UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DESIGN OF NEW MOLECULAR DYNAMICS GLOBAL MINIMUM SEARCH PROTOCOLS FOR MAPPING ENERGY LANDSCAPES AND CONFORMATIONS OF FOLDED POLYPEPTIDES AND MINI-PROTEINS

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

Zunnan Huang

Norman, Oklahoma

2005

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

DESIGNS OF NEW MOLECULAR DYNAMICS GLOBAL MINIMUM SEARCH PROTOCOLS FOR MAPPING ENERGY LANDSCAPE AND CONFORMATIONS OF FOLDED POLYPEPTES AND MINI-PROTEINS

A Dissertation APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY

____
Dr. Ralph A. Wheeler

____
Dr. Michael R. Abraham

____
Dr. Wai Tak Yip

____
Dr. Paul F. Cook

____
Dr. Kieran J. Mullen

For my families
My parents: Shangjing and Wanqun Huang
My wife: Yong Cao
My sisters: Caian, Shuiling, and Bifeng Huang

Acknowledgements

This dissertation work could not have been accomplished without the contributions of many people. First of all, I would like to sincerely thank my thesis adviser, Dr. Ralph Wheeler, for providing me with research assistantship and for helping me with my graduate studies since 2001. He is an invaluable mentor in directing my scientific career. I would also like to thank the other members of my Graduate Committee: Professors Michael R. Abraham, Wai Tak Yip, Paul F. Cook, and Kieran J. Mullen. I sincerely appreciate the time and effort that you have given me over the last five years. I am also grateful to all faculty and staff in the department, especially Dr. LeRoy C. Blank and Ms. Arlene C. Crawford, who have provided me with much assistance since I first entered this country and this university.

Secondly, I wish to thank all present and past members of my research group especially Christopher Adam Hixson, Timothy H., Click, Scott E. Boesch, Zhanyong Guo, and Dr. Asif Rahaman for their contributions to my research projects. Christopher Adam Hixson generously provided his computer code for molecular dynamics simulation. He also gave me instructions to understand his code when I was first involved in programming. Timothy H. Click spent a lot of time helping me with the English grammar for my dissertation. Scott E. Boesch taught me the basics of conducting research in computational chemistry, and he was always nearby to answer any questions concerning use of computer resources, etc. Zhanyong Guo kindly provided his study of PDB 1PEF polypeptide, which is presented as a good example to test my developed algorithms. None of this work would have been possible without the efforts from my colleagues.

Third, I want to express my thankfulness to the University of Oklahoma, the School of Arts and Science, the Department of Chemistry and Biochemistry, the OU Supercomputing for Education and Research (OSCER), the Graduate College, and the International Student Office for their generous support. From these groups and organizations, I am especially grateful to Dr. Henry J. Neeman for his guidance in program coding and Dr. Millie C. Audas for her helpful advice in my social activities.

Finally, and most significantly, I want to offer my deepest thanks to my wife, my parents, my sister and all my friends for all the love, friendship, and support they have given me over the last five years. Without your support, I could not have the opportunity to go abroad and pursue my academic goals.

Table of Contents

List of Figures

List of Tables

Abstract

Molecular dynamics (MD) simulations are widely used for global conformational searches in protein folding. However, conventional canonical ensemble simulations (constant NVT) usually cannot explore biologically active natural structures of proteins because such simulations have extreme difficulty sampling conformational space sufficiently for global energy minimum searches. The work described here delineates the crucial limitations restricting conventional NVT simulations from covering a wide variety of conformations and develops several new MD simulation protocols for efficiently sampling diverse regions of conformational space to search for the global minimum energy structure of polypeptides and mini-proteins.

First, a new MD search strategy called DIvergent Path (DIP) search simulation is developed in which the simulations start with several independent polypeptides having the same initial coordinates and temperatures but different velocity directions, which evolve into different trajectories. The DIP simulations reveal three primary limitations of conventional MD simulations: potential energy traps, free energy traps, and kinetic traps. Among them, kinetic traps are the most limiting factor for MD simulations intended to sample varied conformational space at room temperature. This trap is caused by mechanical equilibrium (when both kinetic and potential energies have reached equilibrium) and can be easily overcome by intervening to reassign atomic velocities and thus randomize simulation trajectories. By combining this trajectory randomization strategy at one temperature with cycles of heating and cooling, the DIsrupted VElocity (DIVE) search simulation is further developed. The DIVE simulations can explore wide ranges of the rugged potential energy surfaces of peptides, sample myriad potential

energy minima, and explore diverse conformations even in a very limited simulation time. Finally, a combined procedure is also built in which the global potential energy minimum and myriad local potential energy minima are explored by using DIVE simulations followed by DIP simulations to search for the global free energy minimum near *in vivo* temperatures.

We performed the new MD simulations for mapping energy landscapes and conformations of a model 13 residues polyalanine peptide Ala13, an amphiphilic octadecapeptide, peptide F, and a 20-residue mini-protein, Trp-cage, using the AMBER force field either *in vacuo* or in a generalized Born/solvent-accessible surface area (GB/SA) implicit solvent for water. The simulation results are also compared with those from several other simulation algorithms including conventional NVT simulations, the replica exchange method (REM), and locally enhanced sampling (LES) molecular dynamics. Our newly developed MD simulation protocols sample the most diverse region of conformational space and complement existing global geometry optimization techniques for predicting 3D protein structures from only primary sequence data.

Chapter 1

Introduction

During the past several decades, molecular dynamics (MD) [1,2] has evolved into an important and widely used computational tool in chemistry, physics, and biology to model the detailed dynamical behavior of many systems, from atomic clusters to large biological molecules. As a computational simulation method, MD simulations give the time evolution, or the trajectory, of an atomic system and further determine the thermodynamics, energetic, structural and dynamical properties of that atomic system [3]. The movement of the atoms, due to the force of their own kinetic energy and the forces exerted upon them by all other atoms in the system, is calculated by integrating the classical equations of motion from Newton using an assumed potential energy function [4,5]. Molecular dynamics can provide ways to simulate costly, dangerous, or experimentally inaccessible systems. Furthermore, simulations can track the time evolution of structural changes at high temporal resolutions [6], which is currently better than experimental methods.

The method of molecular dynamics today is a standard tool to simulate protein folding [7]. In its natural environment, proteins can fold from an extended linear structure to a condensed, compact three-dimensional structure. This automatic folding of proteins generally takes milliseconds to seconds either *in vivo*, or *in vitro* [8-13]. How proteins fold from a primary amino acid sequence into a compact three-dimensional structure in nature remains one of the great mysteries of science. It is generally assumed that the protein sequence contains the information that determines the final three-dimensional structure [8-13]. Therefore, scientists hope that computer simulation methods such as molecular dynamics are able to predict tertiary structure from the amino acid sequence. With this calculation method, one must have a starting extended or random unfolded structure. The potential energy function whose parameters (force field) are derived to reproduce structures and energy trends in various small model systems should be transferable to large proteins [5].

However, some fundamental problems exist when using molecular dynamics to simulate the folding process of proteins. The first problem arises from an astronomical number of possible conformations of a simulated protein. The number of possible conformations for the flexible protein increases exponentially as the number of rotatable bonds increases and rapidly exceeds the number that can be realistically evaluated on even the most powerful supercomputers. In 1969, Levinthal [14] calculated that even if each residue in a 100 amino acid protein is only restricted to three conformations, the total number of structures possible for this 100 residue protein would be approximately 3^{100} (or ~5×10^{47}). Furthermore, if this 100-residue protein took 0.1 ps (10^{-13} s) to convert from one form to another, it would take ~1.6×10^{27} years to sample all the structures. Exploring all possible conformations in search of the native structure requires an exorbitant amount of time, which contradicts the actual folding times of proteins. This is referred to as Levinthal's Paradox [15-18]. With these calculations, Levinthal assumed that there were well-defined pathways to the native state for protein folding [19]. Therefore, in order to find the native structure of proteins within a realistic simulation time, molecular dynamics must circumvent Levinthal's Paradox to explore these well-defined pathways.

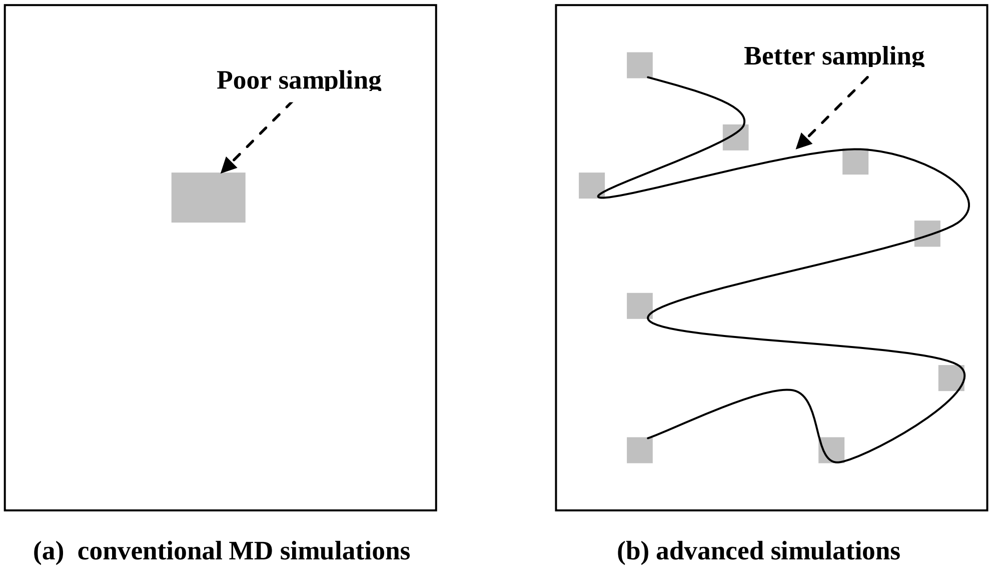
According to Anfinsen's thermodynamic hypothesis [20], the native conformation of a protein is the one in which the free energy of the protein is lowest (global free energy minimum). A protein folds along a path towards this global free energy minimum for conformational stability. This folding helps to limit the number of conformational possibilities that need to be searched in order to find the global free energy minimum [18]. The decrease in the sampling of conformational possibilities in the search for the global minimum is strongly dependent on the free energy surface, which we can build using a force field in computer simulations. Large decreases in free energy only happen if the global free energy well is much deeper and broader than any other local free energy wells.

The search for the global free energy minimum in protein folding constitutes the second difficultly for molecular dynamics while simulating protein folding [12,21-23]. Suppose that we performed an ideal simulation. The potential energy function and force field were perfected to build a very deep and broad global free energy well corresponding to the native structure of a 100-residue protein. The MD simulation was able to search for the native structure on the millisecond to second timescale, in direct comparison with folding experiments. However, the computational time needed to perform this simulation would still be astronomical [21]. Using current technology, if the MD simulation provided a trajectory of 1 nanosecond/day (large overestimate for a solvated 100-residue protein), it would take approximately 2740 years to achieve a millisecond trajectory for protein folding.

The third problem that MD simulations encounter is entrapment in local energy minima. On the folding path toward the global energy minimum, the simulated protein may fall into a deep local energy well. If the local energy well has high barriers to other energy wells (transition state energies) and the protein does not have enough kinetic energy to overcome the barrier, the protein is then caught in a local energy minimum. When local energy minimum entrapment occurs, MD simulations are unable to search additional areas of conformational space. In other words, MD simulations are limited to a local search for a family of similar conformations. Conventional MD simulations fail to find the global energy minimum because simulations at low temperatures tend to get trapped in one of myriad local minimum-energy states [24-33].

The above-described problems imply that molecular dynamics currently suffer limited phase space sampling when used to simulate protein motion. As an aside, limited phase space sampling is universal and exists with any other computer simulations such as Monte Carlo simulations [34] This limited phase space sampling causes molecular dynamics or other computer simulations to have difficultly in predicting the natural three-dimensional (3D) structures of a small protein (or polypeptide) from a random unfolded structure [12,22,23,35]. Currently, the majority of MD simulations are still limited to refinement or confirmation of experimental structures. Attempts to overcome the phase space sampling problem in structure prediction typically involves increasing the system's kinetic energy or decreasing its energy barriers [26-30]. We make an effort to consider these phase space sampling problems at the atomic level and additionally develop new techniques for efficiently sampling conformations of polypeptides and mini-proteins.

In a realistic limited simulation time (such as one microsecond (μs)), any MD simulation can only sample an extremely small fraction of the total conformational space. With the same simulation time and time step, the simulations from different molecular dynamics techniques should always sample the same number of structures (a very small fraction of all the possible conformations). However, the simulation trajectories from different MD techniques may evolve into different regions of conformational space. Figure 1 gives examples of two different simulations. The outside rectangle represents the total conformational space. The enlarged gray regions are sampled by realistic molecular dynamic simulations. A conventional MD simulation at ambient temperature is usually limited to a local search for a family of similar conformations represented by the gray rectangle in Fig. 1.1a. It does not search other regions. If a simulation can sample several regions represented by the smallest gray squares in Fig. 1.1b, the simulation achieves a better sampling of conformational space, and it can explore a wide variety of structures. We aim to develop MD simulation protocols corresponding to Fig. 1.1b.



(a) conventional MD simulations **(b) advanced simulations**

Figure 1.1. Outside rectangle represents complete conformational space. Enlarged gray regions are sampled by MD simulations in a realistic time such as 1μs. All smallest gray squares together with line in **(b)** form a combined region having the same area as the larger gray rectangle in **(a)**.

The new MD simulation techniques illustrated in Fig. 1.1b should give a good balance between the sampling of varied conformations and conformational stability at low temperatures. An MD simulation at very high temperatures may search different conformations quickly, but those sampled conformations are likely to be various unfolded conformations. These conformations represent denatured proteins. It will be difficult to find those well-defined native structures of proteins corresponding to low free energies states near physiological temperatures. Therefore, the new MD simulation techniques should focus on sampling varied conformations at low temperatures. We observed some crucial limitations that restrict conventional MD simulations to a local search for a family of similar conformations. We have developed strategies to overcome these limitations. Here we demonstrate that our newly developed MD simulation protocols are able to sample varied conformations and explore well-defined folding pathways in search of native structures of polypeptides and mini-proteins, even on a limited nanosecond simulation time scale.

Molecular dynamics is a tool for sampling a statistical mechanical ensemble and for determining ensemble averages of thermodynamic quantities or equilibrium properties [36]. However, the limited classical states sampled during realistic MD simulations usually cannot accurately represent the real probability distribution. A simulation of 1 ms is only a very short simulation for sampling all possible conformations; it can only reach an extremely small fraction of the entire conformational space. Therefore, we consider it more important to explore the energy surface and minimum conformations than to calculate thermodynamic properties when MD simulations are limited to partial phase space sampling.

The global free energy minimum and the free energy surface are difficult to determine from computer simulations. However, the potential energy surface is relatively easy to define as a function of the atomic coordinates of the system [37] and, in addition, is temperature independent [38]. Even though the global minimum free energy structure at *in vivo* temperatures may not correspond to the global potential energy minimum, the global free energy minimum is usually either the global minimum or a very low local minimum on the potential energy surface [37]. Therefore, maps of these low potential energy minima and their conformations can be quite valuable. In order to determine the global folded structure, it is necessary to find the minimum potential energy conformations. In addition, the quality of the energy landscapes of proteins (or polypeptides) built from these potential energy minima can test the quality of the potential energy function used, which is usually transferred from small standard molecules.

Therefore, the work described here delineates the crucial limitations restricting conventional constant-temperature simulations from covering a wide variety of conformations and develops several new MD simulation protocols for efficiently sampling many regions of conformational space in search of the global minimum energy structures of polypeptides and mini-proteins.

Chapter 2 of this dissertation gives background for various MD techniques. Relying on the body of conventional MD simulations described first, our own written MD program suite is described, and later several advanced MD methods for enhancing phase space sampling are introduced, including simulated annealing (SA), the replica exchange method (REM), and locally enhanced sampling (LES).

Chapter 3 – 7 focus on our development of new MD techniques, tested by a model polyalanine peptide, Ala13 (13 residues). Chapter 3 describes the DIvergent Path (DIP) search simulations for global free energy minimum searches near *in vivo* temperatures. Simulation results for gas phase Ala13 show three primary limitations of conventional MD simulations, which give rise to limited phase space sampling. Chapter 4 presents the DIsrupted VElocity (DIVE) search simulations for global potential energy minimum searching and demonstrates the advantages of DIVE for sampling conformational space more efficiently than several other simulation algorithms for conformational searches of gas phase Ala13. Chapter 5 presents MD simulations of gas phase Ala13 using these new protocols with different force fields. Comparisons of the sampled secondary structures and their energy minima for this polypeptide highlight likely differences in potential energy landscapes from different force fields. New molecular dynamics simulations of Ala13 in a generalized Born/solvent-accessible surface area (GB/SA) implicit solvent environment for water are further presented in Chapter 6. Chapter 7 addresses the discrepancy of the free-energy minima for solvated Ala13 systems at room temperature simulated by locally enhanced sampling MD vs. conventional MD. Furthermore, it demonstrates that the indirect coupling between copies of the subsystem through the common bath, not a reduction in energy barriers, promotes conformational transitions of the equilibrated system in approximate mean field simulations.

The final Chapters demonstrate the efficiency of DIVE simulations to search for the experimentally determined native structures of two *de novo* designed polypeptides. Chapter 8 addresses DIVE simulations for mapping potential energy landscapes and conformations of an amphiphilic octadecapeptide, peptide F, in implicit water. In addition, a combined procedure is described in which the global potential energy minimum and myriad local potential energy minima are explored by using DIVE simulations followed by DIP simulations to search for the global free energy minimum near *in vivo* temperatures. In chapter 9, the DIVE simulations have been carried out to study the folding conformations and kinetics of a 20-residue mini-protein, Trp-cage, in implicit solvent. In addition to searching diverse conformations and their minimum energy states and thus mapping the potential energy landscape, several folding pathways of this mini-protein are also characterized. This DIVE simulation protocol is a new global optimization technique for predicting 3D protein structures from only sequence data.

Bibliography

(1) Vasquez, M.; Nemethy, G.; Scheraga, H. A. *Chemical Reviews* **1994**, *94*, 2183.

(2) Tuckerman, M. E.; Martyna, G. J. *Journal of Physical Chemistry B* **2000**, *104*, 159.

(3) Beveridge, D. L.; Dicapua, F. M. Free energy via molecular simulation: A primer. In *Computer simulation of biomolecular systems: theoretical and experimental applications*; van Gunsteren, W. F., Weiner, P. K., Eds.; ESCOM: Leiden, 1989; Vol. 1; pp 1.

(4) Frenkel, D.; Smit, B. *Understanding molecular simulaiton: from algorithm to application*; Academic Press: New York, 2002; Vol. 2nd.

(5) Doucet, J. P.; Weber, J. *Computer-aided molecular design: theory and applications*; Academic Press: London, 1996.

(6) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *Journal of Computational Chemistry* **2003**, *24*, 1999.

(7) Hansson, T.; Oostenbrink, C.; van Gunsteren, W. F. *Current Opinion in Structural Biology* **2002**, *12*, 190.

(8) Daggett, V. *Accounts of Chemical Research* **2002**, *35*, 422.

(9) Gruebele, M. *Current Opinion in Structural Biology* **2002**, *12*, 161.

(10) Fersht, A. R. *Proceedings of the National Academy of Sciences of the United States of America* **2000**, *97*, 14121.

(11) Brooks, C. L., III. *Accounts of Chemical Research* **2002**, *35*, 447.

(12) Hardin, C.; Pogorelov, T. V.; Luthey-Schulten, Z. *Current Opinion in Structural Biology* **2002**, *12*, 176.

(13) Lee, M. R.; Tsai, J.; Baker, D.; Kollman, P. A. *Journal of Molecular Biology* **2001**, *313*, 417.

(14) Levinthal, C. *Mossbauer Spectroscopy in Biological Systems*; University of Illinois Press: Urbana, 1969.

(15) Zwanzig, R.; Szabo, A.; Bagchi, B. *Proceedings of the National Academy of Sciences of the United States of America* **1992**, *89*, 20.

(16) Stryer, L. *Biochemistry,* 4th ed.; W. H. Freeman and Company: New York, 1996.

(17) Voet, D.; Voet, J. G. Biochemistry; John Wiley & Sons, Inc.: New York, 1995.

(18) Branden, C.; Tooze, J. *Introduction to protein structure*, 2nd ed.; Garland Publishers, Inc.: New York, 1998.

(19) Levinthal, C. *Journal de Chimie Physique* **1968**, *65*, 44.

(20) Anfinsen, C. B. *Science* **1973**, *181*, 223.

(21) Kollman, P. A.; Merz, K. M., Jr. *Accounts of Chemical Research* **1990**, *23*, 246.

(22) Argos, P.; Abagyan, R. *Computers & Chemistry* **1994**, *18*, 225.

(23) Wales, D. J.; Scheraga, H. A. *Science* **1999**, *285*, 1368.

(24) Kirkpatrick, S.; Gelatt, C. D. J.; Vecchi, M. P. *Science* **1983**, *220*, 671.

(25) Basu, A.; Frazer, L. N. *Science* **1990**, *249*, 1409.

(26) Bassolino-Klimas, D.; Tejero, R.; Krystek, S. R.; Metzler, W. J.; Montelione, G. T.; Bruccoleri, R. E. *Protein Science* **1996**, *5*, 593.

(27) Hornak, V.; Simmerling, C. *Proteins: Structure, Function, and Genetics* **2003**, *51*, 577.

(28) Roitberg, A.; Elber, R. *Journal of Chemical Physics* **1991**, *95*, 9277.

(29) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96.

(30) Okamoto, Y. *Journal of Molecular Graphics & Modelling* **2004**, *22*, 425.

(31) Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **2000**, *329*, 261.

(32) Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **1999**, *314*, 141.

(33) Jarrold, M. F. *Annual Review of Physical Chemistry* **2000**, *51*, 179.

(34) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *Journal of Chemical Physics* **1953**, *21*, 1087.

(35) Lee, J.; Liwo, A.; Scheraga, H. A. *Proceedings of the National Academy of Sciences of the United States of America* **1999**, *96*, 2025.

(36) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, 1987.

(37) Mortenson, P. N.; Wales, D. J. *Journal of Chemical Physics* **2001**, *114*, 6443.

(38) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598.

Chapter 2

Molecular Dynamics (MD) Simulation and Current MD Simulation Methods

2.1. Overview

The material described herein is background for molecular dynamics (MD) simulations. The first part describes different components of conventional MD from which our new MD simulation techniques are derived. The body of the program performing conventional MD simulations mainly includes seven features. MD simulations follow classical mechanics according to Newton's laws of motion [1,2]. Solving the classical equations of motion for a biological system of N atoms requires a numerical integration scheme. We use the velocity Verlet algorithm [1,3,4] to accomplish this. The SHAKE algorithm [5,6] is used to constrain bond stretching with hydrogen in order to use a 1 fs time step in the iterative numerical procedure. In addition, MD simulations require a potential energy function and its corresponding force field definitions to describe the atomic interactions in the system and to evaluate each step's forces acting at each particle position. We use the AMBER potential energy function and force fields [7,8]. The computational algorithms we mentioned so far can constitute an MD simulation in the microcanonical ensemble (NVE). Canonical ensemble (NVT) simulations need to use a temperature coupling scheme to maintain a constant temperature. We use the Nośe-Hoover Chain method [9-11] to maintain constant temperature. Furthermore, MD simulations are usually performed in water or some type of aqueous solution. To save computational cost, we use the generalized Born/surface area (GB/SA) implicit solvent model [12-17] to replace explicit solvent, which would require a large number of water molecules. Finally, we want to remove the translational modes of the entire system.

Fig. 2.1.1 is a schematic diagram showing how we calculate positions and velocities in molecular dynamics. By choosing a set of initial positions and velocities, the new positions and velocities at any time can be calculated in the iterative numerical procedure by calculating the force-derived accelerations from the potential energy function and force field. This constitutes the simplest microcanonical ensemble simulation of gas phase polypeptides. With the addition of the extra forces from the SHAKE algorithm, Nosé-Hoover Chain, and GB/SA implicit solvent, the widely used conventional canonical ensemble simulations of solvated polypeptides or proteins can then be performed.

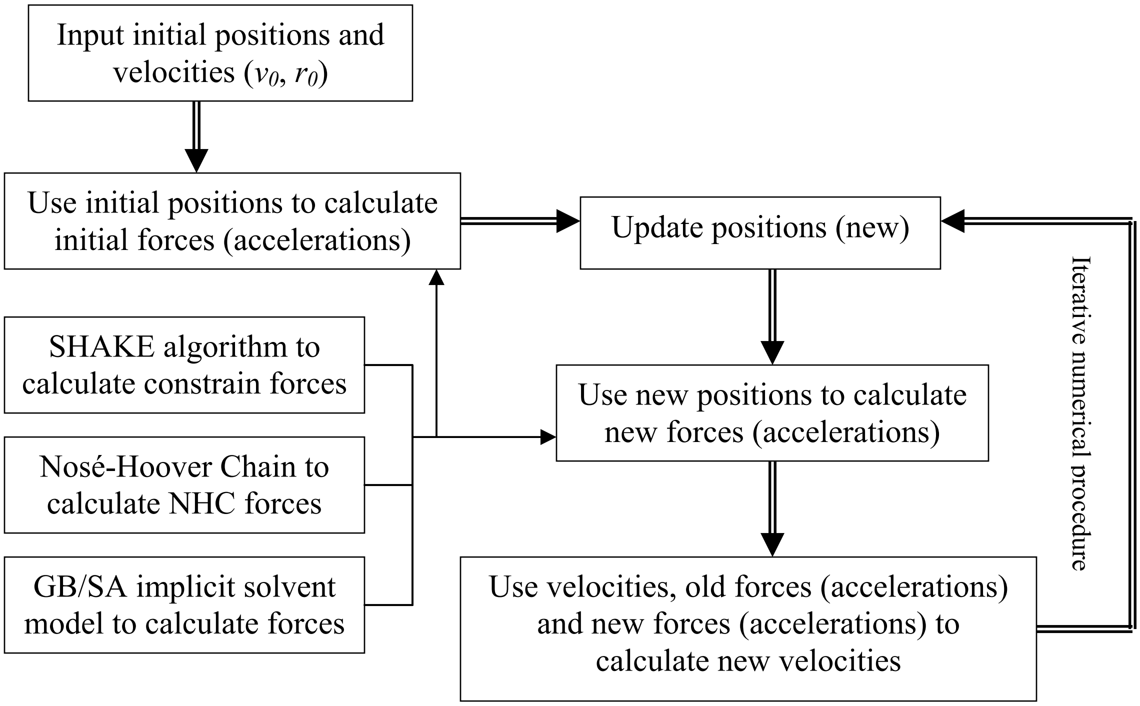

Input initial positions and velocities (v_0, r_0)
Use initial positions to calculate initial forces (accelerations)
Update positions (new)
SHAKE algorithm to calculate constrain forces
Nosé-Hoover Chain to calculate NHC forces
GB/SA implicit solvent model to calculate forces
Use new positions to calculate new forces (accelerations)
Use velocities, old forces (accelerations) and new forces (accelerations) to calculate new velocities
Iterative numerical procedure

Figure 2.1.1. Schematic diagram of position and velocity calculation in molecular dynamics.

Conventional MD simulations are limited in the amount of phase space they sampled. Advanced MD algorithms have been developed to overcome the sampling problem while searching for the global energy minimum in protein folding. The second part gives the introduction of several advanced MD algorithms including simulated annealing (SA) [18], the replica exchange method (REM) [19,20], and locally enhanced sampling (LES) or ensemble mean-field MD [21]. These advanced algorithms enhance phase space sampling during MD simulations.

2.2. Molecular dynamics simulations

2.2.1. Classical mechanics by Newton's law

Molecular dynamics (MD) is a technique used to solve the classical equations of motion governing the microscopic time evolution of a many-body system. Consider a system of N molecules interacting via an assumed potential, the classical motion of the system can be written as the Lagrangian equation [1]:

$$\frac{d}{dt}\left(\frac{\partial\Re}{\partial\dot{q}_k}\right)-\left(\frac{\partial\Re}{\partial q_k}\right)=0 \qquad k=1,2,\ldots N\,, \tag{2.2.1.1}$$

Where the Lagrangian function $\Re(\mathbf{q},\ \dot{\mathbf{q}})$, defined as the subtraction of kinetic energy (K) and potential energy (V), $\Re = K - V$, is a function of the generalized coordinates q_k and their time derivatives $\dot{q}_k$. If an atomic system with Cartesian coordinates $\mathbf{r}_i$ and the usual K and V is defined, then eq. (2.2.1.1) becomes

$$m_i\ddot{\mathbf{r}}_i=\mathbf{f}_i=-\frac{\partial V}{\partial\mathbf{r}_i} \tag{2.2.1.2}$$

where m_i is the mass and $\mathbf{f}_i$ is the force acting on the atom i.

The classical equations of motion (eq. 2.2.1.2) for an atomic system in Newtonian dynamics can also be derived from Hamiltonian mechanics [2]. The Hamiltonian for an N-particle system subject only to an interparticle potential $U(\mathbf{r}_1,\cdots,\mathbf{r}_N)$ is

$$H(\mathbf{p},\mathbf{r}) \equiv H(\mathbf{p}_1,\cdots,\mathbf{p}_N,\mathbf{r}_1,\cdots,\mathbf{r}_N) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}_1,\cdots,\mathbf{r}_N) \tag{2.2.1.3}$$

in which $\mathbf{p}_1,\cdots,\mathbf{p}_N$ are the momenta of the particles given by $\mathbf{p}_i = m_i\mathbf{v}_i$. According to Hamilton's definition,

$$\begin{aligned} \mathbf{r}_i &= \frac{\partial H}{\partial \mathbf{p}_i} = \frac{\mathbf{p}_i}{m_i} \\ \dot{\mathbf{p}}_i &= -\frac{\partial H}{\partial \mathbf{r}_i} = -\frac{\partial U}{\partial \mathbf{r}_i} = \mathbf{f}_i \end{aligned} \tag{2.2.1.4}$$

Then, the classical equations of motion (eq. 2.2.1.2) can be easily obtained.

A unique solution to eq. (2.2.1.2) is achieved by choosing a set of initial spatial positions $\mathbf{r}_i$ and their velocities $\dot{\mathbf{r}}_i$ (or $\mathbf{v}_i$). Although Newton's equations completely determine the full set of 3N positions and velocities for the system as functions of time, an analytical solution to these equations is usually impossible [2]. A standard approximate method for solving eq. (2.2.1.2) is an iterative numerical procedure called a numerical integrator [3]. With the initial positions and velocities of all atoms known at t_0, the new positions and velocities at a later time $t_0 + \delta t$ can be obtained through calculating the force-derived accelerations $\ddot{\mathbf{r}}_i$ (or $\mathbf{a}_i$) from the initial positions using eq. (2.2.1.2). The new positions are then used to calculate the next step's accelerations (i.e. forces) and the equations of motion are numerically solved on a step-by-step basis [1]. The time discretization δt, referred to as the time step, determines the accuracy of the numerical solution [2]. It must be set significantly smaller than the typical time taken for the shortest time for a molecular vibration, in order to achieve a highly accurate microscopic motion picture of the system [1]. After a large number of steps are calculated, the system's equilibrium properties including thermodynamic quantities (pressure, temperature, volume) etc. can be obtained by the statistical average over all time steps.

The classical equations of motion (eq. 2.2.1.2) for an atomic system specify the classical state ($\mathbf{r}_i$ and $\mathbf{v}_i$) of the system at any time in the Newtonian dynamics. The union of all possible classical states of a system defines the phase space, which can also be considered as the complete collection of the full set of particle positions and momenta in the Hamiltonian dynamics [2]. If the limited classical states sampled during the simulations can approximately represent the real distribution of the unlimited points in phase space, the MD calculations are capable of recovering the experimental observables of average thermodynamic quantities.

2.2.2. Velocity Verlet algorithm for the numerical integration schemes

Perhaps one of the most attractive proposed methods of integrating eq. (2.2.1.2) is called the "velocity Verlet" algorithm [1,3,4]. This "velocity Verlet" algorithm takes the form:

$$\begin{aligned} \mathbf{r}(t+\delta t) &= \mathbf{r}(t) + \delta t\,\mathbf{v}(t) + \tfrac{1}{2}\delta t^2 \mathbf{a}(t) \\ \mathbf{v}(t+\delta t) &= \mathbf{v}(t) + \tfrac{1}{2}\delta t[\mathbf{a}(t) + \mathbf{a}(t+\delta t)] \end{aligned} . \tag{2.2.2.1}$$

While the new positions $\mathbf{r}(t+\delta t)$ are calculated from the current positions $\mathbf{r}(t)$, velocities $\mathbf{v}(t)$, and accelerations $\mathbf{a}(t)$, new velocities $\mathbf{v}(t+\delta t)$ are calculated from the current velocities $\mathbf{v}(t)$, accelerations $\mathbf{a}(t)$, and the new force-derived accelerations $\mathbf{a}(t+\delta t)$. This is illustrated in Fig. 2.2.2.1 This algorithm only requires essentially 3N words of storage, and its numerical stability, convenience, and simplicity make it perhaps the most widely used numerical integrator to date [1].

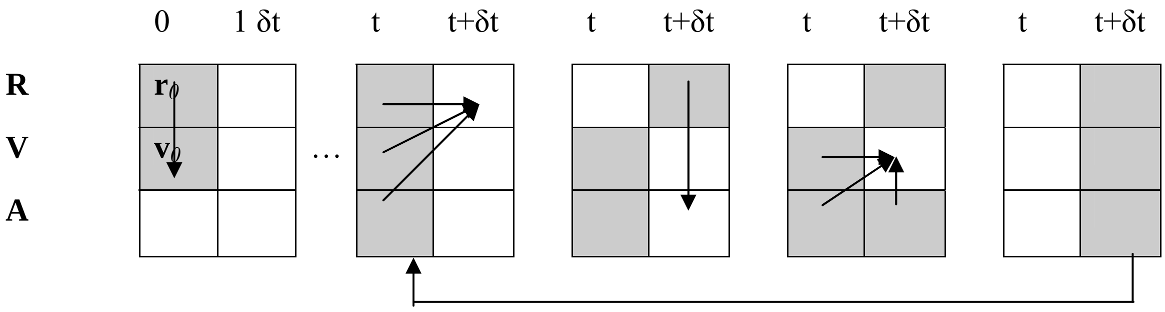


Fig. 2.2.2.1. Flow chart of position and velocity calculations in the velocity Vertlet algorithm [1]. The stored variables are in grey boxes.

2.2.3. Shake algorithm for bond length constraints

In order to obtain a clear microscopic dynamical picture, the time step δt must be set small enough to allow high accuracy in the calculation of atomic positions and velocities from eq. (2.2.2.1). Theoretically, δt should be set on a time scale as small as 10^{-16} second (s), so that the vibration of any bond stretching and angle bending etc. in the system can be simulated appropriately. However, it would take an extremely long computational time for a simulation on the nanosecond (ns, 10^{-9} s) time scale with such a small step to run even on the most powerful supercomputers. Instead, people often use femtosecond (fs, 10^{-15} s) or even longer (2.5 fs, 2 fs) as a typical step size to run MD simulations. At this time level, only the bond stretching involving hydrogen vibrates too fast to be traced accurately. On the other hand, the bond stretching vibrations make little contribution to the conformational change of the whole system. As a result, a constraint procedure is often used to eliminate the bond stretching with hydrogen in normal MD simulations with the time step as large as 1fs.

The SHAKE algorithm [5,6] provides the ability to constrain degrees of freedom including bond stretching in the velocity Verlet numerical integration scheme. For a system that contains only bond length constraints, the SHAKE algorithm makes use of the constraint equations:

$$d_{ij}^2 - \mathbf{r}_{ij}^2 = 0 \qquad (2.2.3.1)$$

Where d_{ij}, and $\mathbf{r}_{ij}$ are the equilibrium bond length and actual bond length at any instant, respectively. In a basic MD calculation generated by integrating eq. (2.2.1.2), the bond distance calculated from the updated positions at a dynamical step usually cannot satisfy the constraints of eq. (2.2.3.1). Therefore, in addition to the original interatomic forces (which are the negative first derivatives of the potential V with respect to the atomic positions), extra forces need to be added into the total forces, so the updated bond distance in the following step meet the criterion. These constraint forces, like the atomic positions and velocities, can only be solved in an iterative numerical procedure.

Suppose the positions $\mathbf{r}_i(t)$, velocities $\mathbf{v}_i(t)$, and total forces $\mathbf{f}_i(t)$, acting on the ith particle at time t are known. The total forces $\mathbf{f}_i(t)$ include both the interatomic forces, $\mathbf{f}_i'(t)$, and the correct constraint forces, $\mathbf{g}_i(t)$, as follows [6]:

$$\mathbf{f}_i(t) = \mathbf{f}_i'(t) + \mathbf{g}_i(t) \qquad (2.2.3.2)$$

According to this assumption, $\mathbf{r}_i(t+\delta t)$ at the next step have already satisfied the constraints of eq. (2.2.3.1), and can be directly calculated from the eq. (2.2.2.1). The calculation of $\mathbf{v}_i(t+\delta t)$, however, requires knowing $\mathbf{a}_i(t+\delta t)$ (i.e. $\mathbf{f}_i(t+\delta t)$) first. In $\mathbf{f}_i(t+\delta t)$, while the $\mathbf{f}_i'(t+\delta t)$ can be easily obtained from the $\mathbf{r}_i(t+\delta t)$, the $\mathbf{g}_i(t+\delta t)$, involving to satisfy the constraint equations require $\mathbf{r}_i(t+2\delta t)$ and difficult to compute, since the calculation of $\mathbf{r}_i(t+2\delta t)$ in turn requires $\mathbf{f}_i(t+\delta t)$ and $\mathbf{v}_i(t+\delta t)$.

The constraint force acting on the *i*th atom of a system involving only bond length constraints of eq. (2.2.3.1) can be written as [6]

$$\mathbf{g}_i(t+\delta t) = \sum_k \lambda_{ik}\mathbf{r}_{ik}(t+2\delta t) \qquad (2.2.3.3)$$

The summation over *k* extends over all atoms that are bonded to atom i and $\mathbf{r}_{ik}$ is their separation vector. The parameters λ_{ik} are symmetric and the equations are not analytically solvable. A set of approximation solutions labeled λ_{ij}^A and an estimate of the difference between λ_{ij}^A and the true λ_{ij} labeled $\delta\lambda_{ij}$, are introduced in the iterative numerical procedure. The quantity $\delta\lambda_{ij}$ is used to update the approximate values $\mathbf{g}_i^A(t+\delta t)$ and $\mathbf{r}_i^A(t+2\delta t)$. The relationship between the updated $\mathbf{r}_i\ (t+2\delta t)$ and the approximation $\mathbf{r}_i^A(t+2\delta t)$ for the coordinates, can be derived from eq. (2.2.2.1) and is as written as [6]

$$\begin{aligned}\mathbf{r}_i(t+2\delta t) &= \mathbf{r}_i^A(t+2\delta t) + \delta\lambda_{ij}\frac{(\delta t)^2}{m_i}\mathbf{r}_{ij}(t+2\delta t)\\ \mathbf{r}_j(t+2\delta t) &= \mathbf{r}_j^A(t+2\delta t) - \delta\lambda_{ij}\frac{(\delta t)^2}{m_j}\mathbf{r}_{ij}(t+2\delta t)\end{aligned} \qquad (2.2.3.4)$$

Using these expressions in the constraint equations eq. (2.2.3.1) and rewriting the form by a Taylor expansion about $\mathbf{r}_{ij}(t+2\delta t)$, we get [6]

$$d_{ij}^2 - |\mathbf{r}_{ij}^A(t+2\delta t)|^2 = 2(\delta t)^2(\frac{1}{m_i}+\frac{1}{m_j})\delta\lambda_{ij}\mathbf{r}_{ij}^A(t+2\delta t)\bullet\mathbf{r}_{ij}(t+2\delta t) + \vartheta((\delta t)^4) = 0$$

$$(2.2.3.5)$$

If we retain terms only to order $(\delta t)^2$, the quantity $\delta\lambda_{ij}$ can be solved and the data can be used to improve $\mathbf{r}_i^A(t+2\delta t)$, $\mathbf{r}_j^A(t+2\delta t)$, $\mathbf{g}_i^A(t+\delta t)$, and $\mathbf{g}_j^A(t+\delta t)$ from eq. (2.2.3.3) and eq. (2.2.3.4).

The iterative numerical procedure to get the initial $\delta\lambda_{ij}$ and update $\delta\lambda_{ij}$ for calculating $\mathbf{g}_i(t+\delta t)$ and $\mathbf{r}_i(t+2\delta t)$ follows. The equation for calculating $\mathbf{r}_i(t+2\delta t)$ can be obtained from eq. (2.2.2.1) by substituting eq. (2.2.3.2) [6]:

$$\begin{aligned}\mathbf{r}_i(t+2\delta t) &= \mathbf{r}_i'(t+2\delta t) + \frac{(\delta t)^2}{m_j}\mathbf{g}_i(t+\delta t) \\ \mathbf{r}_i'(t+2\delta t) &= \mathbf{r}_i(t+\delta t) + \delta t\mathbf{v}_i(t) + \frac{(\delta t)^2}{2m_i}\mathbf{f}_i(t+\delta t) + \frac{(\delta t)^2}{m_i}\mathbf{f}_i'(t+\delta t)\end{aligned} \tag{2.2.3.6}$$

With the initial approximations λ_{ij}=0, the constraints do not exist ($\mathbf{g}_i(t+\delta t)$=0) and $\mathbf{r}_i(t+2\delta t)$ can be computed as in a basic MD calculation since all other variables are already known. These calculated new positions at the time step $t+2\delta t$ serve as $\mathbf{r}_i^A(t+2\delta t)$ to compute the initial $\delta\lambda_{ij}$ from eq. (2.2.3.5). From $\mathbf{r}_i^A(t+2\delta t)$ and $\mathbf{r}_i^A(t+2\delta t)$, $\mathbf{g}_i^A(t+\delta t)$ and $\mathbf{g}_i^A(t+\delta t)$ can be calculated from eq. (2.2.3.4) and eq. (2.2.3.3). The data of $\mathbf{r}_i^A(t+2\delta t)$, $\mathbf{r}_j^A(t+2\delta t)$, $\delta\lambda_{ij}$, $\mathbf{g}_i^A(t+\delta t)$, and $\mathbf{g}_j^A(t+\delta t)$, can be updated by using the above equations eq. (2.2.3.6), eq. (2.2.3.5), eq. (2.2.3.4) and eq. (2.2.3.3) until the tolerance in eq. (2.2.3.7) is reached for all SHAKE bonds or after a certain number of cycles:

$$\varepsilon > | d_{ij} - \mathbf{r}_{ij}^A(t+2\delta t) | \tag{2.2.3.7}$$

SHAKE and other constraint algorithms [7] also have the ability to constrain degrees of freedom related to the motion of the bond angle or torsional angle. For example, in the rigid TIP3P water models used in MD simulations [8], two O-H bonds are constrained, in addition, a third bond between two hydrogen is created and constrained. The third constrained fictitious H-H bond artificially defines the constrained angle H-O-H, in which the bending motion of water is removed. The ability to constrain some degrees of freedom is important in MD simulations. Since their introduction in 1977 [5], constraint algorithms have been used for various applications, such as studying flexible molecules with internal constraints [9], searching for minimum energy conformations of the constrained geometry of the molecule [10], building constant pressure-constant temperature MD [11], performing free energy calculations [12] and so on.

2.2.4. Potential energy function and force field

From Newton's laws of classical mechanics, MD requires an assumed potential as input to describe the interparticle interactions in the atomic system. This interparticle potential, or potential energy function, is used to evaluate accelerations (i.e. forces) in eq. (2.2.2.1) at each step in the iterative numerical procedure from particle positions. One common approach involves the generation of a model using a set of simple equations to describe the complete categories of interactions present between atoms: each equation provides the description of each specific category of interaction. A widely used set of equations for evaluating the interatomic potentials in polypeptide or protein systems used by the AMBER package [13,14] is:

$$
\begin{aligned}
E_{potential} = & \sum_{\text{bonds}} K_r (r - r_{eq})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2 \\
& + \sum_{\text{dihedral}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}} \right]
\end{aligned}
\tag{2.2.4.1}
$$

which consists of terms representing the bond stretching and angle bending terms by a simple harmonic expression, the dihedral angle twisting term by a truncated Fourier series, the van der Waals interaction by a Lennard-Jones potential, and electrostatic interactions by a Coulombic interaction of atom-centered partial charges. In eq. (2.2.4.1),

r_{eq} is the equilibrium bond length and θ_{eq} is the equilibrium bond angle, r and θ correspond to the distance and angle for the actual bond and angle at an instantaneous step. K_r and K_θ are the respective harmonic force constants for the bond stretching and angle bending. V_n, n, ϕ, and γ represent the magnitude, periodicity, dihedral angle, and phase of the torsion respectively. For non-bonded potentials, r_{ij} is the distance between two atoms; A_{ij} and B_{ij} are constants related to van der Waals well depths and radiuses of atom i and atom j. q_i and q_j are atomic charges, and ε represents the dielectric constant of the environment. A full set of the constants (K_r, K_θ , r_{eq}, θ_{eq}, V_n, n, γ, A_{ij}, B_{ij}, q_i, q_j) for all types of atom is commonly referred to as the "force field" and is a major component of MD programs. AMBER force fields are provided in the AMBER package [13,14] or in the literature [15-19,20,21].

The empirical force field is commonly derived from quantum calculations and then is fitted from experimental data. Fundamentally, a good force field is able to model accurately the interaction energy between atoms and to be employed for a large number of biomolecules. Several different force fields are available for protein folding simulations. The familiar ones include the following (not an exhaustive list): AMBER [15-19], CHARMM [22-24], OPLS-AA/L [25,26], and GROMOS [27,28] etc. Developed by different research groups from different level quantum calculations and different experimental data sets, these different force fields may give different MD simulation results for the same polypeptide or protein. Even for the AMBER force fields, seven major generations exist including the AMBER84 [15], 86 [16], 94 [17], 96 [20], 98 [21], 99 [18] and 2003 [19] force fields. These different generations use the same mathematical form (eq. 2.2.4.1) but are characterized by different sets of parameters and may give different simulation results. The quality of the results of an MD simulation strongly depends on the accuracy of the force field.

2.2.5. Canonical ensemble MD and the Nosé-Hoover Chain method

An MD simulation generates detailed information such as atomic positions and velocities etc. Collected from a large number of steps, this detailed information can be converted into macroscopic thermodynamic or kinetic quantities (pressure, temperature etc.) [1]. Therefore, MD is also a tool for sampling from a statistical mechanical ensemble and determining ensemble averages of those thermodynamic quantities or equilibrium properties. In statistical mechanics, statistical ensembles are usually characterized by three fixed values of six thermodynamic variables including energy, E; temperature, T; pressure, P; volume, V; particle number, N; or chemical potential, μ. The macroscopic observables obtained from averaging over a large number of identical systems with each in different microscopic configuration are formulated as ensemble averages [2]. MD simulations in this dissertation involve only two fundamental ensembles: microcanonical or canonical ensembles. The microcanonical ensemble is characterized by constant particle number N, constant volume V, and constant total energy E, and is denoted as the NVE ensemble. The canonical ensemble is the NVT ensemble, which is characterized by constant particle number N, constant volume V, and constant temperature T. In MD, the thermodynamic variables to determine an ensemble can be set as control parameters that specify the conditions under which a simulation is performed [2].

The original MD simulation we describe up to now according to e.q. (2.2.2.1) deals with a microcanonical ensemble. During the simulation, the number of atoms in a polypeptide or protein is fixed to N and no addition or subtraction of atoms happens in the system. The volume is fixed and for an isolated system in standard Newtonian dynamics, the total energy is always conserved. Therefore, a dynamical trajectory of this simulated system is a series of microscopic states having constant *N*, *V*, and *E*, corresponding to a microcanonical ensemble. However, many experimental measurements are taken under the conditions of constant temperature and volume or constant temperature and pressure. Therefore, in order to obtain the equilibrium properties of a system under these conditions, it is necessary to build the corresponding ensemble for MD.

Several schemes have been proposed to design canonical ensemble MD simulations by coupling the simulation to a thermal bath. The oldest method is a momentum scaling procedure, in which the velocities of the particles are rescaled at each time step to exactly maintain a reference temperature [29]. Though its use is straightforward, this method has not been demonstrated to give the correct statistical mechanical values of the canonical ensemble's properties [30]. Berendsen [31] proposed a weak-coupling algorithm for constant temperature MD. In Berendsen coupling, a single velocity scaling factor is only used for all atoms. This algorithm ensures an appropriate total kinetic energy for the desired temperature but does not guarantee an appropriate local temperature distribution. Andersen temperature coupling provides imaginary collisions to randomize the velocities of the particles in a Maxwellian distribution that is able to reproduce the canonical ensemble [32,33]. Because of these sudden stochastic collisions, however, the MD trajectory is discontinuous in phase space.

Perhaps the most widely used thermostating algorithm to generate the canonical ensemble is the Nosé-Hoover chain (NHC) method [30,34,35]. In this method, a set of M thermostats, which successively thermostat each other, act as a heat bath coupled to the system. If the M thermostats have coordinates, η_1, …, η_M, momenta, p_{η_1}, …, p_{η_M}, and masses, Q_1, …, Q_M, the equations of motion can be expressed as [2,35]

$$\begin{aligned}
\dot{\mathbf{r}}_{\mathbf{i}} &= \frac{\mathbf{p}_i}{\mathbf{m}_i} \\
\dot{\mathbf{p}}_i &= \dot{\mathbf{f}}_i - \frac{p_{\eta_1}}{Q_1}\mathbf{p}_i \\
\dot{\eta}_k &= \frac{p_{\eta_k}}{Q_k} \qquad\qquad k = 1,\ldots M \\
\dot{p}_{\eta_k} &= F_k - \frac{p_{\eta_{K+1}}}{Q_{K+1}} p_{\eta_k} \qquad\qquad k = 1,\ldots M-1 \\
\dot{p}_{\eta_M} &= F_M
\end{aligned} \tag{2.2.5.1}$$

and the thermostat forces F_k (k=1, …, M) take the form

$$\begin{aligned}
F_1 &= \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{\mathbf{m}_i} - dNKT \\
F_k &= \frac{p_{\eta_{k-1}}^{\;2}}{Q_{k-1}} - kT \qquad\qquad k = 2,\ldots M
\end{aligned} \tag{2.2.5.2}$$

where d is the dimension of particle's variables. In a Nosé-Hoover chain of M thermostats, only the first thermostat interacts with N particles while others are additional thermostats to form a simple one-dimensional chain with the first. However, all thermostats together control the kinetic energy fluctuation of all degrees of freedom from both particle and thermostat. The demonstration of this Nosé-Hoover chain thermostating algorithm for MD simulations to generate the exact canonical ensemble and the determination of the thermostat masses etc. can be found in the references [2,30,34,35].

Statistical ensembles can impose an extra restriction on the classical microscopic states accessible with the simulated system such as energy conservation in microcanonical ensembles and constant temperature in canonical ensembles. A hypersurface is then defined if the ordinary phase space is extended to include this restricted ensemble [2]. The hypersurface in the NVE ensemble is called the constant energy surface and in the NVT ensemble is the constant temperature surface. The ergodic hypothesis gives the assumption that the simulated trajectory will cover the entire hypersurface of the system in an infinite amount of time [2]. It is worthwhile to point out that most algorithms, such as the Nosé-Hoover chain method, are demonstrated to generate the correct corresponding ensemble based on the assumption of ergodicity for the simulated system.

2.2.6. Solvent environment and generalized Born / surface area (GB/SA) implicit solvent model

2.2.6.1. Solvent environment

Molecular dynamic simulations are generally used to provide atomistic information about biomolecules in solvent on a time scale that is not feasible with current experimental techniques. Since the natural environment for biomolecules such as proteins and DNA is in an aqueous solution, one must consider running most molecular dynamic simulations in water or some type of aqueous solution. In order for a molecular model to be as accurate as possible, many simulations are therefore carried out in the presence of explicit solvent molecules. In these simulations, the biomolecule solute is immersed in a large box of water molecules with periodic boundary conditions. The water molecules are modeled by several types such as the 3-site TIP3P, 4-site TIP4P or 5-site TIP5P, which are parameterized to be used in MD and reproduce many experimental water properties with reasonable computational cost [8,36]. As an example, in the TIP3P water model, two O-H bonds are constrained and a third bond is created between the two hydrogen to form a rigid molecule [8]. The periodic boundary conditions [1,37] eliminate the boundary problem in which water molecules on the surface of an isolated box can experience quite different forces from waters in the bulk. By creating the periodic image of the real box in each direction, the boundary conditions allow the water molecules to move freely between the original box and its neighboring images and preserve the forces acting on all the atoms. Furthermore, cutoff methods are used for truncation of the long-range interactions between remote water molecules and the solute or among water molecules themselves [1]. A radius from an atom imposes an upper distance limit on the calculation of its non-bonded interactions from other atoms and defines a cutoff distance. Usually, this cutoff distance must be long enough (at least 10 Å) in order to obtain an accurate calculation of long-range electrostatic interactions. In a pair-wise potential, the number of these nonbonded interactions increases with the square of the number of atoms. The implementation of the nonbonded cutoff after certain distance can save significant computational costs in an explicit solvent MD simulation.

Although the calculations employing explicit solvent models are accurate enough to reproduce a variety of experimental data, these calculations usually involve thousands of discrete water models and are very computationally demanding. In fact, several orders of magnitude more computer time is generally required in these explicit solvent calculations compared to corresponding gas phase calculations on the same biomolecule [38]. On the other hand, if the same CPU time is considered, these explicit solvent calculations can only be employed to investigate the dynamics of a very small portion of a large biomolecule (e.g. protein). As an example, we simulate a small terminal-blocked polyalanine of 13 residues (Ala13). This polypeptide has 142 atoms and is 48 Å at the fully extended length. A cubic box of water molecules as large as 70 Å ×70 Å ×70 Å must be built in order for a 10 Å cutoff between the surface and any atom of the polypeptide. This results in a system of around 3300 water molecules and 10000 atoms for this small polypeptide Ala13. The computational cost for this system will be equal to that from a gas phase simulation for a very large protein of around 600 residues. Therefore, the explicit solvent simulations are usually lengthy and costly, even with today's powerful supercomputers.

2.2.6.2. Generalized Born / surface area (GB/SA) implicit solvent model

The solution to overcome this limitation of large computational cost is to replace the explicitly represented solvent by continuum models of implicit solvation effects [39-45]. In continuum solvation models, the solvent is usually treated as a continuous medium to surround the solute at its van der Waals surface. A modified or new set of interactions between the solute atoms is constructed to simulate the relevant features (such as electrostatic effects) from the surrounding solvent, and mimic the entire medium having the average properties of the real solvent. Because the implicit solvation models do not include any atom from the bulk solvent, these models can provide approximate solvation effects and dramatically reduce computational time.

A variety of implicit solvation models have been described over the last two decades. The earliest empirical model is proposed by Daggett and coworkers [46,47], who use a distance dependent dielectric function to modify the electrostatic interactions between solute atoms for mimicking the presence of the high dielectric water. This function is included in the potential energy function for AMBER force fields shown as the eq. (2.2.4.1) in which ε can be assign with $\mathbf{r}_{ij}$ or $4\mathbf{r}_{ij}$. This simple distance dependent dielectric model lacks accuracy to reproduce the real solvation effects but shows no extra computational cost compared to the gas phase simulations and has been widely used before [48-51]. On the other hand, the statistical continuum electrostatic theory successfully provides accurate representations of solvation and solvent-mediated interactions through a finite-difference Poisson-Boltzmann (PB) electrostatic equation in a multiple dielectric model [44,45,52-55]. However, the calculations of the numerical solution of PB equation are also too costly to be directly incorporated into MD at each integration step in a routine study [56,57]. Nevertheless, the classical continuum theory serves as the cornerstone for the development of several fast analytic implicit solvent models recently. Among these approximate analytic versions, the generalized Born (GB)/ surface area (SA) solvation model is the most attractive proposed model to date [49-51,58-60]. This model is often able to reproduce the solvation energies and individual charge-charge interactions given by the computationally intensive Poisson-Boltzmann approach for a variety of biomolecules [38,54,61-68]. A little more detailed description from the literature of this GB/SA model as follows.

In the GB/SA model [69], the total energy function E_{total} (eq 2.2.6.2.1) in implicit solvent environment generally includes the energy terms of the solute, $E_{potential}$, and the solvation energy, G_{sol}, for the interaction of the protein with the surrounding solvent. The G_{sol} term is traditionally considered as a sum of a solvent-solvent cavity term (G_{cav}), a solute-solvent van der Waals term (G_{vdW}), and a solute-solvent electrostatic polarization term (G_{pol}) [69].

$$\begin{aligned} E_{total} &= E_{potential} + G_{sol} \\ G_{sol} &= G_{cav} + G_{vdw} + G_{pol} \end{aligned} \qquad (2.2.6.2.1)$$

A combination of the first two terms for the solvation energy is linearly related to solvent-accessible surface area of the atomic types of the solute (eq 2.2.6.2.2). In eq. (2.2.6.2.2), the surface area is determined from the locus of points swept out by the center of the solvent sphere when rolling over the van der Waals surface of a protein [70]. SA_k is the total solvent-accessible surface area of atoms of type k, σ_k is an empirical atomic solvation energy parameter, and the summation extends over all atomic types k.

$$G_{cav} + G_{vdw} = \sum_k \sigma_k SA_k \qquad (2.2.6.2.2)$$

We use the linear combination of pair-wise overlaps (LCPO) approach [71] to calculate the accessible surface area SA_k while a preliminary value of +7.2 [69,72] or +5.0 [54,73] cal/(mol - Å^2) of the surface tension σ_k (dependent upon the GB formula selected) was used for all atom types.

The main equation in the LCPO method used to compute the accessible surface area of atoms *i* (A_i) can be expressed as [71]

$$A_i = P_1 S_i + P_2 \sum_{j \in N(i)} A_{ij} + P_3 \sum_{\substack{j,k \in N(i) \\ k \in N(j) \\ k \neq j}} A_{jk} + P_4 \sum_{j \in N(i)} A_{ij} \left(\sum_{\substack{j,k \in N(i) \\ k \in N(j) \\ k \neq j}} A_{jk} \right) \qquad (2.2.6.2.3)$$

Where *N(i), N (j)* stands for the neighbor list of atoms that overlap with atom i, and atom j, respectively. S_i is the surface area of the isolated sphere *i*. A_{ij} represents the area of sphere *i* buried inside sphere *j*. P_1, P_2, P_3, and P_4 are overlap parameters obtained by multiple linear regressions against numerical surface areas of a set of test compounds. The isolated individual area S_i and the overlap area A_{ij} take the following forms [71,74]:

$$S_i = 4\pi \mathbf{r}_i^2$$
$$A_{ij} = 2\pi \mathbf{r}_i \left(r_i - \frac{d_{ij}}{2} - \frac{r_i^2 - r_j^2}{2d_{ij}} \right) \qquad d_{ij} \le \mathbf{r}_i + \mathbf{r}_j \qquad (2.2.6.2.4)$$

in which $\mathbf{r}_i$ and $\mathbf{r}_j$ are atomic radii (the sum of van der Waals radii and solvent-probe radius of 1.4Å), and d_{ij} is the internuclear distance. A_{ij} is only computed over pairs of overlapping atoms and is set to zero for all pairs of nonoverlapping atoms. On the right of eq. (2.2.6.2.3), the second term, involving summing over A_{ij}, counts the negative effect of pair-wise overlaps of sphere *i* with its all neighbors. The third and fourth terms summing over A_{jk}, compensate for the over-subtraction of overlaps between *i*'s neighbors (*j* and *k*).

The LCPO method utilizes united atoms to compute the solvent-accessible surface areas. The atom types are determined based on atomic number, hybridization, and number of bonded neighbors. The overlap parameters and atomic van der Waals radii of each atom type, as well as the first and second derivatives of the energy with respect to atomic positions can be found in the literature [71].

