p. 141-151.

43. Okur, A., L. Wickstrom, M. Layten, R. Geney, K. Song, V. Hornak, and C. Simmerling, *Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model.* Journal of Chemical Theory and Computation, 2006. **2**(2): p. 420-433.

44. Okur, A., D.R. Roe, G. Cui, V. Hornak, and C. Simmerling, *Improved Convergence of Replica Exchange Simulations through Coupling to a High Tepmerature Structure Reservoir.* Journal of Chemical Theory and Computation, In Press.

45. Halgren, T.A., *Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94.* Journal of Computational Chemistry, 1996. **17**(5-6): p. 490-519.

46. Kaminski, G.A., R.A. Friesner, J. Tirado-Rives, and W.L. Jorgensen, *Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides.* Journal of Physical Chemistry B, 2001. **105**(28): p. 6474-6487.

47. MacKerell, A.D., D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* Journal of Physical Chemistry B, 1998. **102**(18): p. 3586-3616.

48. Garcia, A.E. and K.Y. Sanbonmatsu, *Exploring the energy landscape of a beta hairpin in explicit solvent.* Proteins-Structure Function and Genetics, 2001. **42**(3): p. 345-354.

49. Holm, L. and C. Sander, *Evaluation of Protein Models by Atomic Solvation Preference.* Journal of Molecular Biology, 1992. **225**(1): p. 93-105.

50. Novotny, J., R. Bruccoleri, and M. Karplus, *An Analysis of Incorrectly Folded Protein Models - Implications for Structure Predictions.* Journal of Molecular Biology, 1984. **177**(4): p. 787-818.

51. Park, B. and M. Levitt, *Energy functions that discriminate X-ray and near-native folds from well-constructed decoys.* Journal of Molecular Biology, 1996. **258**(2): p. 367-392.

52. DeBolt, S.E. and J. Skolnick, *Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: Atomic burial position and pairwise non-bonded interactions.* Protein Engineering, 1996. **9**(8): p. 637-655.

53. Godzik, A., A. Kolinski, and J. Skolnick, *De-Novo and Inverse Folding Predictions of Protein-Structure and Dynamics.* Journal of Computer-Aided Molecular Design, 1993. **7**(4): p. 397-438.

54. Liwo, A., P. Arlukowicz, C. Czaplewski, S. Oldziej, J. Pillardy, and H.A. Scheraga, *A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: Application to the UNRES force field.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(4): p. 1937-1942.

55. Maiorov, V.N. and G.M. Crippen, *Contact Potential That Recognizes the Correct Folding of Globular-Proteins.* Journal of Molecular Biology, 1992. **227**(3): p. 876-888.

56. Meller, J. and R. Elber, *Linear programming optimization and a double statistical filter for protein threading protocols.* Proteins-Structure Function and Genetics, 2001. **45**(3): p. 241-261.

57. Simons, K.T., I. Ruczinski, C. Kooperberg, B.A. Fox, C. Bystroff, and D. Baker, *Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins.* Proteins-Structure Function and Genetics, 1999. **34**(1): p. 82-95.

58. Dominy, B.N. and C.L. Brooks, *Identifying native-like protein structures using physics-based potentials.* Journal of Computational Chemistry, 2002. **23**(1): p. 147-160.

59. Cochran, A.G., N.J. Skelton, and M.A. Starovasnik, *Tryptophan zippers: Stable, monomeric beta-hairpins.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(10): p. 5578-5583.

60. Demarest, S.J., R. Fairman, and D.P. Raleigh, *Peptide models of local and long-range interactions in the molten globule state of human alpha-lactalbumin.* Journal of Molecular Biology, 1998. **283**(1): p. 279-291.

61. Demarest, S.J., Y.X. Hua, and D.P. Raleigh, *Local interactions drive the formation of nonnative structure in the denatured state of human alpha-lactalbumin: A high resolution structural characterization of a peptide model in aqueous solution.* Biochemistry, 1999. **38**(22): p. 7380-7387.

62. Case, D.A., T.E. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods, *The Amber biomolecular simulation programs.* Journal of Computational Chemistry, 2005. **26**(16): p. 1668-1688.

63. Ryckaert, J.P., G. Ciccotti, and H.J.C. Berendsen, *Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes.* Journal of Computational Physics, 1977. **23**(3): p. 327-341.

64. Jorgensen, W.L., J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein, *Comparison of Simple Potential Functions for Simulating Liquid Water.* Journal of Chemical Physics, 1983. **79**(2): p. 926-935.

65. Schlitter, J., M. Engels, P. Kruger, E. Jacoby, and A. Wollmer, *Targeted Molecular-Dynamics Simulation of Conformational Change - Application to the T[--]R Transition in Insulin.* Molecular Simulation, 1993. **10**(2-6): p. 291-&.

66. Simmerling, C., R. Elber, and J. Zhang, *MOIL-View - A Program for Visualization of Structure and Dynamics of Biomolecules and STO - A Program for Computing Stochastic Paths*, in *Modelling of Biomolecular Structures and Mechanisms*, A. Pullman et al., Editor. 1995, Kluwer Academic Publishers: Netherlands. p. 241-265.

67. Skelton, N.J., *Personal Communication.*

68. Bonvin, A.M.J.J. and W.F. van Gunsteren, *beta-Hairpin stability and folding: Molecular dynamics studies of the first beta-hairpin of tendamistat.* Journal of Molecular Biology, 2000. **296**(1): p. 255-268.

69. Dyer, R.B., S.J. Maness, E.S. Peterson, S. Franzen, R.M. Fesinmeyer, and N.H. Andersen, *The mechanism of beta-hairpin formation.* Biochemistry, 2004. **43**(36): p. 11560-11566.

70. Klimov, D.K. and D. Thirumalai, *Mechanisms and kinetics of beta-hairpin formation.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(6): p. 2544-2549.

71. Kolinski, A., B. Ilkowski, and J. Skolnick, *Dynamics and thermodynamics of beta-hairpin assembly: Insights from various simulation techniques.* Biophysical Journal, 1999. **77**(6): p. 2942-2952.

72. Lee, J. and S.M. Shin, *Understanding beta-hairpin formation by molecular dynamics simulations of unfolding.* Biophysical Journal, 2001. **81**(5): p. 2507-2516.

73. Ma, B.Y. and R. Nussinov, *Molecular dynamics simulations of a beta-hairpin fragment of protein G: Balance between side-chain and backbone forces.* Journal of Molecular Biology, 2000. **296**(4): p. 1091-1104.

74. Munoz, V., P.A. Thompson, J. Hofrichter, and W.A. Eaton, *Folding dynamics and mechanism of beta-hairpin formation.* Nature, 1997. **390**(6656): p. 196-199.

75. Prevost, M. and I. Ortmans, *Refolding simulations of an isolated fragment of barnase into a native-like beta hairpin: Evidence for compactness and hydrogen bonding as concurrent stabilizing factors.* Proteins-Structure Function and Genetics, 1997. **29**(2): p. 212-227.

76. Sheinerman, F.B. and C.L. Brooks, *A molecular dynamics simulation study of segment B1 of protein G.* Proteins-Structure Function and Genetics, 1997. **29**(2): p. 193-202.

77. Sheinerman, F.B. and C.L. Brooks, *Calculations on folding of segment B1 of streptococcal protein G.* Journal of Molecular Biology, 1998. **278**(2): p. 439-456.

78. Sung, S.S., *Monte Carlo simulations of beta-hairpin folding at constant temperature.* Biophysical Journal, 1999. **76**(1): p. 164-175.

79. Zagrovic, B., E.J. Sorin, and V. Pande, *beta-hairpin folding simulations in atomistic detail using an implicit solvent model.* Journal of Molecular Biology, 2001. **313**(1): p. 151-169.

80. Cochran, A.G., N.J. Skelton, and M.A. Starovasnik, *Tryptophan zippers: Stable, monomeric beta-hairpins (vol 98, pg 5578, 2001).* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(13): p. 9081-9081.

81. Snow, C.D., L.L. Qiu, D.G. Du, F. Gai, S.J. Hagen, and V.S. Pande, *Trp zipper folding kinetics by molecular dynamics and temperature-jump spectroscopy.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(12): p. 4077-4082.

82. Yang, W.Y., J.W. Pitera, W.C. Swope, and M. Gruebele, *Heterogeneous folding of the trpzip hairpin: Full atom simulation and experiment.* Journal of Molecular Biology, 2004. **336**(1): p. 241-251.

83. Yang, W.Y. and M. Gruebele, *Detection-dependent kinetics as a probe of folding landscape microstructure.* Journal of the American Chemical Society, 2004. **126**(25): p. 7758-7759.

84. Wang, T., Y. Xu, D.G. Du, and F. Gai, *Determining beta-sheet stability by Fourier transform infrared difference spectra.* Biopolymers, 2004. **75**(2): p. 163-172.