The G_{pol} term is calculated from the generalized Born (GB) equation. The original from was introduced by Still and coworkers [69]:

$$G_{pol} = -\frac{1}{2}\left(\frac{1}{\varepsilon_p} - \frac{1}{\varepsilon_w}\right)\sum_{ij}\frac{q_i q_j}{f_{gb}}$$

$$f_{gb} = \left[\mathbf{r}_{ij}^2 + \alpha_i \alpha_j \exp\left(-\frac{\mathbf{r}_{ij}^2}{4\alpha_i\alpha_j}\right)\right]^{\frac{1}{2}} \tag{2.2.6.2.5}$$

Where ε_p represents the low dielectric value inside of the protein, ε_w is the water (or other solvent) dielectric constant, r_{ij} is the separation distance of particles i and j, q_i, q_j are their charges, and α_i, α_j are the corresponding effective Born radius [37], respectively. The effective Born radius describes the average distance form a charge to the continuum dielectric boundary. In eq. (2.2.6.2.5), f_{gb} is the complex function of the distance r_{ij} and the effective Born radii α_i and α_j. It interpolates between the α_i (or α_j) of small r_{ij} and r_{ij} itself at large distances.

Several different research groups [38,62,63,65,72,73,75-80] have done a series of works to extend this general formalism and establish a parameterization consistent with widely used force fields such as AMBER [17,18,20], CHARMM [23,24], Jorgensen's OPLS [25], and AM/SM [79,80] etc. During their parameterization, the generalized Born equation and f_{gb} function are all modified a little from the eq. (2.2.6.2.5). Meanwhile, the parameterization is usually capable of reproducing accurate electrostatic salvation free energies based on the target force field compared to the finite difference solution to the PB equation [52]. Here I will briefly describe 4 GB models parameterized for use with the AMBER force field.

Tsui and coworkers proposed a modification that incorporates a Debye-Hückel term $e^{-\kappa f_{gb}}$ in the generalized Born equation to account for salt effects at low salt concentrations [73,75]:

$$
\begin{aligned}
G_{pol} &= -\frac{1}{2}\left(1-\frac{e^{-\kappa f_{gb}}}{\varepsilon_w}\right)\sum_{ij}\frac{q_i q_j}{f_{gb}} \\
f_{gb} &= \left[\mathbf{r}_{ij}^2 + \alpha_i\alpha_j \exp\left(-\frac{\mathbf{r}_{ij}^2}{4\alpha_i\alpha_j}\right)\right]^{\frac{1}{2}}
\end{aligned}
\qquad (2.2.6.2.6)
$$

in which κ is the Debye-Hückel screening parameter. In eq. (2.2.6.2.6), the low dielectric value inside of the protein is considered one and the f_{gb} function remains unchanged. On the other hand, Jayaram and coworkers suggest the modification in f_{gb} function while keeping the original generalized Born equation, taking the forms [72,77]:

$$
\begin{aligned}
& G_{pol} = -\frac{1}{2}\left(1-\frac{1}{\varepsilon_w}\right)\sum_{ij}\frac{q_i q_j}{f_{gb}} \\
& f_{m1gb} = \left[\mathbf{r}_{ij}^2 + \alpha_i\alpha_j \exp\left(-\frac{\mathbf{r}_{ij}^2}{2\alpha_i\alpha_j}\right)\right]^{\frac{1}{2}} \\
& f_{m2gb} = f_{m1gb}\frac{\varepsilon\gamma-\gamma}{\varepsilon\gamma-1} \\
& \textit{where} \\
& \qquad \gamma = 1-\left(\frac{\varepsilon_w-4}{2\varepsilon_w}\right)(\beta^2+2\beta+2)\exp(-\beta) \\
& \qquad \beta = (0.4\mathbf{r}_{ij}+\alpha_{ij})
\end{aligned}
\qquad (2.2.6.2.7)
$$

For a biomolecular solute containing multiply charged particles and an arbitrarily shaped molecular surface, the effective Born radii of any charged atom is dependent upon the positions and volumes of all other atoms in the solute. It is estimated from a numerical integration procedure in the original GB/SA model [69]. Hawkins and coworkers proposed an analytical approximation — a pair-wise descreening procedure—to estimate the effective Born radii from a sum over atom pairs [79,80]:

$$\begin{aligned}
\alpha_i^{-1} &= \rho_i^{-1} - \sum_{j\neq i} g(r_i, r_j, \rho_i, \rho_j) \\
&= \rho_i^{-1} - \frac{1}{2}\sum_{j\neq i}\left[\frac{1}{L_{ij}} - \frac{1}{U_{ij}} + \frac{r_{ij}}{4}\left(\frac{1}{U_{ij}^2} - \frac{1}{L_{ij}^2}\right) + \frac{1}{2r_{ij}}\ln\frac{L_{ij}}{U_{ij}} + \frac{\rho_j^2}{4r_{ij}}\left(\frac{1}{L_{ij}^2} - \frac{1}{U_{ij}^2}\right)\right]
\end{aligned}$$

where

$$\begin{aligned}
L_{ij} &= 1 && \textit{if} && r_{ij} + \rho_j \le \rho_i \\
L_{ij} &= \rho_i && \textit{if} && r_{ij} - \rho_j \le \rho_i < r_{ij} + \rho_j \\
L_{ij} &= r_{ij} - \rho_j && \textit{if} && \rho_i \le r_{ij} - \rho_j
\end{aligned}$$

and

$$\begin{aligned}
U_{ij} &= 1 && \textit{if} && r_{ij} + \rho_j \le \rho_i \\
U_{ij} &= r_{ij} + \rho_j && \textit{if} && \rho_i < r_{ij} + \rho_j
\end{aligned}$$

(2.2.6.2.8)

Here ρ_i is an intrinsic radius for atom *i* and takes the form:

$$\rho_i = S_i(R_i + b_{offset}) \qquad (2.2.6.2.9)$$

where R_i is the van der Waals radius and S_i is the screening parameter. The b_{offset} is not used in the original formulation introduced by Hawkins and coworkers [79,80], nor is it used in the modified GB version by Jayaram and coworkers [72,77]. Tsui and coworkers, however, use this modification. We use GB1 to represent the standard pair-wise generalized Born model described by eq. (2.2.6.2.6) [73,75]. GB2 uses the same eq. (2.2.6.2.6) but introduces an extra packing correction factor λ for the function $g(r_i, r_j, \rho_i, \rho_j)$ in eq. (2.2.6.2.8) [76]. We use GB4 to represent the modified generalized Born model shown in eq. (2.2.6.2.7) using f_{m2gb} function and GB3 using the f_{m1gb} function [72,77]. The input parameters follow the standard values for each GB/SA implicit model developed.

MD procedures require first derivatives of the G_{pol} energy with respect to the Cartesian coordinates of the atoms. The derivatives did not appear in the original literature. It is not a tremendous task to make those derivations, but one still needs to work with effort and caution. The implementation of these GB implicit models in the MD program package is difficult, but vectorized pseudocode for the algorithms to compute G_{pol} and the derivative forces is available [73,81]. The GB/SA implicit models involve two N^2 pair-wise calculations, one to calculate the effective Born radius by eq. (2.2.6.2.8) and another to compute the G_{pol} derivative forces and G_{pol} energy. Under the same conditions, the simulation in the GB/SA implicit solvent model is usually 5 – 6 times slower than the gas phase simulation.

2.2.7. Removal of translational modes and seamless restart format

2.2.7.1. Removal of translational modes

Now, we have written molecular dynamics computer programs and can run conventional MD simulations (NVE and NVT), even for a large, complex biomolecular system. However, we still need to consider several issues. First, the three translation modes of molecular motion should be removed.

The algorithm to remove translation modes of the simulated system is relatively simple. At each step, the translation motions only affect the center of mass position but not the relative positions between atoms of the system. The coordinates of the center of mass at any instant in time can be expressed as [82]:

$$\mathbf{r}_{center} = \frac{\sum_{i}^{N} m_i \mathbf{r}_i}{\sum_{i}^{N} m_i} \qquad (2.2.7.1.1)$$

Therefore, the updated positions of atoms after removal of translational motion are related to the original positions by

$$\mathbf{r}_i^{remove} = \mathbf{r}_i - \mathbf{r}_{center} \qquad (2.2.7.1.2)$$

The translation motions can be removed at each step or after a certain number of steps.

2.2.7.2. Seamless restart format

The second issue is restarting the simulation.

From the numerical integration scheme in MD (Fig. 2.2.2.1), we know that the simulation can be started from a full set of a system's initial positions and velocities. The restarted simulation can also begin from a full set of positions and velocities at any instant in time. This data can be recorded in an output file at the stop step from the former simulation run. Restarted simulations in AMBER use this method. However, this method causes a minor problem. The restarted trajectories are different from the original simulation.

The default initial parameters cause the discrepancy when restarting a simulation from a trajectory. According to the velocity-Verlet integrator (eq. 2.2.2.1), the old forces are needed to update the velocities at the current step allowing for the positions at the next step to be calculated. The old forces are defaulted to zero for the initial step, but they should not be zero at the stop step in the former simulation. If the forces related to the status of the stop step are stored in a binary file and are restored upon restart, the trajectory can be recovered exactly. This restart method has been incorporated within our program.

This restart method is not universally seamless. If a simulation is restarted on another supercomputer with a different operating system, the binary data file cannot be read and the text data file must be used. The text file will cause some round off error problems on the supercomputer, and the restarted trajectory will not be identical.

2.3. Current molecular dynamics simulation methods

2.3.1. Simulated annealing (SA)

Simulated annealing is one of the most powerful algorithms used in various applications for global minimum searches [9,59,83-88]. In the original simulated annealing algorithm proposed by Kirkpatrick etc. [83], a high initial temperature is introduced and gradually reduced when the system moves. The move is accepted if it is downhill in the energy surface. If it is uphill, the move is only accepted when it has a larger acceptance probability than a randomly generated number in the interval (0,.1) [83]. The Metropolis criterion (eq. 2.3.1.1) is used to calculate the acceptance probability $p(\Delta E)$ [89]:

$$p(\Delta E) = \begin{cases} 1, & for\, \Delta E \le 0, \\ \exp(-\dfrac{\Delta E}{k_B T}), & for\, \Delta E > 0, \end{cases} \tag{2.3.1.1}$$

Where k_B is the Boltzmann constant, T is the temperature, and ΔE is the energy change of the move set. The move set can help the system to surmount large energy barriers at high temperatures and force the system to freeze into an energy minimum at low temperatures [84]. The more slowly the temperature decreases, the more likely is the chance of locating the global minimum. Unfortunately, without extremely long simulations with extremely slow cooling procedures, the system often settles into a local minimum rather than the global minimum [84]. As a result, iterations of heating and cooling is necessary for simulated annealing to find the global potential energy minimum of complex system such as proteins.

Generally, simulated annealing can be considered as a technique to use alternate heating and cooling to manipulate the kinetic energy of the system. Increasing the kinetic energy lets high energy barriers be crossed, while decreasing the kinetic energy allows the energy minima to be located. The global energy minimum has the highest probability of trapping the system during repeated heating and cooling.

2.3.2. Generalized-ensemble algorithms and replica exchange method (REM)

2.3.2.1 Generalized-ensemble algorithms

From Hamiltonian mechanics (eq. 2.2.1.3), each state (**r**, **p**) of potential energy E with the Hamiltonian $H(\mathbf{r},\mathbf{p})$ in a canonical ensemble at temperature T is weighted by the Boltzmann factor:

$$W_B(E;T) = \exp(-\beta H(\mathbf{r},\mathbf{p})) \tag{2.3.2.1.1}$$

where β is the inverse temperature defined by $\beta=1/k_BT$ (k_B is the Boltzmann constant). Then, the canonical probability distribution of potential energy, $P_B(E;\, T)$, is given by [90,91]

$$P_B(E;T) \propto n(E)W_B(E;T) \tag{2.3.2.1.2}$$

Here $n(E)$ is the density of states. Since the Boltzmann factor decreases exponentially as E increases but $n(E)$ increases, the canonical ensemble should generate a Gaussian probability distribution that has a maximum at the average potential energy at temperature T [90,92]. On the other hand, each state can be weighted by a non-Boltzmann weight factor $W_{\bar{B}}(E)$ in such a way that a uniform probability distribution of potential energy $P_{\bar{B}}(E)$ is obtained [90,91]:

$$P_{\bar{B}}(E) \propto n(E)W_{\bar{B}}(E) = const \qquad (2.3.2.1.3)$$

This flat probability distribution implies that a random walk in the potential energy space can be realized from this weighting technique.

Generalized-ensemble algorithms utilize this non-Boltzmann probability weighting factor in Monte Carlo or MD simulations to perform random walks in energy space [90,91]. The random walks allow the simulation to cross any energy barrier and then to sample more conformational space than by conventional NVT methods [90,91]. Dependent upon the ensemble characterizing the one-dimensional space in which the random walk is performed, three generalized-ensemble methods exist, including the multicanonical algorithm [93-99], simulated tempering [98-104], and 1/K-sampling [99,105,106]. A simulation using the multicanonical algorithm, which is perhaps the most well known generalized-ensemble method, directly performs a random walk in potential energy space. Simulated tempering and 1/K-sampling perform random walks in temperature and entropy space, respectively; these random walks, however, also induces a random walk in potential energy space and allows the simulation to escape from any local minimum-energy states [90,91].

Unfortunately, the non-Boltzmann weight factor in generalized-ensemble methods is not usually known *a priori* and has to be determined by iterations of short trial simulation runs [90,91]. These preliminary simulations are iterated at a sufficiently high temperature, following a practical procedure to secure a nearly flat potential energy distribution at the reference temperature. The description of this iterative process can be found elsewhere [107,108]; however, this process can be time-consuming and very tedious for systems as complex as proteins with myriad local minimum energy states [90,91].

2.3.2.2. Replica exchange method (REM)

Developed as an extension of simulated tempering, the replica-exchange method (REM) greatly alleviates the difficulty of weight factor determination [90,92,109-113]. In this method, a number of non-interacting copies (or replicas) of the original system are simulated independently and simultaneously at constant temperatures. After a certain number of steps, pairs of replicas exchange temperatures with a Metropolis transition probability. The non-Boltzmann weight factor is just the product of Boltzmann factors for each replica and so it is known [90,91].

A simplified description is given here to show how the replica-exchange method works in MD simulations. Suppose that only two non-interacting replicas (*i*, *j*) of the original system in the canonical ensemble at two different temperatures (T_m, T_n) exist in the generalized ensemble of a REM simulation. Initially, the replica *i* corresponds to T_m, and the replica *j* corresponds to T_n. If we use *x* to label a state in phase space, it is specified by a complete set of coordinates (**r**) and momenta (**p**) of the system's N atoms. Then, the $x_{T_m}^{i}$, $x_{T_n}^{j}$ can be used to represent a state for replica *i* at T_m and replica *j* at T_n. respectively:

$$\begin{aligned} x_{T_m}^{i} &\equiv (\mathbf{r}^{i}, \mathbf{p}^{i})_{T_m} \\ x_{T_n}^{j} &\equiv (\mathbf{r}^{j}, \mathbf{p}^{j})_{T_n} \end{aligned} \qquad (2.3.2.2.1)$$

Let X = ($x_{T_m}^i$, $x_{T_n}^j$) stand for a combined "state" in this generalized ensemble. Because the replicas i and j are non-interacting, the weight factor for the state X can be simply given by the product of Boltzmann factors for each replica [90,92]:

$$W_{REM}(X) = \exp\left\{\beta_{T_m} H(\mathbf{r}^i, \mathbf{p}^i)\beta_{T_n} H(\mathbf{r}^j, \mathbf{p}^j)\right\} \tag{2.3.2.2.2}$$

The exchange of the replica pairs for temperatures in REM indicates a transition

$$X = (x_{T_m}^i, x_{T_n}^j) \rightarrow X^{'} = (x_{T_n}^{i^{'}}, x_{T_m}^{j^{'}}) \tag{2.3.2.2.3}$$

This transition only deals with the momenta exchange of replicas i and j. Meanwhile, the coordinates of the replicas always remain the same. During this transition, the momenta for the replicas are updated by the expressions [90,92]:

$$\begin{aligned} \mathbf{p}^{i^{'}} &\equiv \sqrt{\frac{T_n}{T_m}}\mathbf{p}^i \\ \mathbf{p}^{j^{'}} &\equiv \sqrt{\frac{T_m}{T_n}}\mathbf{p}^j \end{aligned} \tag{2.3.2.2.4}$$

That is, the velocities of all the atoms in the replicas are rescaled uniformly by the square root of the ratio of the two temperatures. In this way, the temperatures are conserved.

In order for this exchange process to converge towards a canonical thermal equilibrium distribution, a two-side balance condition on the transition probability $w(X \rightarrow X^{'})$ is required [91]:

$$\frac{W_{REM}(X)}{Z} w(X \rightarrow X^{'}) = \frac{W_{REM}(X^{'})}{Z} w(X^{'} \rightarrow X) \tag{2.3.2.2.5}$$

where Z is the partition function of the entire system in this generalized ensemble. From the eq. (2.3.2.2.2), (2.3.2.2.4), (2.3.2.2.5) and the Hamiltonian definition eq. (2.2.1.3), we can easily obtain [90,92]:

$$\frac{w(X \to X^{'})}{w(X^{'} \to X)} = \exp(-\Delta),$$
$$where \qquad \Delta = (\beta_{T_m} - \beta_{T_n})\left[(E(\mathbf{r}^j) - E(\mathbf{r}^i)\right] \tag{2.3.2.2.6}$$

The transition balance can be satisfied by using the usual Metropolis criterion [90,92]:

$$w(X \to X^{'}) \equiv w(x_{T_m}^i \mid x_n^j) = \begin{cases} 1, & for\,\Delta \leq 0, \\ \exp(-\Delta), & for\,\Delta > 0, \end{cases} \tag{2.3.2.2.7}$$

Where $w(x_{T_m}^i \mid x_n^j)$ is used to indicate explicitly the pair of replicas to be exchanged in the transition process [90,92]. Note that the transition probability $w(X \to X^{'})$ decreases exponentially with the decrease of the βs, or the increase of the temperatures.

In conventional REM simulations, a large number of *M* non-interacting replicas are usually required (*M* > 2), at *M* different temperatures. The replicas are arranged so that there is always exactly one replica at each temperature [90,92]. However, due to the more complicated multiple exchanges possibly among *M* different temperatures, a permutation function of the temperatures towards the replicas has to be introduced to keep track of the one-to-one correspondence between replicas and temperatures at any step. The general details of REM are described in the literature [90,92].

If M different temperatures follow the order $T_1 < T_2 < \cdots < T_M$, an REM simulation with M non-interacting replicas is then realized by alternately performing the following two steps [90,92]:

1) Each replica in a canonical ensemble is simulated simultaneously and independently at its corresponding temperature for a number of MD steps.
2) Pairs of replicas at neighboring temperatures are exchanged with the transition probability in eq. (2.3.2.2.7).

Note that in step 2 only pairs of replicas corresponding to neighboring temperatures are exchanged. This approach is very reasonable because the acceptance ratio of the exchange decrease exponentially with the increase of the temperature updates [90,92]. Further, an appropriate temperature distribution has to be set for M different temperatures in order to secure the optimal performance of REM for achieving good thermodynamic equilibrium properties. An iterative procedure to obtain these optimal temperatures can be found elsewhere [109].

While the canonical expectation value of a physical quantity such as average potential energy and specific heat etc., at temperature T_m (m=1, …, M), can be easily calculated by the usual arithmetic means [90,92], the expectation value at any intermediate temperature is usually achieved by multiple-histogram reweighing techniques [114,115]. One can find those related mathematical equations to calculate the canonical expectation values from REM simulations in the literature [90,92].

2.3.3. Locally enhanced sampling (LES) and mean field approximations

Since its origin several decades ago, molecular dynamics has been used as a computation technique for solving problems in a wide variety of areas: protein folding [116-121] (the main issue here), DNA combination [122,123], ligand-substrate binding [124-128], ligand diffusion [129-133] and chemical reactions [134-138] etc. Among these different purposes, a common situation likely appears that the primary interest is only a small part of the whole system; the other major part is neither important for solving the problem nor shows major variations of its thermodynamic properties. For example, to find ligand diffusion pathways, the ligand molecule is usually very small, while the bio-molecule substrate (protein) is very large. Further, during the ligand diffusion process, the substrate usually does not change its overall structure although minor alterations are likely to exist in its local structure. For this situation, a decision about how much computational effort is required to solve the problem exists in molecular dynamics simulations. While simulation of the whole system always takes a large amount of computer time and then one cannot efficiently sample the conformational space of the small part of direct interest, the simulation of only this small subsystem with the restraint of the larger subsystem (fixed) may fail to give an accurate solution. Minor variations in the local structure of the larger subsystem usually do have an effect on the behavior of the smaller subsystem.

The locally enhanced sampling (LES) method [139] has been developed to deal with this situation. In LES, the small part of the system of primary interest is copied several times while the rest of the system remains uncopied. The simulation then calculates a bundle of trajectories for the smaller subsystem while generating only a single trajectory for the larger subsystem [139]. In this way, considerable savings in computer time are achieved. The sampling of the interesting part is enhanced multiple times, but the total simulation time increases only a little compare to simulations of one single copy. This approach creates an unphysical system in which the large uncopied subsystem acts as the bath and simultaneously interacts with several copies of the small, interesting subsystem. As a result, it feels the average of the forces (the mean field) contributed by the copied atoms. The copied atoms, on the other hand, feel the same force that the corresponding real atom would feel from the single bath [139].

The LES equations of motion can be generated by using the time-dependent Hartree (TDH) approximation and the Liouville formulation of classical mechanics. In the phase space TDH approximation related to the classical dynamics, it was assumed that the classical phase-space density function $\rho(\mathbf{P}, \mathbf{Q}, t)$ can be approximated as a product of the copied subsystem's density and the bath's density [139]:

$$\rho(\mathbf{P},\mathbf{Q},t) = \rho_s(\mathbf{P}_s,\mathbf{Q}_s,t)\rho_b(\mathbf{P}_b,\mathbf{Q}_b,t) \tag{2.3.3.1}$$

Where **P** and **Q** are, respectively, canonical momenta and coordinates representing the state of all degrees of freedom in the system and *s*, *b* denotes the subsystem of interest and the bath, respectively. In order to derive the individual equations of motion for the subsets *s* and *b*, their density functions $\rho_s(\mathbf{P}_s,\mathbf{Q}_s,t)$ and $\rho_b(\mathbf{P}_b,\mathbf{Q}_b,t)$ are expanded as δ-functions [139]:

$$\begin{aligned} \rho_s(\mathbf{P}_s,\mathbf{Q}_s,t) &= \sum_{k=1}^{C} \omega_{sk}\delta(\mathbf{P}_{sk}(t),\mathbf{Q}_{sk}(t)) \\ \rho_b(\mathbf{P}_b,\mathbf{Q}_b,t) &= \delta(\mathbf{P}_b(t),\mathbf{Q}_b(t)) \end{aligned} \tag{2.3.3.2}$$

Here each delta function represents the sub-state for the various copies or the single bath. The *C* is the number of copies and ω_{sk} is a weighting function, which is generally taken to be 1/*C* in LES. With these phase-space distributions of eq. (2.3.3.1) and (2.3.3.2), the LES equations of motion can then be obtained by using the Liouville equation [140]:

$$\frac{\partial\rho}{\partial t} - \sum_i \left(\frac{\partial H}{\partial \mathbf{Q}_i}\frac{\partial\rho}{\partial \mathbf{P}_i} - \frac{\partial H}{\partial \mathbf{P}_i}\frac{\partial\rho}{\partial \mathbf{Q}_i} \right) = 0 \tag{2.3.3.3}$$

Here *H* is the classical Hamiltonian of the system. The details can be found elsewhere [139]. The resulting equations are [139]:

$$
\begin{aligned}
\dot{\mathbf{q}}_{i,k} &= \frac{\partial H(\mathbf{p}_{i,k},\mathbf{q}_{i,k},\mathbf{P}_j,\mathbf{Q}_j)}{\partial \mathbf{p}_{i,k}} \\
\dot{\mathbf{p}}_{i,k} &= -\frac{\partial H(\mathbf{p}_{i,k},\mathbf{q}_{i,k},\mathbf{P}_j,\mathbf{Q}_j)}{\partial \mathbf{q}_{i,k}} \\
\dot{\mathbf{Q}}_j &= \sum_{k=1}^{C}\omega_k \frac{\partial H(\mathbf{p}_{i,k},\mathbf{q}_{i,k},\mathbf{P}_j,\mathbf{Q}_j)}{\partial \mathbf{P}_i} \\
\dot{\mathbf{P}}_j &= -\sum_{k=1}^{C}\omega_k \frac{\partial H(\mathbf{p}_{i,k},\mathbf{q}_{i,k},\mathbf{P}_j,\mathbf{Q}_j)}{\partial \mathbf{Q}_i}
\end{aligned}
\tag{2.3.3.4}
$$

The index *i* and *j* label the particle in the copied subsystem and bath respectively. The index *k* labels the copy. The lower case variables refer to the copied subsystem *s* and the uppercase variables refer to the bath *b*. A bundle of trajectories corresponds to the s subsystem, each one of which moves in the potential determined by the coordinates of the bath. The single trajectory of the bath is solved by the effective potential obtained by averaging over those simultaneous trajectories of subsystem *s* (ω_k =1/C) [139].

Because of its desirable computational advantage and great practical utility, the LES approach quickly captured the interest of computational scientists. In the years after it was originally proposed, a family of mean field methods have been developed, including conformational matrix (CM), mean-field algorithm (MFA), mean-field theory (MFT), and self-consistent mean field (SCMF) [141]. Over the last decade, researchers have employed these LES or related mean field methods in a wide variety of optimization areas including cofactor-enzyme binding [142-144], non-equilibrium studies [145,146], free energy calculations [147,148], and global minimum searching [48,87,149]. The simulation results from this variety of applications indicate that the methods based on mean field theory are advantageous in classical or quantum dynamics. As an example, when the LES method is applied to model side chains in peptides and proteins for energy minima searches, it is found that LES simulations enhance the sampling of the interesting side chains, and they also facilitate the conformational transition of the whole peptide or protein system [87]. An analysis of the potential energy landscape showed that the energy barriers between potential energy wells are reduced during the LES trajectories [87]. Later, a reduction in energy barrier to make a conformational transition easier became the standard in LES and other mean field approximations [48,88,141,149].

Although LES and related optimization methods based on mean field approximations show several significant advantages compared to conventional molecular dynamics methods, they suffer limitations in their ability to model a Newtonian dynamical process [150-152]. The trajectories generated with such methods do not correspond to physically possible trajectories. Also, when an LES simulation's trajectories are used to calculate thermodynamic properties, the result is found to violate the equipartition of energy theorem due to force averaging from the subsystem copies to the bath [150]. This violation can cause the "temperature disparity problem", which is a failure of the subsystem and bath temperatures to reach the same equilibrium value [151,152]. Furthermore, due to the same force averaging approximation, local minima in LES simulations are not the same as those on the original energy surface, although the global potential energy minimum remains the same [153]. These limitations indicate that the LES mean field approximation should be used carefully in molecular dynamics simulations despite its great practical utility.

2.4. Acknowledgements

The programming work was supported by the Oklahoma Center for the Advancement of Science and Technology through OCAST award No. HR01-148, and by the Chemical Sciences, Geosciences and Bioscience Division, Office of Basic Energy Science, Office of Science, US Department of Energy through Grant no. DE-FG03-01ER15164. I am grateful to Christopher Adam Hixson for his collaboration in writing the software described in this work. We would like to thank Dr. Henry Neeman for programming advice. I would also like to thank Prof. David A. Case, Prof. Carlos Simmerling, and Prof. B. Jayaram for helpful discussions during the programming of GB/SA implicit solvent models.

2.5. Bibliography

(1) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, 1987.

(2) Tuckerman, M. E.; Martyna, G. J. *Journal of Physical Chemistry B* **2000**, *104*, 159.

(3) Verlet, L. *Physical Review* **1967**, *159*, 98.

(4) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *Journal of Chemical Physics* **1982**, *76*, 637.

(5) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *Journal of Computational Physics* **1977**, *23*, 327.

(6) Palmer, B. J. *Journal of Computational Physics* **1993**, *104*, 470.

(7) Andersen, H. C. *Journal of Computational Physics* **1983**, *52*, 24.

(8) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *Journal of Chemical Physics* **1983**, *79*, 926.

(9) Bassolino-Klimas, D.; Tejero, R.; Krystek, S. R.; Metzler, W. J.; Montelione, G. T.; Bruccoleri, R. E. *Protein Science* **1996**, *5*, 593.

(10) Levitt, M. *Journal of Molecular Biology* **1983**, *170*, 723.

(11) Ferrario, M.; Ryckaert, J. P. *Molecular Physics* **1985**, *54*, 587.

(12) Elber, R. *Journal of Chemical Physics* **1990**, *93*, 4312.

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

(14) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham III, T. E.; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Ferguson, D. M.; Radmer, R. J.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *AMBER 5*; University of California-San Francisco: San Francisco, 1997.

(15) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. *Journal of the American Chemical Society* **1984**, *106*, 765.

(16) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *Journal of Computational Chemistry* **1986**, *7*, 230.

(17) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *Journal of the American Chemical Society* **1995**, *117*, 5179.

(18) Wang, J.; Cieplak, P.; Kollman, P. A. *Journal of Computational Chemistry* **2000**, *21*, 1049.

(19) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *Journal of Computational Chemistry* **2003**, *24*, 1999.

(20) Kollman, P.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A. *The development/application of a \"minimalist\" organic/biochemical molecular*

mechanic force field using a combination of ab initio calculations and experimental data, 1997.

(21) Cheatham, T. E., III; Cieplak, P.; Kollman, P. A. *Journal of Biomolecular Structure & Dynamics* **1999**, *16*, 845.

(22) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *Journal of Computational Chemistry* **1983**, *4*, 187.

(23) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *Journal of Physical Chemistry B* **1998**, *102*, 3586.

(24) Neria, E.; Fischer, S.; Karplus, M. *Journal of Chemical Physics* **1996**, *105*, 1902.

(25) Jorgensen, W. L.; Tirado-Rives, J. *Journal of the American Chemical Society* **1988**, *110*, 1657.

(26) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *Journal of Physical Chemistry B* **2001**, *105*, 6474.

(27) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hunenberger, P. H.; Kruger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. Biomolecular simulation; the GROMOS96 manual and user guide; Vdf Hochschulverlag AG an der ETH Zurich: zurich, 1996.

(28) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. *Journal of Computational Chemistry* **2001**, *22*, 1205.

(29) Woodcock, L. V. *Chemical Physics Letters* **1971**, *10*, 257.

(30) Nose, S. *Journal of Chemical Physics* **1984**, *81*, 511.

(31) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *Journal of Chemical Physics* **1984**, *81*, 3684.

(32) Andersen, H. C. *Journal of Chemical Physics* **1980**, *72*, 2384.

(33) Andrea, T. A.; Swope, W. C.; Andersen, H. C. *Journal of Chemical Physics* **1983**, *79*, 4576.

(34) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.

(35) Martyna, G. J.; Klein, M. L.; Tuckerman, M. E. *Journal of Chemical Physics* **1992**, *97*, 2635.

(36) Mahoney, M. W.; Jorgensen, W. L. *Journal of Chemical Physics* **2000**, *112*, 8910.

(37) Born, M.; v. Karman, T. *Physik. Z.* **1912**, *13*, 297.

(38) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *Journal of Physical Chemistry A* **1997**, *101*, 3005.

(39) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199.

(40) Gilson, M. K.; Honig, B. *Proteins: Structure, Function, and Genetics* **1988**, *4*, 7.

(41) Kang, Y. K.; Nemethy, G.; Scheraga, H. A. *Journal of Physical Chemistry* **1987**, *91*, 4105.

(42) Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. *Proceedings of the National Academy of Sciences of the United States of America* **1987**, *84*, 3086.

(43) Warshel, A.; Russell, S. T. *Quarterly reviews of biophysics* **1984**, *17*, 283.

(44) Tomasi, J.; Persico, M. *Chemical Reviews* **1994**, *94*, 2027.

(45) Cramer, C. J.; Truhlar, D. G. *Chemical Reviews* **1999**, *99*, 2161.

(46) Daggett, V.; Kollman, P. A.; Kuntz, I. D. *Biopolymers* **1991**, *31*, 285.

(47) Daggett, V. D. Molecular Modelling of Peptide and Proteins. Doctor of Philosophy, University of California, 1990.

(48) Hornak, V.; Simmerling, C. *Proteins: Structure, Function, and Genetics* **2003**, *51*, 577.

(49) Olson, M. A. *Proteins: Structure, Function, and Bioinformatics* **2004**, *57*, 645.

(50) Arnold, G. E.; Ornstein, R. L. *Proteins: Structure, Function, and Genetics* **1994**, *18*, 19.

(51) Forrest Lucy, R.; Woolf Thomas, B. *Proteins* **2003**, *52*, 492.

(52) Warwicker, J.; Watson, H. C. *Journal of molecular biology* **1982**, *157*, 671.

(53) Honig, B.; Sharp, K.; Yang, A. S. *Journal of Physical Chemistry* **1993**, *97*, 1101.

(54) Sitkoff, D.; Sharp, K. A.; Honig, B. *Journal of Physical Chemistry* **1994**, *98*, 1978.

(55) Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144.

(56) Gilson, M. K.; McCammon, J. A.; Madura, J. D. *Journal of Computational Chemistry* **1995**, *16*, 1081.

(57) Smart, J. L.; Marrone, T. J.; McCammon, J. A. *Journal of Computational Chemistry* **1997**, *18*, 1750.

(58) Xia, B.; Tsui, V.; Case, D. A.; Dyson, H. J.; Wright, P. E. *Journal of Biomolecular NMR* **2002**, *22*, 317.

(59) Liu, Y.; Beveridge, D. L. *Proteins: Structure, Function, and Genetics* **2001**, *46*, 128.

(60) Calimet, N.; Schaefer, M.; Simonson, T. *Proteins: Structure, Function, and Genetics* **2001**, *45*, 144.

(61) Srinivasan, J.; Miller, J.; Kollman, P. A.; Case, D. A. *Journal of biomolecular structure & dynamics* **1998**, *16*, 671.

(62) Srinivasan, J.; Trevathan, M. W.; Beroza, P.; Case, D. A. *Theoretical Chemistry Accounts* **1999**, *101*, 426.

(63) Dominy, B. N.; Brooks, C. L., III. *Journal of Physical Chemistry B* **1999**, *103*, 3765.

(64) Jayaram, B.; Sprous, D.; Young, M. A.; Beveridge, D. L. *Journal of the American Chemical Society* **1998**, *120*, 10629.

(65) Scarsi, M.; Apostolakis, J.; Caflisch, A. *Journal of Physical Chemistry B* **1998**, *102*, 3637.

(66) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *Journal of Physical Chemistry B* **1998**, *102*, 10983.

(67) Rapp, C. S.; Friesner, R. A. *Proteins: Structure, Function, and Genetics* **1999**, *35*, 173.

(68) Edinger, S. R.; Cortis, C.; Shenkin, P. S.; Friesner, R. A. *Journal of Physical Chemistry B* **1997**, *101*, 1190.

(69) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *Journal of the American Chemical Society* **1990**, *112*, 6127.

(70) Lee, B.; Richards, F. M. *Journal of molecular biology* **1971**, *55*, 379.

(71) Weiser, J.; Shenkin, P. S.; Still, W. C. *Journal of Computational Chemistry* **1999**, *20*, 217.

(72) Jayaram, B.; Sprous, D.; Beveridge, D. L. *Journal of Physical Chemistry B* **1998**, *102*, 9571.

(73) Tsui, V.; Case, D. A. *Biopolymers* **2001**, *56*, 275.

(74) Weiser, J.; Weiser, A. A.; Shenkin, P. S.; Still, W. C. *Journal of Computational Chemistry* **1998**, *19*, 797.

(75) Tsui, V.; Case, D. A. *Journal of the American Chemical Society* **2000**, *122*, 2489.

(76) Onufriev, A.; Bashford, D.; Case, D. A. *Journal of Physical Chemistry B* **2000**, *104*, 3712.

(77) Jayaram, B.; Liu, Y.; Beveridge, D. L. *Journal of Chemical Physics* **1998**, *109*, 1465.

(78) Schaefer, M.; Karplus, M. *Journal of Physical Chemistry* **1996**, *100*, 1578.

(79) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Journal of Physical Chemistry* **1996**, *100*, 19824.

(80) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chemical Physics Letters* **1995**, *246*, 122.

(81) Sosa, C. P.; Hewitt, T.; Lee, M. R.; Case, D. A. *Theochem* **2001**, *549*, 193.

(82) Wilson, E. B.; Decius, J. C.; Cross, P. C. *Molecular Vibrations*; Dover Publications, INC.: New York, 1980.

(83) Kirkpatrick, S.; Gelatt, C. D. J.; Vecchi, M. P. *Science* **1983**, *220*, 671.

(84) Huber, G. A.; McCammon, J. A. *Phys. Rev. E* **1997**, *55*, 4822.

(85) Xiang, Y.; Gong, X. G. *Physical Review E: Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **2000**, *62*, 4473.

(86) Basu, A.; Frazer, L. N. *Science* **1990**, *249*, 1409.

(87) Roitberg, A.; Elber, R. *Journal of Chemical Physics* **1991**, *95*, 9277.

(88) Rosenhouse-Dantsker, A.; Osman, R. *Biophysical Journal* **2000**, *79*, 66.

(89) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *Journal of Chemical Physics* **1953**, *21*, 1087.

(90) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96.

(91) Okamoto, Y. *Journal of Molecular Graphics & Modelling* **2004**, *22*, 425.

(92) Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **1999**, *314*, 141.

(93) Hansmann, U. H. E.; Okamoto, Y. *Journal of Chemical Physics* **1999**, *110*, 1267.

(94) Berg, B. A. *Fields Institue Communications* **2000**, *26*, 1.

(95) Yasar, F.; Arkin, H.; Celik, T.; Berg, B. A.; Meirovitch, H. *Journal of Computational Chemistry* **2002**, *23*, 1127.

(96) Mitsutake, A.; Okamoto, Y. *Chem. Phys. Let.* **1999**, *309*, 95.

(97) Mitsutake, A.; Okamoto, Y. *Journal of Chemical Physics* **2000**, *112*, 10638.

(98) Hansmann, U. H. E.; Okamoto, Y. *Physical Review E: Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **1996**, *54*, 5863.

(99) Hansmann, U. H. E.; Okamoto, Y. *Journal of Computational Chemistry* **1997**, *18*, 920.

(100) Li, Y.; Protopopescu, V. A.; Gorin, A. *Physics Letters A* **2004**, *328*, 274.

(101) Kim, J. G.; Fukunishi, Y.; Nakamura, H. *Chemical Physics Letters* **2004**, *392*, 34.

(102) Madras, N. *IMA Volumes in Mathematics and Its Applications* **1998**, *102*, 19.

(103) Marinari, E.; Parisi, G. *Europhysics Letters* **1992**, *19*, 451.

(104) Kim Jae, G.; Fukunishi, Y.; Kidera, A.; Nakamura, H. *Journal of chemical physics* **2004**, *121*, 5590.

(105) Kholmovski, E. G.; Parker, D. L.; Alexander, A. L. *Journal of magnetic resonance imaging: JMRI* **2000**, *11*, 549.

(106) Hesselbo, B.; Stinchcombe, R. B. *Physical Review Letters* **1995**, *74*, 2151.

(107) Berg, B. A.; Celik, T. *Phys. Rev. Lett* **1992**, *69*, 2292.

(108) Okamoto, Y.; Hansmann, U. H. E.; Nakazawa, T. *Chemistry Letters* **1995**, 391.

(109) Hukushima, K.; Nemoto, K. *Journal of the Physical Society of Japan* **1996**, *65*, 1604.

(110) Tesi, M. C.; Janse van Rensburg, E. J.; Orlandini, E.; Whittington, S. G. *J. Stat. Phys.* **1996**, *82*.

(111) Hansmann, U. H. E. *Chemical Physics Letters* **1997**, *281*, 140.

(112) Sugita, Y.; Kitao, A.; Okamoto, Y. *Journal of Chemical Physics* **2000**, *113*, 6042.

(113) Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **2000**, *329*, 261.

(114) Ferrenberg, A. M.; Swendsen, R. H. *Physical Review Letters* **1989**, *63*, 1195.

(115) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *Journal of Computational Chemistry* **1992**, *13*, 1011.

(116) Daggett, V. *Accounts of Chemical Research* **2002**, *35*, 422.

(117) Derreumaux, P.; Mousseau, N. *Trends in Chemical Physics* **2002**, *10*, 235.

(118) Baumketner, A.; Hiwatari, Y. *AIP Conference Proceedings* **2003**, *661*, 195.

(119) Gnanakaran, S.; Nymeyer, H.; Portman, J.; Sanbonmatsu, K. Y.; Garcia, A. E. *Current Opinion in Structural Biology* **2003**, *13*, 168.

(120) Arteca, G. A.; Reimann, C. T.; Tapia, O. *Mass Spectrometry Reviews* **2002**, *20*, 402.

(121) Brooks, C. L., III. *Accounts of Chemical Research* **2002**, *35*, 447.

(122) Olson, W. K. *Oxford Handbook of Nucleic Acid Structure* **1999**, 499.

(123) Miller, J.; Cooney, M.; Miaskiewicz, K.; Osman, R. *ACS Symposium Series* **1998**, *682*, 312.

(124) Guvench, O.; Weiser, J.; Shenkin, P.; Kolossvary, I.; Still, W. C. *Journal of Computational Chemistry* **2002**, *23*, 214.

(125) Glen, R. C.; Allen, S. C. *Current Medicinal Chemistry* **2003**, *10*, 763.

(126) Kouwijzer, M.; Mestres, J. *Pharmacochemistry Library* **2002**, *32*, 57.

(127) Aaqvist, J.; Luzhkov, V. B.; Brandsdal, B. O. *Accounts of Chemical Research* **2002**, *35*, 358.

(128) Rongan, D. *Perspectives in Drug Discovery and Design* **1998**, *9/10/11*, 181.

(129) Czerminski, R.; Elber, R. *Proteins: Structure, Function, and Genetics* **1991**, *10*, 70.

(130) Brunori, M.; Bourgeois, D.; Vallone, B. *Journal of Structural Biology* **2004**, *147*, 223.

(131) Brunori, M. *Biophysical Chemistry* **2000**, *86*, 221.

(132) Case, D. A. *Progress in Biophysics & Molecular Biology* **1988**, *52*, 39.

(133) Kothekar, V. *Indian journal of biochemistry & biophysics* **1996**, *33*, 431.

(134) Sato, H. *Understanding Chemical Reactivity* **2003**, *24*, 61.

(135) Doltsinis, N. L.; Marx, D. *Journal of Theoretical & Computational Chemistry* **2002**, *1*, 319.

(136) Spohr, E. *Solid State Ionics* **2002**, *150*, 1.

(137) Van Speybroeck, V.; Meier, R. J. *Chemical Society Reviews* **2003**, *32*, 151.

(138) Watanabe, H. *Kagaku to Kyoiku* **2000**, *48*, 84.

(139) Elber, R.; Karplus, M. *Journal of the American Chemical Society* **1990**, *112*, 9161.

(140) Rolman, R. C. *The Principles of Statistical Mechanics*Oxford, 1938.

(141) Koehl, P.; Delarue, M. *Current Opinion in Structural Biology* **1996**, *6*, 222.

(142) Caflisch, A.; Miranker, A.; Karplus, M. *Journal of medicinal chemistry* **1993**, *36*, 2142.

(143) Carlson, H. A.; Masukawa, K. M.; McCammon, J. A. *Journal of Physical Chemistry A* **1999**, *103*, 10213.

(144) Miranker, A.; Karplus, M. *Proteins* **1991**, *11*, 29.

(145) Quillin, M. L.; Li, T.; Olson, J. S.; Phillips, G. N., Jr.; Dou, Y.; Ikeda-Saito, M.; Regan, R.; Carlson, M.; Gibson, Q. H.; Li, H. *Journal of molecular biology* **1995**, *245*, 416.

(146) Ulitsky, A.; Elber, R. *Journal of Physical Chemistry* **1994**, *98*, 1034.

(147) Verkhivker, G.; Elber, R.; Nowak, W. *Journal of Chemical Physics* **1992**, *97*, 7838.

(148) Simmerling, C.; Fox, T.; Kollman, P. A. *Journal of the American Chemical Society* **1998**, *120*, 5771.

(149) Simmerling, C.; Lee, M. R.; Ortiz, A. R.; Kolinski, A.; Skolnick, J.; Kollman, P. A. *Journal of the American Chemical Society* **2000**, *122*, 8392.

(150) Straub, J. E.; Karplus, M. *Journal of Chemical Physics* **1991**, *94*, 6737.

(151) Ulitsky, A.; Elber, R. *Journal of Chemical Physics* **1993**, *98*, 3380.

(152) Zheng, W.-M.; Zheng, Q. *Journal of Chemical Physics* **1997**, *106*, 1191.

(153) Stultz, C. M.; Karplus, M. *Journal of Chemical Physics* **1998**, *109*, 8809.

Chapter 3

Divergent Path Search Strategy and Limitations of Conventional Molecular Dynamics Simulations for Protein Folding

3.1. Introduction

During the past several decades**,** molecular dynamics (MD) [1,2] and Monte Carlo (MC) [3] simulation methods have evolved into important and widely used theoretical tools in chemistry, physics, and biology to model the detailed dynamical behavior of many systems, from atomic clusters to large biological molecules. However, a complete and accurate simulation of these systems is difficult due to the problem of quasiergodicity, in which the system is easily trapped in local minima of the energy landscape [4]. A biased non-ergodic trajectory almost always appears on the time scale of the simulation. In order to overcome the quasiergodicity problem by enhanced energy barrier crossings, many techniques including Simulated Annealing (SA) [5-8], Local Enhanced Sampling (LES) [8-11], Multiple-copy methods [12-14]; and Generalized Ensembles (GE) [4,15-30] have been proposed and applied to a variety of sampling and optimization problems, including the protein folding problem.

Because a generalized ensemble simulation performs a random walk in potential energy space over a wide temperature range, generalized ensembles have been shown to be very effective techniques to overcome the quasiergodicity problem in simulations. Many different techniques for using generalized-ensemble algorithms have been developed in the past several years, however, most of these techniques, including multicanonical algorithms [18,19,30], simulated tempering [20,29,30], and the jump-walking algorithm [21-24] have been introduced for MC rather than MD simulations. So far the replica exchange method (REM) is the only generalized ensemble technique widely used within MD [16,17,25-27]. This interesting observation inspired us to re-consider the intrinsic character of MD, especially the quasiergodicity problem in MD simulations, compared to MC simulations. This re-consideration helped us to realize three limitations in conventional MD due to the time scale of the simulation and to develop some strategies to enhance energy barrier crossing and phase space sampling in microcanonical (NVE) or canonical (NVT) ensemble simulations.

In this chapter, we present a new MD protocol, called divergent path (DIP) search simulations. The simulations using this new technique use multiple simultaneous simulations of a polypeptide. Each simulation starts the polypeptide from the same configuration, but directs each polypeptide along a different, independent trajectory. The polypeptides following different trajectories differ only in the directions of their initial atomic velocities, but have the same initial temperature and thus have the same average atomic speed. Comparing results from conventional MD simulations with tests of this divergent path search strategy clearly indicates that potential energy traps, free energy traps, and kinetic traps constitute important limitations of conventional MD simulations used to search for the global potential energy minimum.

Potential energy traps appear when constant energy (constant NVE) simulations become trapped in a local potential energy well, where the kinetic energy is not large enough to overcome high potential energy barriers between local minima. Free energy traps occur when constant temperature (constant NVT) simulations reach thermal equilibrium and become trapped in free energy wells, so that simulations sample portions of phase space having only a certain range of potential energies. The limited sampling of the potential energy surface may arise because of a potential energy trap, but it may also arise because of rapid exchange between kinetic and potential energies. Because of this rapid energy exchange, the thermostat that restricts the range of possible kinetic energies also restricts the system's potential energies to limited regions of phase space. We call the (limited) sampled regions of phase space "thermodynamically accessible regions". Kinetic traps, on the other hand, occur in a constant NVE or NVT simulation that evolves into a large number of equivalent substates within a broad potential energy well or a small number of thermodynamically accessible regions at a given simulation temperature. This causes kinetic problems because it would take an unrealistic amount of time for such a simulation to reach other thermodynamically accessible regions after it finished sampling the large number of thermodynamically accessible substates. Kinetic traps are reached once mechanical equilibrium is reached, once vector velocities of the atoms have reached equilibrium. Thermal equilibrium, when the temperature and average atomic speeds, as well as their fluctuations, are approximately constant, is typically achieved much earlier in a constant NVT simulation than mechanical equilibrium, when vector velocities of the atoms have reached equilibrium.

Polyalanine is the simplest peptide to adopt the α-helical conformation [31], which is the most abundant and important secondary structural element of proteins. Experiments [32-37] show that short alanine-based peptides appear to form an α-helical conformation in aqueous solution and simulations [16,28,31,38-40] of uncharged polyalanines with a sequence length between 10 and 30 indicate that the α-helical conformation is the lowest energy folded conformation for the peptides both *in vacuo* and in aqueous environments. We selected a 13-residue polypeptide of alanine (Ala13) to test the divergent path search strategy in this study.

3.2. Simulation methodology

3.2.1. Simulation protocol and potential energy function

All calculations reported in this work were performed by using our own implementations of conventional constant energy or constant temperature MD, and divergent path (DIP) search simulations. In the DIP simulations, we allow several independent polypeptides of the system to evolve simultaneously such that each polypeptide does not interact with the others and each one follows its own trajectory as in a conventional NVT simulation. Usually, these independent polypeptides share the same initial structure and temperature, differing only in the directions of atomic velocities. The initial velocities of atoms within the first polypeptide were generated from a Gaussian velocity distribution [41] at the canonical temperature, and the velocities of all other polypeptides are re-set by randomly changing the original direction but not the magnitude of the velocity vector for each atom in each polypeptide. The Nosé-Hoover Chain method [42] is used to control the temperature in NVT simulations.

The potential energy function (eq. 3.2.1.1) of the gas-phase polypeptide uses the generic force field with AMBER force field parameters [43].

$$\begin{aligned} E_{potential} = & \sum_{\textbf{bonds}} K_r (r - r_{eq})^2 + \sum_{\textbf{angles}} K_\theta (\theta - \theta_{eq})^2 \\ & + \sum_{\textbf{dihedral}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}{}^{12}} - \frac{B_{ij}}{r_{ij}{}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}} \right] \end{aligned} \qquad (3.2.1.1)$$

The force field consists of bond length stretching and angle bending represented by a simple harmonic expression. The dihedral angle term is represented by a truncated Fourier series, the van der Waals interaction is modeled by a Lennard-Jones potential, and electrostatic interactions are represented by a Coulombic interaction of atom-centered partial charges.

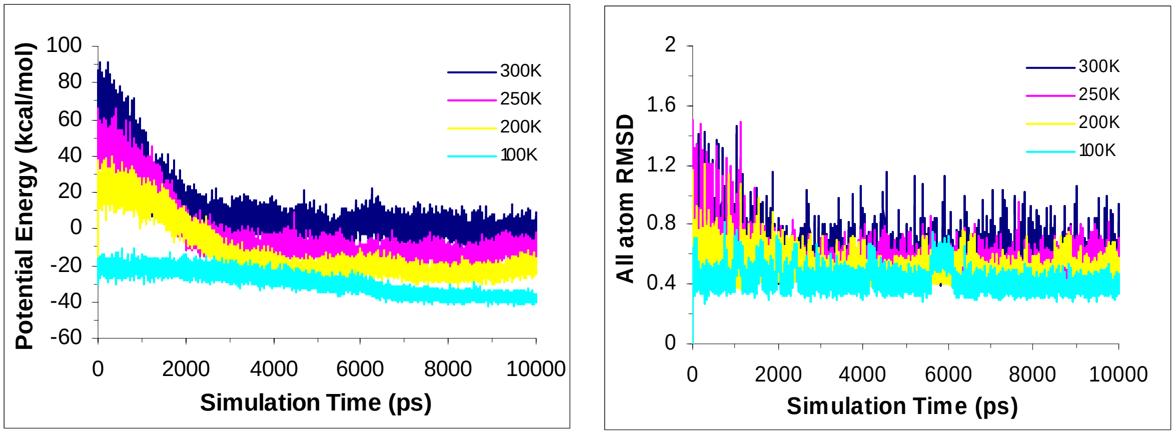
3.2.2. Computational details

The AMBER package [44] was used to generate the initial coordinates of a 13-residue peptide of alanine with the extended structure. The termini of this Ala13 peptide were blocked (acetylated at the N terminus and methyl-amidated at the C terminus) (ACE Acetyl-$(\mathrm{Ala})_{13}$-NME N-methyl). The initial coordinates of the α-helical form of Ala13 were obtained from the Midas program [45] by setting each combination of ϕ, φ torsion angles of the polypeptide backbone to the ideal values: ϕ = -57° and φ = -47° [46]. The final steps necessary to produce suitable input files for our program were accomplished using the Molecular Modeling Toolkit [47]. We used the velocity-Verlet algorithm [2,41,48] to integrate the equations of motion, using forces derived from the AMBER 96 [43]all-atom force field, while the temperature was controlled in canonical ensemble simulations using the Nosé-Hoover Chain (NHC) [42] algorithm. The SHAKE algorithm [49] was used to constrain covalent bond distances involving hydrogen and translation of the entire system was removed at each step. We used a time step of 1 fs, but collected trajectory data (energies and coordinates) at 1 ps intervals in all MD simulations. The total simulation time for each run was at least 10 ns for conventional MD and DIP simulations.

3.3. Simulation facts and derived results

3.3.1. Thermal and mechanical equilibrium in conventional MD simulations

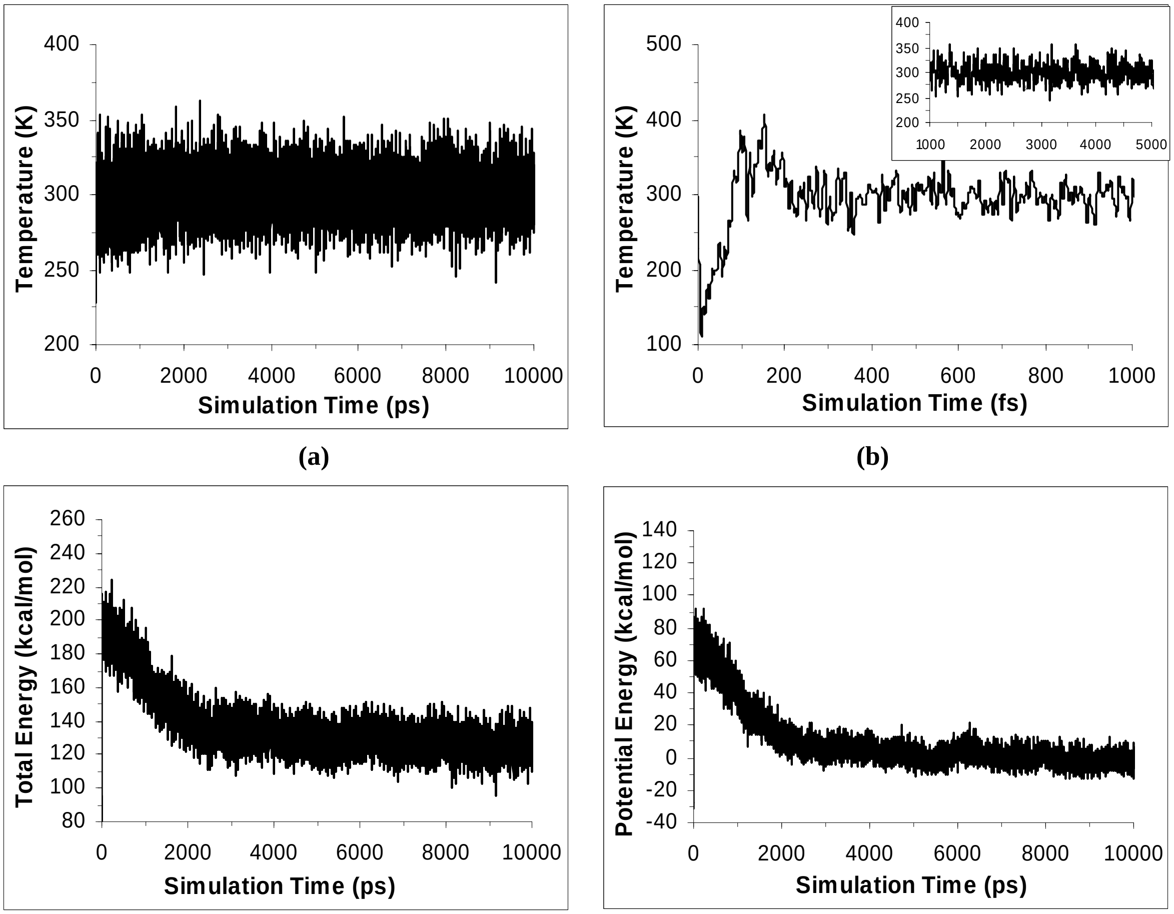
Figure 3.3.1.1 shows the time series of the potential energy, and root-mean-square-deviations (RMSDs) between all atoms of the calculated structures and the ideal α-helical reference structure from four conventional NVT simulations of Ala13 *in vacuo* at temperatures of 300 K, 250 K, 200 K, and 100 K, starting from an ideal α-helix. The plots indicate that while all four simulations locate the same global potential energy well corresponding to an α-helical structure, the simulations at different temperatures sample very different energies. The *virial* theorem [50] indicates that the potential energy of a canonical ensemble is a statistic average ensemble property of the temperature: at higher temperatures, a simulation samples higher potential energies; at lower temperatures, the simulation samples lower potential energies. These temperature-dependent potential energy histories clearly reflect the applicability of the *virial* theorem to the sampled potential energy surface because simulations at the canonical temperature usually do not sample the valleys but instead visit configurations of higher energy on the potential energy surface. Therefore, in molecular dynamics, simulations can reach equilibrium within a certain range of potential energies but not visit in the basins of the potential energy wells. In fact, these basins are usually very difficult to sample in NVT simulations at non-zero temperatures.



(a) **(b)**

Figure 3.3.1.1. Time series of the potential energy and RMSDs from conventional NVT simulations of Ala13 *in vacuo* at different temperatures: 300 K, 250 K, 200 K, and 100 K, starting from an ideal α-helix. **(a)** Potential energy. **(b)** RMSDs between all atoms of the calculated structures and the initial ideal α-helical reference structure. The figure clearly demonstrates the *virial* theorem: at higher temperature, a simulation samples higher potential energies more; at lower temperature, the simulation samples lower potential energies more.

Thermal equilibrium is a well-established scientific principle. Thermalization before data collection is also a standard procedure in MD simulations, especially when trajectory data is to be used to calculate a phase-space average. We did not run any pre-simulation for thermalization in our conventional NVT simulation, but the temperature history (Figure 3.3.1.2a - b) shows that thermal equilibrium is reached very quickly. The temperature is almost equilibrated within the first 1 ps (1000 steps) and then oscillates at 300 ± 50 K under the control of the Nosé-Hoover thermal chain. Figure 3.3.1.2c – d display the corresponding total energy and potential energy histories for conventional NVT simulation of Ala13 at 300 K *in vacuo*. The total energy history is very similar to the potential energy history, differing only in the rapid equilibration of the kinetic energy. While the temperature (kinetic energy) quickly reaches equilibrium, the total energy and potential energy are still far away from equilibrium. These non-equilibrated energies are caused by the lack of equilibration among the magnitudes and especially directions of the 3N velocities. We call this a mechanical non-equilibrium state. On the other hand, mechanical equilibrium indicates that the 3N velocities are universally equilibrated along the whole polypeptide. We can therefore use the potential energy history to evaluate whether mechanical equilibrium has been reached or not. Thus, from the potential energy history (Fig. 3.3.1.2d), this simulation reaches mechanical equilibrium after a time of more than 2 ns.