85. Du, D.G., Y.J. Zhu, C.Y. Huang, and F. Gai, *Understanding the key factors that control the rate of beta-hairpin folding.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(45): p. 15915-15920.

86. Berendsen, H.J.C., J.P.M. Postma, W.F. Vangunsteren, A. Dinola, and J.R. Haak, *Molecular-Dynamics with Coupling to an External Bath.* Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.

87. Hawkins, G.D., C.J. Cramer, and D.G. Truhlar, *Pairwise Solute Descreening of Solute Charges from a Dielectric Medium.* Chemical Physics Letters, 1995. **246**(1-2): p. 122-129.

88. Fersht, A.R., *On the simulation of protein folding by short time scale molecular dynamics and distributed computing.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(22): p. 14122-14125.

89. Zwanzig, R., *Two-state models of protein folding kinetics.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(1): p. 148-150.

90. Thirumalai, D., D.K. Klimov, and S.A. Woodson, *Kinetic partitioning mechanism as a unifying theme in the folding of biomolecules.* Theoretical Chemistry Accounts, 1997. **96**(1): p. 14-22.

91. Matagne, A., S.E. Radford, and C.M. Dobson, *Fast and slow tracks in lysozyme folding: Insight into the role of domains in the folding process.* Journal of Molecular Biology, 1997. **267**(5): p. 1068-1074.

92. Swendsen, R.H. and J.S. Wang, *Replica Monte-Carlo Simulation of Spin-Glasses.* Physical Review Letters, 1986. **57**(21): p. 2607-2609.

93. Tesi, M.C., E.J.J. vanRensburg, E. Orlandini, and S.G. Whittington, *Monte Carlo study of the interacting self-avoiding walk model in three dimensions.* Journal of Statistical Physics, 1996. **82**(1-2): p. 155-181.

94. Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, *Equation of State Calculations by Fast Computing Machines.* Journal of Chemical Physics, 1953. **21**(6): p. 1087-1092.

95. Feig, M., J. Karanicolas, and C.L. Brooks, *MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology.* Journal of Molecular Graphics & Modelling, 2004. **22**(5): p. 377-395.

96. Sanbonmatsu, K.Y. and A.E. Garcia, *Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics.* Proteins-Structure Function and Genetics, 2002. **46**(2): p. 225-234.

97. Karanicolas, J. and C.L. Brooks, *The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: Lessons for protein design?* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(7): p. 3954-3959.

98. Sugita, Y., A. Kitao, and Y. Okamoto, *Multidimensional replica-exchange method for free-energy calculations.* Journal of Chemical Physics, 2000. **113**(15): p. 6042-6051.

99. Kinnear, B.S., M.F. Jarrold, and U.H.E. Hansmann, *All-atom generalized-ensemble simulations of small proteins.* Journal of Molecular Graphics & Modelling, 2004. **22**(5): p. 397-403.

100. Rathore, N., M. Chopra, and J.J. de Pablo, *Optimal allocation of replicas in parallel tempering simulations.* Journal of Chemical Physics, 2005. **122**(2): p. 024111.

101. Fukunishi, H., O. Watanabe, and S. Takada, *On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction.* Journal of Chemical Physics, 2002. **116**(20): p. 9058-9067.

102. Cheng, X., G. Cui, V. Hornak, and C. Simmerling, *Modified Replica Exchange Simulation Methods for Local Structure Refinement.* J. Phys. Chem. B, 2005. **109**(16): p. 8220-8230.

103. Kofke, D.A., *On the acceptance probability of replica-exchange Monte Carlo trials.* Journal of Chemical Physics, 2002. **117**(15): p. 6911-6914.

104. Jang, S.M., S. Shin, and Y. Pak, *Replica-exchange method using the generalized effective potential.* Physical Review Letters, 2003. **91**(5): p. 058305.

105. Mitsutake, A., Y. Sugita, and Y. Okamoto, *Replica-exchange multicanonical and multicanonical replica-exchange Monte Carlo simulations of peptides. II. Application to a more complex system.* Journal of Chemical Physics, 2003. **118**(14): p. 6676-6688.

106. Sugita, Y. and Y. Okamoto, *Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape.* Chemical Physics Letters, 2000. **329**(3-4): p. 261-270.

107. Ghosh, A., C.S. Rapp, and R.A. Friesner, *Generalized born model based on a surface integral formulation.* Journal of Physical Chemistry B, 1998. **102**(52): p. 10983-10990.