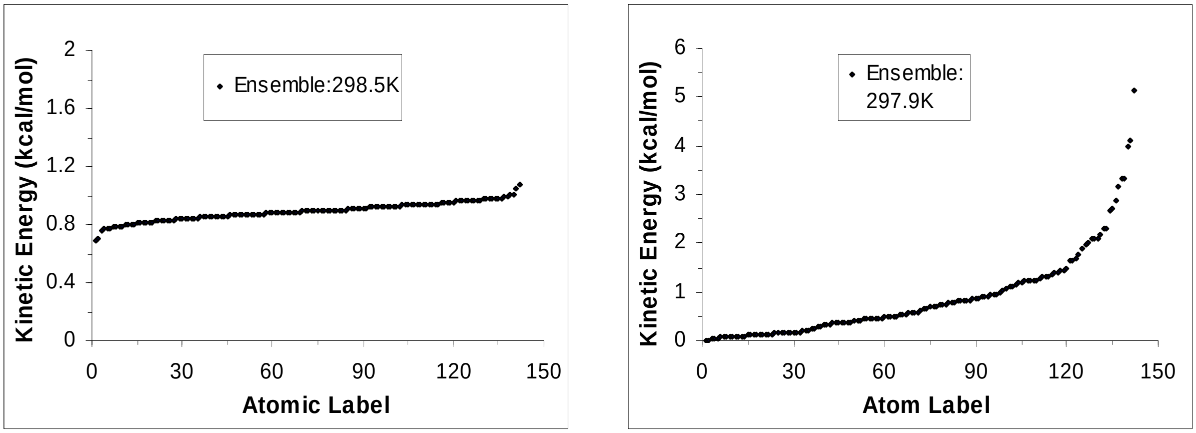
(a) (b)

(c) (d)

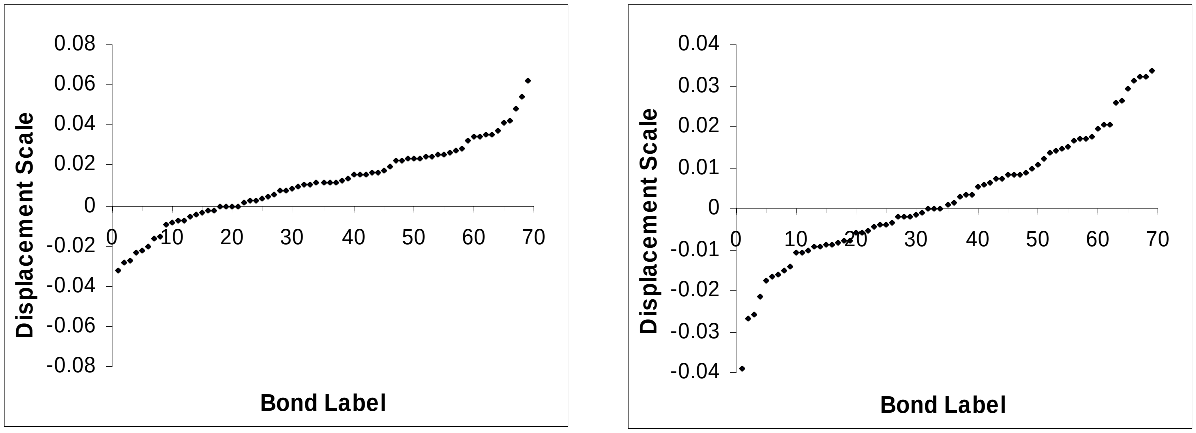
Figure 3.3.1.2. Time series of various properties of the conventional NVT simulation of Ala13 at 300 K *in vacuo*, starting from an ideal α-helix. **(a)** Temperature (ps scale). **(b)** Temperature for the first 1 ps (fs scale). The inset shows temperature between 1 ps and 5ps. **(c)** Total energy. **(d)** Potential energy. The figure indicates that the temperature gets

equilibrated very quickly (ps scale) while the energy (total energy and potential energy) reaches equilibrium relatively slower (ns scale).

We need to consider the initial conditions of the conventional NVT simulations to understand the origin of the mechanical non-equilibrium state. The random initial Gaussian velocity distribution can only fix the initial average canonical-ensemble temperature and velocities but not the equilibrated velocity magnitude and direction for subsets of atoms. Fig. 3.3.1.3a displays the atomic kinetic energy distribution for an ensemble temperature of 298.5 K at the initial step, which is generated from a Gaussian distribution [41] at a mean temperature of 300 K, with a standard deviation of 20 K. The atomic kinetic energy distribution with an ensemble temperature of 297.9 K at the final step is shown in Fig. 3.3.1.3b. Though the direction distribution of the equilibrated velocities is difficult to represent, the equilibrated distribution is far away from this initial Gaussian distribution and highlights the difference between the equilibrated and non-equilibrated velocities. The non-equilibrated individual velocities at the beginning step can cause unbalanced displacement in different parts of the polypeptide in subsequent molecular dynamics steps. Some atoms move very close to each other and others move far away. This biased displacement causes the potential energy to resist equilibration. The unbalanced forces, in turn, cause the simulation to remain thermally non-equilibrated, with a small decrease in energy. Under the perturbation of the external thermostat, thermal equilibrium (involving only the magnitude of atomic velocities) should be reached sooner than mechanical equilibrium, which involves both the magnitude and direction of the atomic velocities. The failure of the direction of the atomic velocities to equilibrate, on the other hand, indicates a mechanical nonequilibration state. This discrepancy between thermal equilibration time and mechanical equilibration time has been demonstrated in the temperature and potential energy histories. Fig. 3.3.1.3c – d display the bond displacement scale (defined as the ratio of bond displacement to bond length) distribution after 1 ns and 10 ns respectively. It is very clear that the latter shows much more evenly distributed displacements than the former. Therefore, the non-equilibrium potential energy history after thermal equilibrium is reached indicates the existence of a mechanical equilibration process caused as the non-equilibrium velocities along the whole polypeptide become more uniform. The mechanical equilibrium state, on the other hand, indicates an equilibrated velocity distribution that will bring a much more uniform coordinate displacement and atomic interactions in the following steps. An observed result for a system that has reached mechanical equilibrium is that the simulation has become equilibrated in a huge number of equivalent energy states with similar conformations. Thus, thermal equilibrium is reached when the temperature or scalar velocities are in equilibrium, while mechanical equilibrium is reached when the more restricted velocity vectors are in equilibrium. Therefore, mechanical equilibrium takes more time to reach.



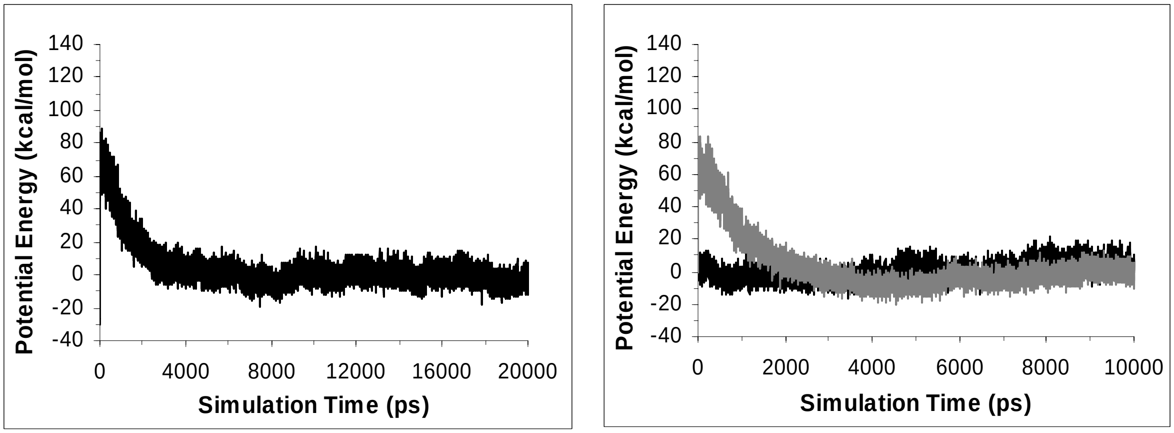
(a) **(b)**



(c) (d)

Figure 3.3.1.3. Atomic kinetic energy and bond displacement scale distribution at several steps from the conventional NVT simulation of Ala13 at 300 K *in vacuo*, starting from an ideal α-helix. **(a)** Atomic kinetic energy distribution of a 298.5 K ensemble temperature at the initial step. **(b)** Atomic kinetic energy distribution of a 297.9 K ensemble temperature after 10 ns. **(c)** Bond displacement scale distribution after 1ns. **(d)** Bond displacement scale distribution after 10 ns.

The phenomena of mechanical equilibrium explains the observation that the re-initialization of the velocity-direction distribution in an MD simulation causes further simulations to leave the original states and evolve into different equilibrium states along a new trajectory. To show this, we ran two conventional NVT simulations of Ala13 restarted from the final structure obtained after a 20 ns conventional NVT simulation. The first simulation (dark plot in Fig. 3.3.1.4b) is assigned the equilibrated velocities but the second simulation (gray plot in Fig. 3.3.1.4b) has velocities re-initialized by changing velocity directions but not their magnitudes (so the temperature remains the same for these two simulations). The first simulation maintains thermal and mechanical equilibrium very well, but the second is out of mechanical equilibrium. Another 2 ns is required before the simulation reaches mechanical equilibrium again.



(a) **(b)**

Figure 3.3.1.4. Time series of the potential energy in the conventional NVT simulation at 300 K *in vacuo*, starting from the ideal α-helix. **(a)** Potential energy for the first 20 ns. **(b)** Potential energy for the following 10 ns with velocities redirection (gray) or without (dark). The figure indicates that the redirection of the system's velocities of NVT simulations will break the original mechanical equilibrium.

Fig. 3.3.1.5 shows the potential energy and temperature histories of conventional NVE simulations of Ala13 at kinetic energies corresponding to 300 K *in vacuo*, initiated from the α-helical conformation. These plots clearly indicate that mechanical equilibrium is reached rapidly in NVE simulations. Thermal equilibrium (evaluated from temperature histories) and mechanical equilibrium (evaluated from potential energy histories) are reached almost simultaneously after 1 ps. We believe that the mechanical equilibrium in NVE simulations is reached more quickly than in NVT simulations because energy flow between the system and the heat bath under control of the Nosé-Hoover thermal chain makes the mechanical equilibration proceed relatively slowly in NVT simulations.

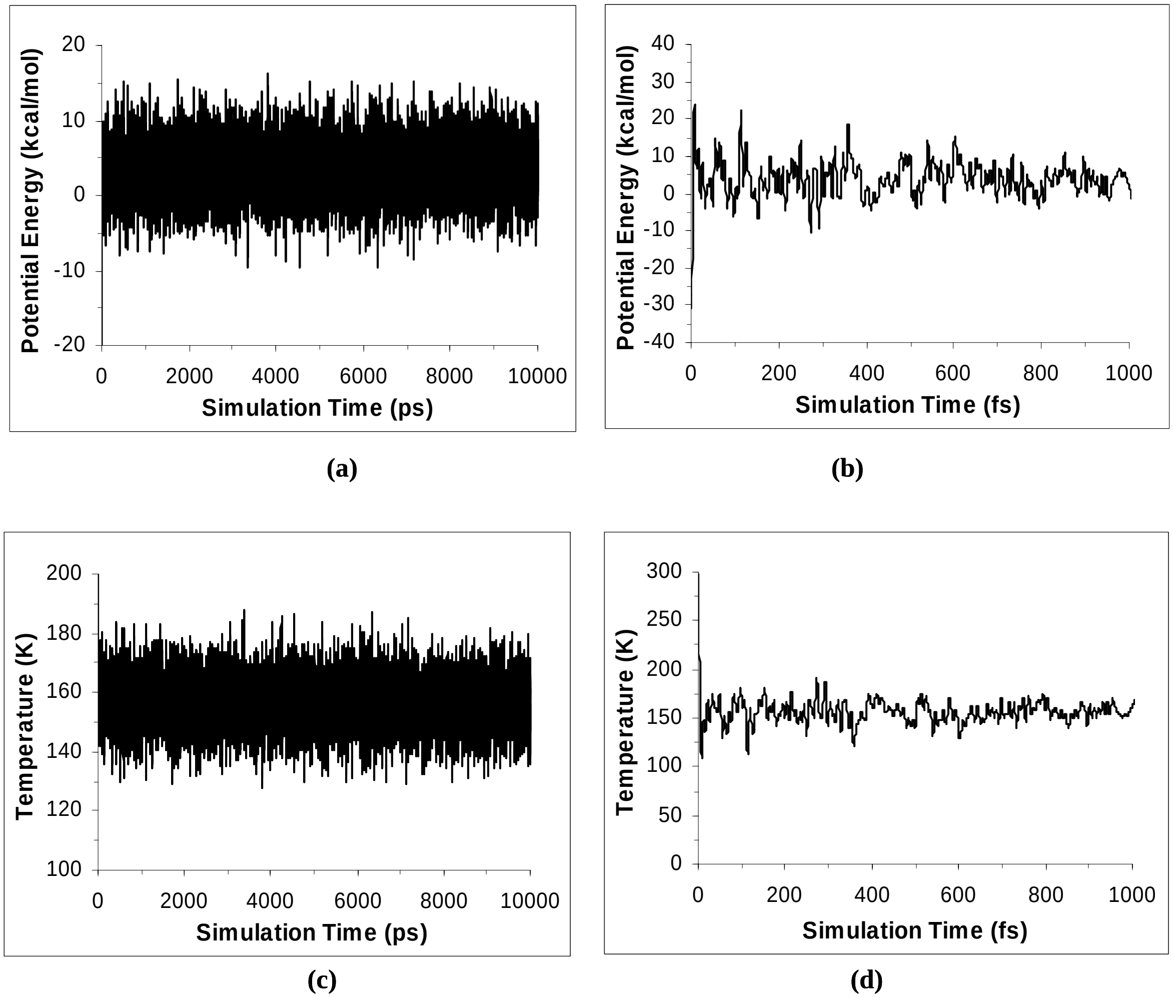
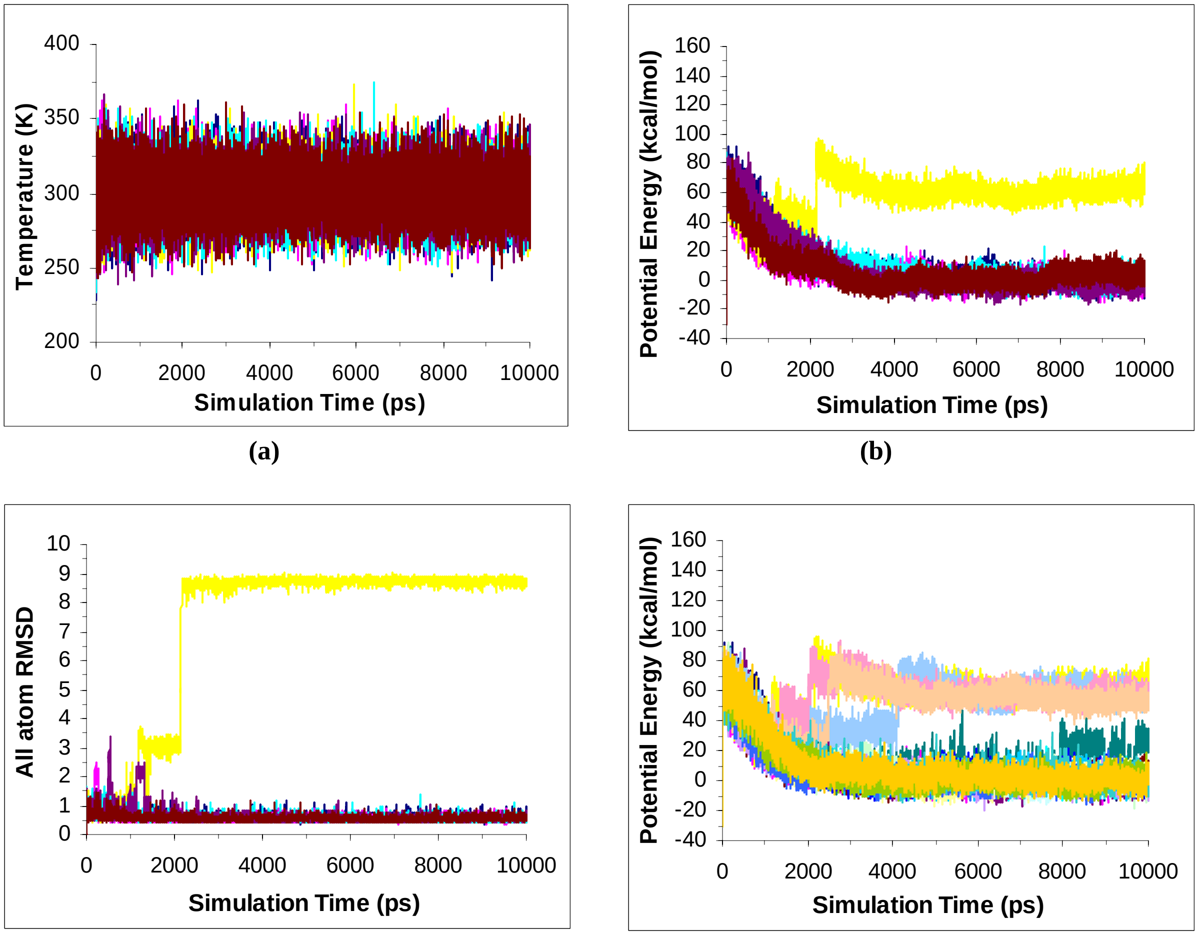


Figure 3.3.1.5. Time series of the potential energy and temperature of the conventional NVE simulation of Ala13 at a 300 K initial temperature *in vacuo*, starting from an ideal α-helix. **(a)** Potential energy (ps scale). **(b)** Potential energy for the first 1 ps (fs scale). **(c)** Temperature (ps scale). **(d)** Temperature for the first 1 ps (fs scale). The figure indicates that both temperature and energy reach equilibrium very quickly (ps scale) in NVE simulations.

3.3.2. Divergent path (DIP) search simulations and three limitations of conventional MD simulations



(a) (b)

(c) (d)

Figure 3.3.2.1. Time series of various properties from a six-path **(a-c)** or twenty-path **(d)**, DIP simulation of Ala13 at 300 K *in vacuo*, starting from an ideal α-helix. **(a)** Temperature. **(b)** Potential energy. **(c)** RMSD between all atoms of calculated structure and an ideal α-helix. **(d)** Potential energy for a twenty-path simulation. The figure indicates that DIP simulation evolving into different trajectories can sample different potential energy states with large conformational differences. **Note**: Blue: 1st, Magenta: 2nd, Yellow: 3rd, Light blue: 4th, Purple: 5th, Brown: 6th trajectory.

DIP simulations were developed to alleviate the limitations of conventional MD simulations for folding polypeptides. Fig. 3.3.2.1a – c displays time series of the temperature, potential energy and RMSD (from an ideal α-helix) in a six-path DIP simulation of Ala13 at 300 K *in vacuo*, starting from an ideal α-helical conformation. While the temperature histories are almost identical for the each polypeptide's simulation

(Fig. 3.3.2.1a), the corresponding potential energies of six polypeptides (Fig. 3.3.2.1b) evolve differently and their minimum energy structures (Fig. 3.3.2.1c) may be located in different potential energy wells, corresponding to different conformations. In fact, five polypeptides maintain the α-helical conformation (Fig. 3.3.2.1c), while one polypeptide displays an unusual increasing potential energy history indicating an unfolding process: near 1.1 ns, the α-helical Ala13 transforms into a 3_{10} helix (RMSD ≈ 3 Å), and then at 2.3ns it entirely unfolds to extended structures (RMSD ≈ 8.5 Å). This unusual trajectory persists in simulations performed at 250 K but disappears at 200 K or below (data not shown here).

The third trajectory from the initial α-helical conformation in the above simulations of blocked Ala13 show the very interesting feature that the global potential energy minimum helical conformation (Chapter 4) can be unfolded into highly energetic extended conformations (RMSD between 8.5 and 9.0). These conformations are stable even after 100 ns (data are not shown). The unfolding of these structures is not due to high temperature conditions; the temperature history for this trajectory oscillates near 300 K ± 50 K (Fig. 3.3.2.1a). Escaping the local potential energy trap, where simulations become trapped in a local potential energy well because the kinetic energy is not large enough to overcome a potential energy barrier between energy wells, argue that the polypeptide should not unfold. The energy barrier from the global minimum α-helical energy well to high energy local minima corresponding to unfolded structures should be higher than that for the folding processes from local wells to the global well. Therefore, if the simulation has enough energy to escape the α-helical well, it should also escape the well corresponding to the extended structure. At room temperature, the kinetic energy of the polypeptide seems large enough to overcome the highest potential energy barriers from the global energy well.

In thermal equilibrium, the range of potential energies sampled in the simulations of Ala13 at 300 ± 50 K *in vacuo* is between approximately – 20 kcal/mol and + 80 kcal/mol. Fig. 3.3.2.1d displays the potential energy histories of the DIP simulations with twenty trajectories of Ala13 at 300 K *in vacuo* from the initial α-helical conformation. It shows that more potential energy states in this region are sampled than in the six-trajectory DIP simulation with Ala13 (Fig. 3.3.2.1b). If we consider other simulations from different initial configurations such as the extended structure, this range of potential energies is fully sampled and extensively overlapped (data not shown).

The observed mechanical equilibration process and equilibrium states in conventional NVT simulations can explain this unusual unfolding simulation easily. While the mechanical equilibration process can make the polypeptide undergo a large conformational change to escape the deeper global energy well of the α-helix due to unbalanced coordinate displacements and interactions, the mechanical equilibrium limits the simulation to a local search for energy states with similar extended conformations. This causes kinetic problems because it would take a long computational time for the simulations to leave this mechanical equilibrium. This system was thus trapped in the high energy extended conformations. In divergent path search simulations, different polypeptides take different times and sample different potential energies to reach mechanical equilibrium (evaluated from the potential energy histories Fig.3.3.2.1b and d). This can also be easily explained: many different energy wells exist on the three dimensional PES. MD simulations beginning from an initial point can evolve into different trajectories if different velocity directions are assigned to each polypeptide in the given trajectory. Different polypeptides evolve into different potential energy wells along different trajectories. The different trajectories need different equilibration times to reach thermal and mechanical equilibrium.

Potential energy traps seem to be a universally cited cause for the inefficient phase space sampling in conventional MD simulations. In other words, the failure to find the global minimum energy state in conventional NVT simulations is always considered to result from simulations getting trapped in one of many local minimum-energy states [5,6,8,17,25,26,51-54]. For canonical ensemble simulations, these traps may be more accurately described as local free energy traps because the energy barriers between local wells should be on the free energy surface (FES) rather than the potential energy surface (PES). However, escaping local free energy traps also cannot account for the observed unfolding, because once again if the system has sufficient energy to escape the free energy minimum corresponding to the α-helices, it should also escape the well corresponding to the extended structures. Local free energy traps are very closely related to thermal equilibrium. The relationship among PES/FES, free energy traps, and thermal equilibrium is described next.

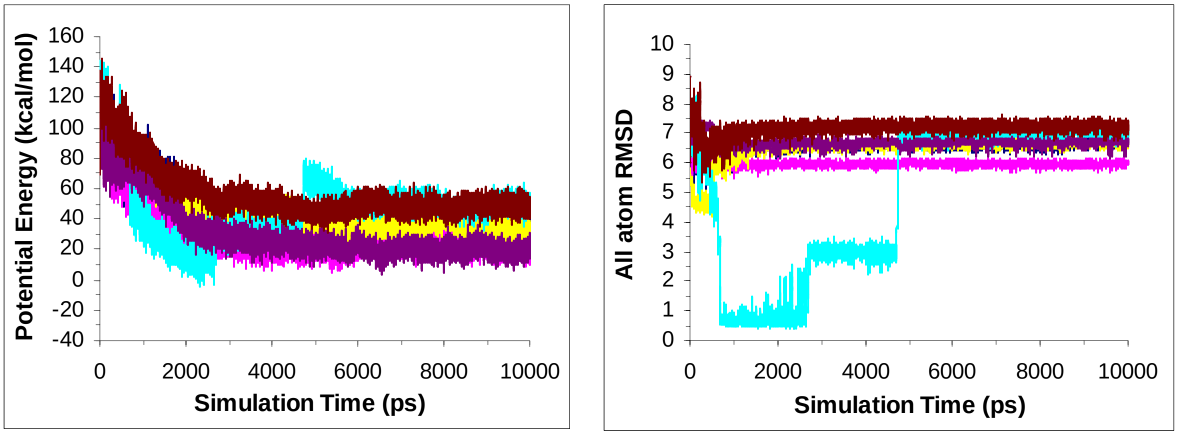
The FES is very difficult to determine, while the PES is relatively easily defined as a function of the atomic coordinates of the system [55] and is independent of temperature, pressure, or any other simulation parameters [56]. Within one minimum on the PES, many more valleys and peaks can appear in the FES (the higher the temperature, the more valleys and peaks appear). The location of these valleys and peaks also shifts along the PES as the temperature changes. Considering the degeneracy of the potential energy states (density of states) in one conformational well, the extra valleys on the FES usually correspond to configurations of higher potential energy than the minimum, while the extra peaks correspond to potential energy minima with low sampling probability. Thermal equilibrium makes simulations become trapped in these free energy minima. In fact, a canonical distribution exists for the potential energies of these free energy minima. The potential energy states of these free energy minima should be the most thermodynamically accessible phase space regions in 3N coordinate dimensions and span a certain range of the PES.

Once mechanical equilibrium is reached, simulations will evolve into a large number of equivalent states at potential energy levels with the same secondary structure. It would take an unrealistic amount of simulation time (at the μs scale or even longer, no one knows *a priori*) to make conformation transitions from one potential well to another potential well in the thermodynamically accessible region of potential energies. This corresponds to kinetic traps. Any small part of a potential well has enormous possible configurations exceeding the number that can be sampled on even the most powerful computers. Even this tremendous number of possible configurations is only a small subgroup of possible conformations. At a low temperature (such as room temperature), once mechanical equilibrium in reached, it is very difficult for further MD simulation to cause a conformational transition to a different secondary structure of the polypeptide. At a high temperature, however, the conformational transition is more likely to happen at mechanical equilibrium due to the stronger interactions (forces) and larger coordinate displacement, which is caused by the higher kinetic energy of each atom.

3.3.3. Global free energy minimum conformation

The above three limitations restricting conformation transitions give a very biased phase sampling in conventional NVT simulations, because sampling is strongly dependent on the initial structures. It is very dangerous to determine the global free energy minimum conformation by assuming it corresponds to the lowest potential energy observed during conventional NVT simulations. Furthermore, the global free energy minimum conformation for a canonical ensemble may not be a single conformation but a distribution of several related conformations, or metastable states [56,57], which can be defined as several nearly isoenergetic conformations [56]. Transitions between such structures can easily happen through protein motions due to small environmental perturbations. This is supported in recent results from the replica exchange method [58] that shows several conformational clusters exist at low temperatures. The dominant conformational clusters are the global free energy minimum conformations.

The results from our DIP simulations indicate that any conformational family's potential energy spans a large range, but different conformations equilibrate at different potential energy levels at the same temperature. This is further supported in our extended DIP simulations starting from a different initial configuration. Fig 3.3.3.1 displays histories of the potential energy and RMSD from an ideal α-helix in a six-path DIP simulation of Ala13 at 300 K *in vacuo*, starting from an extended structure. Five polypeptides show ordinary decreasing potential energy histories (Fig. 3.3.3.1a) but they adopt different conformations (Fig. 3.3.3.1b). However, the fourth polypeptide, displays an unusual potential energy history. In fact, this trajectory indicates a combined folding and unfolding process. Here we will not describe the molecular dynamical picture of helix folding and unfolding in detail, but the analysis of the H-bond formation indicates that the α-helix is initiated by the 1-3 H-bond type β-turn structures [59]. Two β-turns were most likely to cooperate and nucleate a one turn local α-helical conformation which then extends to the whole polypeptide. Table 3.3.3.1 lists the equilibrated average potential energy, average RMSD from an ideal α-helix for the period of 9 ns-10 ns and the final structure after 10 ns for each of six trajectories. Obviously, different conformations equilibrate at different potential energies and the α-helical conformation has the lowest average equilibrated potential energy (approximately 0.5 kcal/mol).



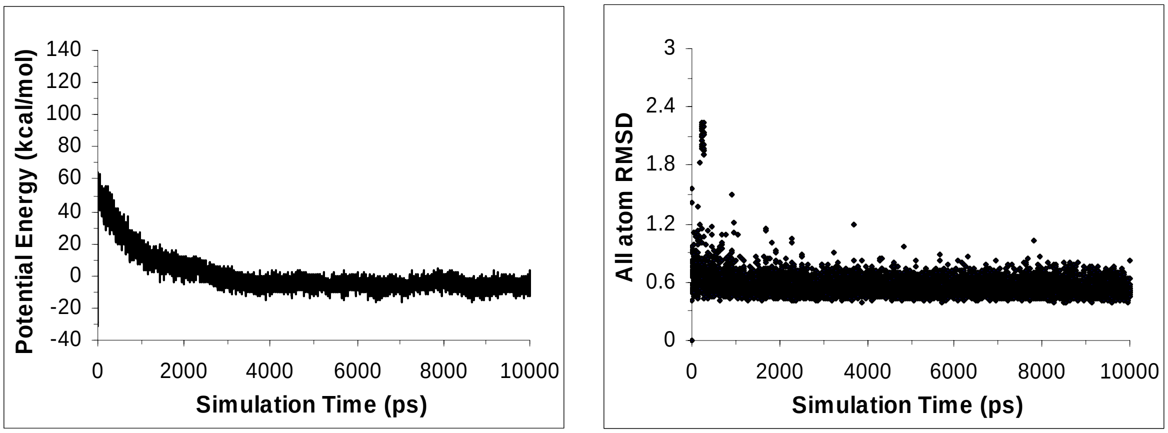
(a) **(b)**

Figure 3.3.3.1. Time series of the potential energy and RMSD from an ideal α-helix in a six-path DIP simulation of Ala13 at 300 K *in vacuo*, starting from an extended structure. **(a)** Potential energy. **(b)** RMSD between all atoms of calculated structure and an ideal α-helix.

Table 3.3.3.1. Equilibrated average potential energy (kcal/mol), average RMSD (Å) from an ideal α-helix for the period of 9 ns – 10 ns and final equilibrated structure after 10 ns for each polypeptide in a six-path DIP simulation *in vacuo* at 300 K, using an extended conformation or an α-helix as initial structure.

Trajectory	MD simulations initiated from the extended structure			MD simulations initiated from the α-helical form		
	Average potential	Ave. RMSD	Equilibrated structure	Average potential	Ave. RMSD	Equil. structure
1st	**38.4**	**6.71**	β-sheets + extended	**-1.74**	**0.54**	α helix
2nd	**20.7**	**5.98**	3-part β-sheets	**-1.74**	**0.54**	α helix
3rd	**33.7**	**6.63**	β-sheets + extended	**63.2**	**8.73**	extended
4th	**49.8**	**7.00**	Almost extend	**-0.84**	**0.55**	α helix
5th	**19.3**	**6.67**	2-part β-sheets	**-2.84**	**0.57**	α helix
6th	**48.7**	**7.23**	Almost extended	**5.00**	**0.57**	α helix

We can analyze the trajectory of the polypeptide with lowest potential energy at each data-collection step to determine the global free energy conformation. If we assume that the number of times an energy minimum is sampled by the copy with lowest potential energy is analogous to a concentration, an equilibrium constant for any conformational change may be defined as the ratio of the number of times the two conformations are sampled. This implies that the global free energy minimum is the minimum most frequently sampled by the polypeptide with lowest potential energy. Fig. 3.3.3.2 displays time series of the potential energy and all-atom RMSD from an ideal α-helix of the polypeptide with lowest potential energy from the twelve trajectories. These twelve trajectories were obtained from the six-path DIP simulations, initiated from a fully extended structure and an ideal α-helix, respectively. The plots indicate that the α-helix is the most frequently sampled structure at the lowest potential energies near 0 kcal/mol. Therefore, the α-helix is the global free energy minimum conformation for blocked Ala13 *in vacuo* at 300 K using the AMBER96 force field. Thus, we can use the trajectory of the polypeptide with lowest potential energies to determine the global free energy minimum conformation.



(a) **(b)**

Figure 3.3.3.2. Time series of the potential energy and RMSD from an ideal α-helix of the polypeptide with the lowest potential energy at each step from the DIP simulations of twelve trajectories of Ala13 *in vacuo* at 300 K. Six trajectories are initiated from a fully extended structure and six trajectories are initiated from an ideal α-helix. **(a)** Potential energy. **(b)** RMSD between all atoms of the calculated structures and an ideal α-helix. **Note**: the RMSD plots here are dot charts.

3.4. Further discussion of three limitations in conventional MD

It is generally accepted that conventional NVT simulations fail to find a global minimum potential energy because simulations at low temperatures tend to get trapped in one of many local minimum-energy states [5,6,8,17,25,26,51-54]. This view must be re-considered because of the difference in global/local minimum states on the FES vs. the PES in canonical ensemble simulations. The quasiergodicity problem is caused by local free energy traps [1] rather than local potential energy traps, but workers in the field primarily consider high energy barriers between different energy wells on the PES [4,60] to describe local traps. Several techniques such as LES have been developed based on the view that potential energy barriers are the primary obstacles to conformational change

[8,10]. However, the potential energy trap is not the only reason for the quasiergodicity problem in time-limited MD simulations; free energy traps and kinetic traps also exist. On limited time scales, simulations fall into free energy traps at thermal equilibrium and sample a relatively narrow range of potential energies, defined as thermodynamically accessible regions corresponding to the system's kinetic energies (and thus temperatures). Furthermore, simulations are limited to a local search for energy states with similar conformations and fall into kinetic traps once mechanical equilibrium is reached. Potential energy traps may not exist but the simulations will still fail to pass across higher potential energy barriers or sample lower potential energy minima.

Currently, two common ways are used to enhance phase space sampling in conventional NVT simulations. One way is to start the system in different initial conformations [61] (Until recently, most published simulations use only two initial conformations, the extended form and the NMR structure [54,62-64]). The second common method is simulated annealing [5-8] to jump energy barriers on the PES. We presented a third way, starting the system with different initial velocity directions, but the same magnitudes (and hence the same temperatures) so that simulations evolve along different trajectories from a single initial configuration. The simulation results for our tests of this divergent path strategy indicate that mechanical equilibrium in conventional MD is the most severe obstacle restricting the simulations from covering a wide variety of conformations. In all, potential energy traps, free energy traps, and kinetic traps constitute three primary limitations of conventional MD simulations and limit the usefulness of conventional NVT/NVE simulations for the predictive protein-folding problem.

3.4.1. Potential energy traps

Potential energy traps occur when the kinetic energy of a system is not large enough to overcome high potential energy barriers between local potential energy minima. When this occurs, simulations become trapped in a local potential energy well. Potential energy traps are actually the least common obstacle to efficient conformational searches in our conventional MD simulations. At room temperature, the kinetic energy available to the system may be large enough to overcome potential energy barriers during a given trajectory. DIP simulations can help to cross potential energy barriers along several different directions and therefore some simulations can reach the global potential energy wells along a folding pathway with lower potential energy-barriers.

3.4.2. Free energy traps, thermal equilibrium and temperature-controlled phase space sampling

Because of the rapid exchange of kinetic and potential energies, limited time simulations at thermal equilibrium can only sample portions of phase space distributed in a relatively narrow range of potential energies corresponding to the system's kinetic energy. We call these portions of the potential energy surface "thermodynamically accessible regions". These thermodynamically accessible regions consist of various substates in different potential energy wells. If the temperature is increased, simulations reach a new state of thermal equilibrium and sample conformations corresponding to higher energy substates in each of the same potential energy wells. This implies that at higher temperatures, only substates of relatively high potential energy are sampled in each well. This is temperature-controlled phase space sampling and can result in a simulation's becoming trapped in local free energy minima at thermal equilibrium, rather than sampling the global potential energy minimum.

Because the most easily accessible states in thermally equilibrated simulations are minimum free energy states, simulations often become trapped in these local free energy minima. Because the free energy is a balance between the enthalpy (including contributions from the potential energy) and the entropy, the local minima in the free energy usually correspond neither to minima nor transition states on the PES. In fact, because the density of states and therefore the entropy is low at potential energy minima, minima on the PES are often peaks on the FES. For an ergodic simulation, these higher free energy states would be accessible if the simulation time were unlimited, but in practice it is very difficult to pass across free energy barriers, so simulations are frequently trapped in local free energy minima and never sample potential energy minima.

In summary, once thermal equilibrium is established, simulations find free energy traps corresponding to neither minima nor transition states on the PES. Such simulations are rarely observed to leave these traps unless the temperature is increased significantly. Then, the structure is no longer equilibrated at the new temperature and the thermal equilibration process begins anew. After thermal equilibrium is achieved, the kinetic energy, but not the potential energy, becomes equilibrated.

3.4.3. Kinetic traps, mechanical equilibrium, and trajectory-controlled phase space sampling

We have observed that after a polypeptide simulation has reached thermal equilibrium, the potential energy continues to decrease, indicating that mechanical equilibrium has not been achieved. During the mechanical equilibration process, the system's potential energy cannot be conserved and local hot spots of high kinetic energy are created. So even though average atomic speeds are uniform after thermal equilibrium is reached, unless mechanical equilibrium has also been achieved, individual atomic velocity vectors are not equilibrated. This may induce large atomic displacements, such as large conformational changes, as well as strong atom-atom interactions within the polypeptide. The common observation that the initial equilibration process in MD simulations are most likely to involve large conformational changes meshes with this interpretation, as does our observation that once mechanical equilibrium is reached, large-scale conformational changes are very rare. When mechanical equilibrium is reached, the individual atomic velocities are equilibrated, making simulations evolve into a large number of equivalent substates in a thermodynamically accessible region. Usually it would take a prohibitively long time before the simulation is able to enter different potential energy wells defining a different thermodynamically accessible region. In conventional NVT molecular dynamics, it is therefore mechanical equilibrium that finally restricts a simulation from sampling a wide variety of different conformations and the simulation occupies a kinetic trap.

While thermal equilibrium limits the thermodynamically accessible regions of potential energy wells that a finite-time NVT simulation may sample at a given temperature, the mechanical equilibration process also determines which substates within the thermodynamically accessible regions are sampled, depending on the starting atomic coordinates and velocities. A single NVT simulation may be trapped in one accessible region even after a very long simulation time. A series of simulations starting from different coordinates and/or velocities can reach many different thermodynamically accessible regions representing diverse conformations. Thus, we can consider the mechanical equilibration process as trajectory-controlled phase space sampling. Because NVT simulations typically reach thermal equilibrium faster than mechanical equilibrium, it is mechanical equilibrium that ultimately determines whether a simulation ends in the global potential energy minimum, a potential energy trap, a free energy trap, or a kinetic trap.

3.5. Summary and discussion

The divergent path search simulations of blocked Ala13 using the AMBER 96 force field support the conclusion that at room temperature the global minimum <u>free energy</u> conformation for this polypeptide *in vacuo* is the α-helical conformation. The DIP simulations from different initial configurations, α-helical and extended forms, give different trajectories but both sets of simulations show that the α-helical conformation occupies the lowest basins on the potential energy surface accessible at 300 K. Independent folding and unfolding processes can be observed from trajectories of some individual polypeptides. An analysis of these folding and unfolding processes supports the idea that the α-helical formations were usually initiated by the β-turn structures with 1 – 3 H-bonds [40,59]. The energy decrease in forming the helical conformation was caused by the Coulombic and van der Waals non-bonded energies [40,65,66]. Torsional energy disfavors the α-helix, contrary to the other simulations based on other force fields [40].

The unfolding trajectory from a stable global potential energy well to a local potential energy well challenges traditional opinions about conventional molecular dynamics. Our simulations indicate the existence of three limitations that prevent conventional MD from making transitions between potential energy well on the time scale of the simulations: (1) potential minimum traps, (2) free energy traps, and (3) kinetic traps. Potential energy traps occur when the kinetic energy is not large enough to overcome high potential energy barriers between local potential energy minima, so the simulations get trapped in the local potential energy well. Free energy traps happen at thermal equilibrium, when the simulations evolve into states in a certain range of potential energies corresponding to the ensemble temperature. These states are thermodynamically accessible during the simulations on a limited time scale, but the extent of their sampling is strongly dependent on the starting structure. Kinetic traps occurs when mechanical equilibrium is reached, indicating that simulations will evolve into a very large number of equivalent substates in some potential energy wells representing a small part of thermodynamically accessible phase space at the simulation temperature. It will take an unreasonable simulation time to sample these myriad substates before making transitions to other potential energy wells. These three limitations constitute serious limitations of conventional NVT molecular dynamics simulations for locating the global potential energy minimum and obtaining accurate phase-space distributions in molecular dynamics.

DIP simulations can alleviate the biased phase space sampling problem in conventional MD simulations and help in achieving both folding or unfolding for some individual polypeptides. While a longer simulation time might not help overcome the three limitations of MD simulations, different initial coordinates should be used with this new strategy to maximize diverse phase space sampling. Recent literature reports accumulated conventional NVT simulations to reproduce the NMR structure of a functional small polypeptide, Trp-cage, in GB/SA implicit solvent starting from many initial conformations [67]. DIP simulations from several different configurations can be useful because a more complete phase space sampling is obtained by simulations beginning from several different conformations combined with several different directions for atomic velocity vectors.

3.6. Bibliography

(1) Vasquez, M.; Nemethy, G.; Scheraga, H. A. *Chemical Reviews* **1994**, *94*, 2183.

(2) Tuckerman, M. E.; Martyna, G. J. *Journal of Physical Chemistry B* **2000**, *104*, 159.

(3) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *Journal of Chemical Physics* **1953**, *21*, 1087.

(4) Calvo, F.; Neirotti, J. P.; Freeman, D. L.; Doll, J. D. *Journal of Chemical Physics* **2000**, *112*, 10350.

(5) Kirkpatrick, S.; Gelatt, C. D. J.; Vecchi, M. P. *Science* **1983**, *220*, 671.

(6) Basu, A.; Frazer, L. N. *Science* **1990**, *249*, 1409.

(7) Huber, G. A.; McCammon, J. A. *Physical Review E: Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **1997**, *55*, 4822.

(8) Roitberg, A.; Elber, R. *Journal of Chemical Physics* **1991**, *95*, 9277.

(9) Elber, R.; Karplus, M. *Journal of the American Chemical Society* **1990**, *112*, 9161.

(10) Stultz, C. M.; Karplus, M. *Journal of Chemical Physics* **1998**, *109*, 8809.

(11) Zheng, W.-M.; Zheng, Q. *Journal of Chemical Physics* **1997**, *106*, 1191.

(12) Hixson, C. A.; Chen, J.; Huang, Z.; Wheeler, R. A. *Journal of Molecular Graphics & Modelling* **2004**, *22*, 349.

(13) Hixson, C. A.; Wheeler, R. A. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics* **2001**, *64*, 026701/1.

(14) Hixson, C. A.; Wheeler, R. A. *Chem.Phys. Lett.* **2004**, *386*, 330.

(15) Mitsutake, A.; Okamoto, Y. *Chemical Physics Letters* **1999**, *309*, 95.

(16) Mitsutake, A.; Okamoto, Y. *Journal of Chemical Physics* **2000**, *112*, 10638.

(17) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96.

(18) Lyubartsev, A. P.; Martsinovskii, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *Journal of Chemical Physics* **1992**, *96*, 1776.

(19) Hukushima, K. *Physical Review E: Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **1999**, *60*, 3606.

(20) Faller, R.; Yan, Q.; de Pablo, J. J. *Journal of Chemical Physics* **2002**, *116*, 5419.

(21) Frantz, D. D.; Freeman, D. L.; Doll, J. D. *Journal of Chemical Physics* **1990**, *93*, 2769.

(22) Frantz, D. D.; Freeman, D. L.; Doll, J. D. *Journal of Chemical Physics* **1992**, *97*, 5713.

(23) Matro, A.; Freeman, D. L.; Topper, R. Q. *Journal of Chemical Physics* **1996**, *104*, 8690.

(24) Curotto, E.; Freeman, D. L.; Doll, J. D. *Journal of Chemical Physics* **1998**, *109*, 1643.

(25) Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **1999**, *314*, 141.

(26) Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **2000**, *329*, 261.

(27) Sugita, Y.; Kitao, A.; Okamoto, Y. *Journal of Chemical Physics* **2000**, *113*, 6042.

(28) Hansmann, U. H. E.; Okamoto, Y. *Journal of Chemical Physics* **1999**, *110*, 1267.

(29) Hansmann, U. H. E. *Chemical Physics Letters* **1997**, *281*, 140.

(30) Hansmann, U. H. E.; Okamoto, Y. *Physical Review E: Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **1996**, *54*, 5863.

(31) Daggett, V.; Levitt, M. *Journal of Molecular Biology* **1992**, *223*, 1121.

(32) Marqusee, S.; Robbins, V. H.; Baldwin, R. L. *Proceedings of the National Academy of Sciences of the United States of America* **1989**, *86*, 5286.

(33) Merutka, G.; Lipton, W.; Shalongo, W.; Park, S. H.; Stellwagen, E. *Biochemistry* **1990**, *29*, 7511.

(34) Chakrabartty, A.; Schellman, J. A.; Baldwin, R. L. *Nature* **1991**, *351*, 586.

(35) Padmanabhan, S.; Marqusee, S.; Ridgeway, T.; Laue, T. M.; Baldwin, R. L. *Nature* **1990**, *344*, 268.

(36) Fiori, W. R.; Miick, S. M.; Millhauser, G. L. *Biochemistry* **1993**, *32*, 11957.

(37) Hudgins, R. R.; Jarrold, M. F. *Journal of the American Chemical Society* **1999**, *121*, 3494.

(38) Son, H. S.; Hong, B. H.; Lee, C. W.; Yun, S.; Kim, K. S. *Journal of the American Chemical Society* **2001**, *123*, 514.

(39) Bertsch, R. A.; Vaidehi, N.; Chan, S. I.; Goddard, W. A., III. *Proteins: Structure, Function, and Genetics* **1998**, *33*, 343.

(40) Takano, M.; Yamato, T.; Higo, J.; Suyama, A.; Nagayama, K. *Journal of the American Chemical Society* **1999**, *121*, 605.

(41) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Clarendon Press: Oxford, 1987.

(42) Martyna, G. J.; Klein, M. L.; Tuckerman, M. E. *Journal of Chemical Physics* **1992**, *97*, 2635.

(43) Kollman, P.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A. *The development/application of a \"minimalist\" organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data*, 1997; Vol. 3.

(44) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Ferguson, D. M.; Radmer, R. J.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. AMBER 5; University of California: San Franciso, 1997.

(45) Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Langridge, R. *Journal of Molecular Graphics* **1988**, *6*, 13.

(46) Voet, D.; Voet, J. G. Biochemistry; John Wiley & Sons, Inc.: New York, 1995; pp 144.

(47) Hinsen, K. *Journal of Computational Chemistry* **2000**, *21*, 79.

(48) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *Journal of Chemical Physics* **1982**, *76*, 637.

(49) Palmer, B. J. *Journal of Computational Physics* **1993**, *104*, 470.

(50) Goldstein, H. Classical Mechanics; Addison-Wesley: MA, 1965; pp 69.

(51) Bassolino-Klimas, D.; Tejero, R.; Krystek, S. R.; Metzler, W. J.; Montelione, G. T.; Bruccoleri, R. E. *Protein Science* **1996**, *5*, 593.

(52) Hornak, V.; Simmerling, C. *Proteins: Structure, Function, and Genetics* **2003**, *51*, 577.

(53) Yoda, T.; Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **2004**, *386*, 460.

(54) Jarrold, M. F. *Annual Review of Physical Chemistry* **2000**, *51*, 179.

(55) Mortenson, P. N.; Wales, D. J. *Journal of Chemical Physics* **2001**, *114*, 6443.

(56) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598.

(57) Li, Z.; Scheraga, H. A. *Proceedings of the National Academy of Sciences of the United States of America* **1987**, *84*, 6611.

(58) Pitera, J. W.; Swope, W. *Proceedings of the National Academy of Sciences of the United States of America* **2003**, *100*, 7587.

(59) Rose, G. D.; Gierasch, L. M.; Smith, J. A. *Advances in Protein Chemistry* **1985**, *37*, 1.

(60) Huber, T.; Torda, A. E.; van Gunsteren, W. F. *Journal of Physical Chemistry A* **1997**, *101*, 5926.

(61) Andersen, H. C. *Journal of Chemical Physics* **1980**, *72*, 2384.

(62) Hudgins, R. R.; Mao, Y.; Ratner, M. A.; Jarrold, M. F. *Biophysical Journal* **1999**, *76*, 1591.

(63) Hudgins, R. R.; Jarrold, M. F. *Journal of Physical Chemistry B* **2000**, *104*, 2154.

(64) Hartings, M. R.; Kinnear, B. S.; Jarrold, M. F. *Journal of the American Chemical Society* **2003**, *125*, 3941.

(65) Scholtz, J. M.; Marqusee, S.; Baldwin, R. L.; York, E. J.; Stewart, J. M.; Santoro, M.; Bolen, D. W. *Proceedings of the National Academy of Sciences of the United States of America* **1991**, *88*, 2854.

(66) Avbelj, F.; Fele, L. *Journal of Molecular Biology* **1998**, *279*, 665.

(67) Snow, C. D.; Zagrovic, B.; Pande, V. S. *Journal of the American Chemical Society* **2002**, *124*, 14548.

Chapter 4

Disrupted Velocity Search Protocols for Mapping Potential Energy Landscapes and Conformations of Polyalanine

4.1. Introduction

Molecular dynamics (MD) simulation methods [1,2] are widely used for global conformational searches in protein folding. Conventional canonical ensemble simulations usually have difficulty in exploring the biologically active natural structures of proteins because these simulations have severe phase space sampling problems whose results depend strongly on the initial structure. To overcome this limited phase space sampling problem, published research focuses on enhancing energy barrier crossings by using specialized MD techniques such as Simulated Annealing (SA) [3-6], Locally Enhanced Sampling (LES) [6-9], and the Replica Exchange Method (REM) [10-14].

The recently developed REM method is a good MD simulation method for addressing the protein-folding problem. This technique facilitates barrier crossings by temperature exchange between different copies of the protein. However, the REM method searches for the global minimum on the free energy surface and is not good at locating the minimum on the temperature-independent potential energy surface (PES) [15]. Furthermore, velocity rescaling by temperature exchange in REM usually only focuses on moving vertically on the energy surface to go over barriers. This may not be enough for a simulation to search for a global minimum in a multidimensional energy surface starting from a random structure. These considerations have spurred our recent research.

A common way to prepare MD simulations is to initialize the system in several different conformations. We consider another alternative — starting multiple simulation trajectories from a single conformation but different randomized velocity directions so the simulations evolve along different paths (Chapter 3). With this divergent path search simulation, we found that peptide folding from high energy local minimum structures and unfolding from the global minimum energy conformation are possible at room temperature during molecular dynamics simulations. These observations made us realize that three primary limitations account for the general failure of global minimum searches in molecular dynamics. Besides local minima traps (potential energy and free energy), a third limitation is kinetic traps, whose chief symptom is all 3N velocities universally equilibrated along the whole polypeptide. The equilibrated velocities, called mechanical equilibrium, usually cause balanced displacements and forces that make conformational transitions of polypeptides difficult. Simulations then oscillate between a large number of substates with similar secondary structures. This mechanical equilibrium constitutes the most restrictive condition preventing diverse phase space sampling in the search for global minimum-energy folded conformations when simulations start from a random, unfolded structure. On the other hand, if mechanical equilibrium can be disrupted, the unbalanced displacements and locally high potential energies caused by strong interactions between atoms can cause the system to explore generally thermodynamically inaccessible states from which conformational transitions can happen much more easily.

In this study, we utilize the equilibration process usually discarded in molecular dynamics and develop the DIsrupted VElocity (DIVE) search simulations by intervening to reassign atomic velocities. Periodic intervention to reassign atomic velocities may partially destroy the Boltzmann distribution and make direct thermodynamics calculation impossible, but we focus on exploring the potential energy surface rather than the temperature-dependent free energy surface. We consider it more important to explore the potential energy surface than to calculate thermodynamic properties when simulations may be limited to partial phase space sampling. Therefore, our technique complements the majority of current MD based simulation protocols designed to enhance phase-space sampling.

In this technique, several polypeptides with different kinetic energies (and therefore different total energies) are simulated independently and simultaneously, and their atomic velocities are reassigned after a fixed time period of NVE (constant number of particles, N, volume, V, and energy, E) molecular dynamics simulations. In this technique, NVE simulations are used rather than NVT because mechanical equilibration is quicker for NVE simulations. We emphasize that velocity(v) reassignment is used here very differently from Andersen dynamics [16] whose purpose in reassigning atomic velocities is to maintain a constant temperature. When we desire to use constant temperature conditions, we use the Nosé-Hoover chain method [17] to maintain a constant temperature.

The v-reassignment algorithm includes two distinct steps: 1) each atom's velocity vector is re-directed but its magnitude does not change, and 2) the magnitude of each atom's velocity is rescaled but its direction does not change. The v-redirection step can disrupt the established mechanical equilibrium of the polypeptide in each trajectory, ensure energy conservation and enhance a polypeptide's chances of going around energy barriers rather than over them. Thus, the v-reassignment algorithm can facilitate conformational transitions, even in low temperature MD simulations. Moreover, the v-rescaling can break the established thermal equilibrium by changing the energy of the polypeptide by multiplying the magnitude of each atom's velocity by a scaling factor, and thus enhancing the atom's ability to surmount energy barriers.

These features make the DIVE simulations free from the primary limitations of conventional MD simulations and enable them to sample diverse regions of conformational space very effectively. Consequently, DIVE simulation identifies potential energy minima quickly and accurately. In our experience, NVE simulations locate potential energy minima faster than NVT simulations because energy regulation by a heat bath makes the mechanical equilibration process much slower in NVT simulations (Chapter 3). Furthermore, the potential energy minima in DIVE are mapped at very low energies corresponding to temperatures near 0 K. At such low kinetic energies, the potential energy minima can be mapped within approximately 1 - 2 kcal/mol.

Though variations of the replica exchange method exist [10-14], the DIVE simulation is different from them in at least two respects. First, the DIVE simulation uses microcanonical ensemble simulations instead of the canonical ensemble simulations used in REM. Second, the v-redirection algorithm is a very important component in the DIVE simulation (Chapter 3). Currently, we focus on locating diverse potential energy minima rather than calculating statistical average ensemble properties. Our experience show that existing methods such as REM sample very small portions of conformational space during limited simulation times and experience repeated structural transitions among several closely related conformations. DIVE also samples small portions of conformational space, but it samples widely-separated portions rather than experiencing repeated structural transitions among related conformations. The DIVE simulations involve segments of microcanonical ensemble simulations but the total energy of the system does not always remain conserved from the beginning to the end. The total energy is conserved only for a fixed time during which the exchange of kinetic energy and potential energy happens quickly, until equilibrium is reached (The classic *virial* theorem indicates that the kinetic energy controls sampling of potential energy states [18]).

Here, we selected a 13-residue polypeptide of alanine (Ala13) to test the DIVE simulation technique. Polyalanine is the simplest peptide to adopt the α-helical conformation [19], which is the most abundant and important secondary structural element of proteins. Experiments [20-25] show that short ala-based peptides appear to form an α-helical conformation in aqueous solution and simulations of uncharged polyalanines with a sequence length between 10 and 30 indicate that the α-helical conformation is the lowest energy folded conformation for the peptides both *in vacuo* and in aqueous environments [11,19,26-29]. We selected polyalanine for the test because its conformations with low potential energies are limited and include well-defined secondary structural elements of proteins such as α-helices and β-sheets.

4.2. Simulation methodology

4.2.1. Simulation algorithms

All calculations reported in this work, except REM simulations, were performed by using our own implementations of conventional canonical MD, disrupted velocity (DIVE) search simulations and an extreme version (EXREM) of REM simulations in which the acceptance ratios of replica exchange are always 100%. In the DIVE simulations described here, we simulate six independent trajectories of Ala13 simultaneously. No polypeptide interacts with the others, so each one follows its own trajectory in a conventional NVE simulation. All of these independent polypeptides are assigned the same initial structure but different initial atomic velocities and therefore different energies and temperatures. Each polypeptide NVE simulation is propagated at constant energy for a fixed time interval (e.g. 50 ps). At the end of each time interval, the polypeptide in each given trajectory has atomic velocities (momenta) rescaled according to the algorithm presented in equation (4.2.1.1):

$$\mathbf{p}' \equiv \sigma^{1/2}\mathbf{p} \tag{4.2.1.1}$$

where **p** and **p'** are the momenta of the particles before and after velocity rescaling, respectively. The scaling parameters σ in equation (4.2.1.1) determine the magnitude of the simulation temperature after velocities are rescaled and may be selected before starting a simulation. $\sigma > 1$ is for heating and $0 < \sigma < 1$ is for cooling. Alternatively, σ can be calculated from the scaling temperature defined as the temperature difference before and after velocity rescaling (ΔT), according to equation (4.2.1.2), or a target temperature defined as the temperature after velocity rescaling (T_{target}), according to equation (4.2.1.3).

$$\sigma \equiv \frac{|T - \Delta T|}{T} \tag{4.2.1.2}$$

$$\sigma \equiv \frac{T_{t\arg et}}{T} \tag{4.2.1.3}$$

While ΔT and T_{target} are input at the start of a simulation, T is the temperature immediately before rescaling. A very low threshold temperature (e.g. 10 K, 5 K, or 1 K) is defined as the temperature below which the system needs to gain energy and above which the system needs to give some energy away. At low threshold temperatures near 0 K, potential energy minima were mapped within approximately 1 - 2 kcal/mol. Practically, cooling occurs several times but heating only occurs once within a heating-cooling cycle and each heating-cooling cycle spans several 50 ps simulation intervals. After every velocity rescaling, velocities are also redirected while their new magnitudes are maintained. The EXREM simulations are different from DIVE by virtue of lacking the v-redirection algorithm and by using the Nosé-Hoover Chain method [17] to maintain a constant temperature after each v-scaling (heating or cooling) for the NVT simulations. We used Amber 8 [30] for our REM simulations and they are described in detail in the next section.

4.2.2. Computational details

The generic force field for the polypeptide in gas phase is shown in equation (4.2.2.1) and consists of terms representing bond stretching and angle bending terms by a simple harmonic expression, dihedral angle twisting by a truncated Fourier series, the van der Waals interaction by a Lennard-Jones potential, and electrostatic interactions by a Coulombic interaction of atom-centered partial charges. The AMBER96 force field [31] was adopted for all parameters in equation (4.2.2.1).

$$\begin{aligned} E_{potential} = {} & \sum_{\text{bonds}} K_r (r - r_{eq})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2 \\ & + \sum_{\text{dihedral}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}} \right] \end{aligned} \qquad (4.2.2.1)$$

The AMBER package [32] was used to generate the initial coordinates of a 13-residue peptide of alanine with the extended structure (all φ = 180°, all ϕ = 180°). The termini of this Ala13 peptide were blocked (acetylated at the N terminus and methyl-amidated at the C terminus). The initial coordinates of the 3_{10}- and α-helical form Ala13 were obtained from the Midas program [33] by setting each combination of ϕ, φ torsional angels of the polypeptide backbone to the ideal values: $\phi = -49°$, $\varphi = -26°$, and $\phi = -57°$, $\varphi = -47°$, respectively [34]. The Molecular Modeling Toolkit [35] was used to convert input files from the AMBER format to that required by our programs. We used the velocity-Verlet algorithm [2,36,37] to integrate the equations of motion, and the SHAKE algorithm [38] was used to constrain the covalent bond distances involving hydrogen. Translation of the center of mass of the entire system was removed at each step.

To run DIVE simulations we simulated six independent trajectories of blocked Ala13 simultaneously with different initial temperatures of 10 K, 50 K, 100 K, 300 K, 600 K, 1000 K. The scaling parameter for cooling was 0.25 and for heating was calculated from the target temperature T_{target} and the temperature (T) at the v-reassignment step by $\sigma = T_{target}/T$. T_{target} was selected to be 1000 K or 1400 K in the gas-phase simulations. The threshold temperature for heating and cooling was 10 K. Thus, during the simulations, the polypeptide in each trajectory was cooled down to ¼ of its temperature T whenever T rose above the threshold temperature. Once the temperature was below 10 K at the v-reassignment time, the polypeptide was heated back to T_{target}. During the simulations, velocity reassignment occurred after every 50 ps. A time step of 1 fs was used and the trajectory data (energies and coordinates) were collected at each ps interval (1000 steps). We used the Carnal program from the AMBER package [32] and our own programs to analyze coordinate data.

The same initial temperatures were used for six trajectories of conventional NVT simulations while the Nosé-Hoover chain method [17] was used to maintain these temperatures. In REM simulations, we used six exponentially distributed temperatures from the original schedule of REM simulations: 239, 286, 342, 409, and 585 K [12,14]. The weak-coupling algorithm [39] was applied to ensure constant temperature. Additional REM simulations with the exchange temperatures 10 K, 50 K, 100 K, 300 K, 600 K, 1000 K were also performed. The acceptance ratios of replica exchange in these simulations were always zero or near zero, depending upon other parameters such as initial temperatures and frequency of replica exchange attempts. When the acceptance ratios were zero, the REM simulations became conventional NVT simulations of six trajectories at different temperatures. The results were not given here because of their inefficient acceptance ratio. The time step was also set to 1 fs and the replica exchange was attempted every 2 ps, with 5000 exchange attempts. Thus, the total simulation time was also 10 ns for each replica. The acceptance ratios of replica exchange are nearly uniform (14% to 18%) and within a conventionally acceptable range (>10%).

The EXREM simulations can be viewed an extension of REM simulations in which the acceptance ratios of replica exchange are always 100%. In fact, the EXREM simulations can achieve the same result as REM [10-14] if several finite temperatures are fixed for cooling and heating and the Metropolis criteria [40] is used for the heating step. For both EXREM and REM simulations, the velocity-rescaling algorithm is the core strategy to enhance phase space sampling, and the NVT ensemble is involved. The

EXREM simulations were further compared with DIVE simulations, in order to delineate clearly the importance of the v-redirection algorithm for diverse conformational space sampling on a very limited time scale. The six replicas of EXREM simulations all had the same simulation parameters as those in the DIVE simulations.

4.3. Simulation results and discussion

4.3.1. Conventional NVT, REM, and EXREM simulations

In order to search for the global α-helical conformation of Ala13, a test case for our DIVE technique, we first performed several different MD simulations, including conventional NVT, simulated annealing (SA) [3-6], locally enhanced sampling (LES) [6-9,41], REM and EXREM. We were unable to achieve an α-helix when performing several simulations starting from a fully extended structure with the above listed MD methods. Here, we will not show the simulation data from SA and LES techniques.

Conventional NVT simulations were performed with six trajectories at different temperatures. Conventional NVT simulations give three different results (Fig. 4.3.1.1). First, the simulations at 10 K, 50 K, and 100 K get trapped immediately in extended conformations. Second, the simulations at 300 K and 600 K show some conformational transitions in the equilibration process during the first 2 ns, followed by the simulations either oscillating among the extended conformational clusters (RMSDs between 8.5 Å – 9 Å) or nearly extended conformational clusters (RMSDs between 7.0 Å – 7.5 Å). Third, the simulation at 1000 K shows a very different trajectory. The RMSDs of the sampled conformations oscillate over a larger range of conformations (RMSDs between 3.0 Å –

9.0 Å) indicating a more diverse conformational space sampling process. Two reasons account for this result. First, the temperature fluctuates over a larger range (more than ±100 K) at 1000 K. The high kinetic energy at this temperature easily causes coordinate displacements large enough to cause conformational transitions. Second, the simulation equilibrates within the potential energy range between 150 and 250 kcal/mol after 2 ns (data not shown here). At these high energies, most of the energy barriers between the wells on the potential energy surface become nearly insignificant compared to the kinetic energy. The sampled conformations are mainly loops, coiled structures or β-sheet-like conformations. Partial α-helical forms can be also sampled, but at this high temperature, they are unstable and disappear very quickly. In a word, at high temperatures, the conformations from the conventional NVT simulations are likely to be flexible but the nearly ideal α-helix and double stranded β-sheets are not located.

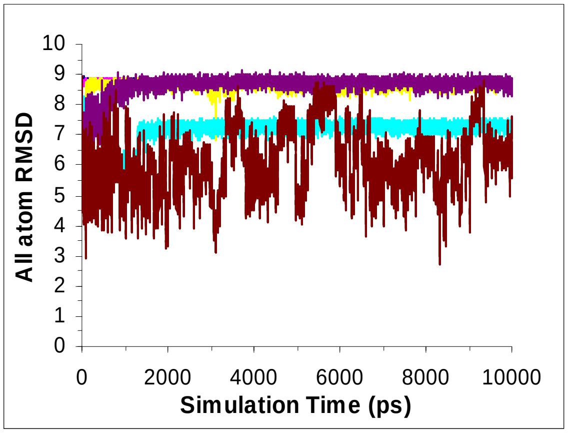
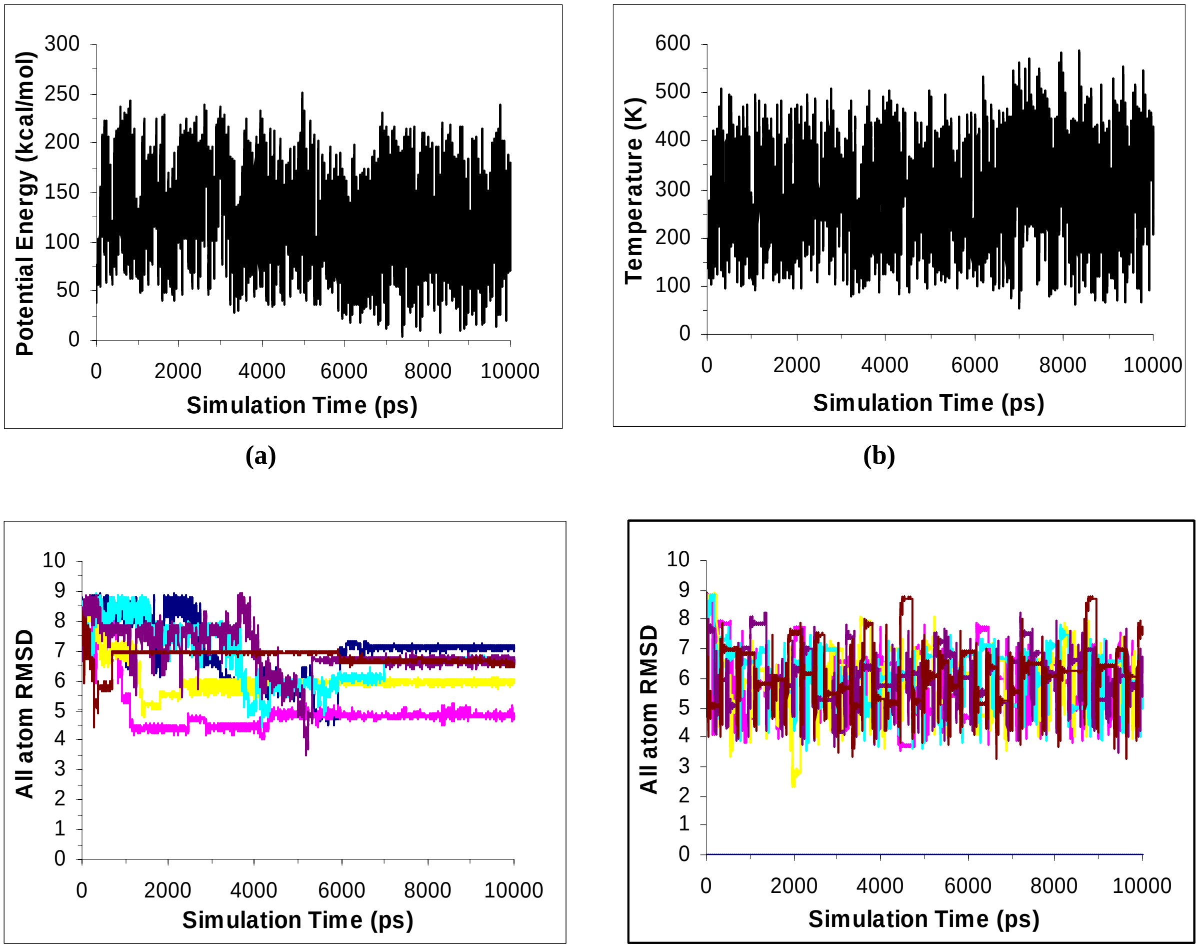


Figure 4.3.1.1. Time series of RMSD between all atoms of the calculated structure and an ideal α-helix of the trajectories from conventional NVT simulations. **Note**: Blue: 1st (10 K), Magenta: 2nd (50 K), Yellow: 3rd (100 K), Light blue: 4th (300 K), Purple: 5th (600 K), Brown: 6th (1000 K) trajectory.

In the REM simulations, each copy of Ala13 *in vacuo* samples a large range of potential energies from nearly 0 to 230 kcal/mol and temperatures ranging from 100 K to 550 K. Fig. 4.3.1.2a – b show the histories of the potential energy and temperature of the first copy as an example. Due to the temperature exchange, the REM simulations can sample a large range of temperatures and potential energies, which allows each replica to move away from any local minimum traps and undergo conformational changes [10,14]. Fig. 4.3.1.2c displays the histories of RMSDs between all atoms of six copies in the REM simulations and an ideal α-helix. Each copy of the REM simulations explores mainly conformations in a range of RMSDs between 4.0 Å – 9.0 Å. The lowest RMSD structure sampled starting from an ideal α-helix is 3.5 Å. By visualizing its conformations along the trajectories, we found that REM simulations only searched several conformational clusters, including partial α-helical structures, loop structures, and loosely extended structures.



(a)

(b)

(c)

(d)

Figure 4.3.1.2. Time series of various properties from a REM or EXREM simulation for six copies of Ala13 *in vacuo*, starting from an extended structure. **(a)** Temperature of the 1st copy in REM simulations. **(b)** Potential energy of the 1st copy in REM simulations. **(c)** RMSD between all atoms of the calculated structure and an ideal α-helix of the six copies in REM simulations. **(d)** RMSD between all atoms of the calculated structure and an ideal α-helix of the six copies in EXREM simulations. **Note**: Blue: 1st, Magenta: 2nd, Yellow: 3rd, Light blue: 4th, Purple: 5th, Brown: 6th replica.

In the EXREM simulations, each copy of blocked Ala13 *in vacuo* samples a large range of potential energies from nearly -55 to 480 kcal/mol and temperatures ranging from several K to 1100 K (data not shown here). Fig. 4.3.1.2d shows the RMSD histories of six copies from the EXREM simulations. The plots indicate that each replica samples mainly conformations with RMSDs between 3.5 Å and 9.0 Å. The lowest RMSD structure sampled is 2.3 Å from an ideal α-helix. By visualizing the conformations, we know that these simulations mainly sample varied coiled and loop structures with only a few samplings of β-sheets and partial helices. A nearly ideal α-helix is also not sampled in these EXREM simulations.

4.3.2. Disrupted velocity (DIVE) search simulations

4.3.2.1. Methodology

Figure 4.3.2.1.1 displays a partial history of total energy from the DIVE simulations of six trajectories of Ala13 *in vacuo* by using the AMBER96 force field, starting from an extended structure. It illustrates three main points. First, each NVE simulation is propagated at constant energy for 50 ps time intervals. After that, the trajectory is interrupted by velocity-rescaling and velocity-redirection algorithms. The velocity-rescaling regulates the polypeptide’s energy because high energy is needed to cross energy barriers and low energy is needed to map potential energy minima. The velocity-redirection directs the trajectory into different paths on the multidimensional potential energy surface. Therefore, we can consider the DIVE simulation protocol as a technique for the dynamical perturbation of trajectory and energy in constant energy MD simulations.

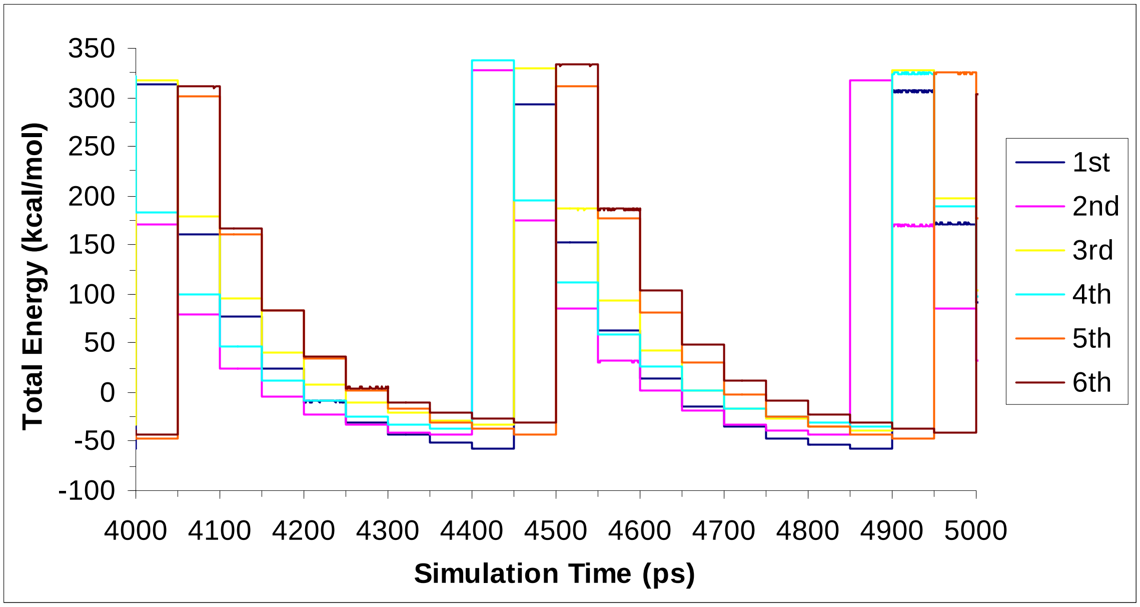
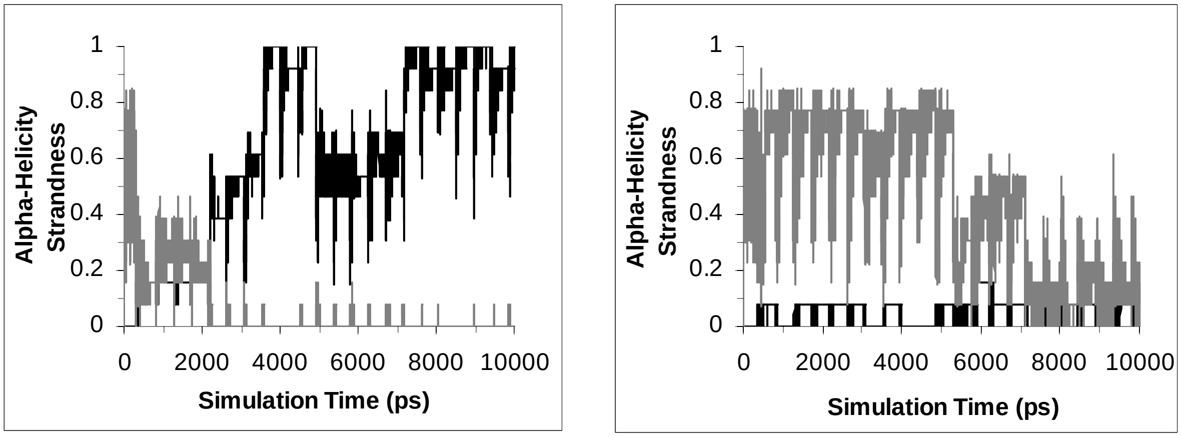


Figure 4.3.2.1.1. Partial history of total energy from a DIVE simulation of six trajectories of Ala13 *in vacuo*, starting from an extended structure.

Second, cooling occurs several times but heating only occurs once within a heating-cooling cycle. We heat once in a heating-cooling cycle because we intended the polypeptide to gain enough kinetic energy to cross high energy barriers. We use the cooling algorithm to locate potential energy minima. However, cooling cannot be accomplished in one step, even if the velocity scaling parameter is very small. A very small scaling parameter can reduce the system's temperature directly below 10 K, but because the system is probably not at a potential energy minimum and the kinetic energy and potential energy are then exchanged quickly according to the *virial* theorem [18]. The system's temperature increases above 10 K and further cooling is needed. An appropriate cooling schedule is therefore needed, resulting in a long simulation time for a slow cooling speed. The simulation usually cannot locate the global potential energy minimum if very fast cooling is used.

Third, different independent polypeptides have their velocities reassigned at different energies and by different amounts. This enhances the conformational search during the simulations. On a multi-dimensional potential energy surface, simulation trajectories are always closely related to the initial conditions. The different initial energy states for different polypeptides may make the simulations sample very different ranges of conformational space.

Figure 4.3.2.1.2 shows time series of the α-helicity and the strandness evaluated from dihedral angles (α-helicity and the strandness are defined in the figure caption). The results clearly demonstrate diverse conformational space sampling of the different trajectories arising from to their different initial energies. Therefore, simultaneous simulations of multiple trajectories can sample much more diverse conformational space than one single trajectory, even if the single trajectory simulation runs much longer. Simulations using different initial momenta and therefore different energies for the polypeptide are similar to starting from several different initial configurations, but we want to emphasize the significance of different trajectories in sampling diverse conformational space even starting from the same coordinates. In the following section, the conformations resulting from different simulations demonstrate the diverse conformational space sampling and effective global minimum search features of the technique.

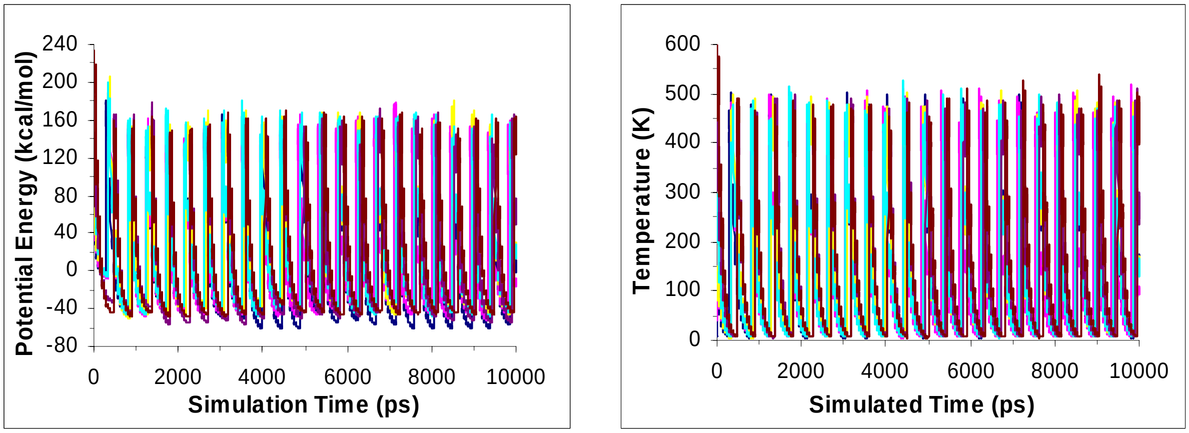


(a) **(b)**

Figure 4.3.2.1.2. Time series of the α-helicity (black) and the strandness (gray) of the polypeptide evaluated from dihedral angles from a DIVE simulation of six trajectories of Ala13 *in vacuo*, starting from an extended structure. **(a)** α-helicity and the strandness of a single polypetide's trajectory **(b)** α-helicity and the strandness of a second polypeptide's trajectory. The α-helicity and standness of each amino-acid is defined as follows: a residue is in the α-helical state when the backbone dihedral angles (ϕ, φ) fall in the range (-57°±30°, -47°±30°), and in the β-strand state if (ϕ, φ) dwell in the range (-119°±30°, -113°±30°) or (-139°±30°, -135°±30°) [27,42]. The α-helicity and standness of the polypeptide is defined as the percent of the residues whose torsional angles fall in the corresponding range.

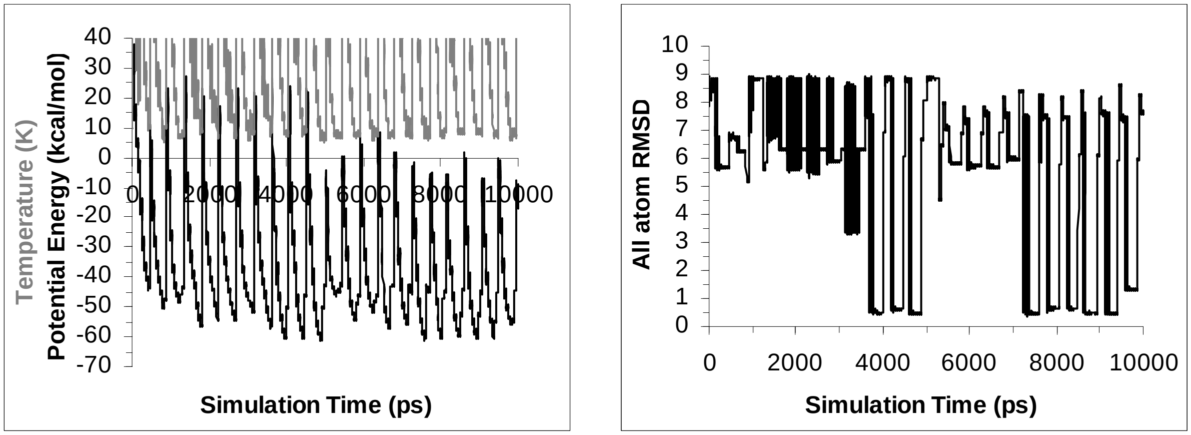
4.3.2.2. Simulation results

The DIVE simulations for blocked Ala13 *in vacuo* sample a large range of potential energies from nearly –60 to 160 kcal/mol and temperatures ranging from several K to almost 500 K (Fig. 4.3.2.2.1) (The simulations involve heating to 1000 K, but after approximately 1 ps (1000 steps), half of the kinetic energy converts to potential energy. Consequently, the high temperature states where data are collected sample temperatures no higher than ~500 K). Fig. 4.3.2.2.2a displays the temperature and potential energy histories of the polypeptide with lowest potential energy at each data-collection step (1 ps intervals). Many local potential energy minima appear along the trajectory at kinetic energies corresponding to a temperature below 10 K. Fig. 4.3.2.2.2b shows the all-atom root-mean-square-deviation RMSD between the polypeptide with the lowest potential energy and an ideal α-helical reference structure. Figures 4.3.2.2.2c – d show the total number of hydrogen bonds (H-bond) and 1-4 H-bonds, respectively. The RMSD and H-bonding plots show that the polypeptide forms a nearly ideal α-helix after approximately 3.8 ns (eleven 1-4 H-bonds and ~ 0.4 Å all-atom RMSD).

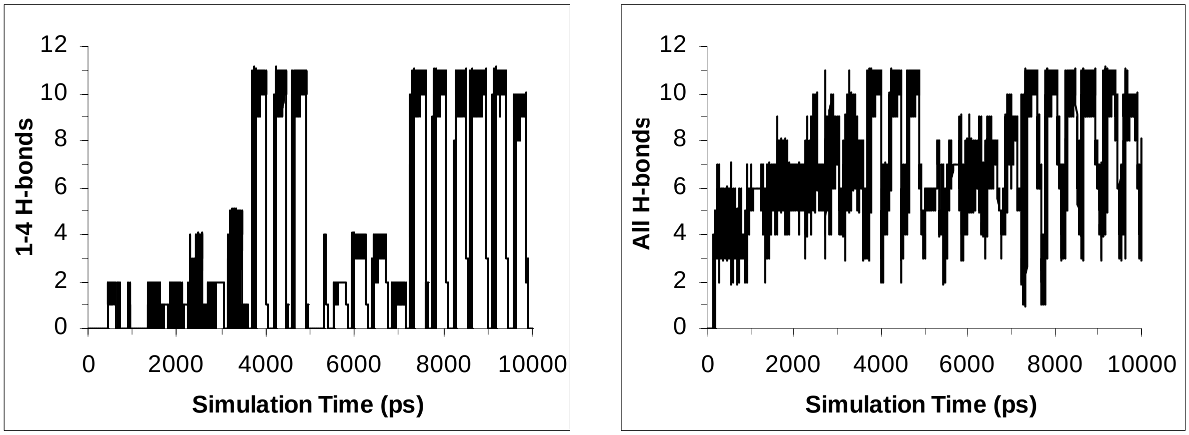


(a) **(b)**

Figure 4.3.2.2.1. Time series of **(a)** potential energy and **(b)** temperature from a DIVE simulation of six trajectories of Ala13 *in vacuo*, starting from an extended structure. **Note**: Blue: 1st, Magenta: 2nd, Yellow: 3rd, Light blue: 4th, Purple: 5th, Brown: 6th trajectory.



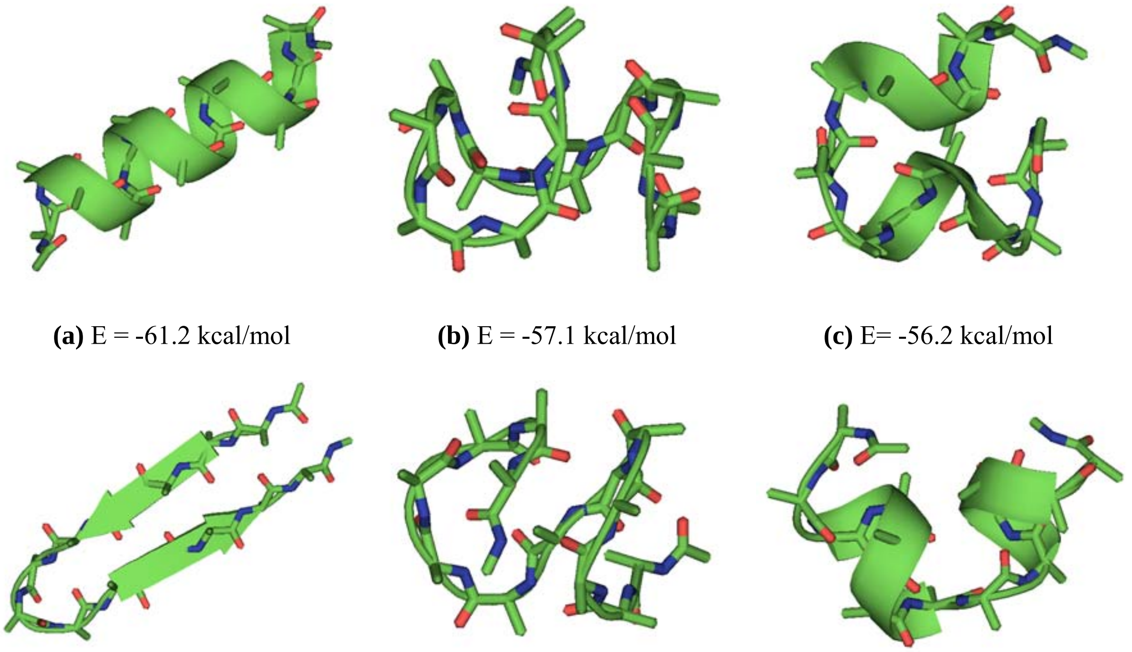
(a) **(b)**



(c) (d)

Figure 4.3.2.2.2. Time series of various properties of the polypeptide with lowest potential energy from a DIVE simulation of six trajectories of Ala13 *in vacuo*, starting from an extended structure. **(a)** Temperature (upper, gray curve) and potential energy (lower, black curve). **(b)** RMSD between all atoms of the calculated structure and an ideal α-helix. **(c)** All hydrogen bonds (A hydrogen bond is identified when the distance between two heavy atoms X is below 3.3 Å and the X-H···X angle differs from 180.0° less than 20.0°). **(d)** 1-4 hydrogen bonds (1-4 H-bonds indicate H-bonds between the carboxyl oxygen of residue i and the amide hydrogen of residue i+4, which are characteristic of an α–helix. An ideal α-helix for Ala13 has eleven 1-4 hydrogen).

We calculated the average temperature in each 50 ps interval and then collected those simulation regions whose average temperature is below 10 K. Forty-nine regions are obtained from the trajectory of the lowest potential energy polypeptide. Within those regions, we selected the lowest potential energy to determine representative minimum potential energy conformations. Some representative structures are shown in Fig. 4.3.2.2.3a - f. The ground state conformation of blocked Ala13 in *vacuo* is a nearly ideal α-helix (Fig. 4.3.2.2.3a) with a global potential energy minimum of approximately –60 kcal/mol. Many local potential energy minima exist between –60 and -40 kcal/mol. Their conformations are diverse but are all well-ordered, with more than 7 H-bonds. Though the energy gap between potential energy minima is relatively small (< 4 kcal/mol between any neighboring minima), the RMSDs between different conformations and an ideal α-helix differ by as much as 9 Å, so the different potential energy minima represent distinctly different structural types.



(a) E = -61.2 kcal/mol

(b) E = -57.1 kcal/mol

(c) E= -56.2 kcal/mol

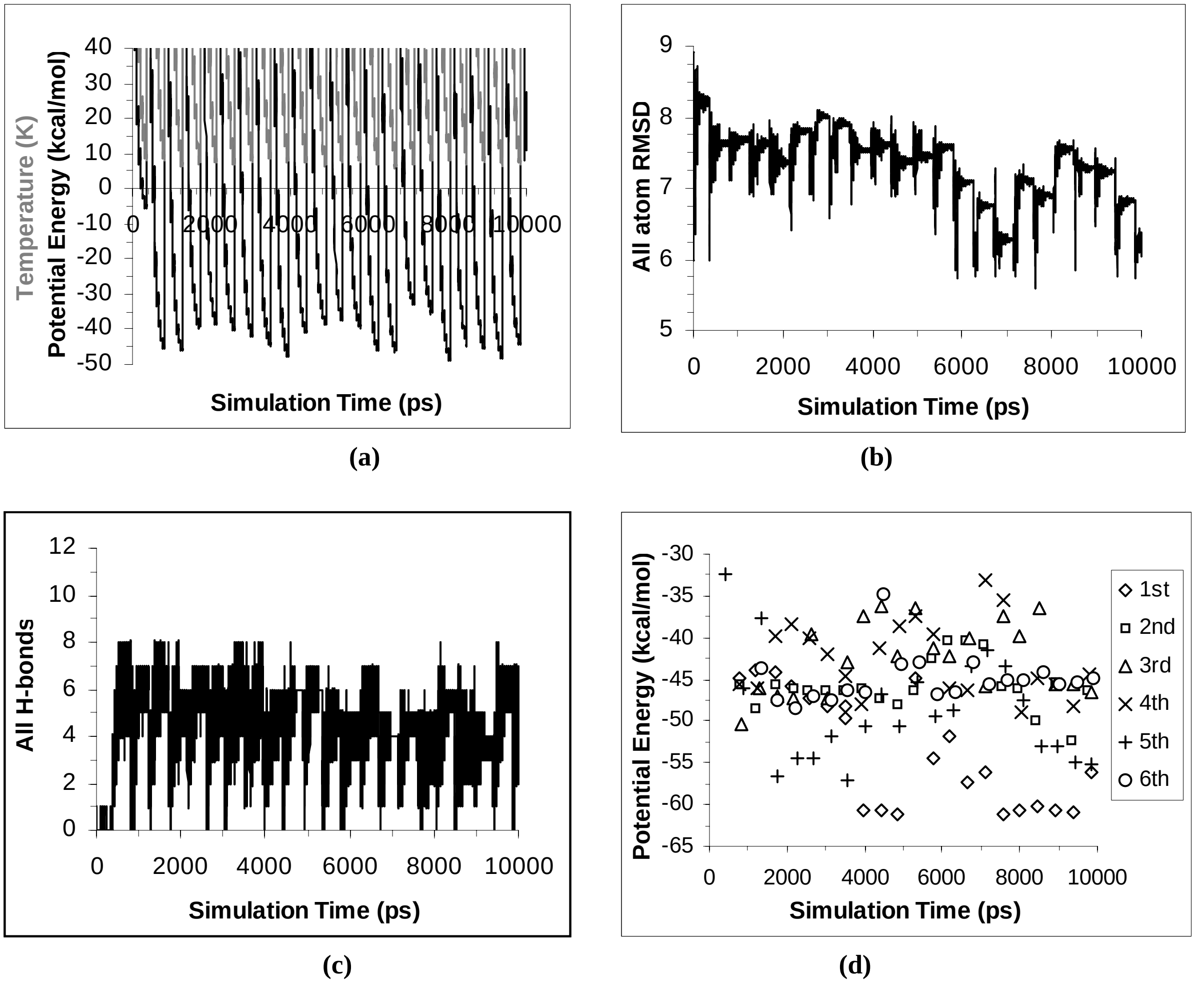
(d) E = -48.4 kcal/mol

(e) E = -46.1 kcal/mol

(f) E = -44.8 kcal/mol

Figure 4.3.2.2.3. Conformations and potential energies (kcal/mol) of the minimum potential energy structures sampled by the polypeptide of the lowest potential energy from a DIVE simulation of six trajectories of Ala13 *in vacuo*, starting from an extended structure. Approximate descriptions of the conformations, RMSD from the ideal α-helix, and number of H-bonds are: **(a)** A nearly ideal α-helix is the global potential energy minimum (RMSD 0.44 Å; eleven hydrogen bonds). **(b)** Compact conformation I, W-shape, having 1-2, 1-3, 1-4, 1-5 and etc. H-bonds (RMSD is 6.3 Å, six hydrogen bonds, plus one heavy atom engaged in two H-bonds). **(c)** Half α-helix and half 3_{10}-helix, linked by a turn (RMSD is 6.0 Å; nine hydrogen bonds). **(d)** A nearly ideal double-stranded β-sheet (RMSD is 8.9 Å; six hydrogen bonds). **(e)** Compact conformation II, S-shape, having 1-2, 1-3, 1-4, 1-5 and etc. H-bonds (RMSD is 5.6 Å, six hydrogen bonds). **(f)** A V-shaped α-helix (RMSD is 4.5 Å; six hydrogen bonds, plus one heavy atom engaged in two H-bonds). A hydrogen bond is identified when the distance between two heavy atoms X is below 3.3 Å and the X-H···X angle differs from 180.0° less than 20.0°. These images were generated by using PyMOL (http://pymol.sourceforge.net/).

A similar analysis of the potential energies of the individual trajectories throughout the course of the simulation gives a total of 131 regions of minimum potential energy (131 – 49 = 82 regions have not been displayed in the lowest potential energy polypeptide trajectory of Fig. 4.3.2.2.2a). Fig. 4.3.2.2.4a shows the temperature and potential energy history of a representative trajectory for which local minima at relatively higher potential energies are sampled, and Fig. 4.3.2.2.4b - c show the corresponding RMSD and H-bonds histories, respectively. Because the figures track a real trajectory for one polypeptide, these plots show very smooth folding processes. The large RMSDs and small number of 1 - 4 H-bonds indicate that this polypeptide samples a region of conformational space far away from helical energy wells. Fig. 4.3.2.2.4d displays the potential energy minima for the 131 regions, except for the extended conformation. Some representative structures with high-energy local minima, sampled from the trajectories of the individual polypeptides (but not sampled by the lowest potential energy polypeptide) are displayed in Fig. 4.3.2.2.5. All sampled potential energy minima and their conformations in the simulations can be summarized as follows. First, the global potential energy minimum is a nearly ideal α-helix with potential energy below –60 kcal/mol. Second, the conformations of the local potential energy minima between –60 kcal/mol to –40 kcal/mol include three large classes. They include 1) compact forms having 1-2, 1-3, 1-4, 1-5, and other H-bonds, 2) a nearly ideal or twisted β-sheet of two or three stands, and 3) a partial α-helix plus (a) extended chains, (b) 3_{10}-helices, (c) π-helices, or (d) another partial α-helix. Third, many different β-sheet-rich conformations as well as looped and coiled structures have local potential energy minima between –40 kcal/mol and –30 kcal/mol. Fourth, the extended conformation, with a potential energy of –7 kcal/mol, represents the highest potential energy minimum sampled in the trajectories. We note that the extended form has its potential energy minimum even lower in energy than the equilibrated average potential energy of an α-helix from constant temperature simulations at 300 K (~ 0 kcal/mol) (Chapter 3). This illustrates one difficulty in using NVT simulations to map potential energy minima or to locate the global potential energy minimum.



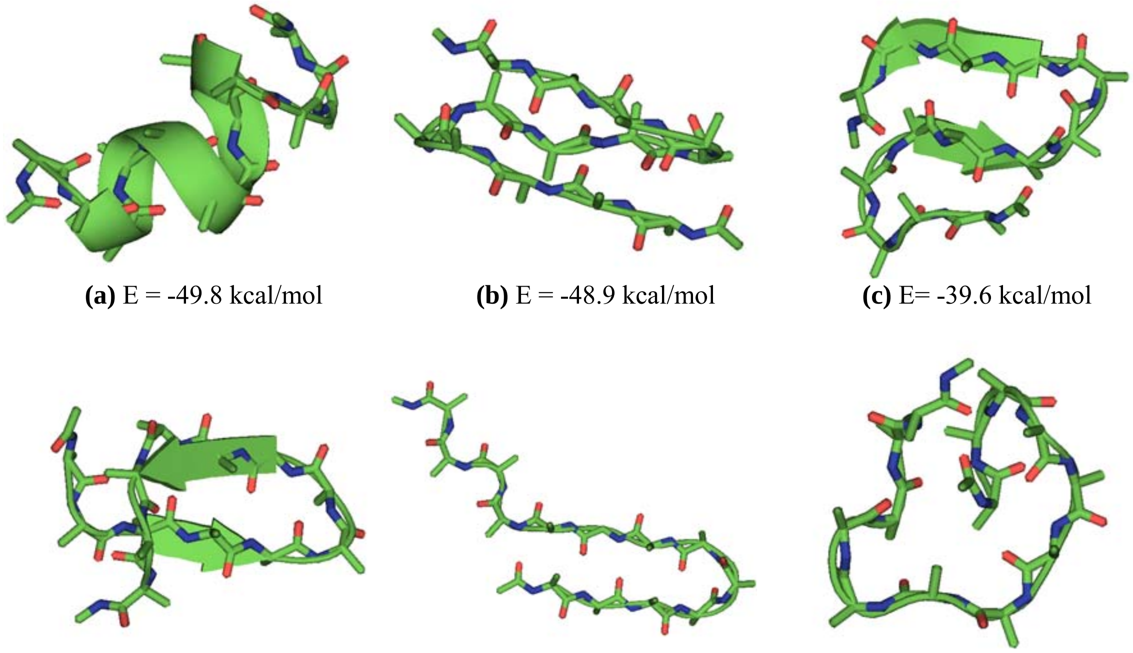
(a)

(b)

(c)

(d)

Figure 4.3.2.2.4. History of various properties from a DIVE simulation of six trajectories of Ala13 *in vacuo*, starting from an extended structure. **(a)** Temperature (upper, gray curve) and potential energy (lower, black curve) of a single polypeptide, which samples the potential energy minima at relatively higher potential energies. **(b)** RMSD between all atoms of the calculated structure and an ideal α-helix of the single polypeptide. **(c)** All hydrogen bonds of the single polypeptide, mostly including 1-2 hydrogen bond type and other types (no more than one of 1-3 or 1-4 hydrogen bonds and zero 1-5 hydrogen bonds). (d) History of potential energy minima sampled from all 6 trajectories, excluding the highest potential energy minimum (E = –7.19 kcal/mol) for the extended conformation.



(a) E = -49.8 kcal/mol **(b)** E = -48.9 kcal/mol **(c)** E= -39.6 kcal/mol

(d) E = -34.7 kcal/mol **(e)** E = -33.2 kcal/mol **(f)** E = -32.5 kcal/mol

Figure 4.3.2.2.5. Conformations and potential energies (kcal/mol) of the minimum potential energy structures sampled by individual polypeptides from a DIVE simulation of six trajectories of Ala13 *in vacuo*, starting from an extended structure. Approximate descriptions of the conformations, RMSD from the ideal α-helix, and number of H-bonds are: **(a)** Mixed helix with 1-3, 1-4, 1-5 H-bonds (RMSD is 3.4 Å; five hydrogen bonds, plus two heavy atoms engaged in two H-bonds). **(b)** Three β-strands (RMSD is 6.9 Å; Five hydrogen bonds). **(c)** An antiparallel β-sheet of three strands (RMSD is 6.4 Å, six hydrogen bonds, plus one heavy atom engaged in two H-bonds). **(d)** Twisted double-stranded β-sheet conformation (RMSD is 8.0 Å; four hydrogen bonds). **(e)** Double-stranded β-sheet with ribbon structure at the C-terminus (RMSD is 6.3 Å, four hydrogen bonds). **(f)** Loop to form a cavity (RMSD is 6.8 Å, six hydrogen bonds). A hydrogen bond is identified when the distance between two heavy atoms X is below 3.3 Å and the X-H···X angle differs from 180.0° less than 20.0°. These images were generated by using PyMOL (http://pymol.sourceforge.net/).

To verify that the global potential energy minimum and the local minima sampled when starting from the extended structure are not artifacts of the initial conditions, we repeated the simulations starting from an ideal 3_{10}-helix and an α-helix. A higher target temperature 1400 K, rather than 1000 K, was used because extra kinetic energy is needed to move away from the starting helical structures. The analysis of the potential energy minima and the conformations below 10 K shows very similar results to those obtained starting from the extended conformation (data not shown). The same α-helical conformation is obtained as the global potential energy minimum and many local potential energy minima having almost identical conformations are found in the three simulations. However, not all local potential energy minima appear when starting from the different conformations. For example, the extended conformation is not sampled when starting from a 3_{10}- or an α-helix, yet the 3_{10}-helix, while located at a relatively higher potential energy than that of a double-stranded β-sheet (-38.4 kcal/mol) (Chapter 5), is not sampled from the extended or the α-helical starting structure. This is reasonable because the DIVE simulation can by no means sample all the myriad local potential energy minima during a very limited simulation time (10 ns for each trajectory, 60 ns total).

4.3.3. DIVE simulations compared to REM, and EXREM simulations

4.3.3.1. Comparisons

It is interesting to compare the results of the DIVE simulations with those of REM simulations (from the standard REM procedure). A range of lower temperatures (several Ks — 100 K) and lower potential energies (~ -60 — 0 kcal/mol) are explored in the DIVE simulations compared to the REM simulations. This clearly indicates that the REM simulations, performed at exchange temperatures much higher than zero K, usually fail to reach the deep minima of the potential energy wells. In contrast, our DIVE simulations easily locate these potential energy minima at temperatures near zero K. On the other hand, slightly higher temperatures (>500 K) and potential energies (> 160 kcal/mol) are explored in the REM simulations compared to those in the DIVE simulations. These higher temperatures and potential energies benefit REM simulations by helping them move away from local minima and enhance phase space sampling. However, the DIVE simulations show a much more efficient sampling ability than REM simulations, since DIVE simulations find a larger range of different conformations: a global α-helix, partial α-helices, double stranded β-sheets, β-sheet-rich conformations, compact conformations, looped and coiled structures. In contrast, many conformations of low potential energies such as a nearly ideal α-helix, a nearly ideal double-stranded β-sheets, and well-ordered compact structures are not sampled in the REM simulations.

Two reasons account for the much more extensive conformational sampling of the DIVE simulations compared with REM simulations. First, the velocity-redirection integrated into DIVE makes simulations undergo conformational transitions easily and thus makes them sample very diverse regions on the multi-dimensional potential energy surface, even in limited simulation times. Second, the DIVE simulations are not as likely as REM simulations to reach equilibrium. In the REM simulations, each replica may equilibrate at one conformation, and then the polypeptide can experience repeated structural transitions among a small number of different conformations. In order to vary conformational space sampling and thus search for the global α-helix, more than six replicas are probably needed in the REM simulations for the temperature range between 200 K and 600 K.

The EXREM simulations sample a much larger range of energies (~-55 — 480 kcal/mol) than the DIVE simulations (~-60 — 160 kcal/mol) because of their larger temperature range (several Ks – 1100 K in EXREM contrasted to several Ks — 500 K in

DIVE). While the NVT simulations easily retain the polypeptide's temperatures because of the heat bath, the NVE simulations reach thermal equilibrium at temperatures very different from the initial or heating temperatures. However, EXREM simulations sample fewer conformations than DIVE simulations, indicating that EXREM simulations are more likely to be restricted to a local region of conformational space. For example, EXREM does not sample the nearly ideal α-helix and many compact conformations of low potential energies. Thus, the trajectory perturbation from the velocity-redirection algorithm incorporated in the DIVE protocol helps ensure more extensive sampling of a multi-dimensional energy surface.

4.3.3.2. Discussion

Disrupted velocity (DIVE) search simulations for blocked Ala13 illustrate that the ground state conformation is a nearly ideal α-helix *in vacuo*. This result is in good agreement with earlier simulations of uncharged polyalanines [11,19,26-29] starting from an α-helix, but the result has rarely been obtained from an extended structure. To our knowledge, no one has previously found an α-helical global potential energy minimum for polyalanine using the AMBER96 force field and starting from an extended structure. Furthermore, we want to emphasize that the energy states for the α-helical conformation sampled in earlier simulations [11,19,26-29] were likely to be far above the minimum potential energies when simulations were performed at the temperatures above 200 K. Though a direct comparison of the potential energies between our simulations and those from earlier simulations with different force fields is impossible, our simulation results clearly indicate that potential energy minima cannot be located in simulations performed at high temperatures.

Our simulation results demonstrate that from among several different simulation techniques the DIVE simulation samples diverse regions of conformational space most effectively. At normal temperatures, conventional NVT simulations easily equilibrate in one conformational cluster. At very high temperatures, these NVT simulations may undergo frequent conformation transitions among several types of unfolded structures, but the global minimum potential energy conformation and other well-ordered structures of low potential energies are nearly impossible to locate. REM simulations and EXREM simulations both utilize a velocity-rescaling algorithm for alternate heating and cooling of the polypeptide. During these simulations, high-temperatures facilitate the conformational transitions of the polypeptide, and low-temperatures trap the polypeptide in structures of lower potential energies. However, velocity rescaling focuses primarily on moving vertically on the energy surface to go over barriers. Therefore, the simulations are likely to experience repeated structural transitions among several similar conformations and the results depend strongly on the initial conditions. In the DIVE simulations, on the other hand, the velocity-redirection step randomizes horizontal motions along the potential energy surface and enhances a polypeptide’s chances of going around energy barriers rather than over them. In this way, the DIVE simulations can sample diverse conformational space even in a very short simulation time.

4.4. Conclusion

Our newly developed disrupted velocity (DIVE) search simulations show that the ground state conformation of blocked Ala13 *in vacuo* for the AMBER96 force field is a nearly ideal α-helix. In contrast, the global minimum potential energy conformation is very difficult to locate by using several popular molecular dynamics methods including simulated annealing [3-6], locally-enhanced sampling [6-9,41] and REM [10-14]. Moreover, the DIVE simulations are able to search a diverse conformational space in a short simulation time (10 ns for each trajectory and 60 ns total), including other partial helices, different kinds of β-sheets, compact, coiled, looped, and extended structures. Compact structures with large numbers of intrapeptide H-bonds are close in energy to the global potential energy minimum. In contrast, β-sheet-rich conformations dominate at relatively higher potential energies. At higher energies than β-sheets are the coiled, looped and extended structures, and they mainly appear during high temperature simulations. For simulations reported here, only results from the DIVE technique are relatively independent of initial structures. In our simulations, the global and local potential energy minima sampled from an initially extended structure are almost identical with those sampled from an initial ideal 3_{10}- or α-helix.

4.5. Bibliography

(1) Vasquez, M.; Nemethy, G.; Scheraga, H. A. *Chemical Reviews* **1994**, *94*, 2183.

(2) Tuckerman, M. E.; Martyna, G. J. *Journal of Physical Chemistry B* **2000**, *104*, 159.

(3) Kirkpatrick, S.; Gelatt, C. D. J.; Vecchi, M. P. *Science* **1983**, *220*, 671.

(4) Basu, A.; Frazer, L. N. *Science* **1990**, *249*, 1409.

(5) Huber, G. A.; McCammon, J. A. *Physical Review E: Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **1997**, *55*, 4822.

(6) Roitberg, A.; Elber, R. *Journal of Chemical Physics* **1991**, *95*, 9277.

(7) Elber, R.; Karplus, M. *Journal of the American Chemical Society* **1990**, *112*, 9161.

(8) Stultz, C. M.; Karplus, M. *Journal of Chemical Physics* **1998**, *109*, 8809.

(9) Zheng, W.-M.; Zheng, Q. *Journal of Chemical Physics* **1997**, *106*, 1191.

(10) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96.

(11) Mitsutake, A.; Okamoto, Y. *Journal of Chemical Physics* **2000**, *112*, 10638.

(12) Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **2000**, *329*, 261.

(13) Sugita, Y.; Kitao, A.; Okamoto, Y. *Journal of Chemical Physics* **2000**, *113*, 6042.

(14) Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **1999**, *314*, 141.

(15) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598.

(16) Andersen, H. C. *Journal of Chemical Physics* **1980**, *72*, 2384.

(17) Martyna, G. J.; Klein, M. L.; Tuckerman, M. E. *Journal of Chemical Physics* **1992**, *97*, 2635.

(18) Goldstein, H. Classical Mechanics; Addison-Wesley: MA, 1965; pp 69.

(19) Daggett, V.; Levitt, M. *Journal of Molecular Biology* **1992**, *223*, 1121.

(20) Marqusee, S.; Robbins, V. H.; Baldwin, R. L. *Proceedings of the National Academy of Sciences of the United States of America* **1989**, *86*, 5286.

(21) Merutka, G.; Lipton, W.; Shalongo, W.; Park, S. H.; Stellwagen, E. *Biochemistry* **1990**, *29*, 7511.

(22) Chakrabartty, A.; Schellman, J. A.; Baldwin, R. L. *Nature* **1991**, *351*, 586.

(23) Padmanabhan, S.; Marqusee, S.; Ridgeway, T.; Laue, T. M.; Baldwin, R. L. *Nature* **1990**, *344*, 268.

(24) Fiori, W. R.; Miick, S. M.; Millhauser, G. L. *Biochemistry* **1993**, *32*, 11957.

(25) Hudgins, R. R.; Jarrold, M. F. *Journal of the American Chemical Society* **1999**, *121*, 3494.

(26) Hansmann, U. H. E.; Okamoto, Y. *Journal of Chemical Physics* **1999**, *110*, 1267.

(27) Son, H. S.; Hong, B. H.; Lee, C. W.; Yun, S.; Kim, K. S. *Journal of the American Chemical Society* **2001**, *123*, 514.

(28) Bertsch, R. A.; Vaidehi, N.; Chan, S. I.; Goddard, W. A., III. *Proteins: Structure, Function, and Genetics* **1998**, *33*, 343.

(29) Takano, M.; Yamato, T.; Higo, J.; Suyama, A.; Nagayama, K. *Journal of the American Chemical Society* **1999**, *121*, 605.

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

(31) Kollman, P.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A. *The development/application of a \"minimalist\" organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data,* 1997; Vol. 3.

(32) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Ferguson, D. M.; Radmer, R. J.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. AMBER 5; University of California: San Franciso, 1997.

(33) Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Langridge, R. *Journal of Molecular Graphics* **1988**, *6*, 13.

(34) Voet, D.; Voet, J. G. Biochemistry; John Wiley & Sons, Inc.: New York, 1995; pp 144.

(35) Hinsen, K. *Journal of Computational Chemistry* **2000**, *21*, 79.

(36) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Clarendon Press: Oxford, 1987.

(37) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *Journal of Chemical Physics* **1982**, *76*, 637.

(38) Palmer, B. J. *Journal of Computational Physics* **1993**, *104*, 470.

(39) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *Journal of Chemical Physics* **1984**, *81*, 3684.

(40) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *Journal of Chemical Physics* **1953**, *21*, 1087.

(41) Hixson, C. A.; Chen, J.; Huang, Z.; Wheeler, R. A. *Journal of Molecular Graphics & Modelling* **2004**, *22*, 349.

(42) Daggett, V. D. Molecular Modelling of Peptide and Proteins. Doctor of Philosophy, University of California, 1990.

Chapter 5

Molecular Dynamics Simulations of Polyalanine *in vacuo* Using Different Force Fields

5.1. Introduction

In Chapter 3, we described a new MD technique called divergent path (DIP) search simulations. In this technique, multiple independent trajectories are simulated. The atomic velocities of each polypeptide have the same magnitude but different directions. The different directions cause the different polypeptides to evolve along different trajectories from a single initial point in configuration space. The results from this DIP strategy indicate that conventional NVT simulations have three primary limitations responsible for their limited sampling in protein folding dynamics: potential energy traps, free energy traps, and kinetic traps. In conventional MD, it is kinetic traps, not potential energy traps, that mostly restrict the simulations from covering a wide variety of conformations. However, the mechanical equilibrium state associate with kinetic traps, can be easily disrupted by the randomization of atomic velocities.

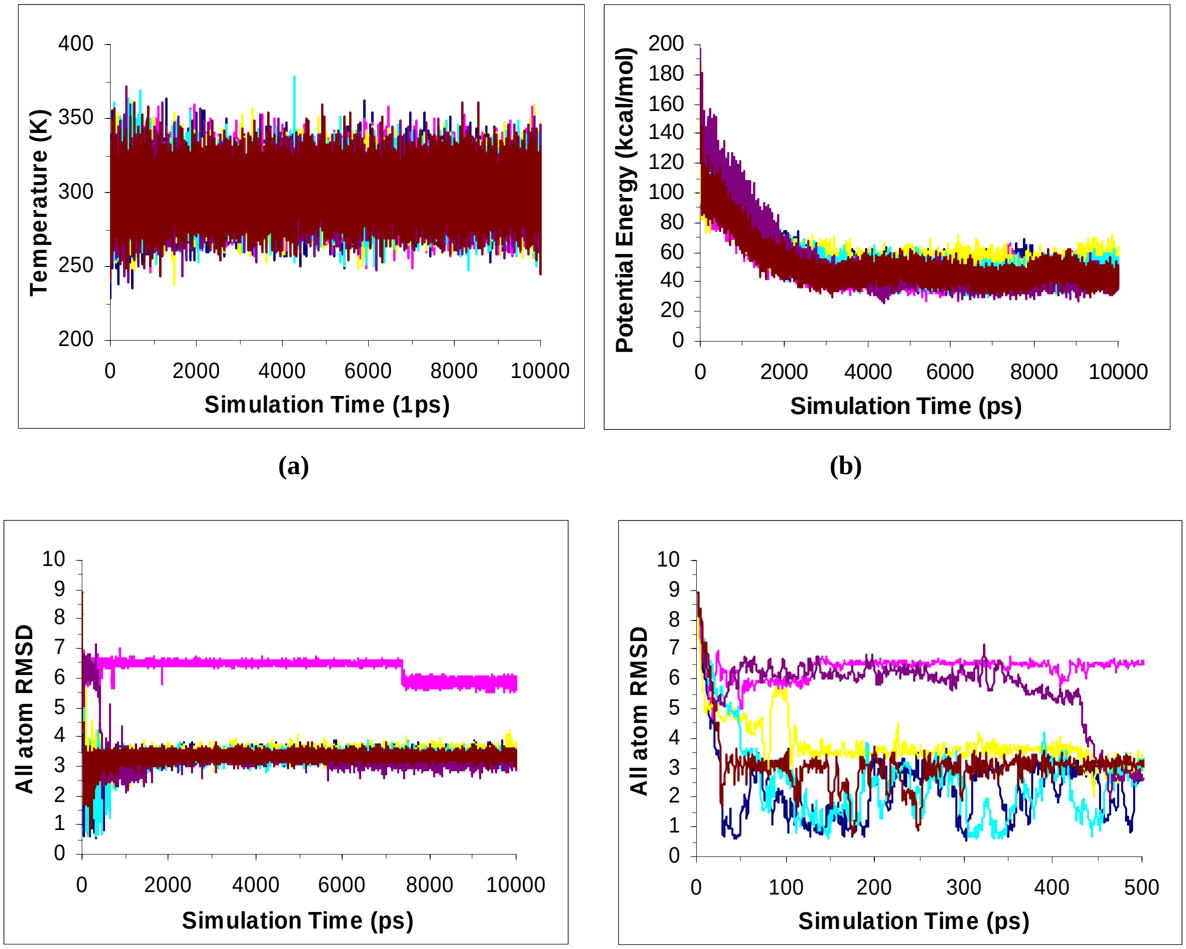
By combining this divergent path strategy at one temperature with cycles of heating and cooling, the disrupted velocity (DIVE) search protocol was developed in Chapter 4. In this technique, several polypeptides with different kinetic energies (and therefore different total energies) are simulated independently and simultaneously. Also, their atomic velocities are reassigned after a fixed time period in NVE molecular dynamics simulations. Because of this frequent perturbation in both trajectory and energy, the DIVE simulations alleviate the three primary limitations of conventional MD and are able to sample diverse regions of conformational space effectively. In addition, this technique accurately identifies potential energy minima at temperatures near 0 K.

Ignoring phase space sampling problems for the moment, the results from molecular dynamics simulations of protein folding are only determined by the amino-acid sequence of the proteins and force fields. The effect of the force fields on MD simulations can be easily understood because different force fields result in different potential energy surfaces for proteins. Recently some studies address the issue of the MD differences due to using different force fields [1,2]. Here we give some comparisons of these new MD simulations protocols, DIP and DIVE for the gas phase polypeptide Ala13 using different AMBER force fields. The comparison that is the most interesting involves using the DIVE protocol to search myriad potential energy minima and conformations.

The contents of the remainder of this chapter are as follows. First, DIP simulations of the polypeptide Ala13 *in vacuo* for AMBER99 [3] force field are presented. Next, we describe results of using the DIVE protocol to investigate the potential energy landscape of Ala13 *in vacuo* for the AMBER99 force field. Finally, the polypeptide's conformations and energetics of the global minimum and local potential energy minima from two different force fields (AMBER96 [4] and AMBER99) are compared. The simulation methodology, including algorithms and computational details, is the same as those described in Chapter 3 for DIP and Chapter 4 for the DIVE techniques.

5.2. DIP simulations of polyalanine using the AMBER99 force field

5.2.1. Simulation results



(a) (b)

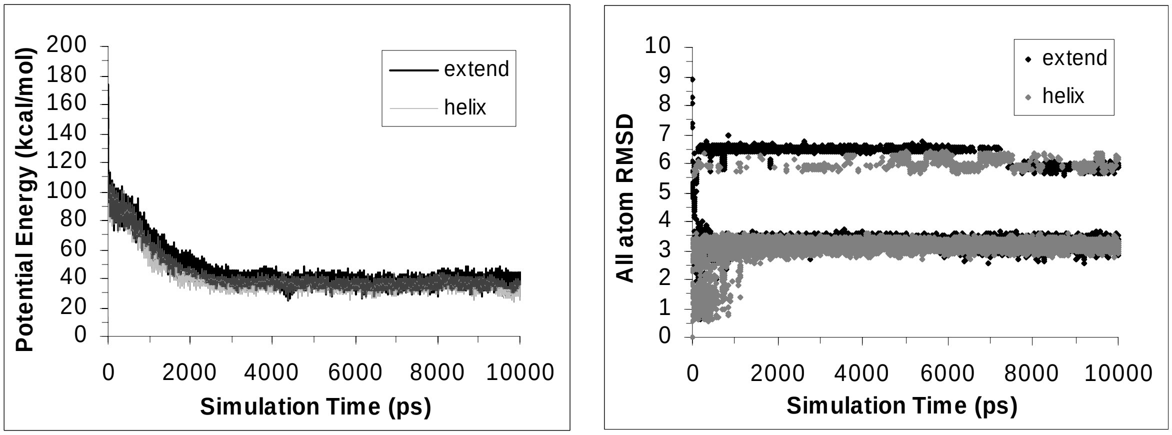
(c) (d)

Figure 5.2.1.1. Time series of various properties from a DIP simulation of six trajectories of Ala13 at 300 K *in vacuo*, starting from an extended structure. **(a)** Temperature, **(b)** Potential energy. **(c)** RMSD between all atoms of the calculated structures and an ideal α-helix. **(d)** RMSD in the first 500 ps. **Note**: Blue: 1st, Magenta: 2nd, Yellow: 3rd, Light blue: 4th, Purple: 5th, Brown: 6th trajectory.

We ran the DIP simulations of Ala13 at 300 K *in vacuo* for the AMBER99 force field. Fig. 5.2.1.1a – b displays the histories of the temperature and potential energy for six trajectories in a DIP simulation, starting from an extended structure. While the temperature histories are nearly identical for each trajectory, the potential energy histories show some variation. In addition, the potential energy histories indicate that the polypeptide in all these trajectories reach mechanical equilibrium after 2ns. Fig. 5.2.1.1c shows that five of six polypeptides equilibrate at the 3_{10}-helical conformation with an average all-atom RMSD of 3.3 Å and one polypeptide equilibrates at the parallel U-shaped 3_{10}-helix with an average 5.9 Å all-atom RMSD. Furthermore, Fig. 5.2.1.1d demonstrates that the conformational transition from the initial extended structure to the equilibrated 3_{10}-helix or U-shaped 3_{10}-helix happens during the first 500 ps in the mechanical equilibration process. A DIP simulation of six trajectories starting from an ideal α-helix gives very similar results as the DIP simulation starting from the extended structure.