108. Srinivasan, J., T.E. Cheatham, P. Cieplak, P.A. Kollman, and D.A. Case, *Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices.* Journal of the American Chemical Society, 1998. **120**(37): p. 9401-9409.

109. Luo, R., L. David, and M.K. Gilson, *Accelerated Poisson-Boltzmann calculations for static and dynamic systems.* Journal of Computational Chemistry, 2002. **23**(13): p. 1244-1253.

110. Sharp, K., *Incorporating Solvent and Ion Screening into Molecular-Dynamics Using the Finite-Difference Poisson-Boltzmann Method.* Journal of Computational Chemistry, 1991. **12**(4): p. 454-468.

111. Alper, H. and R.M. Levy, *Dielectric and Thermodynamic Response of a Generalized Reaction Field Model for Liquid-State Simulations.* Journal of Chemical Physics, 1993. **99**(12): p. 9847-9852.

112. Beglov, D. and B. Roux, *Dominant Solvation Effects from the Primary Shell of Hydration - Approximation for Molecular-Dynamics Simulations.* Biopolymers, 1995. **35**(2): p. 171-178.

113. Beglov, D. and B. Roux, *Finite Representation of an Infinite Bulk System - Solvent Boundary Potential for Computer-Simulations.* Journal of Chemical Physics, 1994. **100**(12): p. 9050-9063.

114. Brooks, C.L., A. Brunger, and M. Karplus, *Active-Site Dynamics in Protein Molecules - a Stochastic Boundary Molecular-Dynamics Approach.* Biopolymers, 1985. **24**(5): p. 843-865.

115. Brooks, C.L. and M. Karplus, *Deformable Stochastic Boundaries in Molecular-Dynamics.* Journal of Chemical Physics, 1983. **79**(12): p. 6312-6325.

116. Kentsis, A., M. Mezei, and R. Osman, *MC-PHS: A Monte Carlo implementation of the primary hydration shell for protein folding and design.* Biophysical Journal, 2003. **84**(2): p. 805-815.

117. King, G. and A. Warshel, *A Surface Constrained All-Atom Solvent Model for Effective Simulations of Polar Solutions.* Journal of Chemical Physics, 1989. **91**(6): p. 3647-3661.

118. Lee, M.S. and M.A. Olson, *Evaluation of poisson solvation models using a hybrid explicit/implicit solvent method.* Journal of Physical Chemistry B, 2005. **109**(11): p. 5223-5236.

119. Lee, M.S., F.R. Salsbury, and M.A. Olson, *An efficient hybrid explicit/implicit solvent method for biomolecular simulations.* Journal of Computational Chemistry, 2004. **25**(16): p. 1967-1978.

120. Lounnas, V., S.K. Ludemann, and R.C. Wade, *Towards molecular dynamics simulation of large proteins with a hydration shell at constant pressure.* Biophysical Chemistry, 1999. **78**(1-2): p. 157-182.

121. Topol, I.A., G.J. Tawa, S.K. Burt, and A.A. Rashin, *On the structure and thermodynamics of solvated monoatomic ions using a hybrid solvation model.* Journal of Chemical Physics, 1999. **111**(24): p. 10998-11014.

122. van der Spoel, D., P.J. van Maaren, and H.J.C. Berendsen, *A systematic study of water models for molecular simulation: Derivation of water models optimized for use with a reaction field.* Journal of Chemical Physics, 1998. **108**(24): p. 10220-10230.

123. Vorobjev, Y.N. and J. Hermans, *ES/IS: Estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model.* Biophysical Chemistry, 1999. **78**(1-2): p. 195-205.

124. Errington, N. and A.J. Doig, *A phosphoserine-lysine salt bridge within an alpha-helical peptide, the strongest alpha-helix side-chain interaction measured to date.* Biochemistry, 2005. **44**(20): p. 7553-8.

125. Groebke, K., P. Renold, K.Y. Tsang, T.J. Allen, K.F. McClure, and D.S. Kemp, *Template-nucleated alanine-lysine helices are stabilized by position-dependent interactions between the lysine side chain and the helix barrel.* Proc Natl Acad Sci U S A, 1996. **93**(9): p. 4025-9.

126. Marqusee, S. and R.L. Baldwin, *Helix stabilization by Glu-...Lys+ salt bridges in short peptides of de novo design.* Proc Natl Acad Sci U S A, 1987. **84**(24): p. 8898-902.

127. Marqusee, S., V.H. Robbins, and R.L. Baldwin, *Unusually stable helix formation in short alanine-based peptides.* Proc Natl Acad Sci U S A, 1989. **86**(14): p. 5286-90.