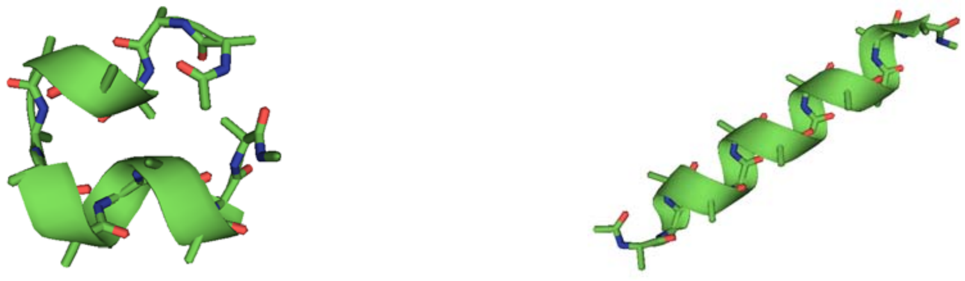
Unlike the AMBER96 force field simulations (Chapter 3), helices were formed very easily with the AMBER99 force field simulations from the initial extended structure. However, the global minimum free energy conformation at ambient temperature is not the α-helix, unlike the AMBER96 force field simulation. The single polypeptide conventional NVT simulations on the time scale as long as 60 ns from both initial extended and α-helical conformations converge to the same result, which indicates the global free energy conformation of Ala13 *in vacuo* at 300 K seems to be a 3_{10}-helix (data not shown). However, the DIP simulations give a slightly different result in that the global free energy conformation appears in clusters of U-shaped 3_{10}-helices and 3_{10}-helices.

Fig. 5.2.1.2 shows the potential energy and RMSD from an ideal α-helix of the polypeptide with the lowest potential energy. These plots show that starting from different conformations the DIP simulations search very similar lowest potential energy conformations at 300 K. In fact, the polypeptide at the lowest potential energies oscillates between a parallel U-shaped 3_{10}-helical (Fig. 5.2.1.3a) and a 3_{10}-helical conformation (Fig. 5.2.1.3b). The dominant conformation is a 3_{10}-helix. In the further supplemental DIP simulation of six trajectories of Ala13 at 300 K *in vacuo* starting from the U-shape 3_{10}-helix (Fig. 5.2.1.3a), three polypeptides maintain the U-shaped 3_{10}-helical conformation and three polypeptides convert into a 3_{10}-helix (data not shown here). The conformational transition from the U-shaped 3_{10}-helix to 3_{10}-helix happens during the mechanical equilibration process, which occurs near the beginning of the simulation. After that, the transition between a U-shaped 3_{10}-helix and a 3_{10}-helix cannot be observed during these simulations. The transition cannot be observed because mechanical equilibrium kinetically traps each polypeptide in a single conformational potential energy well. In spite of this, our simulation results indicate that the global free energy minimum for Ala13 at 300 K *in vacuo* is clusters of U-shaped 3_{10}-helices and 3_{10}-helices.



(a) **(b)**

Figure 5.2.1.2. Time series of the potential energy and RMSD from an ideal α-helix of the polypeptide with the lowest potential energy at each step from a DIP simulation of six trajectories of Ala13 *in vacuo* at 300 K, initiated from the extended (black curve or dot) and α-helical (gray curve or dot) conformation. **(a)** Potential energy. **(b)** RMSD between all atoms of the calculated structures and an ideal α-helix. **Note**: the RMSD plots here are dot charts.



(a): a parallel U-shaped 3_{10}-helix **(b)**: A 3_{10}-helix

Figure 5.2.1.3. The parallel U-shaped 3_{10}-helix and 3_{10}-helix sampled at the final two data-collection steps of the polypeptide with the lowest potential energy from a DIP simulation of six trajectories of Ala13 *in vacuo* at 300 K, starting from an extended structure. **(a)** The parallel U-shaped 3_{10}-helical structure after 9999ps, from the 2nd trajectory. **(b)** The 3_{10}-helical structure after 10 ns, from the 5th trajectory. These images were generated by using PyMOL (http://pymol.sourceforge.net/).

In vacuo, the parallel U-shape 3_{10}-helix is the global potential energy minimum and the 3_{10}-helix is a potential energy minimum above the energy of the α-helical conformation (section 5.3.1). Therefore, the DIP simulation results indicate four points. First, the conformation at global free energy minimum near *in vivo* temperatures is not necessarily the global potential energy minimum conformation at 0 K (section 5.3.1), but can also be a low potential energy local minimum (e.g. 3_{10}-helix *in vacuo*). Second, the global free energy minimum near *in vivo* temperatures may have clusters of several metastable states [5,6] very close in energy (e.g. U-shaped 3_{10}-helix and 3_{10}-helix). So the global free energy minimum is represented by a family of conformations. Third, due to the three primary limitations of MD, single conventional NVT simulation on limited time scales can easily get trapped in a local potential energy well. As a result, the simulation has extreme difficulty in sampling those metastable states in a single trajectory. Fourth, the simulations of several divergent paths can alleviate the very biased phase space sampling in conventional MD. The polypeptide with the lowest potential energy in these

DIP simulations is likely to sample various metastable minimum energy conformations and explore some folding or unfolding processes from the individual trajectories.

5.2.2. Conclusion for DIP simulations

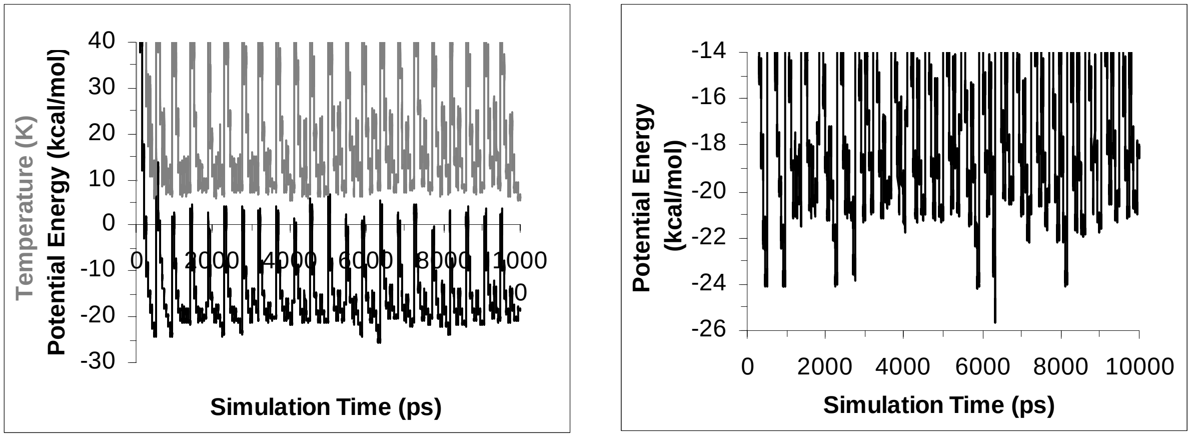
The DIP simulations of this polypeptide using the AMBER99 force field show that the global free energy minimum *in vacuo* at 300 K is not a single helical type but combined 3_{10}-helices. The combined 3_{10}-helices include the U-shaped 3_{10}-helix and nearly ideal 3_{10}-helix, while the 3_{10}-helix is probably the dominant conformation. The 3_{10}-helix is not the global potential energy minimum but a local minimum with its potential energy above the U-shaped 3_{10}-helix and α-helix. This observation supports the idea that the global free energy minimum conformation near *in vivo* temperatures is not necessarily the global potential energy minimum conformation at 0 K but can be a local minimum conformation. A direct oscillation between two lowest free energy minima of a 3_{10}-helix and a U-shaped 3_{10}-helix cannot be observed in a single trajectory throughout the simulations. However, our DIP simulations provide the trajectory of the lowest potential energy polypeptide at each step that clearly displays an oscillation between these two metastable states.

5.3. DIVE simulations of polyalanine using the AMBER99 force field

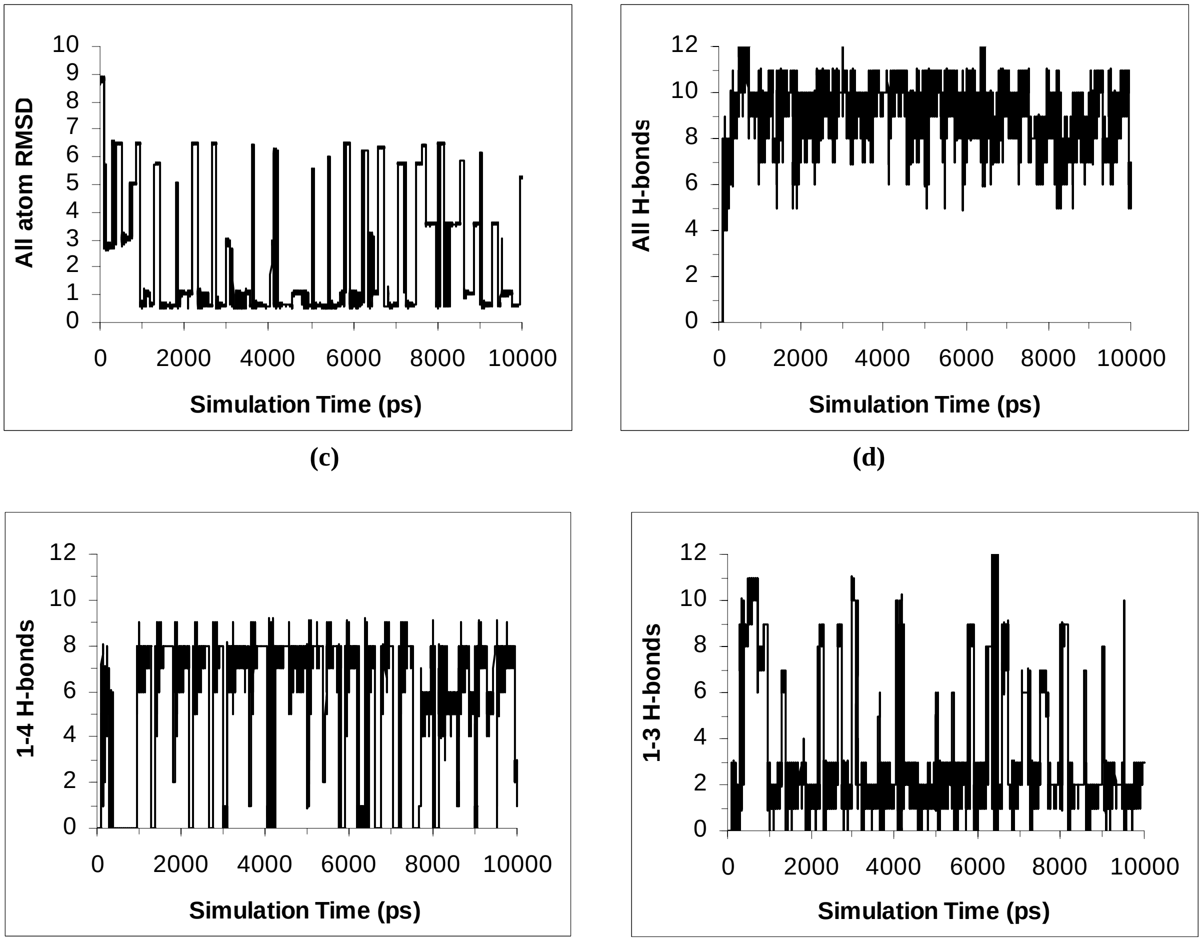
5.3.1. Simulation results

When the AMBER99 force field was used, the DIVE simulations *in vacuo* sample potential energies ranging from nearly –30 to 240 kcal/mol (data not shown). Figure

5.3.1.1a displays the temperatures and potential energies of the polypeptide with the lowest potential energy at each step and Figure 5.3.1.1b enlarges the plot in the range of lowest potential energies to show clearly the different potential energy minima sampled during the simulations. Figure 5.3.1.1c shows the all-atom RMSD between the polypeptide with the lowest potential energy and an ideal α-helical reference structure. Figures 5.3.1.1d – 5.3.1.1f show all hydrogen bonds (H-bond), 1-4 H-bonds and 1-3 H-bonds. The RMSD and H-bonding plots indicate that the polypeptide first forms a nearly perfect 3_{10}-helix after approximately 500ps (1-3 H-bonds: 11, RMSD: approximately 3 Å) and first converts to an α-helix at 1000ps (1-4 H-bonds: 9, RMSD: approximately 0.5 Å). At energies corresponding to a temperature below 10 K, many potential energy minima appear in the trajectory.



(a) **(b)**



(c)

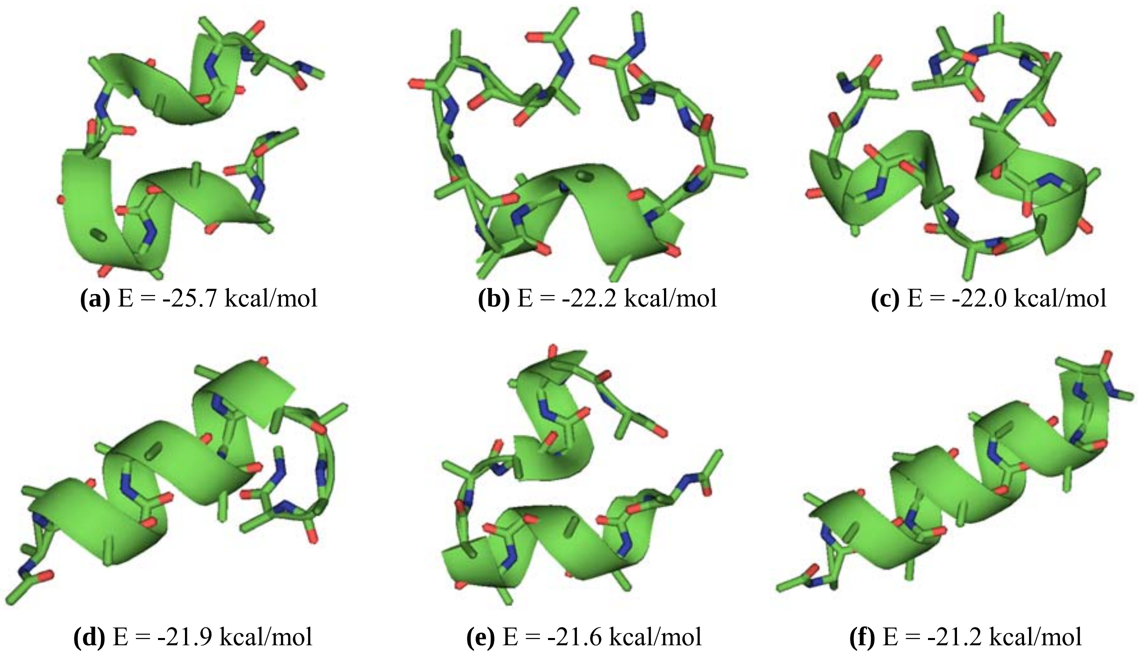
(d)

(e)

(f)

Figure 5.3.1.1. Time series of various properties of the polypeptide with lowest potential energy at each step in a DIVE simulation of six trajectories of Ala13 *in vacuo*, starting from an extended structure. **(a)** Temperature (upper, gray curve) and potential energy (lower, black curve). **(b)** Potential energy displayed on an expanded scale. **(c)** RMSD between all atoms of the calculated structure and an ideal α-helix. The RMSD between a 3_{10}-helix and an ideal α-helix is approximately 3 Å. **(d)** All hydrogen bonds. **(e)** 1-4 hydrogen bonds. **(f)** 1-3 hydrogen bonds. A hydrogen bond is identified when the distance between two heavy atoms X is below 3.3 and the X-H ··· X angle differs from 180° by less than 20.0°. An ideal α-helix has eleven 1-4 hydrogen bonds and a 3_{10}-helix has twelve 1-3 hydrogen bonds.

We calculated the average temperature for each 50ps interval and then collected data from those simulation regions whose average energies correspond to temperatures below 10 K. 63 regions are obtained from the polypeptide with lowest potential energy along this trajectory. Furthermore, we selected the lowest potential energy over each 50ps interval, and determined the conformations of Ala13 in the potential energy minimum for these regions. Representative results are shown in Figure 5.3.1.2. The global potential energy minimum of Ala13 *in vacuo* is a parallel U-shaped 3_{10}-helix (Figure 5.3.1.2a) with a global minimum potential energy of –25.7 kcal/mol. With an RMSD value of approximately 6 Å from the α-helical structure, this conformation was first formed near 2.2ns. There are many local potential energy minima between –23 and -17 kcal/mol. Their conformations are diverse but are all well ordered, with more than 9 H-bonds. Though the difference between potential energy minima of different structural types is relatively small (less than 3 kcal/mol between any two minima closest in energy), the RMSDs between different conformations and an ideal α-helix differ by as much as 6.5 Å.


(a) E = -25.7 kcal/mol

(b) E = -22.2 kcal/mol

(c) E = -22.0 kcal/mol

(d) E = -21.9 kcal/mol

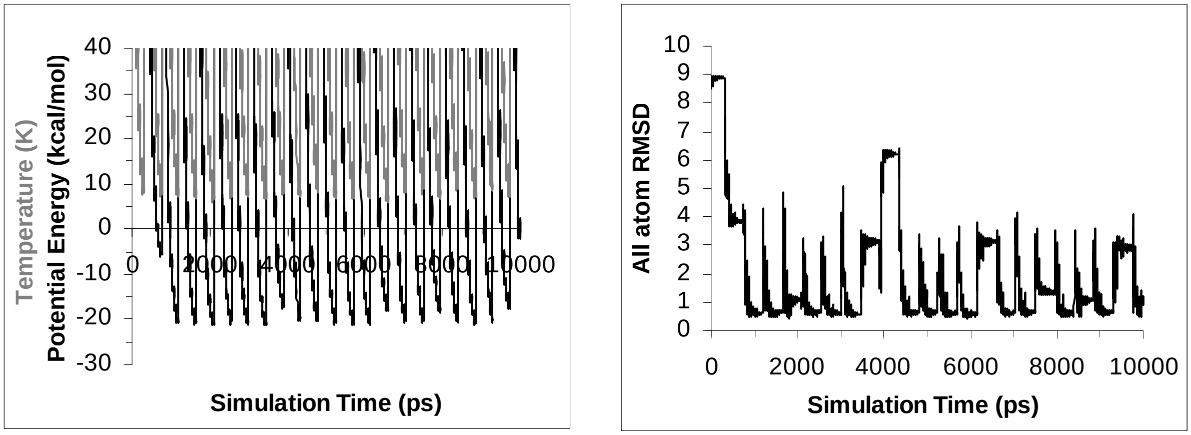
(e) E = -21.6 kcal/mol

(f) E = -21.2 kcal/mol

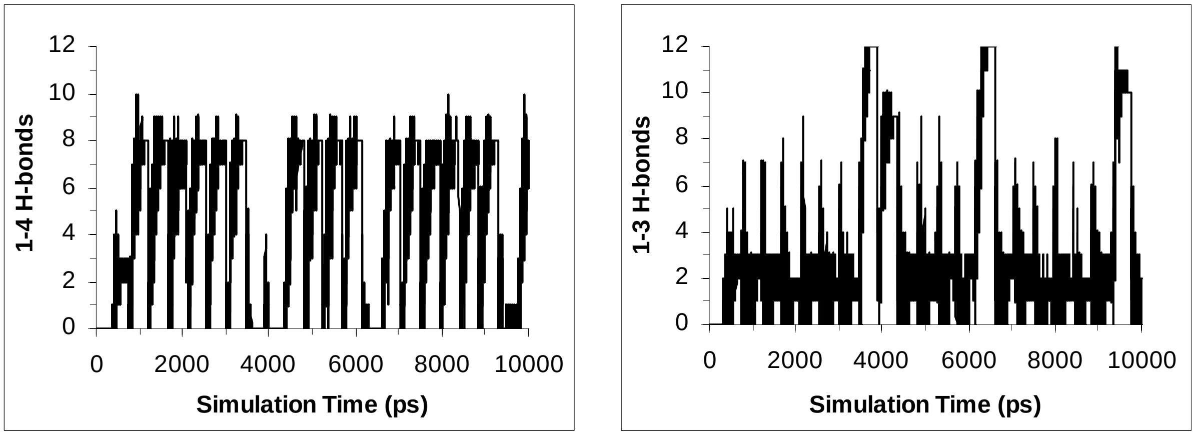
Figure 5.3.1.2. Conformations and potential energies (kcal/mol) of potential energy minima sampled by the polypeptide with lowest potential energy from the DIVE simulation of Ala13 *in vacuo*, starting from an extended structure. Approximate descriptions of the conformations, all-atom RMSD from the ideal α-helix, and number of H-bonds are: **(a)** A parallel U-shaped 3_{10}-helix is the global potential energy minimum (RMSD is 6.2 Å, seven hydrogen bonds, plus one heavy atom engaged in two H-bonds). **(b)** 3_{10}-helix, with N- and C-termini forming hydrogen bonds analogous to a second 3_{10}-helix (RMSD is 5.7 Å; ten hydrogen bonds). **(c)** A U-shaped α-helix (RMSD is 5.9 Å;

seven hydrogen bonds, plus one heavy atom engaged in two H-bonds). **(d)** α-helix, bent at the C-terminus (RMSD is 3.6 Å; eight hydrogen bonds). **(e)** A U-shaped, 1/3 α-helix, 2/3 3_{10}-helix (RMSD is 6.2 Å; eleven hydrogen bonds). **(f)** A nearly ideal α-helix (RMSD is 0.6 Å; eleven hydrogen bonds). A hydrogen bond is identified when the distance between two heavy atoms X is below 3.3 and the X-H ··· X angle differs from 180° by less than 20.0°. These images were generated by using PyMOL (http://pymol.sourceforge.net/).

A similar analysis of the potential energy of the individual trajectories throughout the course of the simulation gives a total of 132 regions of minimum energy (132-63=69 regions have not been displayed in the trajectory for the lowest potential energy polypeptide of Figure 5.3.1.1a). Figure 5.3.1.3a shows an example of the temperature and potential energy histories of the 2nd polypeotide, for which relatively higher, local potential energy minima are sampled. Figures 5.3.1.3b – d display the time histories of the corresponding RMSD, 1-4 H-bonds, and 1-3 H-bonds respectively. For this trajectory, folding to an α-helix occurs more than ten times and folding to a 3_{10}-helix is observed three times. Representative structures sampled from the individual polypeptide trajectories are shown in Figure 5.3.1.4. The highest potential energy minimum sampled in the simulations below 10 K is the extended conformation with a potential energy of approximately 50 kcal/mol.

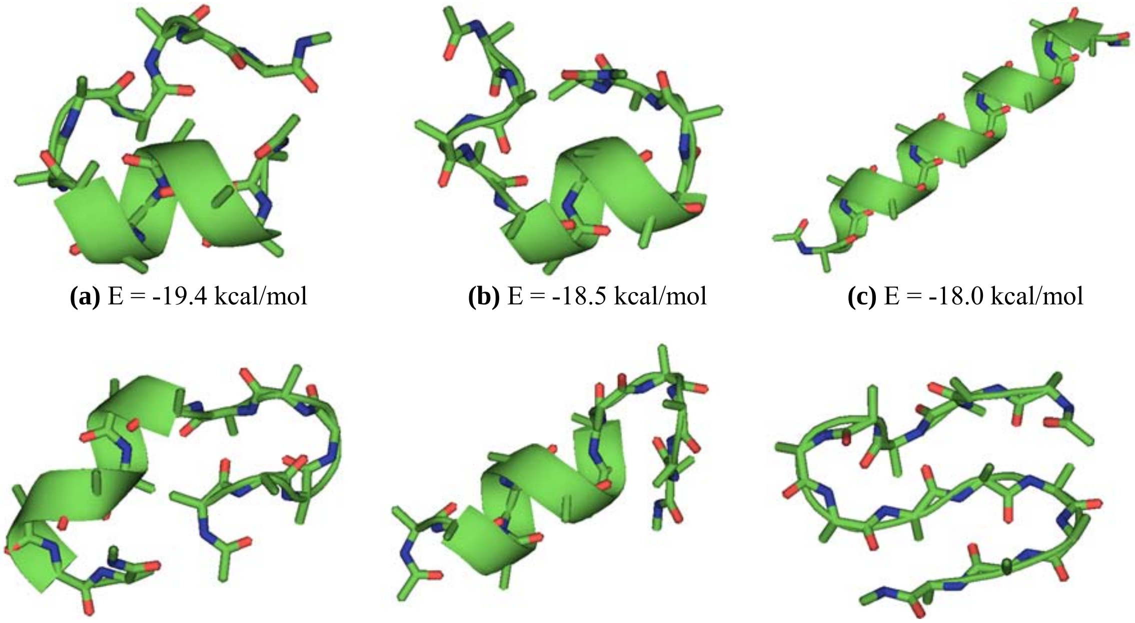


(a) **(b)**



(c) (d)

Figure 5.3.1.3. Time series of various properties of one polypeptide from a DIVE simulation of six trajectories of Ala13 *in vacuo*, starting from an extended structure. **(a)** Temperature (gray curve) and potential energy (black curve). **(b)** RMSD between all atoms of the calculated structure and an ideal α-helix (The RMSD between a 3_{10}-helix and an ideal α-helix is approximately 3 Å). **(c)** 1-4 hydrogen bonds. **(d)** 1-3 hydrogen bonds.



(a) E = -19.4 kcal/mol **(b)** E = -18.5 kcal/mol **(c)** E = -18.0 kcal/mol

(d) E = -12.5 kcal/mol **(e)** E = -6.0 kcal/mol **(f)** E = 3.3 kcal/mol

Figure 5.3.1.4. Conformations and potential energies (kcal/mol) sampled by individual polypeptides from a DIVE simulation of six trajectories of Ala13 *in vacuo*, starting from an extended structure. Approximate descriptions of the conformations, RMSD of all atoms from the ideal α-helix, and number of H-bonds are: **(a)** U-shaped, 2/3 α-helix, 1/3 3_{10}-helix (RMSD is 6.3 Å; eight hydrogen bonds). **(b)** α-helix with frayed C-terminus and ½ turn 3_{10}-helical N-terminus (RMSD is 5.3 Å; seven hydrogen bonds). **(c)** A nearly ideal 3_{10}-helix (RMSD is 3.1 Å; twelve hydrogen bonds). **(d)** 3_{10}-helix, with a large and a

small β-turn at termini (RMSD is 5.3 Å; seven hydrogen bonds, plus one heavy atom engaged in two H-bonds) **(e)** 2/3 the peptide forms an α-helix and 1/3 forms a β-turn (RMSD is 3.8 Å; seven hydrogen bonds). **(f)** Three anti-parallel β-strands (RMSD is 5.7 Å; four hydrogen bonds, plus two heavy atoms engaged in two H-bonds). These images were generated by using PyMOL (http://pymol.sourceforge.net/).

We repeated the simulations starting from an ideal α-helix. Again, the analysis of the potential energy minima and the conformations below 10 K shows very similar results to those obtained starting from an extended chain. The same global potential energy minimum with a parallel U-shaped 3_{10}-helix is obtained and similar local potential energy minima and their conformations are sampled, regardless of the starting structure. The potential energy minima and their conformations in the simulations from these two different initial configurations can be approximately summarized as follows. First, the global potential energy minimum is below –23 kcal/mol and corresponds to a U-shaped 3_{10}-helix. Second, the conformations of local potential energy minima between –23 kcal/mol and –5 kcal/mol include (a) α-helical, (b) 3_{10}-helical, and (c) partially α-helical or partially 3_{10}-helical structures with (i) termini–frayed, (ii) extended, (iii) β-turns with a single 1-3 H-bond, and (iv) another small α-helix/3_{10}-helix. Third, the extended conformation represents the highest potential energy minimum sampled in the trajectories, 48 kcal/mol.

5.3.2. Conclusion for DIVE simulations

From DIVE simulations, the global potential energy minimum of Ala13 *in vacuo* for the AMBER99 force field is a U-shaped 3_{10}-helix. Moreover, the DIVE simulations are able to search a diverse conformational space in a short simulation time including the nearly ideal 3_{10}-helix, α-helix, and other partial helices. The compact structures and β-sheet conformations seem unfavorable in the AMBER99 force field simulations of Ala13. In contrast, their minimum energies are sampled when using the AMBER96 force field (Chapter 4). Results for DIVE simulations of Ala13 are largely independent of initial structures. In our simulations, the global and local potential energy minima sampled from an initially extended structure are almost identical with those sampled from an initial ideal α-helix.

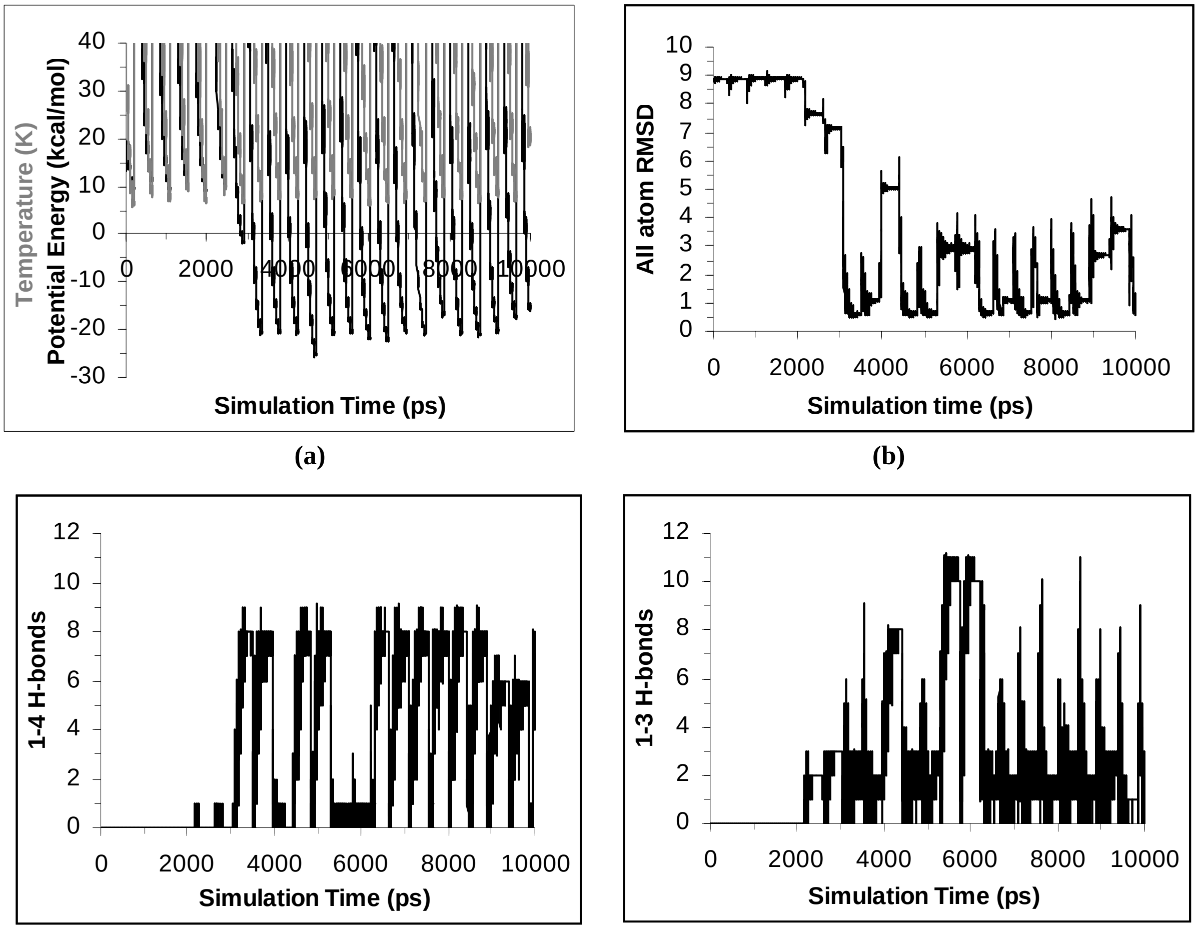
5.4. Comparisons of potential energy landscapes and conformations of polyalanine for different force fields

Different force fields play an important role in phase space sampling. The sampled minimum energy states from different force fields can be significantly different while using the same simulation protocols. For example, the observation of clusters of U-shaped 3_{10}-helices and 3_{10}-helices as the global free energy minimum conformations using the AMBER99 force field contradicts the simulations using the AMBER96 force field where the α-helix is the global minimum conformation for Ala13 *in vacuo* at 300 K. These two different force fields also give different potential energy minima for Ala13 *in vacuo* at 0 K. Moreover, the DIVE simulations from the AMBER96 force field sampled more varied conformations than those from the AMBER99 force field. In the AMBER99 force field simulations of Ala13, the minima for β-sheets and a large number of well-ordered compact conformations could not be searched, while the 3_{10}-helices are easily located.

The different minimum energy states sampled from AMBER96 and AMBER99 force fields reflect the different potential energy landscapes built on these two different force fields. Due to the efficient diverse conformational space sampling for this DIVE technique, we believe that the global potential energy minimum and its conformation is most likely as we reported above, even on limited simulation times (10 ns). On the other hand, we consider that the failure to locate the β-sheet conformational minima from AMBER99 and for the 3_{10}-helical minima from AMBER96 is due to their high potential energies. The observation of some β-sheet-rich conformations in the trajectory from AMBER99 force field simulations at high temperatures provides some evidence. To sample the minimum potential energies for these conformations, we further supplemented our DIVE simulations, starting from double-stranded β-sheets and 3_{10}-helical conformations, by using the same simulation protocols.

5.4.1. Further supplemental DIVE simulations

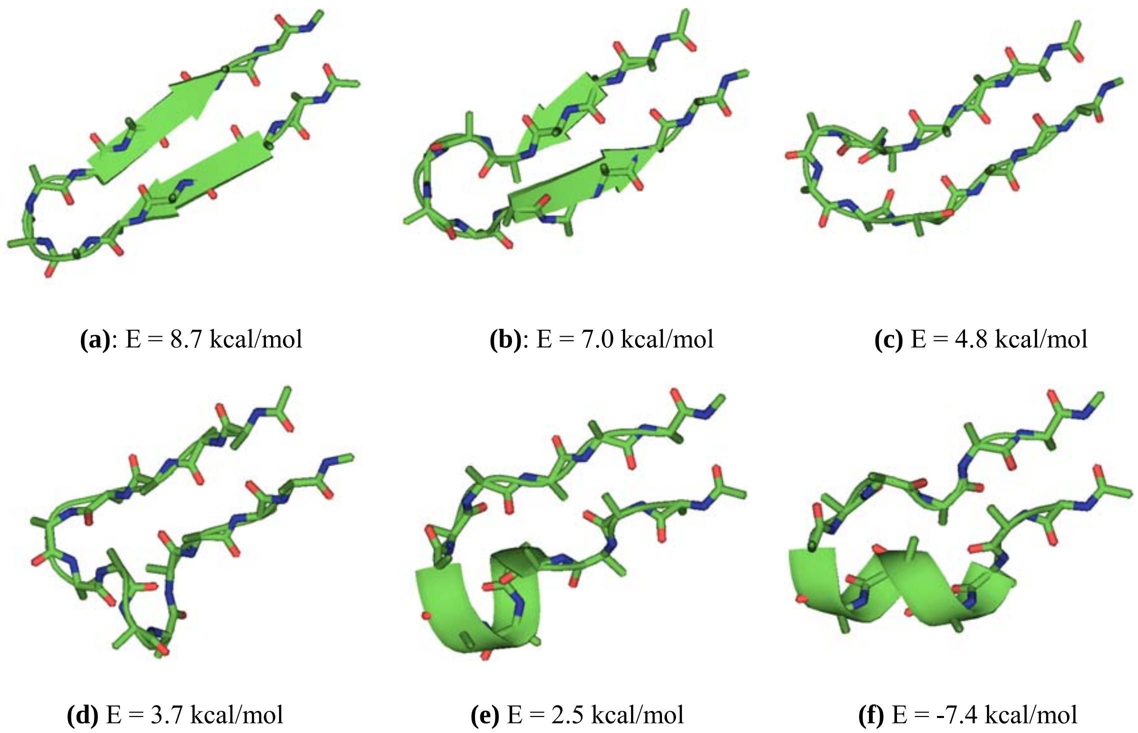
First, we repeated the AMBER 99 force field simulations starting from a nearly ideal double-stranded β-sheet (Fig. 4.3.2.2.3d, Chapter 4) taken from the AMBER 96 force field simulations. Fig. 5.4.1.1a gives an example showing the temperatures and potential energies of a single polypeptide, for which the global potential energy minimum (-25.7 kcal/mol, the U-shaped 3_{10}-helix) is sampled. Fig. 5.4.1.1b – d display the histories of the corresponding RMSD, 1-4 H-bonds, and 1-3 H-bonds respectively. The RMSD and H-bonding plots indicate that the polypeptide almost keeps the β-sheet conformations in the first 2.5ns (RMSD: approximately 9.0 Å) and then converts to helical conformations. The minimum potential energies for β-sheets are approximately 10 kcal/mol. Fig. 5.4.1.2 displays the representative conformations (not shown as the minimum energy states in the simulation starting from the extended structure) sampled from all individual trajectories. Very interestingly, these conformations constitute a reasonable path for the conformational transition from the high potential energy states of β-sheets to the low energy states of helical conformations.



(a) (b)

(c) (d)

Figure 5.4.1.1. Time series of various properties of a single polypeptide from a DIVE simulation of six trajectories of Ala13 *in vacuo*, starting from a double-stranded β-sheet (taken from AMBER96 force field simulation). **(a)** Temperature (upper, gray curve) and potential energy (lower, black curve), global potential energy minimum (-25.7 kcal/mol) was sampled. **(b)** RMSD between all atoms of the calculated structure and an ideal α-helix. The RMSD between the global potential energy minimum U-shaped 3_{10} helical conformation and an ideal α-helix is approximately 6 Å. **(c)** 1-4 hydrogen bonds. **(d)** 1-3 hydrogen bonds.



(a): E = 8.7 kcal/mol

(b): E = 7.0 kcal/mol

(c) E = 4.8 kcal/mol

(d) E = 3.7 kcal/mol

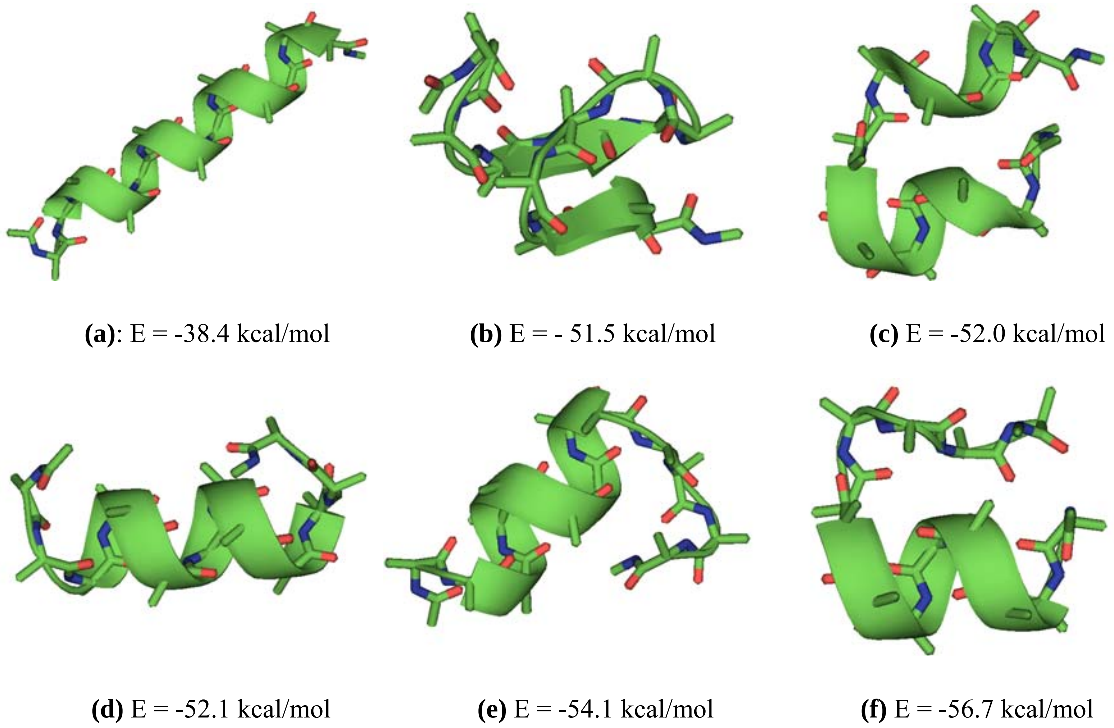
(e) E = 2.5 kcal/mol

(f) E = -7.4 kcal/mol

Figure 5.4.1.2. Conformations and potential energies (kcal/mol) sampled from the DIVE simulations of six trajectories of Ala13 *in vacuo* by using AMBER99 force field, starting from a double-stranded β-sheet structure. Approximate descriptions of the conformations, RMSD of all atoms from the ideal α-helix, and number of H-bonds are: **(a)** A nearly ideal double-stranded β-sheet (RMSD is 8.9 Å; five hydrogen bonds, plus one heavy atom engaged in two H-bonds). **(b)** Double-stranded β-sheet (RMSD is 8.2 Å; six hydrogen bonds). **(c)** Two β-strands (RMSD is 8.1 Å; four hydrogen bonds, plus one heavy atom engaged in two H-bonds). **(d)** Two β-strands, linked by ½ turn 3_{10}-helix (RMSD is 7.6 Å; five hydrogen bonds). **(e)** One turn α-helix, termini forming β-strands (RMSD is 7.7 Å; four hydrogen bonds, plus one heavy atoms engaged in two H-bonds). **(f)** 3_{10}-helix, termini forming β-stands (RMSD is 7.1 Å; ten hydrogen bonds). These images were generated by using PyMOL (http://pymol.sourceforge.net/).

We observed the similarities and differences of the potential energy minima sampled in the DIVE simulations of Ala13 *in vacuo* by using AMBER99 force field, starting from the extended structure, ideal α-helix, and double-stranded β-sheets, respectively (data not shown here). While the sampled minima at low potential energies show a high degree of similarity, those at high potential energies show some differences. This observation indicates that the DIVE simulations might be weakly initial-structure-dependent, if we consider the sampling of high potential energy minima. Two reasons account for the weak dependence on initial structures. First, as we mentioned above, the DIVE simulations of an initial configuration on a limited time scale by no means can sample all the local potential energy minima. Second, high potential energy minima should have much less probability of being sampled at such low temperatures (below 10 K). They only can be searched easily at low temperatures from the direct quenching of their corresponding initial or closely related conformations. In spite of these differences, in all these cases, the same global potential energy minimum U-shaped 3_{10} helical conformation is sampled during the simulations.

The DIVE simulations of Ala13 by using AMBER96 force field starting from a parallel U-shaped 3_{10}-helix and nearly ideal 3_{10}-helix (Fig. 5.3.1.2a and Fig. 5.3.1.4c, taking from the AMBER 99 force field simulations) demonstrate that the nearly ideal α-helix is the global potential energy minimum *in vacuo* for the AMBER96 force field (data not shown). As we expected, the minimum of the 3_{10}-helix is located at high potential energies, above that of the double-stranded β-sheets. Fig. 5.4.1.3 displays the representative conformations (not shown in the simulation starting from the extended structure) sampled from all individual trajectories. Since the energy gaps between the 3_{10}-helices and the β-sheet-rich conformations are small, the history plot of the sampled minimum potential energy states is very similar to those starting from the extended structure and ideal α-helix. However, some conformational minima are still sampled differently from different initial configuration simulations. On the other hand, a target heating temperature of 1000 K does not seem high enough for simulations starting from the α-helix or 3_{10}-helix (which easily converts to an α-helix) to search for β-sheet local minima, but 1200 K is high enough. The observations indicate that the thermalization temperature for complete unfolding of an α-helix *in vacuo* using the AMBER96 force field is between 500 K (=1000 K / 2, since half of the kinetic energy quickly converts into the α-helical potential energy) and 600 K (=1200 K / 2).



(a): E = -38.4 kcal/mol

(b) E = - 51.5 kcal/mol

(c) E = -52.0 kcal/mol

(d) E = -52.1 kcal/mol

(e) E = -54.1 kcal/mol

(f) E = -56.7 kcal/mol

Figure 5.4.1.3. Conformations and potential energies (kcal/mol) sampled from a DIVE simulation of six trajectories of Ala13 *in vacuo* by using AMBER96 force field, starting from a nearly ideal 3_{10}-helix (**a, b, f**) or a parallel U-shaped 3_{10}-helix (**c, d, e**). Approximate descriptions of the conformations, RMSD of all atoms from the ideal α-helix, and number of H-bonds are: **(a)** A nearly ideal 3_{10}-helix (RMSD is 2.7 Å; twelve hydrogen bonds). **(b)** Two β-strands (RMSD is 5.3 Å; five hydrogen bonds). **(c)** A U-shaped 3_{10}-helix (RMSD is 6.2 Å; ten hydrogen bonds, plus one heavy atom engaged in two H-bonds). **(d)** α-helix, termini frayed (RMSD is 3.2 Å; eight hydrogen bonds). **(e)** α-helix, a β-turn at the C-terminus (RMSD is 3.9 Å; nine hydrogen bonds). **(f)** U-shaped, 2/3 α-helix, 1/3 extend, linked by a turn (RMSD is 5.6 Å; nine hydrogen bonds). These images were generated by using PyMOL (http://pymol.sourceforge.net/).

5.4.2. Discussion

A direct comparison of the potential energy minima and their different conformations is necessary for us to clarify the distinct potential energy landscapes from different force fields. Table 5.4.2.1 shows the sampled minimum potential energies of several representative conformations from AMBER96 and AMBER99 force field simulations. Fig. 5.4.2.1a – b display the comparison of the potential energy gap of these conformations between these two force fields. The larger energy gap (73.8 kcal/mol) of the extended structure from the global minimum in AMBER99 compared to that in AMBER96 (54 kcal/mol), indicates a much flatter potential energy surface of Ala13 built from the AMBER96 force field (shown in Fig. 5.4.2.1a). Furthermore, the energy gap among these conformations, excluding the extended structure, is more closely and evenly distributed in AMBER96, while it clearly splits two regions of β-sheets and 3_{10}/α-helical conformations in AMBER99 (Fig. 5.4.2.1b). The large energy gap between sheets and helices accounts for the preference of helical conformations for the AMBER99 force field. Yoda, Takao, *et al.* demonstrate that G-peptide, known to form β-hairpin structures in aqueous solutions, was even shown to adopt some helical conformations in the AMBER99 explicit solvent simulations using generalized-ensemble algorithms [7].

Table 5.4.2.1. Comparison of the sampled minimum potential energies, potential energy gap from the global minimum, and potential energy gap from the minimum immediately below of several representative conformations for two different force fields.

Structure	Min. potential energy (kcal/mol)		Energy gap from global minimum (kcal/mol)		Energy gap from min. immediately below (kcal/mol)	
	AMBER		AMBER		AMBER	
	96	99	96	99	96	99
1: Extended	-7.2	48.1	54.0	73.8	31.2	39.4
2: Double-stranded β-sheets	-48.4	8.7	12.8	34.4	3.6	26.7
3: 3_{10}-helix	-38.4	-18.0	22.8	7.7	10.0	3.2
4: Parallel U-shaped 3_{10}-helix	-52.0	-25.7	9.2	0	9.2	0
5: α-helix	-61.2	-21.2	0	4.5	0	4.5
Global potential energy minimum	-61.2	-25.7				

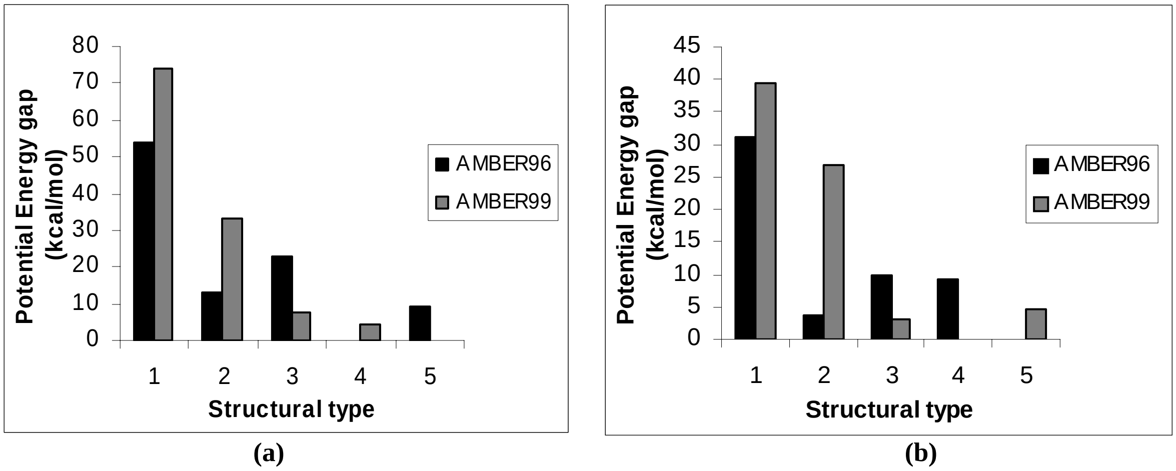


Figure 5.4.2.1. The comparison of **(a)** potential energy gap between the global minimum and the structure types in Table 5.4.2, and **(b)** potential energy gap between minima listed in Table 5.4.2 and the potential energy minimum immediately below each minimum. The Figure compares results for the AMBER96 and AMBER99 force fields.

The distinct features of the potential energy landscapes show two side effects of molecular dynamics simulations. First, when starting from an extended structure, an efficient simulation technique like the DIVE protocol can sample much more diverse regions of conformational space in AMBER96 than AMBER99. Second, the less efficient conventional NVT simulations of the polypeptide can be easily trapped in local β-sheet conformations in AMBER96 (Chapter 3) but is able to sample the global minimum helical conformation in AMBER99. It is difficult to evaluate accurately which force field is better for use in molecular dynamics, since here we test them only by using a model polypeptide of Ala13 in the gas phase. However, the simulation results may indicate some undesirable features of potential energy landscapes for this polyalanine built from both force fields. In the AMBER96 force field, the potential energy minima of many compact structures are close to those of α-helices and β sheets. As a result, the sampling of these diverse structures is likely to reduce the efficient sampling of those important secondary structural elements of proteins in a realistic MD simulation. On the other hand, in the AMBER99 force field, the minimum potential energies of β-sheets are too high compared to those of helices. As a result, these large energy differences can bias the sampling of the natural folded structure from β-sheets to helices.

The global potential energy minimum, and low potential energy local minima sampled starting from several very different conformations give almost identical results. These indicate that the disrupted velocity (DIVE) search simulation is an effective optimization technique for predicting 3D protein structures from only sequence data. In the DIVE simulations, we currently focus on locating potential energy minima rather than calculating statistical average ensemble properties. We think that it is meaningless to consider statistical averages if only a very small part of phase space can be sampled. Unfortunately, this condition known as a limited phase space sampling is usually true in most computational simulations in the last decade. Therefore, our strategy to develop this technique is to enhance phase space sampling as much as possible on the limited simulation time. We also believe that the DIVE technique can illuminate some effects of different force fields on MD simulations and the protein folding problem. This is because different force fields may reverse the energy levels of different minima, but are unlikely to exclude particular conformations. As an example, in the AMBER99 force field, the α-helix, known as the conformation of polyalanine [1,12-16], is not favored for Ala13 *in vacuo* at room temperature from the conventional NVT simulations (or DIP simulations), but it can be easily sampled in DIVE simulations.

The large differences in simulation results from different force fields may indicate that the current force fields are not yet capable of predicting the naturally polypeptide conformations. The free energy difference from the unfolded extended structure to the naturally folded conformation built from these force fields cannot drive the protein into folding automatically. However, our technique has the ability to sample diverse regions of conformational space efficiently to locate the naturally folded structure, though it may be neither the global potential energy minimum nor the free energy minimum. Polypeptides spanning tens of residues, rather than dipeptides and tetrapetides, are necessary to test molecules for further development of mechanical force fields. The DIVE simulation technique provides an effective way to help build large portions of the potential energy landscape of polypeptides for testing the next generation of force fields.

5.5. Bibliography

(1) Takano, M.; Yamato, T.; Higo, J.; Suyama, A.; Nagayama, K. *Journal of the American Chemical Society* **1999**, *121*, 605.

(2) Price, D. J.; Brooks, C. L., III. *Journal of Computational Chemistry* **2002**, *23*, 1045.

(3) Wang, J.; Cieplak, P.; Kollman, P. A. *Journal of Computational Chemistry* **2000**, *21*, 1049.

(4) Kollman, P.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A. *Computer Simulation of Biomolecular Systems* **1997**, *3*, 83.

(5) Li, Z.; Scheraga, H. A. *Proceedings of the National Academy of Sciences of the United States of America* **1987**, *84*, 6611.

(6) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598.

(7) Yoda, T.; Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **2004**, *386*, 460.

(8) Jarrold, M. F. *Annual Review of Physical Chemistry* **2000**, *51*, 179.

(9) Hudgins, R. R.; Mao, Y.; Ratner, M. A.; Jarrold, M. F. *Biophysical Journal* **1999**, *76*, 1591.

(10) Hudgins, R. R.; Jarrold, M. F. *Journal of Physical Chemistry B* **2000**, *104*, 2154.

(11) Hudgins, R. R.; Jarrold, M. F. *Journal of the American Chemical Society* **1999**, *121*, 3494.

(12) Hansmann, U. H. E.; Okamoto, Y. *Journal of Chemical Physics* **1999**, *110*, 1267.

(13) Son, H. S.; Hong, B. H.; Lee, C. W.; Yun, S.; Kim, K. S. *Journal of the American Chemical Society* **2001**, *123*, 514.

(14) Mitsutake, A.; Okamoto, Y. *Journal of Chemical Physics* **2000**, *112*, 10638.

(15) Daggett, V.; Levitt, M. *Journal of Molecular Biology* **1992**, *223*, 1121.

(16) Bertsch, R. A.; Vaidehi, N.; Chan, S. I.; Goddard, W. A., III. *Proteins: Structure, Function, and Genetics* **1998**, *33*, 343.

Chapter 6

Molecular Dynamics Simulations of Polyalanine in GB/SA Implicit Solvent

6.1. Introduction

The three limitations of potential energy traps, free energy traps, and kinetic traps cause a very biased phase space trajectory in conventional MD simulations, strongly dependent upon the initial coordinates. Divergent path (DIP) search simulations can alleviate this biased phase space sampling problem in MD simulations and help in attaining both folding and unfolding processes for some individual trajectories. A more complete phase space sampling can be further obtained with a number of DIP simulations, starting from several different conformations. In this way, the global free energy minimum near *in vivo* temperatures are more likely to be sampled. On the other hand, Disrupted velocity (DIVE) search simulations are performed by combining divergent path at one temperature by using cycles of heating and cooling. As a result, DIVE simulations can effectively sample diverse conformational space in a short simulation time. In addition, the technique can accurately map the global potential energy minimum and myriad local potential energy minima at temperatures near 0 K.

The previous three chapters described these new molecular dynamics simulation techniques to study Ala13 *in vacuo*. The results are strongly dependent on the force fields used in the simulations. For the AMBER 96 [1] force field, the conformation for both the global potential energy minimum and the global free energy minimum at room temperature is identified as an α-helix. For the AMBER 99 [2] force field, the global potential energy minimum is a U-shaped 3_{10}-helix while the family of global free energy minimum conformations at room temperature are clusters of U-shaped 3_{10}-helices and 3_{10}-helices. In our simulations, dependence on the initial conformation is minimal because the same simulation results are obtained from several very different starting structures. The occurrence of different global minimum energy conformations are artifacts of the different force fields. The fact that the global free energy and potential energy minimum conformation of the gas phase Ala13 is not the α-helix in the AMBER 99 force field aroused our interest in testing new simulations of this polypeptide in water using the same force field.

Polyalanine is known to form the α-helical conformation. Experiments [3-11] show that short alanine-based peptides may appear to form the 3_{10}-helix, rather than only α-helical conformations in aqueous solution, especially near their termini [4,9]. Simulations [12-18] of uncharged polyalanines with a sequence length between 10 and 30 indicate that the α-helical conformation is the global minimum energy conformation for the peptides both *in vacuo* and in solvent environments. It should be noted, however, that these simulations were based on earlier force fields (e.g. Amber 91 [15,17], ECEPP/2 [14,18]) and/or less accurate implicit solvent models (e.g.. distance dependent dielectric models [16,19]). On the other hand, most of the simulations were initiated from the α-helical conformation [13-15,17]. A direct observation of the folding process to global helical conformations from the initial extended conformation or the unfolding process from the initial α-helical conformation at ambient temperatures is much more valuable. Therefore, in this chapter, we present these two new molecular dynamics simulations (DIP and DIVE) in order to investigate the global free energy and potential energy minimum conformations of Ala13 in the GB/SA implicit solvent model for water using the AMBER 99 force field.

6.2. Simulation methodology

6.2.1. Potential energy function and GB/SA implicit solvent models

The potential energy function (eq 6.2.1.1) of the solute polypeptide uses the AMBER force field model [20,21] consisting of bond length stretching and angle bending represented by a simple harmonic expression. The dihedral angle twisting term is represented by a truncated Fourier series, the van der Waals interaction is modeled by a Lennard-Jones potential, and electrostatic interactions are represented by a Coulombic interaction of atom-centered partial charges. The Amber 99 force field [2] was adopted for all parameters in equation (6.2.1.1).

$$\begin{aligned} E_{potential} &= \sum_{\textbf{bonds}} K_r (r - r_{eq})^2 + \sum_{\textbf{angles}} K_\theta (\theta - \theta_{eq})^2 \\ &+ \sum_{\textbf{dihedral}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}{}^{12}} - \frac{B_{ij}}{r_{ij}{}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}} \right] \end{aligned} \tag{6.2.1.1}$$

In the GB/SA model [22], the total energy function E_{total} (eq 6.2.1.2) in implicit solvent environment includes the energy terms of the solute, $E_{\text{potential}}$, and the solvation energy, G_{sol}, for the interaction of the protein with the surrounding solvent. The G_{sol} term is traditionally considered as a sum of a solvent-solvent cavity term (G_{cav}), a solute-solvent van der Waals term (G_{vdW}), and a solute-solvent electrostatic polarization term (G_{pol}) [22].

$$\begin{aligned} E_{total} &= E_{potential} + G_{sol} \\ G_{sol} &= G_{cav} + G_{vdw} + G_{pol} \end{aligned} \tag{6.2.1.2}$$

A combination of the first two terms for the solvation energy is linearly related to solvent-accessible surface area (SA) of the atom types of the solute (eq 6.2.1.3), where SA_k is the total solvent-accessible surface area of atoms of type k, σ_k is an empirical atomic solvation energy parameter, and the summation extends over all atom types k. We follow Ref. [23] to calculate the accessible surface area SA_k while a preliminary value of +7.2 [22,24] or +5.0 [25,26] cal/(mol - Å^2) for the surface tension σ_k is used for all atom types.

$$G_{cav} + G_{vdw} = \sum_k \sigma_k SA_k \tag{6.2.1.3}$$

The G_{pol} term is calculated from the generalized Born (GB) equation. The original from was introduced by Still and co-workers [22] for the OPLS force field [27]:

$$\begin{aligned} G_{pol} &= -\frac{1}{2}\left(\frac{1}{\varepsilon_p} - \frac{1}{\varepsilon_w}\right)\sum_{ij}\frac{q_i q_j}{f_{gb}} \\ f_{gb} &= \left[\mathbf{r}_{ij}^2 + \alpha_i \alpha_j \exp\left(-\frac{\mathbf{r}_{ij}^2}{4\alpha_i \alpha_j}\right)\right]^{\frac{1}{2}} \end{aligned} \tag{6.2.1.4}$$

Where ε_p represents the low dielectric value inside of the protein, ε_w is the water (or other solvent) dielectric constant, r_{ij} is the separation distance of particles i and j, q_i, q_j are their charges, and α_i, α_j are the corresponding effective Born radii, respectively. The f_{gb} function interpolates between the α_i (or α_j) of small r_{ij} and r_{ij} itself at large distances.

Here we used two modified versions for which the parameterization is consistent with the AMBER force field. The first version incorporates a Debye-Hückel term $e^{-\kappa f_{gb}}$ in the generalized Born equation to account for salt effects at low salt concentrations [26,28]:

$$
\begin{aligned}
G_{pol} &= -\frac{1}{2}\left(1-\frac{e^{-\kappa f_{gb}}}{\varepsilon_w}\right)\sum_{ij}\frac{q_i q_j}{f_{gb}} \\
f_{gb} &= \left[\mathbf{r}_{ij}^2 + \alpha_i\alpha_j \exp\left(-\frac{\mathbf{r}_{ij}^2}{4\alpha_i\alpha_j}\right)\right]^{\frac{1}{2}}
\end{aligned}
\qquad (6.2.1.5)
$$

where κ is the Debye-Hückel screening parameter. This is the standard pairwise generalized Born model described by Tsui and Case [26,28]. We use GB1 to represent it. The second version modifies the f_{gb} function while keeping the original generalized Born equation, and takes the form [24,29]:

$$
\begin{aligned}
G_{pol} &= -\frac{1}{2}\left(1-\frac{1}{\varepsilon_w}\right)\sum_{ij}\frac{q_i q_j}{f_{gb}} \\
f_{gb} &= \left[\mathbf{r}_{ij}^2 + \alpha_i\alpha_j \exp\left(-\frac{\mathbf{r}_{ij}^2}{2\alpha_i\alpha_j}\right)\right]^{\frac{1}{2}}\left(\frac{\varepsilon\gamma-\gamma}{\varepsilon\gamma-1}\right) \\
&\textit{where} \\
\gamma &= 1-\left(\frac{\varepsilon_w-4}{2\varepsilon_w}\right)(\beta^2+2\beta+2)\exp(-\beta) \\
\beta &= (0.4\mathbf{r}_{ij}+\alpha_{ij})
\end{aligned}
\qquad (6.2.1.6)
$$

This is called the modified generalized Born model derived by Jayaram, Sprous and Beveridge [24,29]. We use GB4 to represent it. In both versions, the low dielectric value inside of the protein is considered one. We will not describe in detail the procedure to calculate the effective Born radius α values in these two different models, because it can be found elsewhere [24,26,28,29]. In the following simulations, the input parameters follow the standard values for each GB/SA implicit model. In GB1/SA, we set the water dielectric constant ε_W 78.5 and the surface tension parameter σ is +0.005 kcal/mol. In GB4/SA, we set the water dielectric constant ε_W = 80 and the surface tension parameter σ is +0.072 kcal/mol.

6.2.2. Computational details

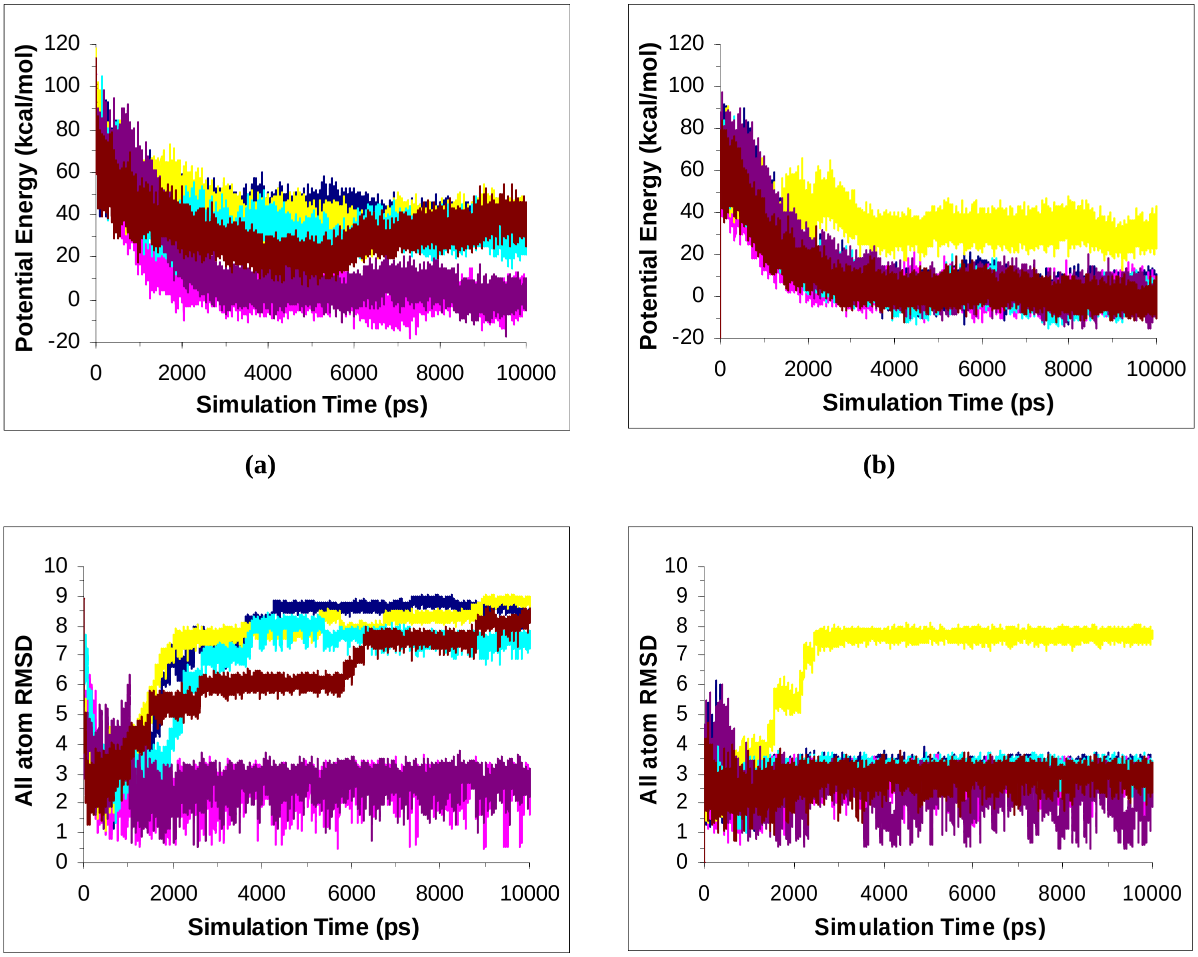
To run DIP simulations, we simulated six trajectories of Ala13 simultaneously and each polypeptide in a given trajectory was simulated independently for every step. These independent polypeptides were assigned the same initial structure (extended structure or α-helix) and temperature (300 K), but the same atom in each polypeptide had a different direction for its velocity. This was done in two steps: first, the same initial coordinates and velocities were assigned to the different polypeptides; then the velocities of the other five polypeptides were randomly re-set by changing the original direction but not the magnitude of the velocity vector for each atom in each polypeptide. The initial velocities of the first polypeptide were generated from a Gaussian velocity distribution [30] at 300 K. We used a time step of 1 fs but collected the trajectory data (energies and coordinates) at every 1000 time steps (1 ps intervals) in all MD simulations. The simulation time for each run was 10 ns for DIP simulations and 60 ns for conventional NVT simulations.

To run DIVE simulations, we simulated six trajectories of Ala13 simultaneously with different initial energies corresponding to temperatures of 10 K, 50 K, 100 K, 300 K, 600 K, and 1000 K. The scaling parameter for cooling was 0.25 and the scaling parameter for heating was calculated from the target temperature T_{target} (corresponding to the target energy) and the temperature (energy) T at the velocity reassignment step by $\sigma = T_{target}/T$. T_{target} was selected to be 1000 K. The threshold temperature for heating and cooling was 10 K. Thus, during the velocity reassignment, each polypeptide was cooled to 0.25 of its temperature T whenever its temperature rose above the threshold temperature. Once the temperature was below 10 K at the reassignment time, the polypeptide was heated back to T_{target}. Velocity reassignment occurred after every 20 ps. We used a time step of 1 fs but collected the trajectory data (energies and coordinates) at every 500 time steps (0.5 ps intervals) for 2 ns of implicit solvent simulations.

6.3. DIP Simulations for folding and unfolding studies of solvated polyalanine

6.3.1. Simulation results in GB1/SA implicit solvent model

Fig. 6.3.1.1 displays the time series of the potential energy and all-atom RMSD from an ideal α-helix from the DIP simulations of six trajectories of Ala13 at 300 K in GB1/SA implicit solvent. The simulations were started from the extended and α-helical conformations. While the temperature histories (not shown) are very similar for the simulations begun in either extended or α-helical conformations (oscillating between 300 K ± 50 K for any polypeptide), the corresponding potential energies (Fig. 6.3.1.1a - b) evolve differently, and the structures equilibrate in different potential energy wells. These different potential energy wells may include very different conformational states. The RMSD plots (Fig. 6.3.1.1c - d) clearly show that two polypeptides of the six (the 2nd and 5th polypeptide) undergo a folding process from the initial extended conformation (From the plots, RMSDs stabilize at values below 3.5 Å. so an RMSD less then 3.5 Å of the conformations is considered a folded structure). One polypeptide (the 3rd polypeptide) experiences an unfolding process from the initial α-helical conformation. A dynamic view of the seven folded trajectories in the Moil-view molecular graphics display programs [31] shows that the folded structures are either α-helical, 3_{10}-helical conformations, or clusters of mixed α/3_{10}-helices.



(a)

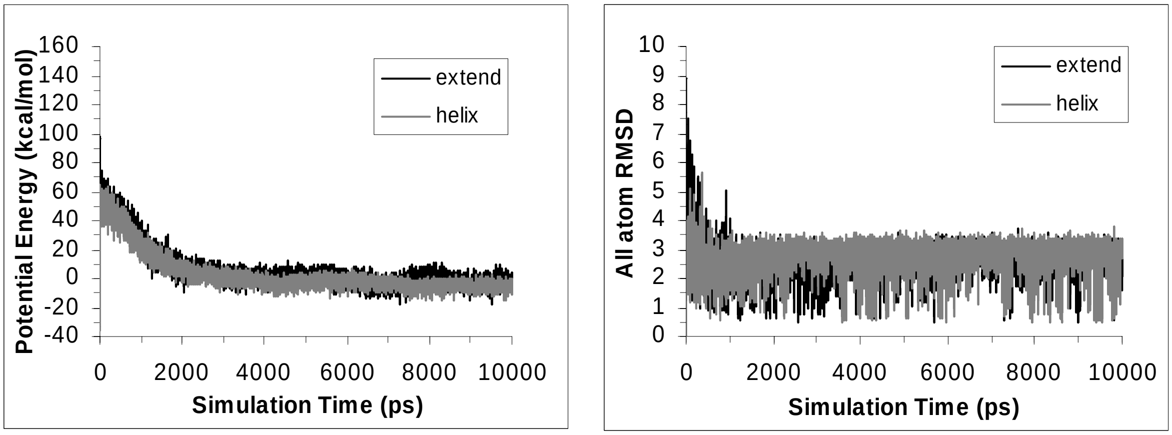
(b)

(c)

(d)

Figure 6.3.1.1. Time series of the potential energy and all-atom RMSD from an ideal α-helix in a six-trajectory, DIP simulation of Ala13 at 300 K in GB1/SA implicit solvent. **(a)** Potential energy, starting from an extended structure. **(b)** Potential energy, starting from an ideal α-helix. **(c)** RMSD between all atoms of the calculated structure and an ideal α-helix, starting from an extended structure. **(d)** RMSD between all atoms of the calculated structure and an ideal α-helix, starting from an ideal α-helix. The RMSD for an ideal 3_{10}-helix compared to the ideal α-helical reference structure is approximately 3 Å. **Note**: Blue: 1st, Magenta: 2nd, Yellow: 3rd, Light blue: 4th, Purple: 5th, Brown: 6th trajectory.

In conventional NVT simulations, however, different simulation results are usually obtained from the different initial conformations. The NVT simulation from the initial α-helix always gets trapped in the α-helical potential energy well, while that of the initial extended structure gets trapped in some local energy well [32,33]. Our conventional NVT simulations of the 1st polypeptide in both simulations on the time scale as long as 60 ns also support this phenomenon (data not shown here). Of course, the folding and unfolding processes of the helical conformations are not observed during the conventional NVT simulations.



(a) (b)

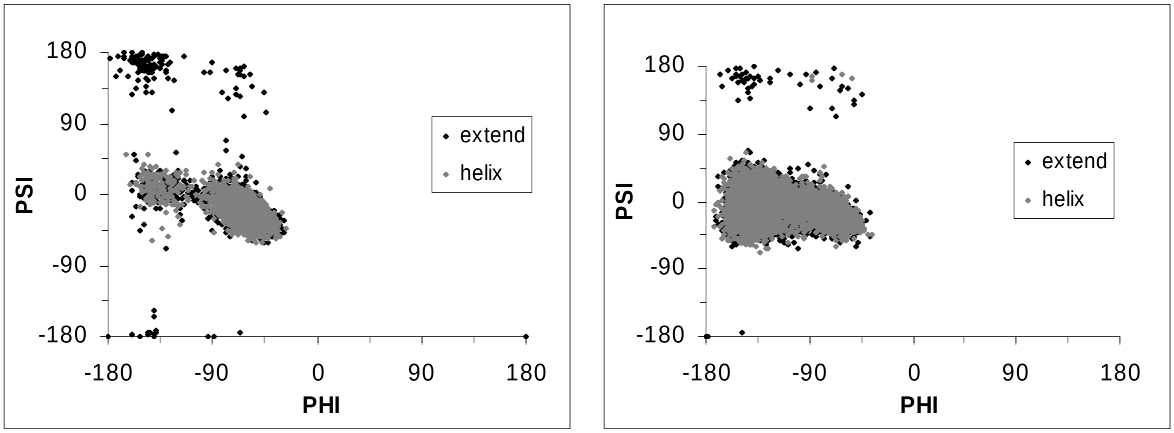
Figure 6.3.1.2. Time series of various properties of the polypeptide with lowest potential energy from a six-trajectory, DIP simulation of Ala13 at 300 K in GB1/SA implicit solvent, starting from an extended structure (black curve) or an ideal α-helix (gray curve), respectively. **(a)** Potential energy. **(b)** RMSD between all atoms of the calculated structure and an ideal α-helix.

Fig. 6.3.1.2a - b compare the histories of the potential energy and all-atom RMSD from an ideal α-helix of the polypeptide with the lowest potential energy at each data-collection step (ps interval) in DIP simulations starting from the initial extended or α-helical structures. Interestingly, the histories of the same type plot (potential energy or all-RMSD) from the different initial conformations are very similar to each other. The

RMSD histories of the polypeptide with the lowest potential energy from DIP simulations indicate that at 300 K in GB1/SA implicit solvent the Ala13 oscillates between the α-helix and 3_{10}-helix while the latter is the dominant helical form.

This conclusion is also supported in the plots of the hydrogen bonds (data not shown here). The number of hydrogen bonds (H-bond) for simulations starting from extended or α-helical structures is nearly identical after the first 500 ps of the simulations. Moreover, in both trajectories of the polypeptide with the lowest potential energy, 1-3 H-bonds dominate over 1-4 H-bonds, while other H-bonds are very rarely formed. The data indicate that in a solvent environment this polyalanine peptide oscillates between α-helical and 3_{10}-helical conformations. Thus, the global free energy conformations are not only limited to the α-helix at room temperature. In other words, the free energy minimum conformations for Ala13 at 300 K in solvent are clusters of similar helical conformations in which α-helices and 3_{10}-helices are metastable states. Metastable states can be defined as several nearly isoenergetic conformations [34,35]. Transitions between them can easily happen through protein motions due to small environmental perturbations. The results are different from the earlier simulation reports of uncharged polyalanines with a sequence length between 10 and 30 [12-17] which imply that the α-helical conformation predominates. Nevertheless, they are in good accord with experiments for short alanine-based peptides in aqueous solution [3,5,6,10,33,36].

It was reported from earlier simulations [17,37] and experiments [11,38] that the helical structure near the C-terminus is more fragile than that near the N terminus. Our simulations support this point. We first analyzed the histories of the first torsional angle φ (psi) (H-CT-C-N) in the ACE terminus (N-terminus of Ala13) and the last torsional angle ϕ (phi) (C-N-CT-H) in the NME terminus (C-terminus of Ala13) of the lowest potential energy trajectory. From both initial extended and α-helical structures, the last torsional angle ϕ at the C-terminus samples more angular space than the first torsional angle φ at the N-terminus of the Ala13 (data not shown here). This indicates that the frayed C-terminus of the α-helix should be more stable than the frayed N-terminus.



(a)

(b)

Figure 6.3.1.3. Ramachandran plots of the polypeptide with lowest potential energy from a six-trajectory, DIP simulation of Ala13 at 300 K in GB1/SA implicit solvent, starting from an extended structure (black dots) and ideal α-helix (gray dots). **(a)** ϕ, φ plot for the N-terminal alanine residue. **(b)** ϕ, φ plot for the C-terminal alanine residue. In both plots, large numbers of black dots are obscured by gray dots.

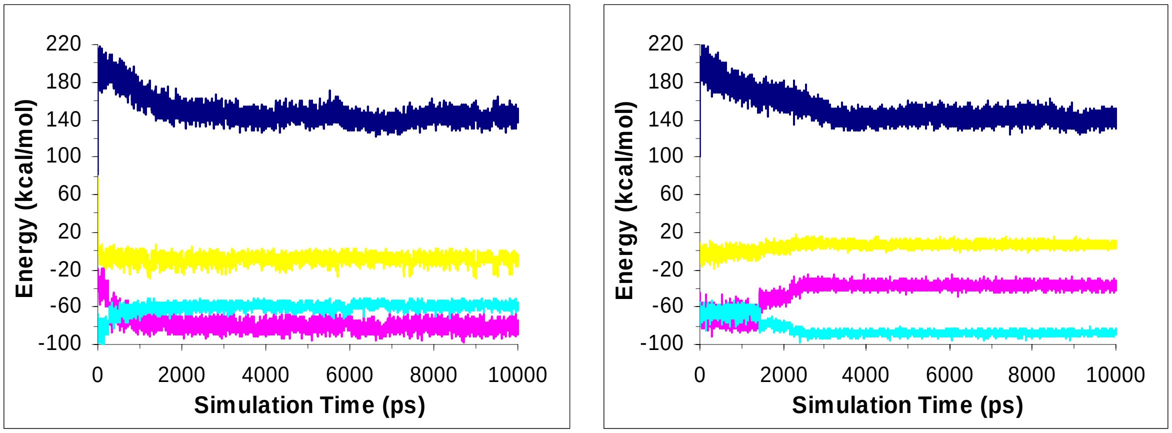
We also analyzed the histories of the ϕ, φ angle in the first alanine residue near the N-terminus and the last alanine residue near the C-terminus. Fig. 6.3.1.3 shows the Ramachandran plots of these ϕ, φ angles for polypeptide with the lowest potential energy from two different initial conformations. The ϕ, φ angles near the C-terminus (Fig. 6.3.1.3b) samples more conformations than those near the N-terminus do (Fig. 6.3.1.3a). Again, this is in good accord with the more fragile helical structure near the C-terminus than at the N-terminus. On the other hand, the ϕ, φ angles from the initially extended structure (black dots) samples a little more phase space than those from the initial α-helical form (gray dots). It is reasonable because the trajectory of the lowest potential energy polypeptide from the initial extended structure must undergo an extra folding process to reach the helical conformation. The observation that the N-terminal local helical conformation is more stable than the C-terminal local helical conformation is also consistent with the conformational potential energy minima. For Ala13 the minimum energy of the C-terminal frayed (or bent) helix is 2-3 kcal/mol less than that of the N-terminal frayed (or bent) helix (section 6.4.1).

Although it is controversial to make conclusions based on breaking the total energy into components[17,39-41], it is nonetheless interesting to consider what components of the energy contribute most to the total energy decrease upon forming the folded helical conformation. By analyzing each energy term, we found that the dominant energies that favor the helical conformation were the Coulombic and the van der Waals non-bonded energies. Throughout the simulations from two different initial configurations, the equilibrated unfolded extended conformations and folded helical conformations are both obtained from individual trajectories. From the initial extended conformation, the 2nd polypeptide folds and equilibrates by oscillating between α-helical and 3_{10}-helical conformations. The 4th polypeptide partially folds and unfolds, then equilibrates in a mostly extended conformation. From the initial α-helical conformation, the 3rd polypeptide unfolds and also equilibrates at a mostly extended conformation. The 5th polypeptide partially unfolds and refolds, then equilibrates by oscillating between the α-helical and 3_{10}-helical conformations. We calculated the equilibrated energies and the differences between energies from the unfolded extended conformations and the folded helical conformations in the simulation using the last 1 ns (9 ns - 10 ns) from both initial extended and α-helical conformations (See Table 6.3.1.1). The gas-phase Coulombic interaction (ΔE ≈ -44 kcal/mol), which includes H-bonds (11 1-4 H-bonds for α-helix and 12 1-3 H-bonds for 3_{10}-helix) appears to be a dominant factor in stabilizing the folded helical conformations in solvent. This opinion was also expressed by others [17,42,43]. The electrostatic solvation energy *G*pol (ΔE ≈ 29 kcal/mol) disfavors the global helical conformations. The net sum of the Coulombic energy and *G*pol energy is still decreased by 15 kcal/mol after folding. The net decrease indicates that formation of the intrapeptide H-bonds overcomes the increase of the solute-solvent polarization interaction. The van der Waals energy change (ΔE ≈ -18 kcal/mol) is another helix-stabilizing factor in our simulations. This is also consistent with the result from the published simulations of a 15-residue polyalanine (Ala15) in explicit solvent [17]. It should be noted, however, that a contradictory observation exists between our simulations and those reported by Takano, et al. in their explicit solvent simulations of Ala15 [17]. Our simulations clearly indicate that the geometric energy change, especially the torsional energy change, is not a helix-stabilizing factor — it disfavors the α-helix but favors the 3_{10}-helix a little compared to the extended conformation (data not shown here). This may be explained by the different force fields used in two different simulations: we use AMBER 99 and they use the AMBER 91 force field.

Table 6.3.1.1. The difference of average energies for each term and average potential energies (kcal/mol) between the equilibrated extended and folded helical conformations during the simulation for the last 1 ns (9 ns - 10 ns).

Energy Term (kcal/mol)	MD simulations from the initial extended conformation			MD simulations from the initial α-helical conformation		
	Mostly Extended (4th, 7.47[a])	Combined Helices (2nd, 2.37[a])	Energy Difference (ΔE)	Mostly Extended (3rd, 7.69[a])	Combined Helices (5th, 1.88[a])	Energy Difference (ΔE)
VdW	6.26	-9.77	**-16.0**	6.50	-13.2	**-19.7**
Coulombic	-37.0	-81.1	**-44.1**	-37.1	-80.8	**-43.7**
Geometric	140.2	142.6	**2.4**	139.7	146.3	**6.6**
Gpol	-87.5	-57.9	**29.6**	-87.4	-58.1	**29.3**
SA	7.13	6.02	**-1.12**	7.18	5.93	**-1.25**
Potential	29.2	-0.12	**-29.3**	28.8	-0.14	**-28.9**

a: The average value of RMSD (Å) from the ideal α-helix between 9ns and 10 ns.



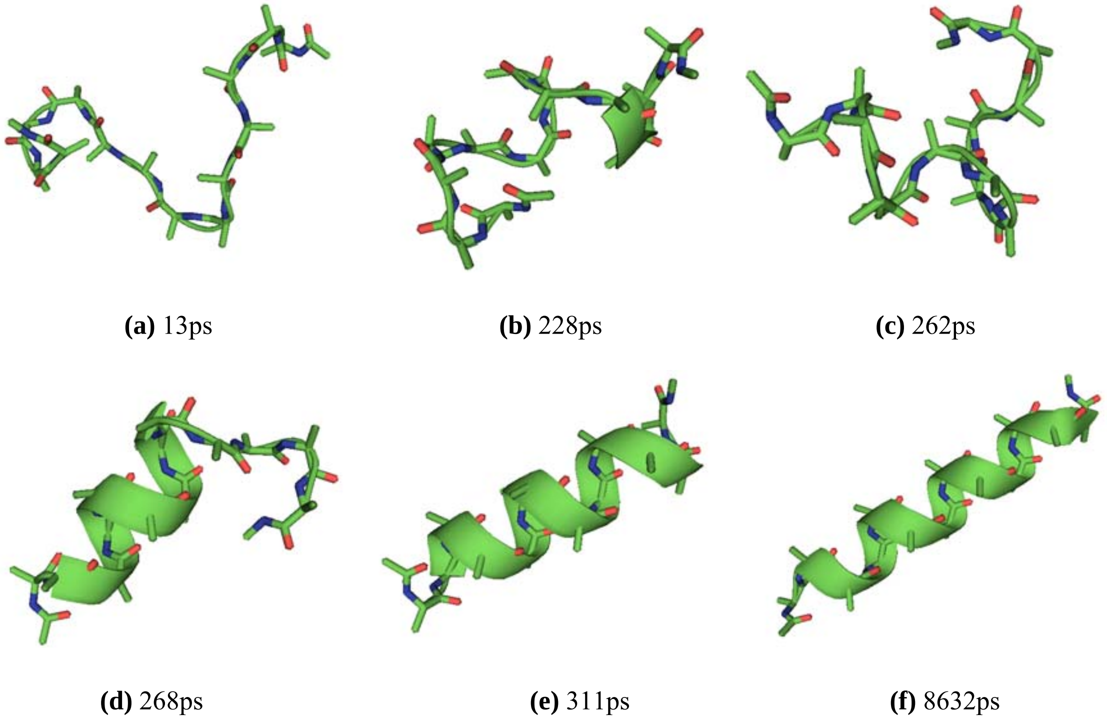
(a) **(b)**

Figure 6.3.1.4. Time series of the Geometric, Coulomb, van der Waals, and *G*pol energy terms of **(a)** of the 2nd polypeptide involved in α/3_{10}-helix folding process, starting from an extended structure, and **(b)** the 3rd polypeptide involved in helix unfolding process, starting from an ideal α-helix in a six-trajectory, DIP simulation at 300 K in GB1/SA implicit solvent. **Note**: Blue: Geometric, Magenta: Coulomb, Yellow: van der Waals, Light blue: *G*pol.

The helix-stabilizing factors can also be obtained from an analysis of the folding and unfolding dynamics process. Fig. 6.3.1.4a - b show the time series of the Geometric, Coulombic, van der Waals, and *G*pol energy terms of the 2nd polypeptide, starting from the initial extended conformation, and of the 3rd polypeptide, starting from the initial α-helical conformation. If during the first 3 ns of the simulations involving the helical folding or unfolding process, the histories of the Coulomb, van der Waals, and *G*pol energies favor the process, they will disfavor the reverse process, and vice versa. The history of the geometric energy, however, indicates that this energy always decreases no matter whether the simulation is involved in a folding or an unfolding process. We explain this unusual decrease of the geometric energy involving both helical folding and unfolding processes next.

The initial random Gaussian velocity distribution can only guarantee the appropriate canonical-ensemble average temperature; it cannot guarantee the appropriate velocity magnitude and direction for each individual atom. Inappropriately distributed individual velocities can induce very unbalanced coordinate displacements and interactions in different parts of the polypeptide. Some atoms move too close so that their potential energies become very large. These biased coordinate displacements cause unequilibrated potential energy (and total energy), indicating that mechanical equilibrium has not been reached. The mechanical non-equilibrium states are indicated by a large SHAKE energy from the beginning of the simulations. The states sampled during this equilibration process are normally located in the thermodynamically accessible regions at high temperatures when mechanical equilibrium is reached. The large decrease in total energy during the mechanical equilibration process, therefore, indicates the transition from high-energy states of folded or unfolded conformations to low-energy states of unfolded or folded conformations. The geometric energy, not the long-distant non-bonded interactions, is the main contribution to the non-equilibrium energies for the conformation sampled. Its normal decrease reflects the transition between these energy states. The mechanical equilibration process also account for the unusual observation that simulations can exit the global minimum energy helical conformation but get trapped in the local energy wells of the extended structure (e.g. the 3rd polypeptide trajectory from the initial α-helix simulations). The geometric energy obtained from the mechanical equilibration process may not be accurate enough to analyze the energy difference between folded and unfolded structures. The detailed explanation of mechanical equilibration process and equilibrium states is in Chapter 3.

The all-atom RMSD histories of the two polypeptides folding from the initially extended conformation (Fig. 6.3.1.1c) show that the helix folding process happens within the first several hundred picoseconds. Fig. 6.3.1.5 shows snapshots of the conformational change in the 2nd trajectory. The formation of helical structures can be described as follows. First, an extended Ala13 forms quickly into a random-coil conformation. This random coil has two or three turns with no more than one H-bond (Fig 6.3.1.5a). After 20 ps, several H-bonds characteristic of helices were formed in these turns in which the 1-3 type (β-turn [44]) predominated. Then, the polypeptide oscillated among these random-coil states with the migration of some turns by breaking and forming H-bonds. At 228 ps, a conformation with three β-turns in appropriate positions was formed (Fig. 6.3.1.5b). These β-turns were most likely to cooperate and to convert into three local half-turn α-helical structures (Fig. 6.3.1.5c). Next, a partial α-helical conformation spanning the middle and N-terminus was formed (Fig. 6.3.1.5d). Finally, the partial helix was folded to an entire α-helix at 311 ps (Fig. 6.3.1.5e). In the last 9.7 ns, Ala13 oscillates between α-helical and 3_{10}-helical conformations. Fig. 6.3.1.5f shows a nearly ideal 3_{10}-helix at 8632 ps. The unfolding trajectory of the 3rd polypeptide from the initial α-helix indicates that the 1-3 H-bond conformations also appear and may be transition states. In addition, the unfolding process happened in a more obviously cooperative way (figure not shown). A one turn 3_{10}-helix unfolds easily but refolds quickly. The substantial unfolding process happened when the two-turn 3_{10}-helix near the N-terminus unfolds into an extended structure. The partially extended form was then propagated. Finally, a nearly fully extended structure was obtained. The folding and unfolding studies of this polypeptide supports the ideal that α-helical conformations are initiated by 1-3 H-bonding within β-turns [17,44]. It should be noted, however, that the folding and unfolding processes show some differences in conformational transitions near the fully unfolded structures.



(a) 13ps **(b)** 228ps **(c)** 262ps

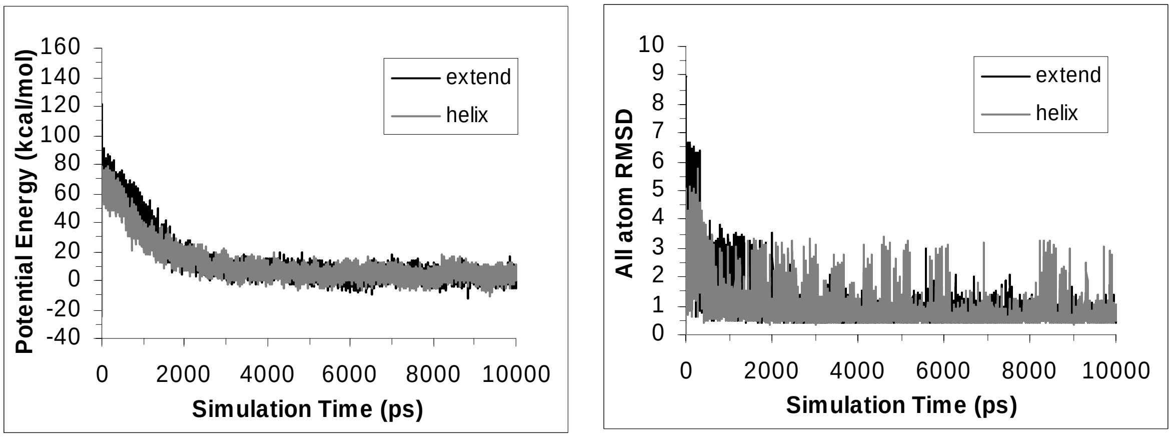
(d) 268ps **(e)** 311ps **(f)** 8632ps

Figure 6.3.1.5. Snapshots of the structural change in the 2nd trajectory upon folding, after starting from an extended structure. All hydrogen atoms are removed. **(a)** Three turns, no H-bonds. **(b)** Three 1-3 β-turns formed. **(c)** 1-4 H-bonds formed in turns. **(d)** Partial α-helix formed. **(e)** An α-helix. **(f)** An 3_{10}-helix. The images were generated by using PyMOL (http://pymol.sourceforge.net/).

6.3.2. Simulation results in GB4/SA implicit solvent model

We also ran a six-trajectory, DIP simulation of Ala13 at 300 K in GB4/SA implicit solvent. Most of the simulation results for the GB4/SA model are very similar to those of the GB1/SA model. For example, the simulations in GB4/SA also display folding trajectories from the initial extended structure and unfolding trajectories from the initial α-helix. The helical structure on the C-terminus is more fragile than that on the N-terminus. The global folded conformations are also helical clusters including both α-helices and 3_{10}-helices. The dominant conformation in the global folded conformations, however, is different for the simulations from these two different GB/SA implicit solvent models.

Figs. 6.3.2.1a - b show the time series of the potential energy and all-atom RMSD from an ideal α-helix of the polypeptide with the lowest potential energy from the initial extended and α-helix simulations, respectively. Again, the histories of the same type plot (potential energy or all-RMSD) from the different initial conformations are very similar to each other. The plots further show that the polypeptide with the lowest potential energy oscillates between an α-helix and a 3_{10}-helix (all-atom RMSDs < 3.5 Å). Notwithstanding, compared to the simulations in the GB1/SA model, α-helices rather than 3_{10}-helices are the dominant conformations during the trajectories using the GB4/SA model. From all the data, it is difficult to determine whether the α-helical or 3_{10}-helical conformation is the global minimum free energy conformation for Ala13 in solvent at room temperature. Instead, all simulation results indicate that the global minimum free energy conformation of Ala13 at 300 K should not be a single conformation but clusters of α-helices and 3_{10}-helices.



(a) **(b)**

Figure 6.3.2.1. Time series of the potential energy and all-atom RMSD from an ideal α-helix of the polypeptide with the lowest potential energy from a six-trajectory, DIP simulation of Ala13 at 300 K in GB4/SA implicit solvent, starting from the extended (black curve) and α-helical (gray curve) conformations. **(a)** Potential energy. **(b)** RMSD between all atoms of the calculated structures and an ideal α-helix.

The different dominant conformations for the lowest potential energy trajectory in GB4/SA and GB1/SA reflect the difference between the *G*pol and SA energies in these two different GB/SA models. The difference in *G*pol energies between these two different GB models is large. Table 6.3.2.1 lists the energy difference between GB1/SA and GB4/SA implicit solvent models for the initial extended and α-helical conformations. While the *G*pol for the extended structure from GB4 is as much as 17 kcal/mol higher than that from GB1, *G*pol for the ideal α-helix is only 7.8 kcal/mol higher than that from GB1. An analysis of the *G*pol energies from these two GB models throughout the simulations gives very similar results. The SA energy shows little difference between these two GB models, which is caused by the different values of the surface tension parameter σ. A high σ in GB4 also helps to increase the energy difference between the folded and unfolded conformations. All the effects combine to give a larger energy gap between the minima of different conformational potential wells in GB4 compared to

GB1. This is also supported by our DIVE simulations for mapping potential energy landscapes and minimum energy conformations of Ala13 (section 6.4.2).

Table 6.3.2.1. Energy differences for the initial extended and α-helical conformations between GB1/SA and GB4/SA implicit solvent models.

Energy terms	Initial extended conformation (Energy Unit: kcal/mol)			Initial α-helical conformation (Energy Unit: kcal/mol)		
	GB1/SA	GB4/SA	ΔE	GB1/SA	GB4/SA	ΔE
Gpol	-68.8	-51.8	-17.0	-54.7	-46.9	-7.8
SA	7.13	10.3	-3.17	5.61	8.07	-2.46
Potential	58.8	78.9	-20.2	-35.6	-25.2	-10.3

The similarities and differences between these two different GB/SA implicit solvent models can be further illustrated by simulation results for the individual polypeptides. Tables 6.3.2.2 and 6.3.2.3 show the equilibrated average potential energy, all-atom RMSD from an ideal α-helix, and conformations of the last 1 ns trajectory (9 ns - 10 ns) for six polypeptides of Ala13 at 300 K in GB1/SA and GB4/SA implicit solvent, starting from extended and α-helix conformations, respectively. The majority of the simulation results in these two different GB/SA models are similar and comparable with each other for the same polypeptide. In general, the GB1 model prefers unfolded states in contrast to the GB4 model. The corresponding all-atom RMSD in GB1 is higher than that in GB4 (with one exception). An interesting observation involves the structural differences between the equilibrated helical conformations in the corresponding polypeptides between these two different GB models. These structural differences include (1) 3_{10}/α-helix (GB1) vs α-helix (GB4) for the 2nd polypeptide and the lowest potential energy polypeptide in Table 6.3.2.2 as well as the 5th polypeptide in Table 6.3.2.3; (2) 3_{10}-helix (GB1) vs 3_{10}/α-helix (GB4) for the 5th polypeptide in Table 6.3.2.2, the 2nd polypeptide and the 6th polypeptide in Table 6.3.2.3. These results support the point that the GB4 model modified the *G*pol energies and thus enlarges the energy gap between the ground state α-helix and other conformations. Thus, the GB4 model makes the α-helix more stable than the GB1 model throughout the simulations.

Table 6.3.2.2. Equilibrated potential energy (kcal/mol), all-atom RMSD (Å) from an ideal α-helix, and conformations of the last 1 ns trajectory (9 ns - 10 ns) for six polypeptides of Ala13 in GB1/SA and GB4/SA implicit solvent starting from the initial extended structure.

Polypeptide Order Num.	MD simulations from the initial extended structure in GB1/SA			MD simulations from the initial extended structure in GB4/SA		
	Potential Energy	RMSD	Conform.	Potential energy	RMSD	Conform.
1st	35.3	8.53		52.6	7.61	
2nd	-0.12	2.37	3_{10}/α-helix	4.50	0.63	α-helix
3rd	39.5	8.86		53.1	8.20	
4th	29.2	7.47		42.0	6.50	
5th	1.54	2.74	3_{10}-helix	8.35	0.92	3_{10}/α-helix
6th	38.4	8.10		48.0	7.47	
Lowest	-1.90	2.45	3_{10}/α-helix	3.38	0.62	α-helix

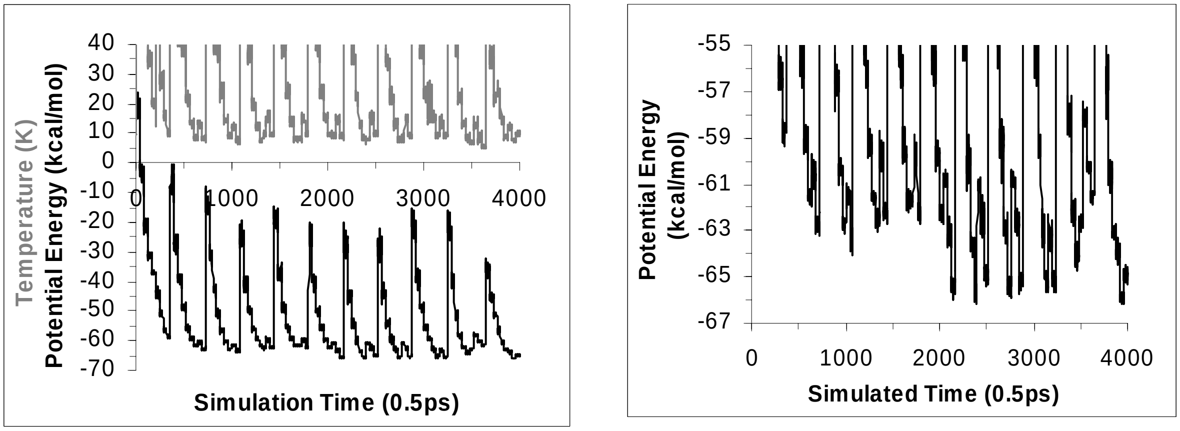
Table 6.3.2.3. Equilibrated potential energy (kcal/mol), all-atom RMSD (Å) from an ideal α-helix, and conformations of the last 1 ns trajectory (9 ns - 10 ns) for six polypeptides of Ala13 in GB1/SA and GB4/SA implicit solvent starting from the initial α-helix.

Polypeptide Order Num.	MD simulations from the initial α-helix in GB1/SA			MD simulations from the initial α-helix in GB4/SA		
	Potential Energy	RMSD	Conform.	Potential energy	RMSD	Conform.
1st	2.95	3.16	3_{10}-helix	18.6	3.06	3_{10}-helix
2nd	1.94	2.95	3_{10}-helix	17.9	2.27	3_{10}/α-helix
3rd	28.8	7.69		54.6	8.91	Extended
4th	-0.91	3.11	3_{10}-helix	18.5	3.06	3_{10}-helix
5th	0.14	1.88	3_{10}/α-helix	3.67	0.59	α-helix
6th	-1.64	2.91	3_{10}-helix	12.8	1.46	3_{10}/α-helix
Lowest	-4.99	2.64	3_{10}/α-helix	3.15	0.60	3_{10}/α-helix

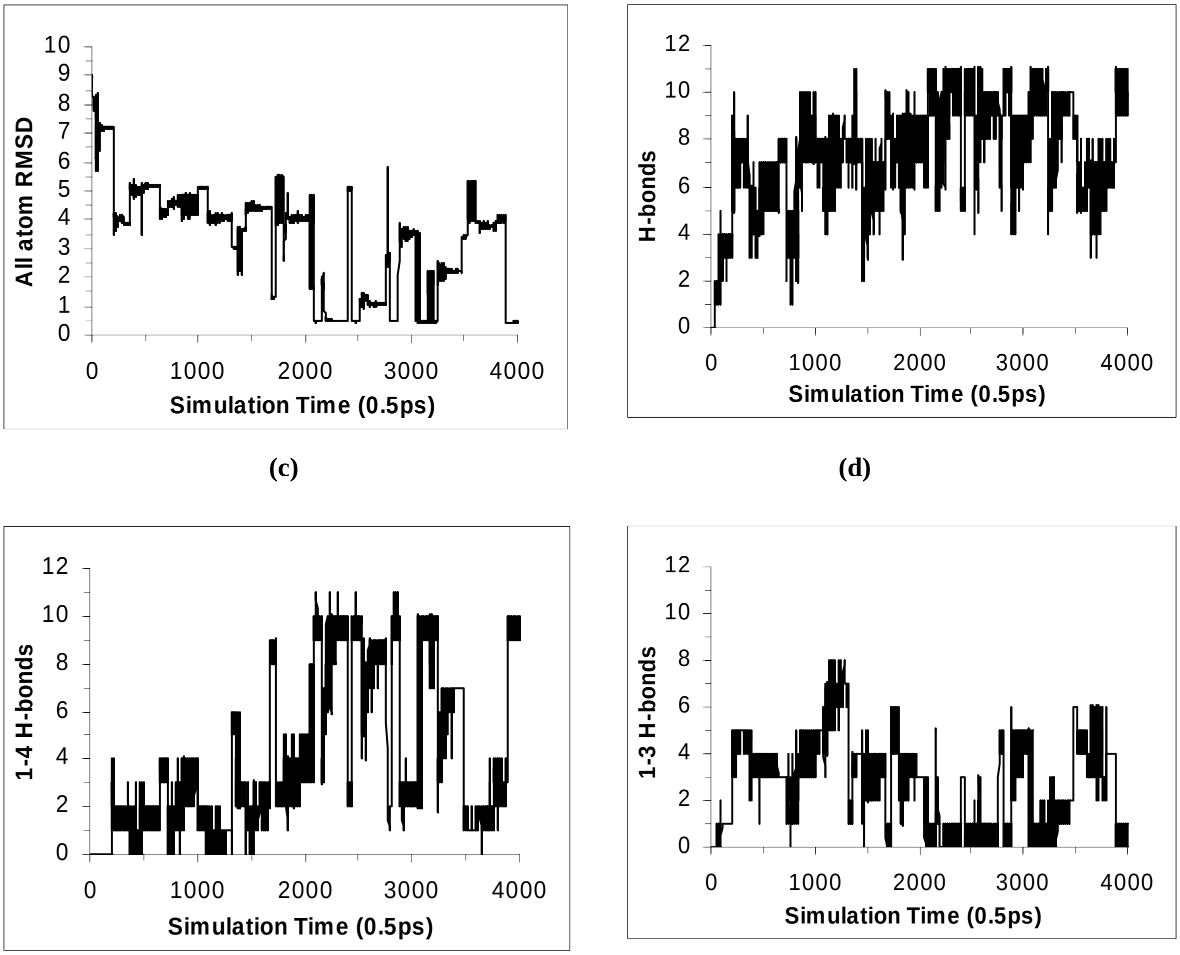
6.4. DIVE simulations for mapping potential energy landscapes and conformations of solvated polyalanine

6.4.1. Simulation results in GB1/SA implicit solvent

Fig. 6.4.1.1a displays the temperature and potential energy histories of the polypeptide with the lowest potential energy for six trajectories of Ala13 from DIVE simulations in the GB1/SA implicit solvent model, starting from the extended structure. Figure 6.4.1.1b enlarges the potential energy history in the low potential energy region to show more clearly the different potential energy minima sampled during the simulations. Fig. 6.4.1.1c - f show the corresponding RMSD from the ideal α-helix, and all H-bonds, 1-4 H-bonds, 1-3 H-bonds, respectively. 27 regions of local potential energy minima are obtained from the lowest potential energy history and 65 regions in all are obtained from the individual potential energy histories of six trajectories of Ala13.



(a) (b)



(c)

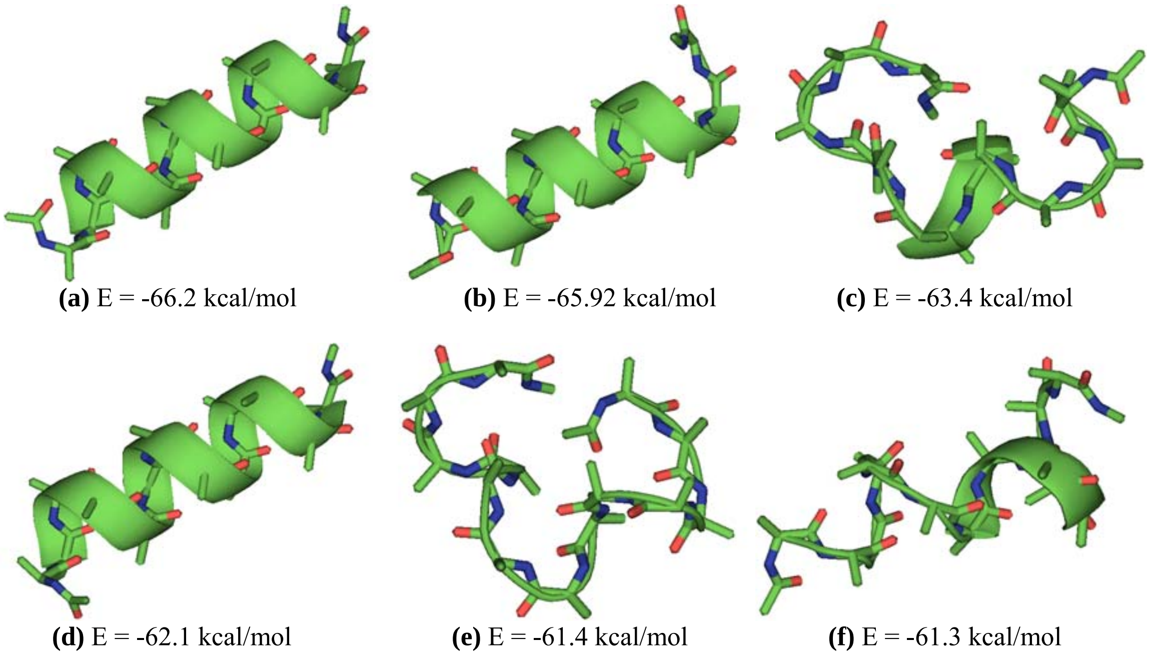
(d)

(e)

(f)

Figure 6.4.1.1. Time series of various properties of the polypeptide with lowest potential energy from a DIVE simulation of six trajectories of Ala13 using the GB1/SA implicit solvent model for water, starting from an extended structure. **(a)** Temperature (upper, gray curve) and potential energy (lower, black curve). **(b)** Potential energy displayed on an expanded scale. **(c)** RMSD between all atoms of the calculated structure and an ideal α-helix (The RMSD between a 3_{10}-helix and an ideal α-helix is approximately 3 Å). **(d)** All hydrogen bonds. **(e)** 1-4 hydrogen bonds. **(f)** 1-3 hydrogen bonds. A hydrogen bond is identified when the distance between two heavy atoms X is below 3.3 Å and the X-H ··· X angle differs from 180° by less than 20.0°.

Figure 6.4.1.2 shows representative conformations corresponding to the potential energy minima, taken from the lowest potential energy trajectory. The ground state conformation is nearly a perfect α-helix with a potential energy of approximately –65 kcal/mol. Many local potential energy minima were found to be only 1-5 kcal/mol higher than the global potential energy minimum. The conformations of these local potential energy minima seem to be closely related to the ground state α-helical conformation. For example, the C-terminal frayed α-helix of Figure 6.4.1.2b is 2-3 kcal/mol more stable than the N-terminal frayed helix of Figure 6.4.1.2d. This indicates that the helical structure near the C-terminus should be more frayed than that near the N terminus, which is consistent with experiments [11,38] and other simulations [17,37]. In addition, even though the conformation in Figure 6.4.1.2e appears very different from an α-helix, a transition from the α-helix to this loop structure through the intermediate structures of Figs. 6.4.1.2f and 6.4.1.2c is observed in the individual trajectories.

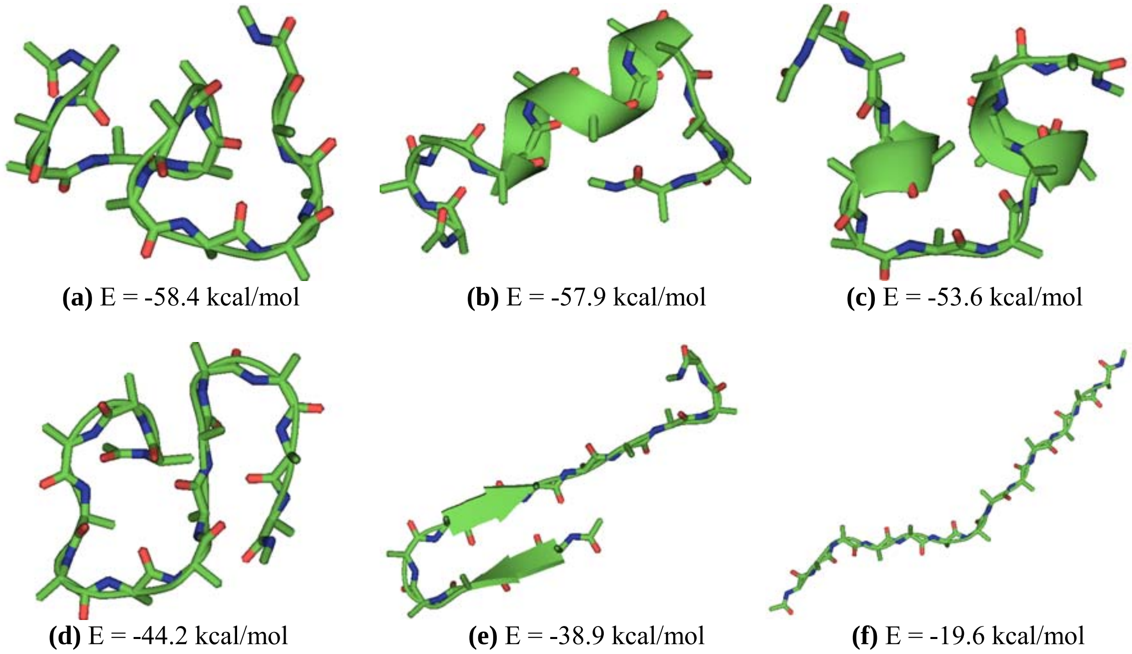

(a) E = -66.2 kcal/mol **(b)** E = -65.92 kcal/mol **(c)** E = -63.4 kcal/mol

(d) E = -62.1 kcal/mol **(e)** E = -61.4 kcal/mol **(f)** E = -61.3 kcal/mol

Figure 6.4.1.2. Representative conformations and potential energies (kcal/mol) of potential energy minima sampled by the polypeptide with lowest potential energy in a six-trajectory, DIVE simulation of Ala13 using the GB1/SA implicit solvent model for water, starting from an extended structure. Approximate descriptions of the conformations, all-atom RMSD from the ideal α-helix, and number of H-bonds are: **(a)** Nearly ideal α-helix is the global potential energy minimum (RMSD is 0.5 Å; eleven hydrogen bonds). **(b)** α-helix with frayed C-terminus (RMSD is 1.0 Å; eight hydrogen bonds, plus one heavy atom engaged in two H-bonds). **(c)** Combined 3_{10}-helix and α-helix, frayed at the C-terminus (RMSD is 4.1 Å; two hydrogen bonds, plus three heavy

atoms engaged in two H-bonds). **(d)** α-helix with frayed N-terminus (RMSD is 1.3 Å; ten hydrogen bonds). **(e)** Three β-turns form a cavity (RMSD is 5.0 Å; three hydrogen atoms engaged in two H-bonds). **(f)** Mixed helix with 1-3, 1-4, and 1-5 H-bonds, an intermediate between several minima shown in this figure (RMSD is 2.7 Å; four hydrogen bonds, plus one heavy atoms engaged in two H-bonds and one heavy atom engaged in three H-bonds). A hydrogen bond is identified when the distance between two heavy atoms X is below 3.3 Å and the X-H ··· X angle differs from 180° by less than 20.0°. These images were generated by using PyMOL (http://pymol.sourceforge.net/).

Almost as interesting as the most stable structures observed during the DIVE simulations are several higher-energy structures that were less frequently detected. Although a U-shaped 3_{10}-helical and ideal 3_{10}-helical conformation found to be global or local minima *in vacuo* (Chapter 5) did not appear in simulations with the GB1/SA solvent model, partial 3_{10} helices (at the N or C terminus) are frequently sampled and occupy very low potential energy minima (less than 2 kcal/mol higher than the global potential energy minimum, e.g. Figure 6.4.1.2c). This is in good accord with experiments that imply a significant fraction of 3_{10}-helix exists near the termini for short alanine-based peptides in aqueous solution[4,9]. In addition, β-sheet structures are rare conformations with the GB1/SA implicit solvent model. We found only a few partial β-sheet-like conformations (occupying relatively high potential energy minima) in all local potential energy minimum regions from six individual trajectories. They are shown in Figure 6.4.1.3. A very interesting observation is that the potential energy of the extended form, the highest potential energy minimum sampled in the simulations below 10 K, is 20 kcal/mol lower than the equilibrated average potential energy of the helical conformation for Ala13 at 300 K (around 0 kcal/mol) (section 6.3.1). The six trajectories of Ala13 in DIVE simulations using the GB1/SA implicit solvent model starting from the initial α-helical structure give very similar results, so they are not shown here.



(a) E = -58.4 kcal/mol **(b)** E = -57.9 kcal/mol **(c)** E = -53.6 kcal/mol

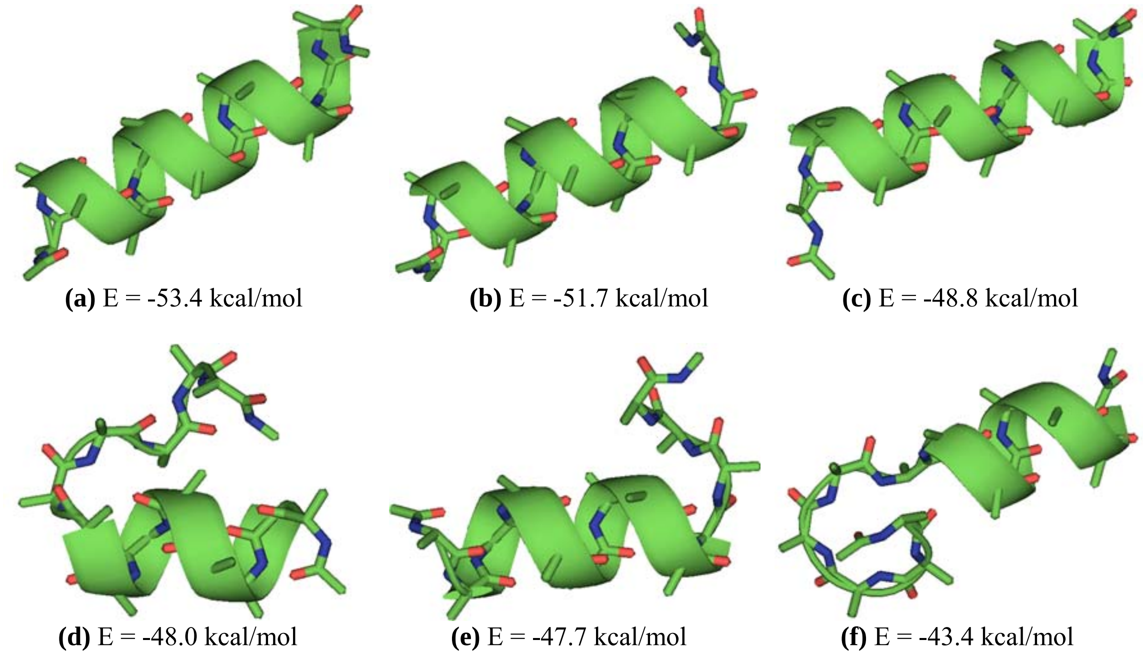
(d) E = -44.2 kcal/mol **(e)** E = -38.9 kcal/mol **(f)** E = -19.6 kcal/mol

Figure 6.4.1.3. Representative conformations and potential energies (kcal/mol) sampled by individual polypeptides in a six-trajectory, DIVE simulation of Ala13 using the GB1/SA implicit solvent model for water, starting from an extended structure. Approximate descriptions of the conformations, all-atom RMSD from the idea α-helix, and number of H-bonds are: **(a)** Three β turns forming a double U-shaped loop (RMSD is 5.0 Å; five hydrogen bonds, plus one heavy atom engaged in two H-bonds). **(b)** Helix with 1-3 and 1-4 H-bonds, frayed at the C-terminus (RMSD is 3.8 Å; six hydrogen bonds, plus one heavy atom engaged in two H-bonds). **(c)** Two partial 3_{10}-helices with a turn, frayed N- and C-termini (RMSD is 5.1 Å; three hydrogen bonds, plus one heavy atom engaged in two H-bonds). **(d)** Two perpendicular extended chains with a small β-turn at the N-terminus and a larger β-turn at the C-terminus (RMSD is 6.1 Å; four hydrogen bonds). **(e)** Half extended with a small β-turn at the C-terminus and a larger β-turn at the N-terminus (RMSD is 6.3 Å; four hydrogen bonds). **(f)** Extended structure (RMSD is 8.0 Å; zero hydrogen bonds). These images were generated by using PyMOL (http://pymol.sourceforge.net/).

Two observations indicate solvation reduces the high barriers between different potential energy wells *in vacuo* and enhances energy barrier crossings in searching for the ground state conformation in solvent. First, only 2 polypeptides of 6 trajectories *in vacuo* formed a U-shaped 3_{10}-helix, the global energy minimum *in vacuo* (Chapter 5), while 4 polypeptides out of 6 trajectories evolved into the α-helical ground state conformation in solvent environment. Second, the energy difference between the lowest potential energy minimum of the α-helix and the highest local potential energy minimum of the extended-conformation in GB1/SA (approximately –45 = -65 – (-20) kcal/mol) is 25 kcal/mol less than that *in vacuo* (around –70 = -20 – (50) kcal/mol). Thus, a flatter potential energy surface exists in solvent compared to that *in vacuo*. Solvation should thus make folding and unfolding processes easier for Ala13. The flatter PES and the smaller energy gap between potential energy minima indicate that NVT simulations in solvent environment are much more likely to oscillate between several metastable conformations than they are *in vacuo*. *In vacuo*, Ala13 is more easily equilibrated in one conformational potential energy well (Chapter 5).