128. Maison, W., E. Arce, P. Renold, R.J. Kennedy, and D.S. Kemp, *Optimal N-caps for N-terminal helical templates: effects of changes in H-bonding efficiency and charge.* J Am Chem Soc, 2001. **123**(42): p. 10245-54.

129. Heitmann, B., G.E. Job, R.J. Kennedy, S.M. Walker, and D.S. Kemp, *Water-solubilized, cap-stabilized, helical polyalanines: calibration standards for NMR and CD analyses.* J Am Chem Soc, 2005. **127**(6): p. 1690-1704.

130. Chen, K., Z. Liu, and N.R. Kallenbach, *The polyproline II conformation in short alanine peptides is noncooperative.* Proc Natl Acad Sci U S A, 2004. **101**(43): p. 15352-7.

131. McColl, I.H., E.W. Blanch, L. Hecht, N.R. Kallenbach, and L.D. Barron, *Vibrational Raman optical activity characterization of poly(l-proline) II helix in alanine oligopeptides.* J Am Chem Soc, 2004. **126**(16): p. 5076-7.

132. Shi, Z., C.A. Olson, G.D. Rose, R.L. Baldwin, and N.R. Kallenbach, *Polyproline II structure in a sequence of seven alanine residues.* Proc Natl Acad Sci U S A, 2002. **99**(14): p. 9190-5.

133. Asher, S.A., A.V. Mikhonin, and S. Bykov, *UV Raman demonstrates that alpha-helical polyalanine peptides melt to polyproline II conformations.* J Am Chem Soc, 2004. **126**(27): p. 8433-40.

134. Kentsis, A., M. Mezei, T. Gindin, and R. Osman, *Unfolded state of polyalanine is a segmented polyproline II helix.* Proteins, 2004. **55**(3): p. 493-501.

135. Garcia, A.E., *Characterization of non-alpha helical conformations in Ala peptides.* Polymer, 2004. **45**(2): p. 669-676.

136. Mezei, M., P.J. Fleming, R. Srinivasan, and G.D. Rose, *Polyproline II helix is the preferred conformation for unfolded polyalanine in water.* Proteins, 2004. **55**(3): p. 502-507.

137. Onufriev, A., D. Bashford, and D.A. Case, *Exploring protein native states and large-scale conformational changes with a modified generalized born model.* Proteins-Structure Function and Bioinformatics, 2004. **55**(2): p. 383-394.

138. Onufriev, A., D. Bashford, and D.A. Case, *Modification of the generalized Born model suitable for macromolecules.* Journal of Physical Chemistry B, 2000. **104**(15): p. 3712-3720.

139. Kofke, D.A., *On the acceptance probability of replica-exchange Monte Carlo trials (vol 117, pg 6911, 2002).* Journal of Chemical Physics, 2004. **120**(22): p. 10852-10852.

140. Darden, T., D. York, and L. Pedersen, *Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems.* Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.

141. Bondi, A., *Van Der Waals Volumes + Radii.* Journal of Physical Chemistry, 1964. **68**(3): p. 441-&.

142. Tsui, V. and D.A. Case, *Molecular dynamics simulations of nucleic acids with a generalized born solvation model.* Journal of the American Chemical Society, 2000. **122**(11): p. 2489-2498.

143. Ponder, J.W. and F.M. Richards, *An Efficient Newton-Like Method for Molecular Mechanics Energy Minimization of Large Molecules.* Journal of Computational Chemistry, 1987. **8**(7): p. 1016-1024.

144. Liu, P., B. Kim, R.A. Friesner, and B.J. Berne, *Replica exchange with solute tempering: A method for sampling biological systems in explicit water.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(39): p. 13749-13754.

145. Shi, Z., R.W. Woody, and N.R. Kallenbach, *Is polyproline II a major backbone conformation in unfolded proteins?* Adv Protein Chem, 2002. **62**: p. 163-240.

146. Mitsutake, A., Y. Sugita, and Y. Okamoto, *Replica-exchange multicanonical and multicanonical replica-exchange Monte Carlo simulations of peptides. I. Formulation and benchmark test.* Journal of Chemical Physics, 2003. **118**(14): p. 6664-6675.

147. Wickstrom, L., A. Okur, K. Song, V. Hornak, D.P. Raleigh, and C.L. Simmerling, *The Unfolded State of the Villin Headpiece Helical Subdomain: Computational Studies of the Role of Locally Stabilized Structure.* Journal of Molecular Biology, 2006. **360**(5): p. 1094-1107.