6.4.2. Simulation results in GB4/SA implicit solvent

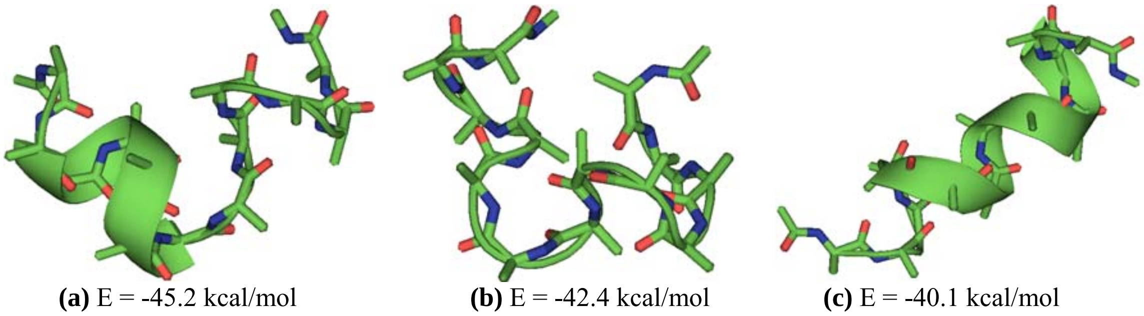
We also performed DIVE simulations using six trajectories of the extended Ala13 and the GB4/SA implicit solvent model. The simulation results are very similar to those obtained from the GB1/SA model (data not shown here). Figures 6.4.2.1 and 6.4.2.2 show conformations of minimum potential energy sampled by the lowest potential energy trajectory and by the individual trajectories, respectively. The ground state is α-helical (Fig. 6.4.2.1a). The C-terminal frayed α-helices (Figure 6.4.2.1b or Figure 6.4.2.1e) are 2-5 kcal/mol more stable than the N-terminal frayed α-helices (Figure 6.4.2.2c or Figure 6.4.2.2f). The 3_{10}-helix is also observed near the ends of Ala13 for some low potential energy conformations, and local potential energy minima with β-sheet conformations are very rare.



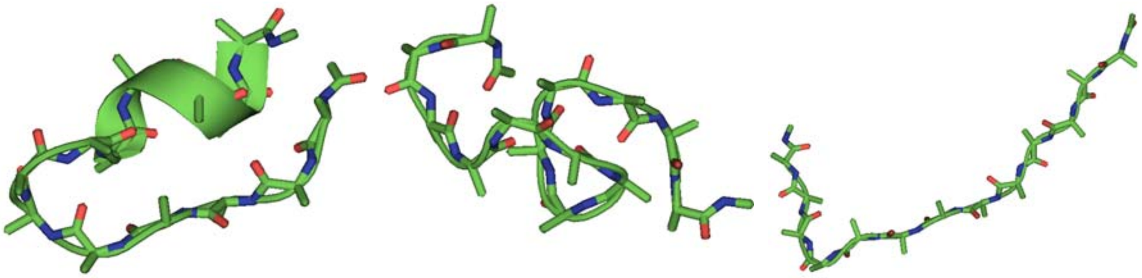
(a) E = -53.4 kcal/mol **(b)** E = -51.7 kcal/mol **(c)** E = -48.8 kcal/mol

(d) E = -48.0 kcal/mol **(e)** E = -47.7 kcal/mol **(f)** E = -43.4 kcal/mol

Figure 6.4.2.1. Representative conformations and potential energies (in kcal/mol) of potential energy minima sampled by the polypeptide with lowest potential energy in a six-trajectory, DIVE simulation of Ala13 using the GB4/SA implicit solvent model for water, starting from an extended structure. Approximate descriptions of the conformations, all-atom RMSD from the ideal α-helix, and number of H-bonds are: **(a)** Nearly ideal α-helix is the global potential energy minimum (RMSD is 0.4 Å; eleven hydrogen bonds). **(b)** α-helix with frayed C-terminus (RMSD is 1.0 Å; ten hydrogen bonds). **(c)** α-helix with frayed N-terminus (RMSD is 1.3 Å; ten hydrogen bonds). **(d)** α-helix plus 3_{10}-helix, linked by a turn (RMSD is 5.5 Å; nine hydrogen bonds). **(e)** α-helix plus C-terminal β-turn (RMSD is 3.5 Å; ten hydrogen bonds). **(f)** α-helix plus N-terminal β-turn (RMSD is 3.6 Å; six hydrogen bonds, plus one heavy atom engaged in three H-bonds). These images were generated by using PyMOL (http://pymol.sourceforge.net/).



(a) E = -45.2 kcal/mol **(b)** E = -42.4 kcal/mol **(c)** E = -40.1 kcal/mol



(d) E = -35.5 kcal/mol **(e)** E = -26.1 kcal/mol **(f)** E = -3.0 kcal/mol

Figure 6.4.2.2. Representative conformations and potential energies (in kcal/mol) sampled by individual polypeptides in a six-trajectory, DIVE simulation of Ala13 using the GB4/SA implicit solvent model for water, starting from an extended structure. Approximate descriptions of the conformations, all-atom RMSD from the ideal α-helix, and number of H-bonds are: **(a)** V-shaped, twisted α-helix (RMSD is 4.4 Å; six hydrogen bonds, plus two heavy atoms engaged in two H-bonds). **(b)** Four β-turns looped to form a cavity (RMSD is 5.1 Å; five hydrogen bonds, plus one heavy atoms engaged in two H-bonds). **(c)** Mixed helix with 1-3, 1-4, and 1-5 H-bonds, an intermediate between several minima shown in this figure (RMSD is 2.1 Å; two hydrogen bonds, plus four heavy atoms engaged in two H-bonds). **(d)** 3_{10}-helix at C-terminus, extended chain at N-terminus, linked by a turn (RMSD is 7.3 Å; six hydrogen bonds). **(e)** Three β-turns forming a U-shaped loop (RMSD is 5.0 Å; one hydrogen bond, two heavy atoms engaged in two H-bonds). **(f)** V-shaped, extended structure (RMSD is 7.1 Å; zero hydrogen bonds). These images were generated by using PyMOL (http://pymol.sourceforge.net/).

The energetic analysis for global α-helix formation gives similar results as that for the GB1/SA model. However, the potential energy gap between the global and local minima or between different local minima is slightly larger in GB4/SA than in GB1/SA. This is supported by contrasting the history of all sampled potential energy minima between these two GB/SA models (data not shown here). The potential energy excluding the SA energy shows the same result and indicates that the different GB algorithms are primarily responsible for this difference. At the global minimum, the SA energy for GB4 (σ = 0.0072 kcal/ mol-Å^2) is 2-3 kcal/mol higher than that for GB1 (σ = 0.005 kcal/ mol-Å^2), but the total potential energy in GB4/SA is approximately +13-14 kcal/mol higher than that in GB1/SA at the global minimum. Therefore, G_{pol} of GB4 is around 11 kcal/mol higher than that of the GB1 algorithm at low potential energies. A similar analysis for the extended structure indicates that the G_{pol} of GB4 is approximately 16 kcal/mol higher than that of the GB1 algorithm for unfolded Ala13. A relatively large energy gap and a large energy difference between folded and unfolded structures are in accord with known features of GB4: it was developed for modeling the electrostatic states in the interior of folded conformations more accurately than the GB1 algorithm [24,29].

6.5. Discussion and summary

Divergent path search simulations of Ala13 in GB/SA implicit solvent using the AMBER 99 force field clearly illustrate that near 300 K the global minimum free energy conformation for this polypeptide in aqueous solution is not a single helix but clusters of mixed $\alpha/3_{10}$-helical conformations. This result is obtained from two different GB/SA implicit solvent models. However, the GB4 model gives a larger *G*pol energy difference between different types of folded helices, making the ground state α-helix more stable near *in vivo* temperatures. The combined $\alpha/3_{10}$-helical conformations near the global free energy minimum are different from earlier simulation reports of uncharged polyalanines with a sequence length between 10 and 30 which reported an α-helix folded conformation [12-17]. However, our simulation results are in good accord with experiments for short alanine-based peptides in aqueous solution [3,5,6,10,33,36]. The DIP simulations starting from extended and α-helical forms give very similar lowest potential energy trajectories and similar folded conformations of the polypeptide. In addition, independent folding and unfolding processes can be observed from trajectories of some individual polypeptides. A conventional NVT simulation for a single polypeptide starting from the extended structure evolves into unfolded structures while the simulation starting from the α-helix maintains helical conformations, even for a longer simulation time. This clearly indicates that conventional MD simulations sample very biased phase space trajectories, and the simulation results are strongly dependent on the initial conformations.

An analysis of the folding and unfolding of Ala13 in GB/SA implicit solvent from the DIP simulations supports the point that the α-helix formation is usually preceded by formation of a short β-turn (1-3 H-bond) structure [17,44]. The energy decrease upon forming the helical conformation is mainly caused by the Coulombic and van der Waals non-bonded energies. Torsional energy disfavors the α-helix, contrary to other simulations based on earlier force fields [17]. It should be noted, however, that the determination of these contributions from different energy components in the global helix folding is made from indirect comparisons between the unfolded and folded equilibrated energies from the polypeptides in different trajectories. For a single trajectory, the polypeptide folding or unfolding process happens during mechanical equilibration. Thus, decomposing the energy to search for energy terms that favor the folded structures may not be accurate enough to make definitive conclusions.

Disrupted velocity (DIVE) search simulations of Ala13 clearly illustrate that its ground state conformation is a nearly perfect α-helix in water, different from the U-shaped 3_{10}-helix *in vacuo* (chapter 5). This result is obtained from two different GB/SA implicit solvent models. *In vacuo*, the U-shaped 3_{10}-helix, ideal 3_{10}-helix, α-helix and their combined conformations dominate at low potential energies, whereas in a solvent environment, the U-shaped 3_{10}-helix and the ideal 3_{10}-helix are not sampled. Thus, solvation raises the potential energies for the 3_{10}-helical conformations far above the α-helical portion(s) of the surface. Implicit solvent models also make the potential energy surface flatter and enhance barrier crossings and phase space sampling. Furthermore, solvation decreases the energy gap between the global minimum and some local minima from 3-5 kcal/mol *in vacuo* to 1-3 kcal/mol in implicit solvent. These local minima *in vacuo* can also become metastable states[34,35] in solvent. These simulation results imply a very important biological role for solvation: it smoothes potential energy surfaces so proteins may fold or unfold more easily or oscillate among several metastable conformational states.

DIVE simulations show several important aspects of the potential energy landscape of the Ala13 polypeptide. First, the relatively large energy gap between potential minima for structures with similar conformations can cause the initial structure to have a large impact on simulation results. This is a fact well-documented in the literature [32,33], and supported by our own conventional NVT simulations of single Ala13 *in vacuo* or in GB/SA implicit solvent models at 300 K. Second, the potential energy differences between minima representing members of a family with similar secondary structures may be very large (up to 100 kcal/mol), but the energy difference between minima representing different structural families is very small (a few kcal/mol). This observation implies serious challenges for MD or MC simulations designed to locate potential energy minima. The small energy gap between different potential energy minima can cause real difficulties in determining the global potential energy minimum. On the other hand, because NVT simulations sample <u>free energy</u> minima, they may sample only conformations with relatively high potential energies [45], but not locate true potential energy minima. In contrast to NVT simulations, the DIVE simulations can accurately sample the range of potential energy minima within 1-2 kcal/mol of the global minimum, at energies corresponding to temperatures near 0 K. Thus, the DIVE technique is an effective way to determine the global potential energy minimum and its conformation. It is a new global optimization protocol for predicting 3D protein structures from only sequence data.

6.6. Bibliography

(1) Kollman, P.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A. *The development/application of a \"minimalist\" organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data*, 1997; Vol. 3.

(2) Wang, J.; Cieplak, P.; Kollman, P. A. *Journal of Computational Chemistry* **2000**, *21*, 1049.

(3) Padmanabhan, S.; Marqusee, S.; Ridgeway, T.; Laue, T. M.; Baldwin, R. L. *Nature (London, United Kingdom)* **1990**, *344*, 268.

(4) Miick, S. M.; Martinez, G. V.; Fiori, W. R.; Todd, A. P.; Millhauser, G. L. *Nature (London, United Kingdom)* **1992**, *359*, 653.

(5) Merutka, G.; Lipton, W.; Shalongo, W.; Park, S. H.; Stellwagen, E. *Biochemistry* **1990**, *29*, 7511.

(6) Marqusee, S.; Robbins, V. H.; Baldwin, R. L. *Proceedings of the National Academy of Sciences of the United States of America* **1989**, *86*, 5286.

(7) Marqusee, S.; Baldwin, R. L. *Proceedings of the National Academy of Sciences of the United States of America* **1987**, *84*, 8898.

(8) Hudgins, R. R.; Mao, Y.; Ratner, M. A.; Jarrold, M. F. *Biophysical Journal* **1999**, *76*, 1591.

(9) Fiori, W. R.; Miick, S. M.; Millhauser, G. L. *Biochemistry* **1993**, *32*, 11957.

(10) Chakrabartty, A.; Schellman, J. A.; Baldwin, R. L. *Nature* **1991**, *351*, 586.

(11) Chakrabartty, A.; Kortemme, T.; Baldwin, R. L. *Protein Science* **1994**, *3*, 843.

(12) Son, H. S.; Hong, B. H.; Lee, C. W.; Yun, S.; Kim, K. S. *Journal of the American Chemical Society* **2001**, *123*, 514.

(13) Hansmann, U. H. E.; Okamoto, Y. *Journal of Chemical Physics* **1999**, *110*, 1267.

(14) Mitsutake, A.; Okamoto, Y. *Journal of Chemical Physics* **2000**, *112*, 10638.

(15) Daggett, V.; Levitt, M. *Journal of Molecular Biology* **1992**, *223*, 1121.

(16) Bertsch, R. A.; Vaidehi, N.; Chan, S. I.; Goddard, W. A., III. *Proteins: Structure, Function, and Genetics* **1998**, *33*, 343.

(17) Takano, M.; Yamato, T.; Higo, J.; Suyama, A.; Nagayama, K. *Journal of the American Chemical Society* **1999**, *121*, 605.

(18) Mitsutake, A.; Okamoto, Y. *Chem. Phys. Let.* **1999**, *309*, 95.

(19) Mortenson, P. N.; Wales, D. J. *Journal of Chemical Physics* **2001**, *114*, 6443.

(20) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *Journal of Computational Chemistry* **1986**, *7*, 230.

(21) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *Journal of the American Chemical Society* **1995**, *117*, 5179.

(22) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *Journal of the American Chemical Society* **1990**, *112*, 6127.

(23) Weiser, J.; Shenkin, P. S.; Still, W. C. *Journal of Computational Chemistry* **1999**, *20*, 217.

(24) Jayaram, B.; Sprous, D.; Beveridge, D. L. *Journal of Physical Chemistry B* **1998**, *102*, 9571.

(25) Sitkoff, D.; Sharp, K. A.; Honig, B. *Journal of Physical Chemistry* **1994**, *98*, 1978.

(26) Tsui, V.; Case, D. A. *Biopolymers* **2001**, *56*, 275.

(27) Jorgensen, W. L.; Tirado-Rives, J. *Journal of the American Chemical Society* **1988**, *110*, 1657.

(28) Tsui, V.; Case, D. A. *Journal of the American Chemical Society* **2000**, *122*, 2489.

(29) Jayaram, B.; Liu, Y.; Beveridge, D. L. *Journal of Chemical Physics* **1998**, *109*, 1465.

(30) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Clarendon Press: Oxford, 1987.

(31) Simmerling, C.; Elber, R.; Zhang, J. *Jerusalem Symposia on Quantum Chemistry and Biochemistry* **1995**, *27*, 241.

(32) Hudgins, R. R.; Jarrold, M. F. *Journal of Physical Chemistry B* **2000**, *104*, 2154.

(33) Hudgins, R. R.; Jarrold, M. F. *Journal of the American Chemical Society* **1999**, *121*, 3494.

(34) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598.

(35) Li, Z.; Scheraga, H. A. *Proceedings of the National Academy of Sciences of the United States of America* **1987**, *84*, 6611.

(36) Fiori, W. R.; Miick, S. M.; Millhauser, G. L. *Biochemistry* **1993**, *32*, 11957.

(37) Young, W. S.; Brooks, C. L., III. *Journal of Molecular Biology* **1996**, *259*, 560.

(38) Lyu, P. C.; Liff, M. I.; Marky, L. A.; Kallenbach, N. R. *Science* **1990**, *250*, 669.

(39) Wang, L.; O'Connell, T.; Tropsha, A.; Hermans, J. *Biopolymers* **1996**, *39*, 479.

(40) Yang, A. S.; Honig, B. *Journal of molecular biology* **1995**, *252*, 351.

(41) Braxenthaler, M.; Avbelj, F.; Moult, J. *Journal of Molecular Biology* **1995**, *250*, 239.

(42) Scholtz, J. M.; Marqusee, S.; Baldwin, R. L.; York, E. J.; Stewart, J. M.; Santoro, M.; Bolen, D. W. *Proceedings of the National Academy of Sciences of the United States of America* **1991**, *88*, 2854.

(43) Avbelj, F.; Fele, L. *Journal of Molecular Biology* **1998**, *279*, 665.

(44) Rose, G. D.; Gierasch, L. M.; Smith, J. A. *Advances in Protein Chemistry* **1985**, *37*, 1.

(45) Simmerling, C.; Fox, T.; Kollman, P. A. *Journal of the American Chemical Society* **1998**, *120*, 5771.

Chapter 7

Relaxation Simulation of a Multiple-Copy Region in Locally Enhanced Sampling Indicates Large Change of a Free Energy Surface through Mean-Field Approximations

7.1. Introduction

In 1991, Elber and Karplus [1] developed a multiple-copy molecular dynamics (MD) method, called locally enhanced sampling (LES), to hasten the diffusion of carbon monoxide in a simulation of a large myoglobin protein. This method creates a bundle of non-interacting copies of a small subsystem of primary interest and allows a larger subsystem, the bath, to interact with each copy of the subsystem. Each copied atom feels the same force that the corresponding real atom would feel, while an atom in the uncopied bath experiences the average of the forces contributed by the copied atoms (the mean field).

Due to this mean field approximation, LES and other related methods were considered to be able to reduce energy barriers on the potential energy surface (PES) [2]. In these simulations, varied conformations of the copied section of interest are usually obtained, and thus the energy barriers between the copied sub-system and bath is seen to be overcome compared to a conventional MD simulation, which can easily be trapped in a local energy minimum. Because the number of degrees of freedom in the LES system is greatly reduced compared to the situation where the bath atoms are also copied, using LES allows considerable savings in computer resources. In the last decade, the LES method has been widely used in a variety of optimization problems including cofactor-enzyme binding [3-5], non-equilibrium studies [6-8], free energy calculations [9,10], and global minimum searching [2,10,11].

Although LES and other optimization methods based on mean field theory show desirable computational advantage and great practical utility, they suffer limitations in the ability to model a realistic Newtonian dynamical process [12-16]. The trajectories generated with such methods do not always correspond to physically possible trajectories. In the simulations, several uncertainties are caused by approximating the average forces from the copies of the sub-system on the bath [17-20]. For example, the data obtained from LES violates the equipartition of energy theorem [17]. This violation causes the "temperature disparity problem" [17-19], which is a failure of the sub-system and bath temperatures to achieve the same equilibrium value if they are initiated differently. Our current view goes further. Since the free energy is a temperature-dependent property for a canonical ensemble, the "temperature disparity problem" can be restated as a severe "free energy problem", in which the free energy minima for the same conformations may be sampled very differently on the LES free energy surface (FES) and on the free energy surface of the real system.

A crucial point in the application of LES mean-field molecular dynamics for global optimization is the conservation of potential energy minima for the real system. The global or local energy minima on the potential energy surface (PES) of the real system correspond to global or local energy minima on the LES potential energy surface [20]. However, the native folded conformation of a protein at *in vivo* temperatures is the global free energy minimum rather than the global potential energy minimum [21]. The molecular dynamics of a real system at the temperature of interest samples the free energy surface but not the temperature-independent potential energy surface. Our previous simulations illustrate that the potential energy minima are usually not the free energy minima but are likely to be states of higher free energies (Chapter 3). In this work, we investigate the properties of the LES free energy surface and demonstrate that the global and local free energy minima usually do not correspond to the global and local free energy minima respectively of the real system.

In this chapter we present a different LES computer algorithm than AMBER LES [22] to provide the correct LES behavior for the interactions between pairs of atoms from the sub-system and the bath. In this algorithm, the whole protein is copied but bath copies of the same particles always have the same velocities and positions. Therefore, the copied bath particles can be considered to constitute a bath of "pseudo-single particles". This algorithm realizes the mean-field method in two steps. First, each replica executes an independent canonical ensemble simulation (NVT) so pairs of atoms from the same region (sub-system or bath) experience the correct interactions. Second, the assignment of the same averaged forces for each identical particle from sub-bath copies provides the mean field interactions of bath particles from the copied sub-system. Because of this property, we can run relaxation simulations after LES, defined by removing the restriction that identical bath particles experience average forces due to atoms of the copied subsystem. Thus, after relaxation, each replica subsystem plus bath, executes an NVT simulation independent of all other replicas. At the initial step, if the copied subsystem is extended to the whole system (zero bath particles), the LES mean field method becomes a divergent path (DIP) search technique, in which the simulations start with several independent trajectories of proteins at the same initial temperature but different velocities (Chapter 3). This causes the independent simulated proteins to evolve along different trajectories from a single initial configuration.

We selected the 13-residue peptide of alanine (Ala13) as a simple test polypeptide for LES mean field and relaxation simulations to investigate the change of the free energy surface for LES mean field MD compared to NVT molecular dynamics. The GB1/SA [23] implicit solvent model with the AMBER 99 force field [24] is used here. We have already investigated Ala13 by using our new MD techniques and found that the global minimum potential energy conformation at 0 K is an α-helix. The global minimum free energy conformation with implicit solvent at 300 K is identified as clusters of 3_{10}/α-helices (Chapter 6). The relaxation simulations of copied regions for Ala13 indicate a large change in the free energy minima of the same conformations because of the mean-field approximation.

7.2. Simulation methodology

7.2.1. LES Program

In the locally enhanced sampling (LES) mean field MD program, we use a different format than that in AMBER [22]. In the AMBER LES format, if the copied part has M atoms and C copies, and the bath has N atoms, its input topology file includes (CM+N) atoms. The algorithm to obtain the LES forces follows. All of the force field parameters in the copied region have been scaled by 1/C for C copies (in the topology input file). This provides the correct LES behavior for the interactions between pairs of atoms in the bath, because of no force field scaling in the bath region, and for the interactions between pairs of atoms from the copied part and the bath part, respectively [22]. However, it is wrong for the interactions between pairs of atoms in the same copy. Therefore, the SanderLES program dynamically scales up these interactions in the same copy by a factor of C as the program runs. In this way, AMBER corrects the LES behavior for all interactions.

In our LES format, if the copied part has M atoms, the bath part has N atoms and C copies are made, our input topology file includes the original (M+N) atoms. But the program makes C copies of these (M+N) atoms in memory. We do not scale the force field parameters in the copied region. It is very simple to provide the correct LES behavior by averaging the forces in the C copies of the bath atoms. In fact, the time-consuming non-bonded force calculation in the bath part is performed only one time for the first copy and then distributed to other copies, because bath particles are always at same location; geometric forces(bond, angle and torsional) are calculated individually and the average force is distributed. Although this format uses more memory and a little more CPU time during the simulations than the AMBER LES format, it has two advantages. It makes possible dynamic migration — automatic variation of the copied region and bath during the simulations. This "migration technique" was our original impetus to develop the new format. In addition, it makes relaxation possible. The relaxation technique is defined as removing the restraint of using average forces for identical bath atoms after a certain simulation time. This makes the LES mean-field simulations convert to independent canonical ensemble simulations of several trajectories. In contrast to the binary collision modified LES (cLES) [7,18] and ensembles extracted from atomic coordinate transformations (the EXACT approximation) [25-27], this relaxation process provides a simple way to completely erase incorrect *virials* and energy partitioning introduced by the LES approximation.

7.2.2. Computational Details

In the LES simulations, three copied regions were located at the N-terminus, the middle part, and the C-terminus, as shown in Figure 7.2.2.1. At the N-terminus, 34 atoms were included in the copied region while the other 108 atoms were in the bath. At the C-terminus, 28 atoms were contained in the copied region, and the other 114 atoms were in the bath. In the middle part, three alanine residues constitute the copied region, and the remaining 112 atoms form the bath. Carbon-carbon bonds in the backbone were selected as the separations between the copied and bath regions because carbon-carbon bonds should have the largest flexibility. In the following relaxation simulations, the restriction of using average forces on the bath atoms from atoms in the copied region was removed. Then, the pseudo-single bath was changed into several copies. These bath copies combined with the original copied regions of the interesting subsystem to generate several independent NVT simulations of different trajectories.

CH3-CO-[NH-CH(CH$_3$)-CO]$_2$-NH-CH(CH$_3$)-CO-[NH-CH(CH$_3$)-CO]$_{10}$-NH-CH$_3$
N-terminal LES region

Middle LES region
CH3-CO-[NH-CH(CH$_3$)-CO]$_4$-NH-CH(CH$_3$)-**CO-[NH-CH(CH$_3$)-CO]$_2$-NH-CH(CH$_3$)**
-CO-[NH-CH(CH$_3$)-CO]$_5$-NH-CH$_3$

CH3-CO-[NH-CH(CH$_3$)-CO]$_{10}$-NH-CH(CH$_3$)-**CO-[NH-CH(CH$_3$)-CO]$_2$-NH-CH$_3$**
C-terminal LES region

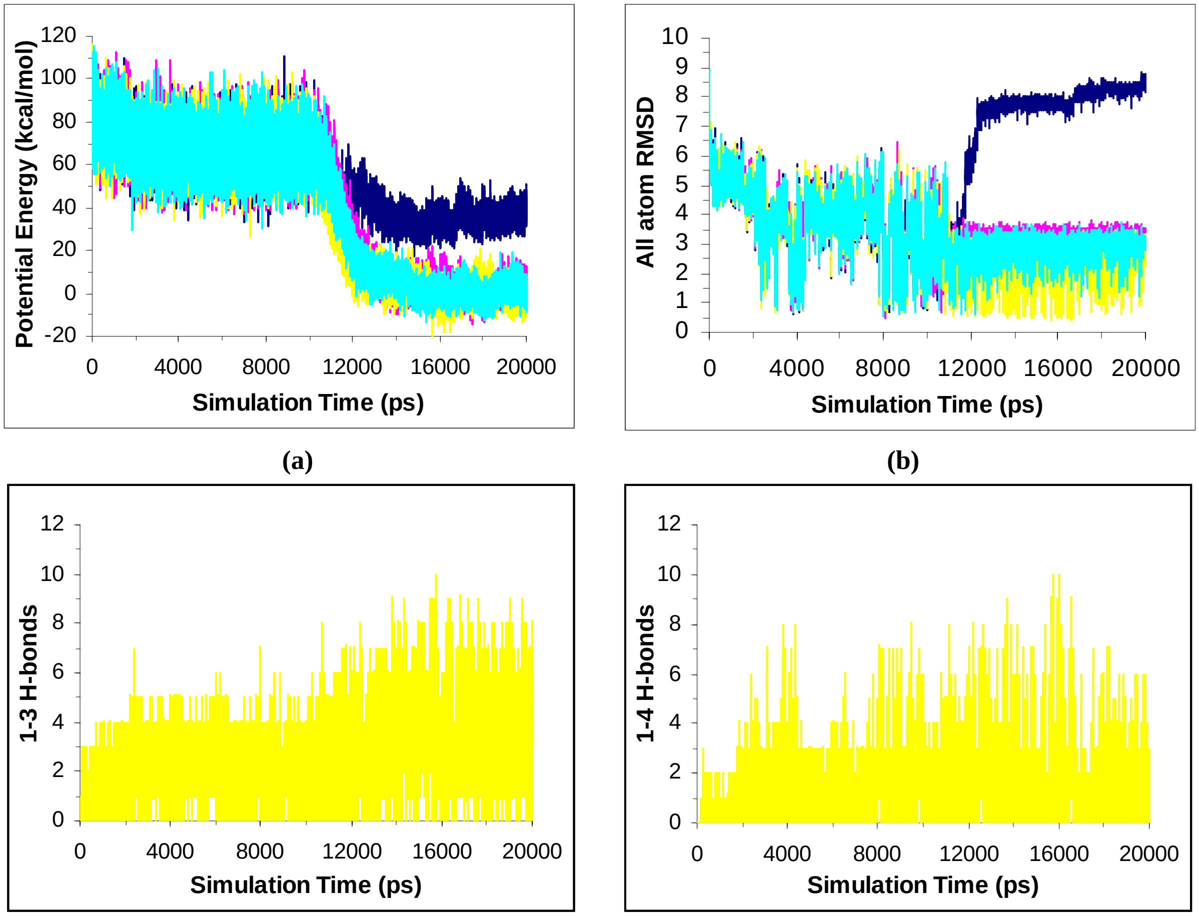
Figure 7.2.2.1. The copied region at the N-terminus (34 bolded atoms), middle part (30 bolded atoms), and C-terminus respectively (28 bolded atoms).

We used a time step of 1 fs but the trajectory data (energies and coordinates) were collected at every 1000 steps (1 ps intervals) in all simulations. The total simulation time for each run was 10 ns for each independent DIP simulation. 20 ns were split evenly between LES and subsequent relaxation simulations (10 ns for each).

7.3. Simulation results

First, we briefly describe the results from the six-trajectory DIP simulations of Ala13 at 300 K in GB/SA implicit solvent using the AMBER99 force field. The simulations indicate that helical clusters of $3_{10}/\alpha$-helices (mainly 3_{10} helices with ~ 3 Å RMSD from an ideal α-helix) are the most populated conformations at the lowest equilibrated average potential energies near 0 kcal/mol. These conformations we consider as the global free energy minimum for Ala13 at room temperature with implicit solvation. A detailed analysis can be found in the Chapter 6 (section 6.3.1).

In LES and the following relaxation simulations, the temperature histories are independent of the replica number, copied region, and simulation protocol, but the histories of other ensemble properties display large differences. Fig. 7.3.1a - b show the time series of the potential energy and all-atom RMSD from an ideal α-helix from the C-terminus LES and relaxation simulations. During the LES simulations, the polypeptide forms helical conformations (RMSDs < 3.5 Å, Chapter 6) located on the potential energy surface above 30 kcal/mol and corresponding to the global free energy minimum. During the relaxation process, three polypeptides maintain folded helical conformations while one polypeptide unfolds into extended conformations (RMSDs ~ 8 Å). However, the potential energies for these polypeptides all decrease. Three polypeptides forming equilibrated helical conformations have the lowest average potential energies near 0 kcal/mol, which are consistent with the DIP simulations. Fig. 7.3.1c - d show the histories of 1–3 H-bonds and 1–4 H-bonds of the replica in yellow, confirming that the polypeptide oscillates between 3_{10}-helical and α-helical conformations. A further 10 ns simulation of the copied part in the LES simulations did not decrease the potential energies of the helical conformations sampled (data not shown). The large change of the sampled potential energy between two different molecular dynamics processes implies that the temperature-dependent free energy surface changes substantially in the LES simulations compared to conventional NVT simulations (each polypeptide in a given trajectory executes conventional NVT simulations in DIP simulations). On the other hand, the transitions between folded and partially unfolded structures are more frequently observed during the LES simulations than during the conventional NVT simulations. It should be noted, however, that we believe that the indirect coupling between the copies through the common bath, not a reduction in energy barriers, promotes conformational transitions of the equilibrated whole system in LES simulations. In section IV, we will analyze LES molecular dynamics in detail.

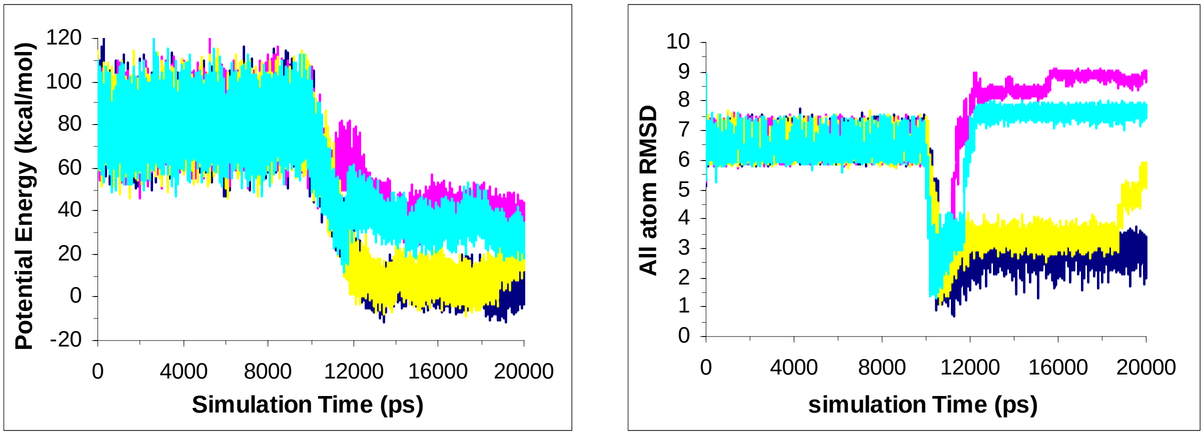


(a) **(b)**

(c) **(d)**

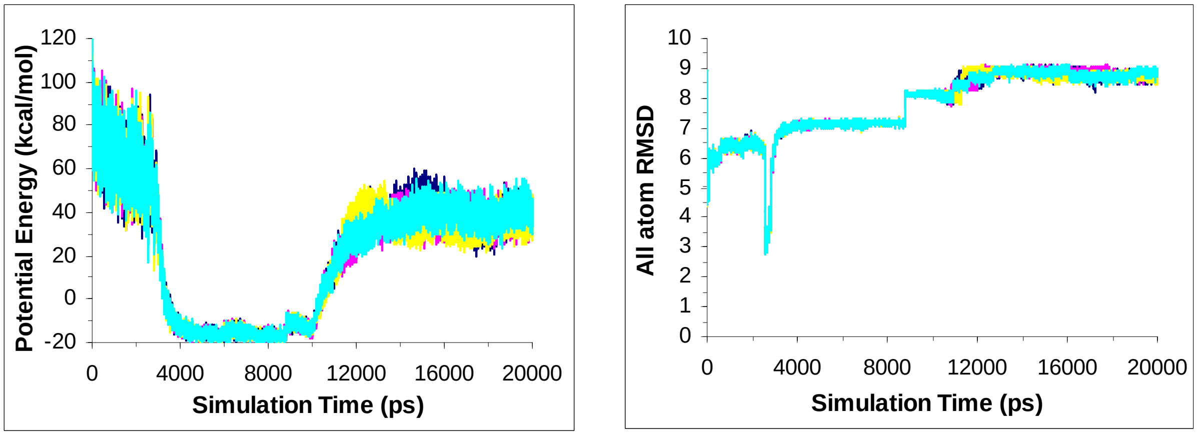
Figure 7.3.1. Time series of various properties in four-copy LES and relaxation simulations of a fully extended Ala13 in GB/SA implicit solvent at 300 K. 0 - 10000 ps is the history of LES simulations, where only the "C-terminal LES region" was replicated. 10000 – 20000 ps is the history of NVT relaxation simulations. **(a)** Potential energy. **(b)** RMSD between all atoms of the calculated structure and an ideal α-helix (The RMSD between a 3_{10}-helix and an ideal α-helix is ~ 3 Å). **(c)** 1-3 hydrogen bonds. **(d)** 1-4 hydrogen bonds. A hydrogen bond is identified when the distance between two heavy atoms X is below 3.3 Å and the X-H ··· X angle differs from 180° by less than 20.0°. An ideal α-helix has eleven 1-4 hydrogen bonds and a 3_{10}-helix has twelve 1-3 hydrogen bonds.

Figures 7.3.2 – 7.3.3 display the time series of potential energy and all-atom RMSD from an ideal α-helix for the LES and relaxation simulations where the "N-terminal LES region" or "middle LES region" was replicated, respectively. While the N-terminus LES simulations (Fig. 7.3.2) cannot make the polypeptide form completely helical conformations, the relaxation process helps some polypeptides search for the global free energy minimum (all-atom RMSD < 3.5 Å) while others unfold into extended conformations (all-atom RMSD > 7 Å). What is more interesting, in the middle region LES simulations (Fig. 7.3.3), a conformation with ~7.0 Å RMSD has the lowest ensemble average potential energies sampled — approximately –15 kcal/mol. At first sight, the results seem very strange but they are reasonable. Usually, at nonzero temperatures, the free energy minima are not the potential energy minima and they are usually located on higher energy parts of the potential energy surface. The ground state conformation at 0 K is the α-helix with the global potential energy minimum below –65.0 kcal/mol (Chapter 6, section 6.4.1). At 300 K, the global free energy minimum corresponds to a potential energy near 0 kcal/mol (Chapter 6, section 6.3.1), far above the global potential energy minimum. Therefore, the LES simulations change the potential energies of the free energy minima and can make the local potential energy minima into the global free energy minimum conformation.



(a) (b)

Figure 7.3.2. Time series of the potential energy and all-atom RMSD from an ideal α-helix in four-copy LES and relaxation simulations of a fully extended Ala13 in a GB/SA implicit solvent at 300 K. 0 - 10000 ps is the history of LES simulations, where only the "N-terminal LES region" of Fig. 7.2.2.1 was replicated. 10000 – 20000 ps is the history of NVT relaxation simulations. **(a)** Potential energy. **(b)** RMSD between all atoms of the calculated structure and an ideal α-helix.



(a) (b)

Figure 7.3.3. Time series of the potential energy and all-atom RMSD from an ideal α-helix in four-copy LES and relaxation simulations of a fully extended Ala13 in a GB/SA implicit solvent at 300 K. 0 - 10000 ps is the history of LES simulations, where only the "middle LES region" of Fig. 7.2.2.1 was replicated. 10000 – 20000 ps is the history of NVT relaxation simulations. **(a)** Potential energy. **(b)** RMSD between all atoms of the calculated structure and an ideal α-helix.

LES can sample diverse regions of phase space starting from the fully extended structure depending upon the set of regions that are copied. In addition, diverse phase space sampling is also observed in LES simulations starting from an ideal α-helix. Here we will not show data from LES and relaxation simulations for three segments of an ideal α-helical Ala13 in a GB/SA implicit solvent at 300 K. The same results are achieved in LES simulations starting from the α-helix as those in the extended LES simulations. For example, in the LES simulations, where the "C-terminal LES region" or "N-terminal LES region" of Fig. 7.2.2.1 was replicated, global free energy minimum conformations (helical conformations) in NVT simulations become highly energetic local minima in the LES simulations. It was also observed that the transition between folded and partially unfolded structures happens frequently during LES simulations. In the LES simulations, where only the "middle LES region" of Fig. 7.2.2.1 was replicated, high energy structures in NVT simulations have the lowest equilibrated energies in the LES simulations. Therefore, it is commonly observed that the free energy minimum states in the LES simulations do not correspond to those in conventional NVT simulations. For the canonical ensemble simulations near *in vivo* temperatures, the match of the potential energy minima in NVT and LES is not sufficient for LES simulations to replicate the results of conventional MD.

At 0 K, the free energy surface is identical to the potential energy surface. An ideal MD simulation with zero kinetic energy should always find in the (potential) free energy minima. Therefore, we think that near 0 K the potential energies sampled in LES and conventional NVT simulations should not show much difference since they should converge to the same single potential energy and free energy minimum states. Our extended LES and relaxation simulations near 0 K support this assumption (data not shown here). The trajectories from the LES and subsequent relaxation NVT simulations are almost identical. Both are trapped in the potential energy minima of their respective initial conformational energy wells. Those simulations indicate that the potential energy minimum for the fully extended structure is approximately – 23 kcal/mol and for the ideal α-helix is approximately – 68 kcal/mol. These results are in good agreement with those from disrupted velocity search (DIVE) simulations designed to search for diverse potential energy minima in a realistic simulation time (Chapter 6, section 6.4.1).

7.4. Molecular dynamics using the LES method.

It is generally thought that the more frequently observed transitions between the folded and unfolded structures in LES compared to conventional NVT MD simulations is caused by the reduction of energy barriers between the copied region and bath [2,7,18,28]. We think that this reduction of energy barriers between two different regions is questionable due to the function of the averaging operation in the LES algorithm. While the interactions inside the same regions in LES simulations are not affected compared to those in NVT simulations, the cross interactions between the copied regions and bath are averaged at each step. This force averaging cannot increase the overall cross interactions to help overcome the energy barriers between these two regions. For individual replicas, the situation is different. For some replicas which have a small cross interaction, crossing energy barriers is really enhanced due to the increased averaged forces. Other replicas which have a large cross interaction have a reduced crossing capability. However, the energy barriers for any individual replica remain unchanged, because we consider the "real" potential energy surface for individual replicas. This view is also in good agreement with the fact that the collision modified LES seems to show no ability to lower energy barriers [7,18] compared to conventional MD simulations.

It is commonly believed that LES reduces energy barriers, an idea that arises from a plausible, logical deduction [2] rather than from a strict mathematical derivation. For example, Roitberg, *et al.* [2] built a model state as a premise to consider the transitions happening in the LES simulations. This model state described two unrealistic situations during the LES trajectory: only two minima existed, or all other minima always had lower energies than the transition states connecting these two minima of interest. Each individual copy can alternately undergo transitions from the unfolded conformation to the folded conformation. Therefore at any transition state, Roitberg, *et al.* argue that the calculated barrier height will always be reduced on the "effective" averaged potential energy surface, which is built from a sum of "real" potential energies calculated from different copies and multiplied by a constant normalization factor (usually 1/C, C is the copied number) [2].

In reality, many different minima exist on the potential energy surface and most of them are located at higher energies than that those of the transition states leading to the global minimum. Different copies may be distributed into many different energy states with different conformations. As a result, the barrier height on the "effective" potential energy surface is averaged from the barriers of the transition states sampled by individual copies. We therefore think that LES is likely to reduce the energy barriers in the trajectories of some replicas but increase the energy barriers for others, rather than to reduce the overall energy barriers between the copied and bath regions.

The fundamental mechanism by which LES enhances phase space sampling in contrast to conventional NVT is caused by the different initial velocities or temperatures for the atoms in the copied region. Our DIP simulations imply that the independent system can evolve into different trajectories due to different initial velocity directions. Similarly, in the LES method, phase space sampling is enhanced because the copies of the partial polypeptide follow different trajectories. These different trajectories undergo different equilibration processes, initiated from different velocity or temperature conditions. This explanation is more reasonable than the energy barrier reduction if we consider an extreme situation in LES simulations. If the atoms from the copied region are starting with the same initial velocities and coordinates, the LES simulation is identical to a single-copy, conventional NVT simulation and phase space sampling cannot be enhanced.

DIP simulations made us realize three limitations of conventional NVT simulations: potential energy traps, free energy traps, and kinetic traps (Chapter 3). Potential energy traps occur when a system becomes trapped in a local potential energy well where its kinetic energy is not large enough to overcome the high potential energy barriers between energy wells. Free energy traps occurs at thermal equilibrium when simulations sample portions of phase space corresponding only a certain range of potential energies (which also correspond to the system's kinetic energies). We call these regions of phase space thermodynamically accessible regions. Kinetic traps happen when the simulation evolves into a large number of nearly equivalent substates on a small part of the thermodynamically accessible regions at the simulation temperature. It would take an unrealistic amount of time for such a simulation to reach other parts of the thermodynamically accessible regions after it finishes sampling a large number of substates. In other words, kinetic traps make the transition from one free energy minimum to another minimum very time-consuming in conventional MD simulations. In contrast, those parts of the thermodynamically accessible regions corresponding to other free energy minima can be easily reached from a different trajectory, or starting from a different initial structure.

We can further clarify the change of a free energy surface and enhanced transitions in LES simulations by analyzing the sampling of states on the PES. If we consider the potential energies for each single copy rather than the average effective potential energies for the combined LES configuration, the potential energy surface for each copy in LES simulations is the same as that in NVT simulations. In conventional NVT simulations, once kinetic trapping occurs, simulations will evolve into a large number of equivalent substates at potential energy levels corresponding to the ensemble temperature. In an LES simulation, the simulation involves a larger system. This larger system, including copies of subsystem of interested and one bath, show a further difference from the original system because of indirect coupling between the copies through the common bath [18]. These differences make the trajectories of an LES simulation in kinetic traps able to sample a larger number of nearly equivalent substates. As a result, the transitions between two different potential wells are much more likely to happen during LES molecular dynamics. On the other hand, because the probability distribution from a conventional NVT simulation is also altered from the indirect coupling between the copies, the locations of the potential energy wells are different in LES and NVT molecular dynamics.

7.5. Discussion and Summary

A new computer algorithm was used to implement LES molecular dynamics. Instead of scaling the force field in the input files, we used copies of the whole system. Only the “locally enhanced region” is differentiated by using individual coordinates for atom copies. “Copies” of bath atoms always have identical velocities and coordinates during the simulations. By using this format, we also used a relaxation algorithm that removes the restraint of identical velocities and coordinates of identical bath atoms in each copy and thus converted the LES mean-field simulations into independent canonical ensemble simulations of several polypeptides.

The LES and subsequent relaxation simulations of Ala13 in GB/SA implicit solvent at 300 K display very different potential energy histories, even though these different simulations may evolve into similar trajectories. This observation indicates a large change of the free energy surface in LES simulations compared to the more physical NVT simulations. At nonzero temperatures, the canonical ensemble simulations do not sample the valleys but sample higher energy conformations on the potential energy surface. These configurations usually correspond to free energy minima. The LES simulations change the location of these conformations on the potential energy surface, indicating the migration of the free energy minima on the PES and a change of the free energy surface at room temperature. However, near 0 K, the LES simulations do not show much difference from NVT simulations and they both show physically reasonable descent into potential energy minima.

In contrast to conventional NVT simulation, LES molecular dynamics can enhance phase space sampling in two respects: (1) an improved statistical sampling of alternative conformations for the copied region(s). (2) an increased frequency of conformational transitions [28]. It should be noted, however, that the enhanced phase space sampling is not due to energy barrier reduction between the copied and bath regions but is caused by the different initial velocities or temperatures of the copied region. While the gain in sampling of diverse conformations for the copied region can be attributed to their different trajectories, the more frequent conformational change of the entire system arises from larger fluctuation in potential energies of individual copies, even after they are in kinetic traps. Considering the simulation of the single copy (several copies have the same bath atoms) in LES, the averaging interaction energies acting on the bath due to the copies can help some replicas to cross energy barriers but impede other replicas. In a phase space with a large number of minima, both LES and conventional NVT molecular dynamics are usually limited to a narrow region of phase space due to kinetic traps. However, LES can span a larger range of potential energies and more easily oscillates between two energy wells, compared to NVT simulations.

The violation of the equipartition of energy theorem not only causes the "temperature disparity problem" in LES but also a "free energy problem". It indicates that LES simulations should be used very carefully in protein optimization near *in vivo* temperatures. In general, copying an interior segment of a polypeptide is not recommended for global minimum searching. Hierarchical LES [28] may be useful to generate many diverse equilibrated conformations. However, LES relaxation simulations should be further applied to investigation minimum energy conformations near *in vivo* temperatures from individual potential energy histories. The relaxation simulations can not only solve the free energy problem of reversing the energies of structures obtained from LES simulations but also generate a reliable probability distribution of different conformations. Energy minimization can be used to find the true potential energy minimum, but cannot help to generate a reliable probability distribution. For a complex system including a large protein and a small molecule, such as substrate-ligand [1,7,18] or cofactor-enzyme [3-5] complexes, LES may be a good approximation for building a reference distribution or diffusion pathway of small molecules [18,8].

7.6. Bibliography

(1) Elber, R.; Karplus, M. *Journal of the American Chemical Society* **1990**, *112*, 9161.

(2) Roitberg, A.; Elber, R. *Journal of Chemical Physics* **1991**, *95*, 9277.

(3) Miranker, A.; Karplus, M. *Proteins* **1991**, *11*, 29.

(4) Carlson, H. A.; Masukawa, K. M.; McCammon, J. A. *Journal of Physical Chemistry A* **1999**, *103*, 10213.

(5) Caflisch, A.; Miranker, A.; Karplus, M. *Journal of medicinal chemistry* **1993**, *36*, 2142.

(6) Quillin, M. L.; Li, T.; Olson, J. S.; Phillips, G. N., Jr.; Dou, Y.; Ikeda-Saito, M.; Regan, R.; Carlson, M.; Gibson, Q. H.; Li, H. *Journal of molecular biology* **1995**, *245*, 416.

(7) Ulitsky, A.; Elber, R. *Journal of Physical Chemistry* **1994**, *98*, 1034.

(8) Czerminski, R.; Elber, R. *Proteins: Structure, Function, and Genetics* **1991**, *10*, 70.

(9) Verkhivker, G.; Elber, R.; Nowak, W. *Journal of Chemical Physics* **1992**, *97*, 7838.

(10) Simmerling, C.; Fox, T.; Kollman, P. A. *Journal of the American Chemical Society* **1998**, *120*, 5771.

(11) Simmerling, C.; Lee, M. R.; Ortiz, A. R.; Kolinski, A.; Skolnick, J.; Kollman, P. A. *Journal of the American Chemical Society* **2000**, *122*, 8392.

(12) Huber, G. A.; McCammon, J. A. *Physical Review E: Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **1997**, *55*, 4822.

(13) Huber, T.; van Gunsteren, W. F. *Journal of Physical Chemistry A* **1998**, *102*, 5937.

(14) Huber, T.; Torda, A. E.; van Gunsteren, W. F. *Biopolymers* **1996**, *39*, 103.

(15) Huber, T.; Torda, A. E.; van Gunsteren, W. F. *Journal of Physical Chemistry A* **1997**, *101*, 5926.

(16) Maranas, C. D.; Floudas, C. A. *Journal of Chemical Physics* **1994**, *100*, 1247.

(17) Straub, J. E.; Karplus, M. *Journal of Chemical Physics* **1991**, *94*, 6737.

(18) Ulitsky, A.; Elber, R. *Journal of Chemical Physics* **1993**, *98*, 3380.

(19) Zheng, W.-M.; Zheng, Q. *Journal of Chemical Physics* **1997**, *106*, 1191.

(20) Stultz, C. M.; Karplus, M. *Journal of Chemical Physics* **1998**, *109*, 8809.

(21) Anfinsen, C. B. *Science* **1973**, *181*, 223.

(22) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Ferguson, D. M.; Radmer, R. J.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. AMBER 5; University of California: San Franciso, 1997.

(23) Tsui, V.; Case, D. A. *Biopolymers* **2001**, *56*, 275.

(24) Wang, J.; Cieplak, P.; Kollman, P. A. *Journal of Computational Chemistry* **2000**, *21*, 1049.

(25) Hixson, C. A.; Wheeler, R. A. *Chem.Phys. Lett.* **2004**, *386*, 330.

(26) Hixson, C. A.; Wheeler, R. A. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics* **2001**, *64*, 026701/1.

(27) Hixson, C. A.; Chen, J.; Huang, Z.; Wheeler, R. A. *Journal of Molecular Graphics & Modelling* **2004**, *22*, 349.

(28) Hornak, V.; Simmerling, C. *Proteins: Structure, Function, and Genetics* **2003**, *51*, 577.

Chapter 8

Molecular Dynamics Simulations of an Amphiphilic Octadecapeptide in GB/SA Implicit Solvent

8.1. Introduction

Many conventional NVT simulations fail to find a global potential energy minimum, and it is generally accepted that this due to simulations at low temperatures becoming trapped in one of myriad local minimum-energy states [1-10]. However, this explanation must be re-examined in light of the difference between global and local minima on the free energy surface (FES) and on the potential energy surface (PES) in canonical ensemble simulations. The quasiergodicity problem [11] is caused by local free energy minimum traps [12] , not local potential energy minimum traps, but the PES [11,13] is most often used to describe the local minimum traps. Several techniques such as Locally Enhanced Sampling (LES) [5,14-16] have recently been developed, taking into account that potential energy barriers, not free energy barriers, are the limiting factor. Because the temperature-independent PES, not the temperature-dependent FES, represents the conformational energy landscape, we consider local minimum traps as potential energy traps. Potential energy traps describe the situation where a simulation becomes trapped in a local potential energy well, and its kinetic energy is not large enough to overcome the high potential energy barrier to move the system to another minimum on the potential energy surface. This shows that the potential energy trap is not the only reason for the quasiergodicity problem in time-limited MD simulations.

Other factors, such as free energy traps and kinetic traps, also affect simulations. Free energy traps occur at thermal equilibrium, reflected in small fluctuations in kinetic energy at constant temperature, so that MD simulations will only sample certain ranges of potential energies on a realistic simulation time. Regions of phase space sampled at thermal equilibrium are termed thermodynamically accessible regions. Similarly, kinetic traps happen at mechanical equilibrium, reflected by small fluctuations in potential energy when the velocity (vector) is in equilibrium. In this case, simulations are further limited to a local search for energy states with similar conformations. These two factors combined mean that realistic simulations usually sample only a very narrow range of phase space. Potential energy traps may not exist, but the simulations will still often fail to pass over higher potential energy barriers or fail to sample lower potential energy minima. Practically, this produces a biased phase space trajectory, which strongly depends on the initial coordinates. To reduce simulation dependence on initial coordinates, many workers start simulations from different conformations. The corresponding technique of using different initial energies, and therefore different momenta, for each simulation is similar in spirit to starting from different initial conformations. The work described here disrupts the long-scale equilibria associated with these traps in MD simulations, alleviates the biased phase space sampling problem and helps in modeling both folding and unfolding processes during short simulations.

According to Anfinsen's thermodynamic hypothesis [17], the native folded conformation of a protein is the global free energy minimum structure at *in vivo* temperatures. Unfortunately, the global free energy minimum is difficult to determine, as is the free energy surface, in both MD and MC simulations. In contrast, the potential energy surface (PES) is relatively easy to define as a function of the atomic coordinates of the system [18] and in addition is temperature independent [19]. In most cases, it is assumed that the global free energy minimum conformation corresponds to the conformation of lowest potential energy sampled during the trajectory. This assumption does not necessarily hold true for MD, in light of the high-energy barriers and long-scale equilibrium that restrict conformational transitions, bias phase space sampling, and trap the simulation in local energy minima. Even though the global minimum free energy structure at *in vivo* temperatures may not correspond to the global potential energy minimum, the global free energy minimum is usually either the global minimum or a very low local minimum on the potential energy surface [18]. Therefore, maps of these low potential energy minima and their conformations can be quite valuable. Many minimization algorithms based on Monte Carlo (MC) simulations have been developed recently to create these maps [18,20-23]. The disrupted velocity (DIVE) search simulation technique can provide an alternative, simpler way to identify global minima. Further, by starting with global or locally optimized conformations, divergent path (DIP) search simulations can be used to carry out global relaxation dynamics, and identify the lowest energy conformation near physiological temperatures.

This chapter describes our molecular dynamics simulations of a synthetic, amphiphilic octadecapeptide peptide F (EQLLKALEFLLKELLEKL)[24] in a GB/SA implicit solvent model for water. This *de novo* designed polypeptide has a high apolar-polar amino acid ratio and can self-associate into hexamers in aqueous solution (the helical ribbon in Fig. 8.1.1). First, the results of using the DIVE technique to investigate potential energy landscapes and myriad conformations of the polypeptide are presented.

During the DIVE simulations, both the global potential energy minimum conformation and the experimentally determined X-ray structure are located quickly (in several nanoseconds). Next, we give a series of DIP simulations of peptide F, conducted at 300 K from several different starting conformations. The global free energy minimum conformation at room temperature is identified in an aggregate simulation time of 360 ns. The simulation results indicate that both global minimum potential energy and free energy structures, sampled by molecular dynamics for an isolated polypeptide with implicit water solvation, may deviate from the experimental X-ray structures. However, the family of structures resembling the helical X-ray structure has potential energies very close to those of the global potential and free energy minima. The X-ray structure in fact displays three salt bridges between polypeptides, which serve to stabilize the helix in the crystal.

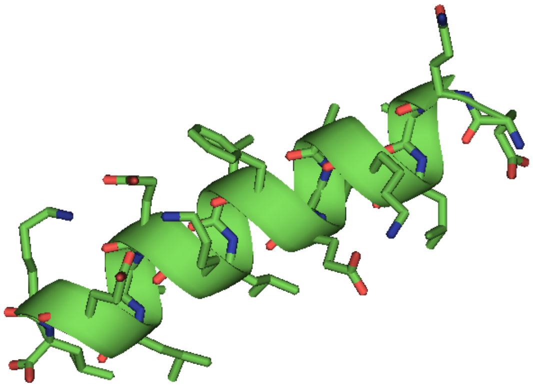


Figure 8.1.1. Native X-ray structure of peptide F (PDB 1PEF). The image was generated using PyMOL (http://pymol.sourceforge.net/).

8.2. Simulation methodology

8.2.1. Algorithms and force field

The simulation algorithm used was that described in Chapter 3 for the DIP strategy and in Chapter 4 for the DIVE protocol. The AMBER99 force field [25] was used, and the GB1/SA[26] implicit model was used for the solvent environment. In GB1, the water dielectric constant of $\varepsilon_w = 78.5$ was used, and the dielectric constant inside the protein was set to $\varepsilon_p = 1$. These parameters follow the standard values for GB1/SA implicit solvent.

8.2.2. Computational details

The AMBER 8 software package [27] was used to generate the initial coordinates for the extended structure of the 18-residue peptide F. The Molecular Modeling Toolkit [28] was used to convert input files from the AMBER format to that required by our programs. We used the velocity-Verlet algorithm [29-31] to integrate the equations of motion and the SHAKE algorithm [32] to constrain the covalent bond distances involving hydrogen. At each step, the translational motion of the center of mass of the entire system was removed. The Nosé-Hoover Chain method [33] was used to control the temperature in conventional NVT simulations. A preliminary, conventional NVT simulation was performed for the initially extended peptide F in *vacuo* at 10 K in 5 ps with a time step of 0.01 fs, because the original extended structure disintegrated when using a time step of 1 fs. After this pre-simulation, the extended conformation remained intact.

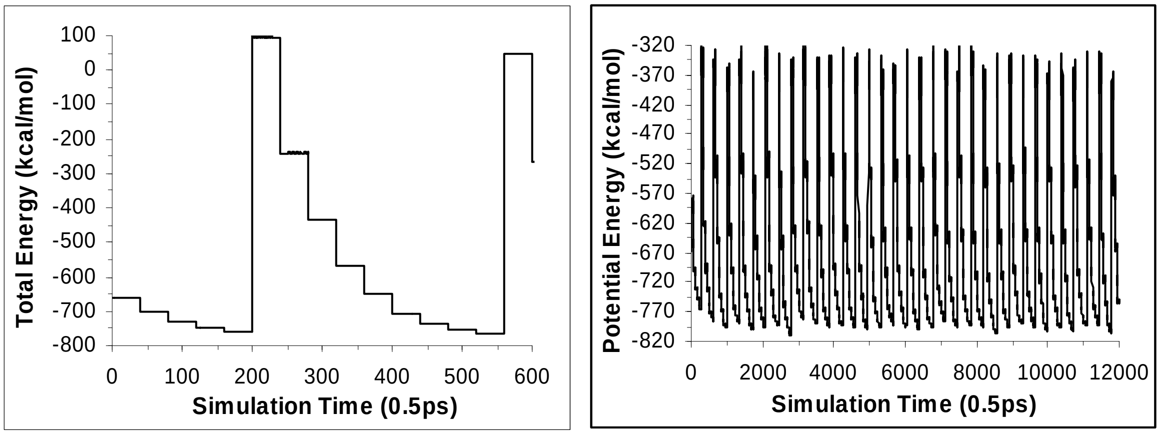
For the DIVE simulations, we simulated six independent trajectories of peptide F simultaneously with initial temperatures of 10 K, 50 K, 100 K, 300 K, 600 K, and 1000 K. The scaling parameter for cooling was 0.25 and the scaling factor for heating was calculated from $\sigma = T_{target}/T$, where T_{target} is the target temperature and T is the temperature at that perturbation step. The threshold temperature for heating and cooling was 10 K. Thus, during the simulations, each polypeptide was cooled down to ¼ of its temperature whenever it rose above the threshold temperature. Once the temperature was below 10 K, the polypeptide was heated back to T_{target}. T_{target} was selected to be 1000 K. During these simulations, velocity reassignment occurred every 20 ps. A time step of 1 fs was used and the trajectory data (energies and coordinates) were collected at 0.5 ps intervals (500 steps). The simulation time for each trajectory was 6 ns and the aggregate time for the DIVE simulations of six trajectories was 36 ns.

To run DIP simulations, six trajectories of peptide F were simulated simultaneously, and each polypeptide was simulated independently at a constant temperature for every step. Six diverse conformations were selected as initial structures for DIP simulations. These different conformations were the global potential energy minimum conformation, α-helix, V-shaped α-helix, a mixed helical intermediate, a coiled structure, and an extended structure. Except for the extended structure, these were the potential energy minimum conformations as sampled from the DIVE simulations. Each independent polypeptide was assigned the same initial structure and a temperature of 300 K, but corresponding atoms in each polypeptide had different directions for their velocities. Because each initial structure was modeled by six different trajectories, thirty-six trajectories in total were obtained in all DIP simulations. We also used a time step of 1fs and collected the trajectory data every 500 time steps. The simulation time for each DIP simulation was 10 ns and the aggregate time for 36 trajectories was therefore 360 ns. We used the PTRAJ program from the AMBER 8 package [27], MMTSB tool set [34], and custom programs to analyze the coordinate data.