148. Zhang, W., C. Wu, and Y. Duan, *Convergence of replica exchange molecular dynamics.* Journal of Chemical Physics, 2005. **123**(15): p. -.

149. Zuckerman, D.M. and E. Lyman, *A Second Look at Canonical Sampling of Biomolecules Using Replica Exchange Simulation*. 2006. p. 1200-1202.

150. Zhang, J., M. Qin, and W. Wang, *Folding mechanism of beta-hairpins studied by replica exchange molecular simulations.* Proteins-Structure Function and Bioinformatics, 2006. **62**(3): p. 672-685.

151. Lednev, I.K., A.S. Karnoup, M.C. Sparrow, and S.A. Asher, *alpha-helix peptide folding and unfolding activation barriers: A nanosecond UV resonance raman study.* Journal of the American Chemical Society, 1999. **121**(35): p. 8074-8086.

152. Matagne, A., M. Jamin, E.W. Chung, C.V. Robinson, S.E. Radford, and C.M. Dobson, *Thermal unfolding of an intermediate is associated with non-arrhenius kinetics in the folding of hen lysozyme.* Journal of Molecular Biology, 2000. **297**(1): p. 193-210.

153. Oliveberg, M., Y.J. Tan, and A.R. Fersht, *Negative Activation Enthalpies in the Kinetics of Protein-Folding.* Proceedings of the National Academy of Sciences of the United States of America, 1995. **92**(19): p. 8926-8929.

154. Segawa, S.I. and M. Sugihara, *Characterization of the Transition-State of Lysozyme Unfolding .1. Effect of Protein Solvent Interactions on the Transition-State.* Biopolymers, 1984. **23**(11): p. 2473-2488.

155. Cavalli, A., P. Ferrara, and A. Caflisch, *Weak temperature dependence of the free energy surface and folding pathways of structured peptides.* Proteins-Structure Function and Genetics, 2002. **47**(3): p. 305-314.

156. Ferrara, P., J. Apostolakis, and A. Caflisch, *Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations.* Journal of Physical Chemistry B, 2000. **104**(20): p. 5000-5010.

157. Frantz, D.D., D.L. Freeman, and J.D. Doll, *Reducing Quasi-Ergodic Behavior in Monte-Carlo Simulations by J-Walking - Applications to Atomic Clusters.* Journal of Chemical Physics, 1990. **93**(4): p. 2769-2784.

158. Zhou, R.H. and B.J. Berne, *Smart walking: A new method for Boltzmann sampling of protein conformations.* Journal of Chemical Physics, 1997. **107**(21): p. 9185-9196.

159. Andricioaei, I., J.E. Straub, and A.F. Voter, *Smart darting Monte Carlo.* Journal of Chemical Physics, 2001. **114**(16): p. 6994-7000.

160. Brown, S. and T. Head-Gordon, *Cool walking: A new Markov chain Monte Carlo sampling method.* Journal of Computational Chemistry, 2003. **24**(1): p. 68-76.

161. Lyman, E., F.M. Ytreberg, and D.M. Zuckerman, *Resolution exchange simulation.* Physical Review Letters, 2006. **96**(2): p. -.

162. Lyman, E. and D.M. Zuckerman, *Resolution exchange simulation with incremental coarsening.* Journal of Chemical Theory and Computation, 2006. **2**(3): p. 656-666.

163. Case, D.A., T.A. Darden, T.E. Cheatham, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, B. Wang, D.A. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J.A. Caldwell, W.S. Ross, and P.A. Kollman, *AMBER 8*. 2004, University of California, San Francisco.

164. Schenck, H.L. and S.H. Gellman, *Use of a designed triple-stranded antiparallel beta-sheet to probe beta-sheet cooperativity in aqueous solution.* Journal of the American Chemical Society, 1998. **120**(19): p. 4869-4870.

165. Syud, F.A., J.F. Espinosa, and S.H. Gellman, *NMR-based quantification of beta-sheet populations in aqueous solution through use of reference peptides for the folded and unfolded states.* Journal of the American Chemical Society, 1999. **121**(49): p. 11577-11578.

166. Fesinmeyer, R.M., E.S. Peterson, R.B. Dyer, and N.H. Andersen, *Studies of helix fraying and solvation using 13C' isotopomers*. 2005. p. 2324-2332.

167. Neidigh, J.W., R.M. Fesinmeyer, and N.H. Andersen, *Designing a 20-residue protein.* Nature Structural Biology, 2002. **9**(6): p. 425-430.