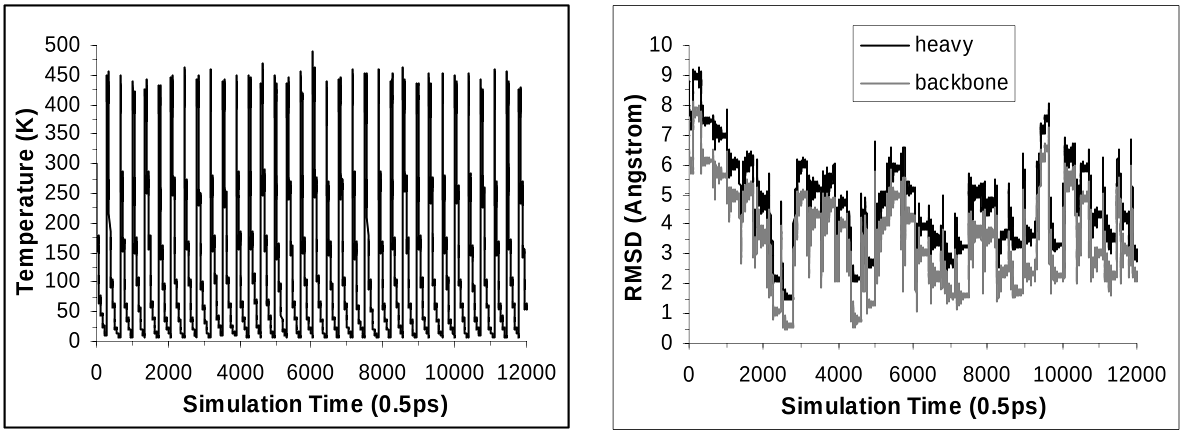
Conformational clustering was conducted on the snapshots sampled from the simulations. A step-wise optimizing fixed radius clustering algorithm [34-37] was used to sort the simulation conformations into different clusters. This algorithm optimizes cluster assignment using a restraint on cluster radius, such that any member of a cluster is closer than a specified distance from the cluster center. The distance between the cluster center and its members was measured by the Cartesian coordinate RMSD of the heavy atoms. An iterative minimization procedure is used to minimize the distance between the cluster center and its members. The algorithm iterates to a solution from the initial cluster assignment until the cluster centers differ by less than a specific error tolerance. In this optimal clustering algorithm, the cluster radius and the error tolerance are crucial, as they determine both the size and the number of the formed clusters. In this case, a 3 Å cluster radius and a 0.5 Å error tolerance were used. We also performed cluster analysis by using other radius and error tolerance values, but the selection of a 3 Å cluster radius and 0.5 Å error tolerance gave a good balance between the number of clusters and structural diversity.

8.3. DIVE simulations for mapping potential energy landscapes and conformations of peptide F

Figure 8.3.1 displays the time series of various properties for the polypeptide at an initial temperature of 300 K from the six-trajectory DIVE simulation of peptide F in a GB/SA implicit solvent, starting from an extended structure. The partial history of total energy (Fig. 8.3.1a) indicates that the conventional NVE simulation remains at constant energy for a fixed time interval (20 ps), then the simulation is perturbed by trajectory and energy, and subsequently falls to a new total energy level (The trajectory and energy perturbation are accomplished by velocity redirection and rescaling, respectively). The DIVE simulation for peptide F in implicit solvent samples an enormous range of potential energies from nearly -300 to -820 kcal/mol (Fig. 8.3.1b) and temperatures ranging from near 0 K to almost 450 K (Fig. 8.3.1c) during the course of the simulation. Although the simulation involves heating to 1000 K, after approximately 1 ps (1000 steps), nearly half of the kinetic energy converts to potential energy. Consequently, the high temperature states where data are collected sample temperatures below ~500 K. Fig. 8.3.1d displays both the all-heavy atom and backbone atom RMSD between the calculated structures and the X-ray reference structure. The RMSD plots show that the simulation continues to search diverse conformations of the backbone with the RMSD, ranging from 0.39 Å to 8.10 Å, while the polypeptide forms the structure closest to the native α-helix after approximately 1.5 ns (heavy RMSD = 1.45 Å and backbone RMSD = 0.39 Å).



(a) **(b)**

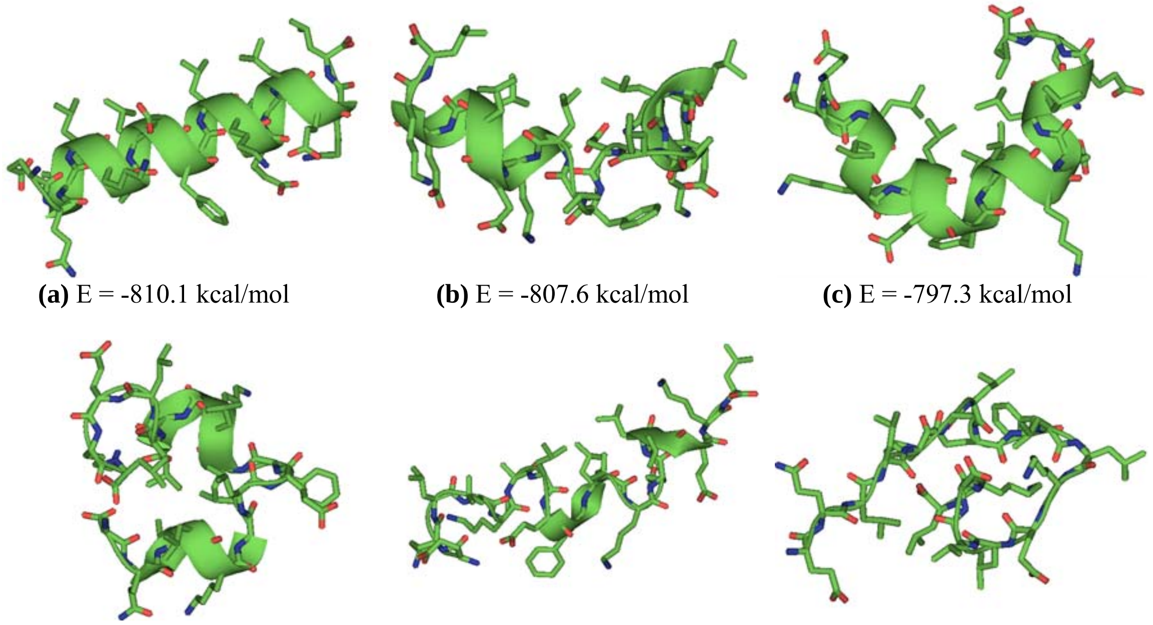


(c) (d)

Figure 8.3.1. Time series of various properties for one polypeptide (T_0 = 300 K) from a six-trajectory DIVE simulation of peptide F in a GB/SA implicit solvent, starting from an extended structure. **(a)** Total energy for the first 300 ps. **(b)** Potential energy. **(c)** Temperature. **(d)** RMSD between all heavy atoms (dark) or backbone atoms (gray) of the calculated structure and the X-ray structure.

Many local potential energy minima appear along the trajectory at kinetic energies corresponding to temperatures below 10 K. To investigate these structures, the average temperature in each 20 ps interval was calculated, and then the simulation regions whose average temperature was below 10 K were collected. Thirty-three regions were obtained. For each region, 40 potential energies were collected in 20 ps intervals. In each region, the energies usually oscillated within approximately ±1.5 kcal/mol and RMSDs oscillated within approximately ± 0.05 Å. Within these 40 regions, the lowest potential energy was selected as a representative minimum potential energy conformation. Representative results are shown in Fig. 8.3.2a – f. The lowest energy conformation of peptide F sampled is the native α-helical structure (Fig. 8.3.2a) with a potential energy minimum of approximately –810 kcal/mol. There are also many local potential energy minima between –810 and -765 kcal/mol. While the conformations are diverse, they are all well-ordered, with 10 or more hydrogen bonds (data not shown).

Though the energy gap between potential energy minima is relatively small (less than 4 kcal/mol between any two neighboring minima), the backbone RMSDs between these conformations and the α-helical X-ray structure differ by as much as 8 Å, so the different potential energy minima represent distinctly different structural types. Remarkably, the diverse conformations of low energies (e.g. Fig. 8.3.2a - e) always have an apolar hydrophobic interface (side chain of Leu residue) at one side and a polar hydrophilic interface (side chain of Glu and Lys residues) on the other side. This result illustrates the significance of hydrophobic and hydrophilic interactions in constructing stable conformations for this amphiphilic polypeptide in a solvent environment.



(a) E = -810.1 kcal/mol **(b)** E = -807.6 kcal/mol **(c)** E = -797.3 kcal/mol

(d) E = -797.3 kcal/mol **(e)** E = -795.7 kcal/mol **(f)** E = -766.2 kcal/mol

Figure 8.3.2. Representative conformations and potential energies (kcal/mol) of minima sampled for the polypeptide at initial temperature of 300 K, from a six-trajectory DIVE simulation of peptide F in GB/SA implicit solvent, starting from an extended structure. Approximate descriptions of the conformations, all-heavy and backbone atom RMSD from the X-ray structure: **(a)** the lowest potential energy minimum conformation sampled is an α-helix (backbone RMSD = 0.61 Å and heavy RMSD = 1.54 Å). **(b)** A V-shaped α-helix (backbone RMSD = 3.04 Å and heavy RMSD = 4.14 Å). **(c)** A narrower V-shaped α-helix (backbone RMSD = 4.70 Å and heavy RMSD = 5.56 Å). **(d)** A U-shaped, 1/3 3_{10} helix, and 2/3 α-helix with frayed termini (backbone RMSD = 6.63 Å and heavy RMSD

= 7.58 Å). **(e)** A mixed helical intermediate having 1-2, 1-3, 1-4, 1-5 and etc. hydrogen bonds (backbone RMSD = 1.73 Å and heavy RMSD = 3.36 Å). **(f)** A coiled structure (backbone RMSD = 7.73 Å and heavy RMSD = 8.91 Å). These images were generated by using PyMOL (http://pymol.sourceforge.net/).

Though the native structure has the lowest potential energy in the trajectory of the polypeptide at an initial temperature of 300 K, it is not the global potential energy minimum conformation from the all simulations. Compared to the conformational search of the single trajectory, the six-trajectory DIVE simulation of the polypeptides at initial temperatures of 10 K, 50 K, 100 K, 300 K, 600 K, and 1000 K extend the diverse conformational space sampling in a very limited simulation time (6 ns). The corresponding technique of using different initial energies, and therefore different momenta, for the polypeptide in each trajectory is similar to starting from different initial conformations, but these studies show the significance of using different perturbed trajectories in sampling diverse conformational space, even when starting from the same coordinates. Table 8.3.1 shows the lowest potential energies (in kcal/mol) and lowest backbone RMSDs of all sampled structures, when compared to the X-ray structure of peptide F in the six-trajectory DIVE simulation during the simulation times of 0 – 3 ns, 3 ns – 6 ns, and 0 – 6 ns respectively. The data indicates that different initial conditions can cause different sampling trajectories even when using short simulation times. However, the results usually converge as the simulation time extends when using this perturbed technique. Perturbation conditions can always disturb simulation trajectories into sampling different regions of conformational space, thus avoiding the limitations of conventional MD simulations that sample local regions. In these DIVE simulations, diverse conformational space regions were sampled and 198 different minimum regions in total were obtained from six trajectories.

Table 8.3.1. The lowest potential energies sampled (in kcal/mol) and lowest backbone RMSDs of the calculated structures compared to the X-ray structure of peptide F in a six-trajectory DIVE simulation during the simulation times of 0 – 3 ns, 3 ns – 6 ns, and 0 – 6 ns, respectively. **Note**: Generally, the lowest sampled potential energies and lowest RMSDs do not correspond precisely.

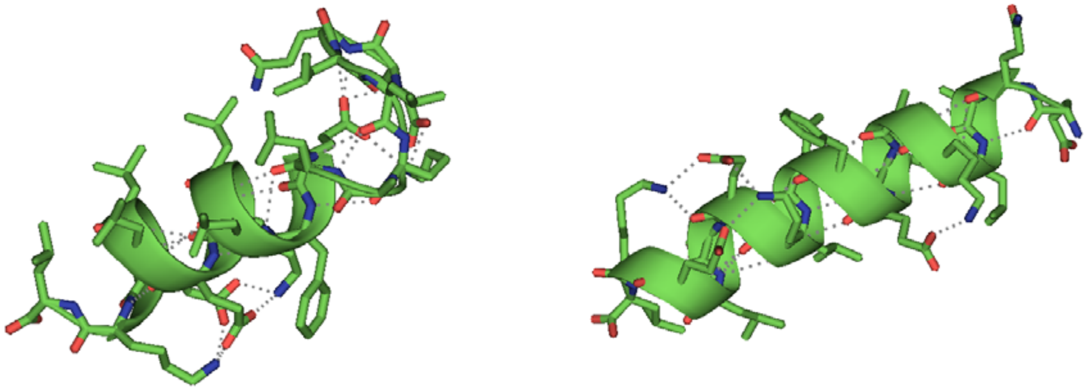
Time	0 - 3ns		3ns - 6ns		0 - 6ns	
	min. RMSD	Min. E	min. RMSD	min. E	Min. RMSD	min. E
Trajectory	Å	kcal/mol	Å	kcal/mol	Å	kcal/mol
1	1.60	-803.3	0.37	-808.6	0.37	-808.6
2	0.82	-809.2	0.55	-815.9	0.55	-815.9
3	0.71	-802.0	0.64	-811.2	0.64	-811.2
4	0.46	-810.1	1.08	-807.6	0.46	-810.1
5	0.87	-810.5	0.67	-804.1	0.67	-810.5
6	0.44	-809.2	0.39	-812.3	0.39	-812.3

Table 8.3.2 displays minimum potential energies (in kcal/mol) of the lowest ten energy minima and their RMSDs from the X-ray structure of peptide F in the six-trajectory DIVE simulation. The first line shows the data for the original structure. The second line shows the relaxed structure, equilibrated for 100 ps in a conventional NVT simulation at 2 K, starting from the experimental X-ray structure. This relaxed X-ray structure (2 K) has a slightly higher backbone atom RMSD from the reference structure than the sampled native structure (5) in our DIVE simulations, while the latter has the same secondary structure as the reference structure. Many similar folded structures (RMSD between 0.4 Å and 2.0 Å) were sampled during the simulations. These structures were distributed into different energy levels on a very rugged multi-dimensional potential energy surface, which exhibits no single deep well.

Table 8.3.2. Minimum potential energies (in kcal/mol) of the ten lowest energy minima, and their RMSDs (heavy and backbone) between the calculated structures and the X-ray structure of peptide F in a six-trajectory DIVE simulation. Also shown is the secondary structure assignment from the (ϕ, ψ) torsion angles of each residue (A = α-helix, 3 = 3_{10}-helix, S = sheet, P = polyglycine II or poly-L-proline II helix, C = Collagen, H = π-helix, R = 2.2_7 ribbon, O = others except for the above types, (±20°, ±20°) window from their respective standard point [38]). The first two lines give the data for the original and relaxed X-ray structure at 2 K, respectively.

Order	Energy (kcal/mol)	RMSD (Å)		Secondary structure
		Heavy	Backbone	
X-ray	-699.9	—	—	OAAAAAAAAAAAAAA3OO
2 K	-811.3	1.05	0.72	O3AAAAAAAAAAAA3OOA
1	-815.9	4.74	3.35	OOOCA3OAAAAAAOOAOC
2	-812.3	5.82	4.83	OA3OA3O33OAAAAOO3O
3	-811.2	3.83	3.05	OAAA3OO33OAAAOOOOA
4	-810.5	2.46	1.52	O3AAAAAAAAAAA3OOOA
5	-810.1	1.54	0.61	OAAAAAAAAAAAAAA3OO
6	-809.2	2.60	1.81	OA3333OAAAAAAA3OOA
7	-809.2	1.86	0.78	O3AAAAAAAAAAAAAAOA
8	-809.1	3.40	2.63	O3AAOOOA3OAAAAAAOA
9	-808.6	2.01	0.99	O3AAOAAAAAAAAA3OOC
10	-808.4	5.49	4.61	OOO333OA3OOA3OOOOS

The lowest potential energy state sampled in the DIVE simulation was not the native α-helical structure, but instead was a structure with a partial α-helix in the middle of the polypeptide (Fig 8.3.3a). This structure is most likely the global potential energy minimum conformation, verified by unpublished supplementary simulations starting from different initial configurations (data not shown). The potential energy for the relaxed X-ray structure at 2 K (Fig. 8.3.3b) also supports this hypothesis. With approximately 4 - 5 kcal/mol lower potential energies than the experimentally derived helical structures, this global minimum conformation has more intrapolypeptide salt bridges on the hydrophilic side, which accounts for the energy decrease from the helical X-ray structures. Three intrapolypeptide salt bridges are observed in the original X-ray structure Lys5 – Glu8, Lys12 – Glu13, and Lys12 – Glu16 [24]. While the relaxed structure (Fig. 8.3.3b) has one extra salt bridge of Glu13 – Lys17, the sampled native structure (Fig. 8.3.2a) has this salt bridge instead of Lys12 – Glu13. (The additional or changed salt bridge seen in the MD simulations correctly reflects a more stable pattern of intromolecular ion pairing in single, noninteracting helices [39]). However, the global potential energy minimum conformation has six salt bridges: Glu1 – Lys5, Lys5 – Glu8, Lys12 – Glu13, Lys12 – Glu16, Glu13 – Lys17, Glu16 – Lys17, which largely stabilize the conformation. In addition, comparing the relaxed X-ray structure at 2 K with the global minimum conformation, the decreased Coulombic energy overcomes the increased Gpol and SA energies (Fig. 8.3.4). Therefore, this conformation may reflect the real global potential energy minimum structure for an isolated polypeptide with implicit solvation, but may deviate from the experimental structures observed in a more complex biological environment.



(a) E = -815.9 kcal/mol

(b) E = -811.3 kcal/mol

Figure 8.3.3. Conformations and potential energies (kcal/mol) of the global potential energy minimum and the relaxed X-ray structure at 2 K in GB/SA implicit solvent. Also shown are the intrapolypeptide polar contacts. **(a)** The global potential energy minimum conformation is a structure with partial α-helix in the middle (backbone RMSD = 3.35 Å and heavy atom RMSD = 4.74 Å). **(b)** The relaxed X-ray α-helical structure at 2 K (backbone RMSD = 0.72 Å and heavy atom RMSD = 1.05 Å). These images were generated by using PyMOL (http://pymol.sourceforge.net/).

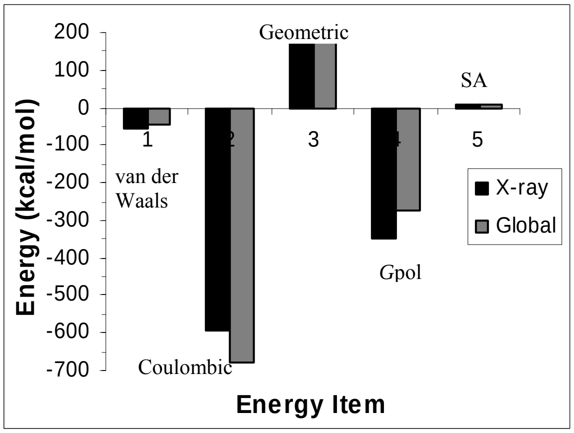
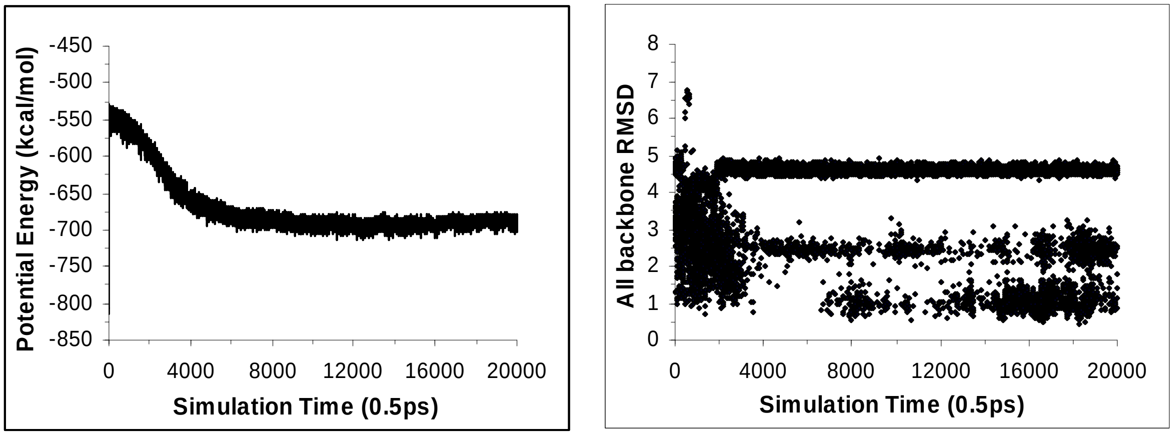


Figure 8.3.4. Energy components of the potential energy minima in the global and relaxed X-ray structures at 2 K. While the G_{pol} and Van der Waals energies favor the X-ray α-helical conformation, the Coulombic energy favors the global minimum energy conformation.

Finally, cluster analysis was conducted to evaluate diverse conformations sampled by our simulations. A total of 605 clusters were identified for 72,000 structures. Interestingly, the most-populated cluster included 3,464 structures, corresponding to the experimentally demonstrated native α-helices. Every other cluster contained less than 1,000 structures. Most of them (408 out of 604) were sparsely populated, and contained less than 100 structures. Among the more populated clusters (196), 18 clusters contained more than 500 structures. As is the case for many DIVE simulations, the size of each cluster is small but the number of clusters is large. This occurs because these simulations sample discrete diverse regions of conformational space instead of being restrained to a local search of many similar conformations.

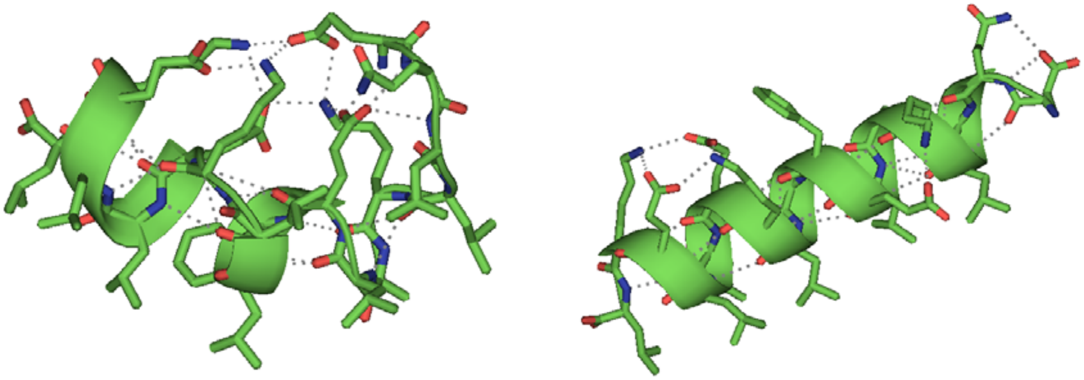
8.4. DIP simulations for sampling the global minimum free energy conformation of peptide F at 300 K

In addition, DIP simulations of peptide F in GB/SA implicit solvent at 300 K were conducted to locate the global free energy minimum at 300 K. Thirty-six trajectories were studied, obtained from the six-path DIP simulations. Figures 8.4.1a and b show the time series of the potential energy and all-backbone RMSD between the X-ray structure and the polypeptide with the lowest potential energy at each data-collection step (in 0.5 ps intervals) from the thirty-six trajectories. The potential energy history indicates that mechanical equilibrium was reached after 4 ns. In the following 6 ns trajectory, a total of 12,000 structures were collected. Cluster analysis revealed six distinct groups. Figure 8.4.2 shows the representative conformations and potential energies of the two largest clusters. The most-populated cluster contains 80% of the total structures, and displays a structure (Fig. 8.4.2a) that is closely related to the global minimum potential energy conformation (Fig. 8.3.3a). In fact, we observed one trajectory start from the initial global potential energy minimum conformation and convert directly to this conformational cluster, The conversion thus occurred from the global potential energy minimum to potential energy states approximately 120 kcal/mol higher than the global minimum. At room temperature, the oscillating kinetic energies control sampling of these potential energy states while the polypeptide equilibrates in the many conformations within this cluster. The second most-populated cluster shows the experimentally demonstrated native α-helical conformation (Fig. 8.4.2b), and includes 12% of the total structures. The other clusters are sparsely populated and include 3_{10}-helices, V-shaped α-helices, terminal frayed α-helices and partial α-helices. Subsequently, we subjected the entire 10 ns trajectory to further cluster analysis. The number of the identified clusters increased to 37 clusters from 20,000 structures. The number of clusters sampled from molecular dynamics increased significantly, mainly because the conformational change happens frequently during the mechanical equilibration process. However, the two most populated clusters were identical to those from the initial analysis.



(a) **(b)**

Figure 8.4.1. Time series of the **(a)** potential energy and **(b)** backbone RMSD from the X-ray α-helix of the lowest potential energy polypeptide from 36 trajectories in the DIP simulations of peptide F in GB/SA implicit solvent at 300 K.



(a) E = -700.8 kcal/mol, Ē = -691.4 kcal/mol

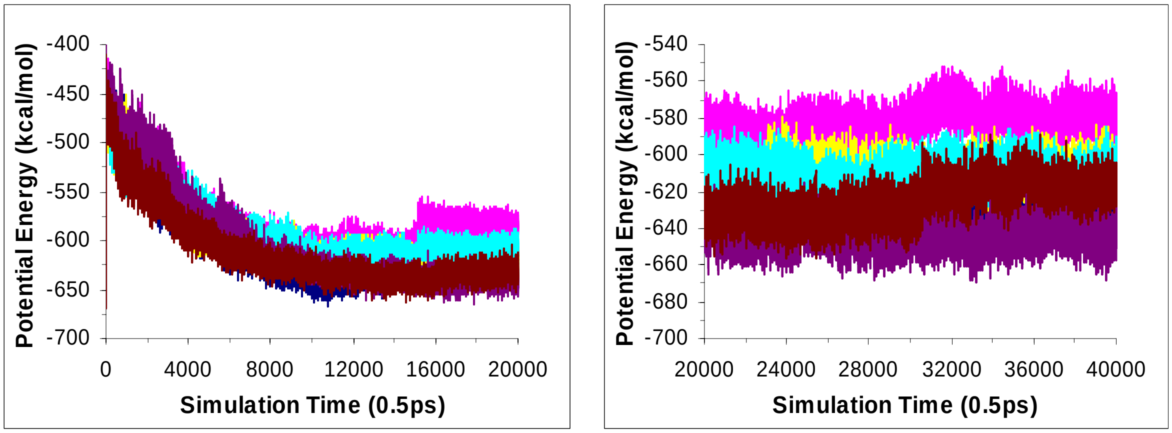
(b) E = -699.9 kcal/mol, Ē = -690.6 kcal/mol

Figure 8.4.2. Representative conformations and potential energies (kcal/mol) of the two largest clusters sampled from the lowest potential energy polypeptide of 36 trajectories. Also shown are the intrapolypeptide polar contacts. The representative structure for each cluster is the one nearest the cluster center. E is the potential energy of the representative conformations, and Ē is the averaged potential energy of the corresponding clusters. **(a)** The sampled global free energy minimum conformation is a structure closely related to the global potential energy minimum conformation. (2.50 Å backbone RMSD from the global potential energy minimum conformation, and 4.60 Å backbone RMSD from the X-ray structure). **(b)** α-helix (3.64 Å backbone RMSD from the global potential energy

minimum conformation, and 0.94 Å backbone RMSD from the X-ray structure). These images were generated by using PyMOL (http://pymol.sourceforge.net/).

After thorough exploration of the energy surface by using different initial coordinates and trajectories, the global minimum free energy state is the conformation of the lowest equilibrated potential energy with the largest population. Therefore, the structure shown in Fig. 8.4.2a is likely to be the global minimum free energy conformation for peptide F at 300 K in GB/SA with implicit solvent. This global conformation has an approximately 0.8 – 0.9 kcal/mol lower energy than the experimentally derived α-helical structure when both are sampled at room temperature. The global minimum free energy conformation has seven salt bridges: Glu1 – Lys5, Glu1 – Lys12, Lys5 – Glu8, Lys5 – Glu13, Lys12 – Glu16, Glu13 – Lys17, Glu1 – Lys17, which largely stabilize its conformation. In contrast, the α-helix shown in Fig. 8.4.2b has five salt bridges: Lys5 – Glu8, Lys12 – Glu17, Lys12 – Glu16, Glu13 – Lys17, Glu16 – Lys17. The extra intrapolypeptide salt bridges on the hydrophilic side of the global minimum free energy conformation accounts for its energy decrease from the helical X-ray structures at room temperature. In fact, the decreased Coulombic energy barely overcomes the increased Gpol and SA energies when comparing energies of the native α-helical structure with those of the global minimum free energy conformation near *in vivo* temperature (data not shown). Therefore, the global free energy minimum conformation, similar to the global potential energy minimum conformation, was sampled using molecular dynamics, and deviated from the experimental structure exhibited in a much more complex biological environment. Two possible reasons exist for the difference between the global minimum and X-ray structure. First, implicit solvent artificially favors salt bridges because no water molecules are available to solvate charged amino acid sidechains. Second, in the X-ray structure, three addition inter-polypeptide salt bridges appear between helices and thus further decrease its potential and free energies. *In vivo*, inter-polypeptide salt bridges may also appear in the global helical-structure by forming helical bundles and thus further decrease its potential and free energies.

After we determined the global minimum free energy conformation from the polypeptide with the lowest potential energy of thirty-six trajectories, we examined the entire DIP simulations. After mechanical equilibrium was reached, all trajectories were limited mainly to sampling potential energies in the range between -570 kcal/mol and -700 kcal/mol. These potential energies are the thermodynamically accessible regions corresponding to temperatures of 300 K ± 50 K. Different equilibrated potential energy levels always symbolize different conformations, and correspond to different free energy minima at 300 K. The transition from one free energy minimum to another free energy minimum is always time-consuming after mechanical equilibrium is reached in molecular dynamics. Therefore, the global free energy minimum is difficult to find in a conventional NVT simulation, especially if the simulation starts from a random initial structure. In addition, insufficient simulations due to limited simulation time or limited trajectories can easily bias the simulation results. For example, Fig. 8.4.3a displays the time series of potential energies from the six-path DIP simulation of peptide F at 300 K in GB/SA implicit solvent for 10 ns, starting from the extended structure. From these six trajectories, the equilibrated potential energies cannot decrease to -670 kcal/mol, even in 10 ns or longer simulations (Fig. 8.4.3b). The simulations sample high potential energy states. Both the global free energy minimum conformation and X-ray α-helix cannot be reached. Instead, only partial α-helices are explored during the mechanical equilibration process (data not shown).



(a)

(b)

Figure 8.4.3. Time series of the potential energy in a six-path DIP simulation of peptide F in GB/SA implicit solvent at 300 K, starting from the initial extended structure. **(a)** Potential energy in the first 10ns. **(b)** Potential energy displayed on an expanded scale in the following 10 ns. Different colors refer to different trajectories.

8.5. Conclusion

This chapter has discussed molecular dynamics simulations of an amphiphilic octadecapeptide peptide (1PEF) in the GB/SA implicit solvent model for water. This *de novo* designed polypeptide has a high apolar-polar amino acid ratio and can self-associate into hexamers in aqueous solution. Due to its complex energy surface as compared to polyalanine, a combined procedure is described in which the global potential energy minimum and myriad local potential energy minima are explored by using DIVE simulations followed by DIP simulations to search for the global free energy minimum near *in vivo* temperatures.

The disrupted velocity (DIVE) search simulations were started from an extended structure. With this technique, we find that the X-ray structure (at 1.5 Å resolution) of the folded polypeptide was quickly reproduced after a very short simulation time (backbone RMSD 0.39 Å, heavy atom RMSD 1.45 Å after 1.5 ns). However, even though many similar folded structures are sampled during the simulations (backbone RMSD between 0.4 Å and 2.0 Å), these structures do not correspond to the lowest energy states sampled on the very rugged multi-dimensional potential energy surface. Indeed, the potential energy surface exhibits no single deep well. On the other hand, the lowest energy state sampled (backbone RMSD = 3.35 Å) corresponds to a structure with a partial α-helix in the middle of the polypeptide. This conformation has approximately 4 kcal/mol lower potential energy than the experimentally derived helical structures. Like those X-ray structures, it has an apolar hydrophobic interface on one side and a polar hydrophilic interface on the other side.

In addition, divergent path (DIP) search simulations of peptide F at 300 K were performed to identify the global free energy minimum conformation. Thirty-six trajectories in total were obtained using the six-path DIP simulations of six different starting conformations. The initial conformations were the global potential energy minimum structure, an α-helix, a V-shaped α-helix, a mixed helical intermediate, a coiled structure, and an extended structure, respectively. We identified the global free energy minimum conformation from these thirty-six trajectories during a 360 ns aggregate simulation time. The global free energy minimum conformation exhibited the lowest equilibrated potential energy with the largest population. The X-ray α-helix appears to occupy the second lowest free energy minimum. Therefore, the simulation results indicate that both the global potential energy and the global free energy minimum structures, sampled by using molecular dynamics for an isolated polypeptide in water, may deviate from the experimental structures in a more complex biological environment. *In vivo*, several more inter-polypeptide salt bridges can appear for the global helical-structure by forming helical bundles and thus further decrease the potential and free energies as in the X-ray structure.

8.6. Bibliography

(1) Kirkpatrick, S.; Gelatt, C. D. J.; Vecchi, M. P. *Science* **1983**, *220*, 671.

(2) Basu, A.; Frazer, L. N. *Science* **1990**, *249*, 1409.

(3) Bassolino-Klimas, D.; Tejero, R.; Krystek, S. R.; Metzler, W. J.; Montelione, G. T.; Bruccoleri, R. E. *Protein Science* **1996**, *5*, 593.

(4) Hornak, V.; Simmerling, C. *Proteins: Structure, Function, and Genetics* **2003**, *51*, 577.

(5) Roitberg, A.; Elber, R. *Journal of Chemical Physics* **1991**, *95*, 9277.

(6) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96.

(7) Yoda, T.; Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **2004**, *386*, 460.

(8) Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **2000**, *329*, 261.

(9) Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **1999**, *314*, 141.

(10) Jarrold, M. F. *Annual Review of Physical Chemistry* **2000**, *51*, 179.

(11) Calvo, F.; Neirotti, J. P.; Freeman, D. L.; Doll, J. D. *Journal of Chemical Physics* **2000**, *112*, 10350.

(12) Vasquez, M.; Nemethy, G.; Scheraga, H. A. *Chemical Reviews* **1994**, *94*, 2183.

(13) Huber, T.; Torda, A. E.; van Gunsteren, W. F. *Journal of Physical Chemistry A* **1997**, *101*, 5926.

(14) Elber, R.; Karplus, M. *Journal of the American Chemical Society* **1990**, *112*, 9161.

(15) Stultz, C. M.; Karplus, M. *Journal of Chemical Physics* **1998**, *109*, 8809.

(16) Zheng, W.-M.; Zheng, Q. *Journal of Chemical Physics* **1997**, *106*, 1191.

(17) Anfinsen, C. B. *Science* **1973**, *181*, 223.

(18) Mortenson, P. N.; Wales, D. J. *Journal of Chemical Physics* **2001**, *114*, 6443.

(19) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598.

(20) Sali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248.

(21) Sali, A.; Shakhnovich, E.; Karplus, M. *Journal of Molecular Biology* **1994**, *235*, 1614.

(22) Li, H.; Helling, R.; Tang, C.; Wingreen, N. *Science* **1996**, *273*, 666.

(23) Miller, M. A.; Wales, D. J. *Journal of Chemical Physics* **1999**, *111*, 6610.

(24) Taylor, K. S.; Lou, M. Z.; Chin, T. M.; Yang, N. C.; Garavito, R. M. *Protein science: a publication of the Protein Society* **1996**, *5*, 414.

(25) Wang, J.; Cieplak, P.; Kollman, P. A. *Journal of Computational Chemistry* **2000**, *21*, 1049.

(26) Tsui, V.; Case, D. A. *Biopolymers* **2000**, *56*, 275.

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

(28) Hinsen, K. *Journal of Computational Chemistry* **2000**, *21*, 79.

(29) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Clarendon Press: Oxford, 1987.

(30) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *Journal of Chemical Physics* **1982**, *76*, 637.

(31) Tuckerman, M. E.; Martyna, G. J. *Journal of Physical Chemistry B* **2000**, *104*, 159.

(32) Palmer, B. J. *Journal of Computational Physics* **1993**, *104*, 470.

(33) Martyna, G. J.; Klein, M. L.; Tuckerman, M. E. *Journal of Chemical Physics* **1992**, *97*, 2635.

(34) Feig, M.; Karanicolas, J.; Brooks, C. L. *Journal of Molecular Graphics & Modelling* **2004**, *22*, 377.

(35) Carpenter, G. A.; Grossberg, S. *Applied Optics* **1987**, *26*, 4919.

(36) Karpen, M. E.; Tobias, D. J.; Brooks, C. L., III. *Biochemistry* **1993**, *32*, 412.

(37) Pao, Y. H. *Adaptive Pattern Recognition and Neural Networks*; Addison Wesley: New York, 1989.

(38) Voet, D.; Voet, J. G. Biochemistry; John Wiley & Sons, Inc.: New York, 1995; pp 144.

(39) Marqusee, S.; Baldwin, R. L. *Proceedings of the National Academy of Sciences of the United States of America* **1987**, *84*, 8898.

Chapter 9

Exploring Trp-cage Mini-protein Conformations and Folding Pathways by Disrupted Velocity Search Simulations

9.1. Introduction

In recent years, impressive advances in the artificial design of small, fast-folding proteins have been made by experimental scientists. These small proteins, on the order of 20 to 35 residues, can exhibit tertiary structural properties similar to natural single-domain proteins. Their globular folded structures are well defined, and can contain multiple secondary structural elements, a compact hydrophobic core, and tertiary contacts [1,2]. Examples include the 35-residue villin headpiece subdomain studied by McKnight and coworkers [3,4], the 28-residue ββα motif studied by Dahiyat and Mayo [5,6], and the 20-residue Trp-cage motif studied by Neidigh *et al.* [7,8]. These mini-proteins provide excellent model systems for folding simulations.

The 20-residue mini-protein Trp-cage ($NL_2YIQWLK_8DGG_{11}PSS_{14}GRP_{17}PP_{19}S$) was derived from the C-terminal fragments of a 39-residue exendin-4 peptide [9], and encapsulates the Trp-6 side chain in a "cage". It folds spontaneously and cooperatively from an extended structure to a highly defined 3D structure of a Trp-cage motif in approximately 4 microseconds [10]. The Trp-cage folds with two-state kinetics [7,10]. The unfolded structure under native conditions displays some elements of the tertiary structure, especially in the early association of Trp-6 and Pro-12 [7]. Experimental data suggest that the folding of the Trp-cage motif is highly cooperative [7,10]. In addition, Neidigh et al. [7] reported the partial mechanism for the Trp cage folding in the case of a 39-residue exendin-4 peptide EX4. In that case, the Trp cage formation corresponded to the docking of the three consecutive prolines in the C-terminus onto an exposed Trp indole ring of a preformed helix [7]. The Trp-cage mini-protein represents the smallest known cooperatively folding protein-like molecule identified to date [7]. The small size, high stability, and fast folding time make Trp-cage an ideal model system to provide direct comparison between simulation and experiments [11].

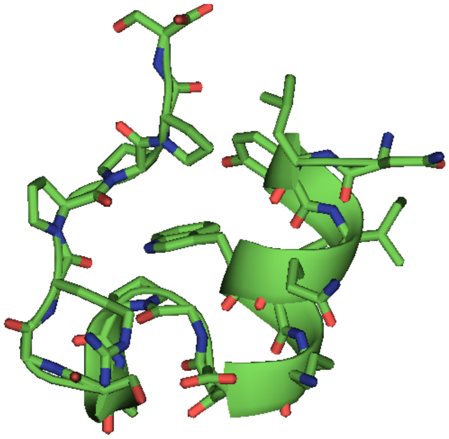


Figure 9.1.1. Native NMR structure of Trp-cage (the first one of the 38 NMR structures in 1L2Y pdb file). The image was generated using PyMOL (http://pymol.sourceforge.net/).

The NMR structure of this Trp-cage motif in aqueous solution (1L2Y, Fig. 9.1.1) reveals a well-defined tertiary structure with multiple secondary structures held together by a tightly packed hydrophobic core. Starting from the amino terminus, the secondary structure elements of this mini-protein include an α-helix extending from residues Leu-2 through Lys-8 (the helical ribbon in Fig. 9.1), followed by a 3_{10}-helix spanning four residues (Gly-11 to Ser-14), and finally a rigid polyproline II helix (Pro-17 to Pro-19) at the C-terminus. The compact hydrophobic core surrounding the aromatic ring of Trp-6 is formed by the side chains of three proline residues (Pro-12, Pro-18, and Pro-19), a glycine (Gly-11), and a tyrosine (Tyr-3). Stabilizing interactions for this globular folded structure also include a salt bridge between the side chains of Asp-9 and Arg-16, a hydrogen bond between the NH of Gly-11 and the backbone carbonyl oxygen of Trp-6, and a hydrogen bond between the NH$^{\varepsilon 1}$ of the Trp-6 indole ring and the backbone carbonyl oxygen of Arg-16 [1,2].

Several molecular dynamics simulations have been published on this Trp-cage mini-protein. Simmerling *et al* [12] performed the first simulations, in which a series of 20 ns – 100 ns molecular dynamics simulations were conducted with a modified AMBER99 force field. The blind structure prediction simulations produced a structure within 0.97 Å C$^{\alpha}$ atom RMSD and 1.4 Å heavy atom RMSD (excluding the terminal residues: Asn–1 and Ser–20, and the long side chains of Leu-2, Lys-7, and Arg-16 of the experimental NMR structure). Unfortunately, this paper did not include detailed information about attempted simulations, including any failed simulations at 300 K and successful simulations at 325 K etc. [12] Instead, they identified potential native structures from many low potential energy structures [11]. Snow et al.[13] ran thousands of stochastic dynamics simulations on the multi-nanosecond time scale (1 ns – 80 ns) with the united atom OPLS force field. Over 1000 simulations extended to 30 ns, and of these, eight trajectories contained structures less than 2 Å C$^{\alpha}$ RMSD from the NMR structure. The lowest C$^{\alpha}$ RMSD conformation searched shows a 1.4 Å RMSD. Pitera and Swope [14] made use of the parallel replica-exchange method to run MD simulations using the AMBER94 force field. Twenty-three replicas over a range of temperatures from 250 K to 630 K were simulated for 4 ns and the aggregate simulation time was 92 ns. They found a 1.0 Å C$^{\alpha}$ RMSD structure from the NMR structure. Finally, Chowdhury et al [1,2] reported 77 simulations of 100 ns each, using the newly developed AMBER 2003 force field. Five of the simulation trajectories yielded structures with main-chain RMSDs of 1.0 - 2.0 Å from the native NMR structure (not including the four terminal residues of Asn-1, Leu-2, Pro-19, and Ser-20). The lowest heavy-atom RMSD was approximately 2 Å from the native NMR structure.

Other computational methods have been applied to the Trp-cage system. Schug *et al.* [15] reported Trp-cage folding with the stochastic tunneling method by using their free-energy protein force field (PFF01) to reproduce a native-like conformation of 1.8 Å backbone atom RMSD, compared to the NMR structures. Nikiforovich and colleagues [16,17] used stepwise elongations of the peptide chain and the ECEPP force field to model possible locally driven folding pathways, and were able to generate fragments similar to residues 3 – 18 at about 1.5 Å C^{α} RMSD. Carnevali and co-workers [18] employed Monte Carlo simulations to reproduce the fast folding of the Trp-cage with the Amber94 force field. RMSD data for the calculated native-like structure were not reported, but they indicate that the structure was slightly lower quality than that reported by Simmerling *et al* [12]. All the studies described above used either the fully extended structure or rapidly relaxed structure. The generalized Born continuum solvent model [19-21] or the more general constant dielectric continuum solvent models were used to reduce computational complexity. Of these simulations, only Chowdhury et al. [1,2] and Nikiforovich et al. [16,17] were able to explore folding kinetics, the area of most interest to experimentalists.

In this chapter, we discuss simulations of the Trp-cage conformations and folding pathways, using disrupted velocity (DIVE) search simulations. Surprisingly, the Trp-cage motif can be folded in only 2 ns simulations with the best RMSDs identified to date (0.9 Å, 1.0 Å, and 1.6 Å considering all backbone, C^{α}, and heavy atoms, respectively), when compared to the experimentally determined native structure. Additionally, many different folding and unfolding trajectories were observed in a total aggregate of 36 ns simulation time, providing significant data on possible protein folding mechanisms. The simulations described not only reproduced the conformations and folding pathways explored by other laboratories, but also suggest additional features not previously identified in computational studies that are consistent with experiments.

9.2. Simulation methodology

9.2.1. Simulation algorithms and force field

In the DIVE simulation described below, six independent trajectories of mini-protein Trp-cage were simulated simultaneously. The polypeptide in each given trajectory did not interact with any others, so each followed its own perturbed trajectory in a conventional NVE simulation. Each independent polypeptide was assigned the same initial structure, but different initial atomic velocities and therefore different energies and temperatures. For each DIVE trajectory, an NVE simulation was propagated for a fixed time interval (e.g. 10 ps). At the end of the time interval, each polypeptide had its atomic velocities reassigned in one of two ways. In the first method, the reassignment was accomplished in two steps: first, the velocities of each atom in the polypeptide were rescaled, then the atomic velocities in each polypeptide were randomly redirected but their magnitudes were unchanged. This is the standard method used in all DIVE simulations discussed in earlier chapters. Alternatively, a second method of reassignment was used wherein the velocities of each atom in the polypeptide were reinitialized from a Gaussian distribution [22]. The rescaled temperature of the distribution was the mean temperature and the standard deviation was 0 K, so each atom was reassigned the same speed.

The AMBER99 force field [23] was used, along with the GB4/SA [24,25] implicit model for the solvent environment. For the modified generalized Born model employed here, the only parameters modified from literature [25] values were used as screening parameters. These were taken from gb_solvate (provided by Prof. B. Jayaram), and used to correct systematic errors arising from the pair-wise screening approximation used to calculate the effective Born radius. In GB4, the water dielectric constant was $\varepsilon_w = 80.0$ and that inside of the protein was $\varepsilon_p = 1$, following the standard values for the model developed.

Table 9.2.1.1. The screening parameters used were from gb_solvate (provided by Prof. B. Jayaram).

	MGB in gb_solvate	MGB from literature [25]
H	0.8847	0.8846
C	0.9182	0.9186
O	0.8833	0.8836
N	0.8728	0.8733
S	0.9292	0.9323

9.2.2. Computational details

The AMBER 8 software package [26] was used to generate initial coordinates for an extended structure 20-residue peptide of Trp-cage (1L2Y). The Molecular Modeling Toolkit [27] was used to convert input files from AMBER format to that required by our programs. The velocity-Verlet algorithm [22,28,29] was used to integrate the equations of motion and the SHAKE algorithm [30] was used to constrain the covalent bond distances between hydrogen and heavy atoms. Translation of the center of mass of the entire system was removed at each step. The Nosé-Hoover Chain method [31] was used to control the temperature in the conventional NVT simulations. A preliminary conventional NVT simulation was performed for the initially extended gas phase Trp-cage with a time step of 0.01 fs, because the original extended structure disintegrated when using a time step of 1fs either in *vacuo* or in a GB/SA implicit solvent simulations. The simulation was performed at 300 K for 20 ps, resulting in a random-coil structure with 7.57 Å heavy atom RMSD and 6.39 Å backbone atom RMSD, compared to the averaged NMR structure (The averaged NMR structure was derived from 38 NMR structures reported in the PDB). This coiled structure was used as the initial structure for DIVE simulations.

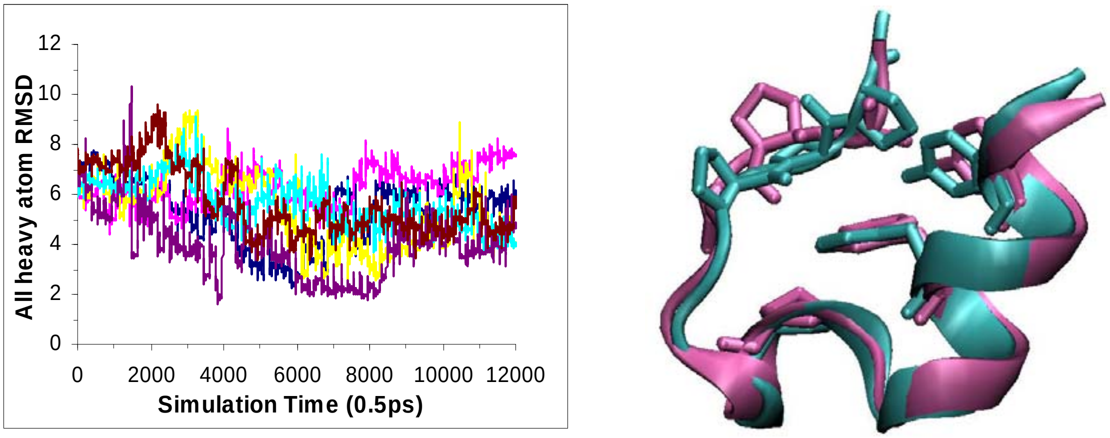
For DIVE simulations, six independent trajectories of Trp-cage were simultaneously simulated, with initial temperatures of 10 K, 50 K, 100 K, 300 K, 600 K, and 1000 K. The scaling parameter for cooling was 0.25, and the scaling factor for heating was $\sigma = T_{target}/T$, where T_{target} is the target temperature, and T is the temperature at the perturbation step. The threshold temperature for heating and cooling was 10 K. Thus, during the simulations, each polypeptide was cooled down to ¼ of its temperature T, whenever T rose above the threshold temperature. Once the temperature was below 10 K at the perturbation time, the polypeptide was heated back to T_{target}. T_{target} was always 1000 K. During the simulations, velocity reassignment occurred each 10 ps. A time step of 1 fs was used, and the trajectory data (energies and coordinates) were collected at 0.5 ps intervals (500 steps). The PTRAJ program from AMBER 8 [26], MMTSB tool set [32], and custom programs were used to analyze coordinate data.

Molecular mechanics (MM) energy minimization was performed using AMBER 7 [33]. The screening parameters in the topology file of Trp-cage generated by the TLEAP program were changed as shown in Table 9.2.1.1. NMR structures were energy minimized using 2000 cycles of steepest descent (SD) followed by conjugate gradient (CG) using the GB/SA implicit solvent model. A minimum potential energy was reached after a few hundred cycles of CG. The energy minimization for the average NMR structure was calculated in two stages to maintain structural integrity. First, the structure was minimized in *vacuo* with 2000 cycles of SD and 3000 cycles of CG; then the minimized gas phase structure was minimized in GB/SA implicit solvent as described previously. Full conjugate gradient minimization was used on structures sampled from our DIVE simulations. A slightly better minimum potential energy was quickly located after hundreds of cycles of CG minimization.

The six-trajectory DIVE simulations of Trp-cage for 6 ns produced 72,000 sets of coordinates. Conformational clustering was conducted on these snapshots. The simulation conformations were sorted into clusters using a step-wise optimizing fixed radius clustering algorithm [32,34-36] (see also Chapter 8). A cluster radius of 3 Å and a 0.5 Å error tolerance were used. We also performed the cluster analysis by using other radius and error tolerance values, but the selection of a 3 Å cluster radius and 0.5 Å error tolerance gave a good balance between the number of clusters and structural diversity. The distance between a cluster center and its members was measured by Cartesian coordinate RMSD of heavy atoms.

9.3. Simulation results

9.3.1. Potential energy landscapes and minimum energy conformations



(a)

(b)

Figure 9.3.1.1. (a) Time series of RMSDs between all heavy atoms of the calculated structure and the averaged NMR structure from a six-trajectory DIVE simulation of Trp-cage in GB/SA implicit solvent, starting from a randomly coiled structure. **(b)** The overlays of calculated (mauve) and averaged NMR structures (cyan). Backbone is shown in ribbon diagram. The Trp, Tyr and 5 Pro residues are shown in Licorice (stick) model. This calculated structure was sampled at 161.2 K. The image was generated using VMD [37].

The DIVE simulations for Trp-cage in GB/SA implicit solvent sampled an enormous range of potential energies from nearly -700 to -220 kcal/mol, with temperatures ranging from slightly above 0 to 480 K. Fig. 9.3.1.1a displays the history of RMSD between the heavy atoms of the calculated structure and the averaged NMR structure from all six trajectories. The RMSD plots show that the simulations continue to search diverse conformations with heavy atom RMSDs between 1.6 Å and 10.3 Å, while the mini-protein forms the structure closest to the native structure only after 2 ns (heavy-atom RMSD = 1.6 Å ,backbone RMSD = 0.9 Å). Fig 9.3.1.1b shows the overlay of the calculated conformation (mauve) and the averaged NMR structure (cyan). These two structures are almost fully overlapped, except for a slightly different orientation of the stretched PPII helix from three consecutive prolines. The calculated structure has virtually the same backbone conformation and tertiary topology as the experimentally observed native conformation. For the calculated cage structure, the indole ring of Trp-6 is securely sandwiched between the proline rings, and the phenyl ring of Tyr-3 tightly closes up the cage. In addition, two characteristic hydrogen bonds and the salt bridge from the Trp-cage motif were observed.

Further analysis of the calculated structure with the lowest RMSD from each of the six polypeptides in the 6 ns trajectory was conducted and the results are shown in Table 9.3.1.1. Three of six trajectories form folded structures with RMSDs less than 2.5 Å for all heavy atoms, and backbone atoms RMSDs less than 1.5 Å. These limits are reasonable, because the pair-wise all heavy atoms RMSDs of the 38 models in the NMR ensembles range from 0.8 Å to 2.3 Å, and all backbone atoms RMSDs range from 0.2 Å to 1.4 Å.

Table 9.3.1.1. The lowest heavy atom RMSD, backbone atom RMSD and C^{α} atom RMSD of the calculated structures sampled for each trajectory in 6 ns from a six-trajectory DIVE simulation of Trp-cage, compared to the averaged NMR structure. Three of the six trajectories searched native structure with all heavy atom RMSDs less than 2.5 Å and backbone atom RMSDs less than 1.5 Å from the native structure.

Trajectory	All heavy atoms RMSD	All backbone atoms RMSD (C, CA, N)	All C^{α} RMSD (CA)
1	2.25	1.16	1.22
2	4.39	2.85	3.10
3	2.32	1.26	1.29
4	3.58	2.63	2.84
5	1.62	0.89	0.98
6	3.33	2.18	2.35

From the temperature and potential energy histories, it can be seen that many local potential energy minima appear along the trajectories at kinetic energies corresponding to a temperature below 10 K. The average temperature was calculated for each 10 ps interval, and simulation regions with an average temperature below 10 K were identified. Sixty-six regions were obtained from one 6 ns trajectory, and 396 regions were obtained from all six trajectories. For each region, 20 potential energies were collected at 10 ps intervals. These energies usually varied within ± 0.5 kcal/mol , while the RMSDs of their conformations varied within ± 0.02 Å. Within these regions, the lowest potential energy structure was used as a representative potential energy minimum conformation.

Table 9.3.1.2 shows the minimum potential energies (in kcal/mol) of the ten lowest energy structures and the RMSDs between the calculated structures and the averaged NMR structure of Trp-cage in the six-trajectory DIVE simulations. Also shown are the secondary structure assignment (from the (ϕ, ψ) torsion angles) of each residue in the minimum energy conformations. The energy gap between potential energy minima is small (less than 2 kcal/mol between any two neighboring minima). Many similar folded structures (backbone RMSDs between 1 Å and 1.5 Å) were sampled during the simulations. They are distributed into different energy levels, along with other diverse conformations on the very rugged, multi-dimensional potential energy surface. The potential energy surface exhibits no single deep well.

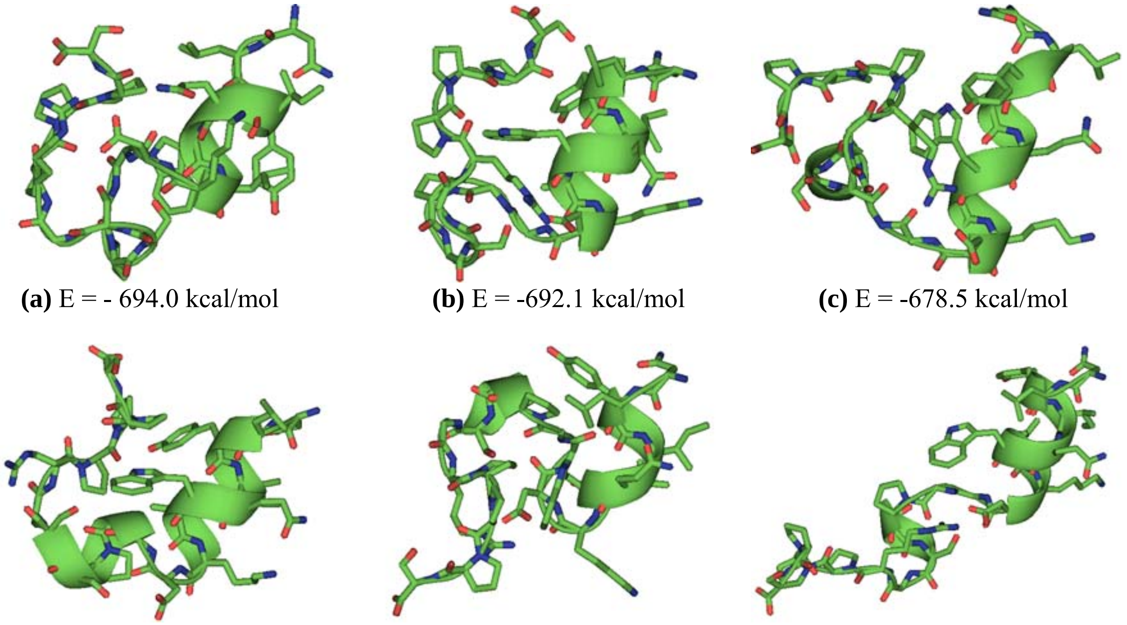
In fact, the minimized potential energies of the 38 models in the NMR ensembles range from -600 kcal/mol to -629 kcal/mol. Their RMSDs from the average NMR structure range from 0.8 Å to 1.4 Å for all heavy atoms, and from 0.3 Å to 0.8 Å for all backbone atoms. However, if the NMR structures are minimized first in *vacuo*, then in the GB/SA implicit solvent, their minimized potential energies range from -624 kcal/mol to -663 kcal/mol and RMSDs range from 1.2 Å to 2.1 Å for all heavy atoms and from 0.8 Å to 1.3 Å for all backbone atoms. By comparison, the minimum potential energies of the sampled structures closest to the average NMR structure range from -675 kcal/mol to -693 kcal/mol and RMSDs range from 1.6 Å to 2.5 Å for all heavy atoms and from 1 Å to 1.5 Å for all backbone atoms.

Table 9.3.1.2. Minimum potential energies (in kcal/mol) of the ten lowest energy structures, and their RMSDs between the calculated structures and the averaged NMR structure in a six-trajectory DIVE simulation of Trp-cage. Also shown is the secondary structure assignment from the (ϕ, ψ) torsion angle of each residue for the minimum energy conformations (A = α-helix, 3 = 3_{10}-helix, S = sheet, P = polyglycine II or poly-L-proline II helix, C = Collagen, H = π-helix, R = 2.2_7 ribbon, O = others except for the above types, (±20°, ±20°) window from their respective standard point [38]). The first two lines refer to the averaged NMR structure and the first NMR structure from the database of 38 NMR structures, respectively.

	Minimum potential energy (kcal/mol)	RMSD (Å)		Secondary Structure
		Heavy atoms	Backbone atoms only	
Averaged NMR structure	-660.00	—	—	O3AAAAAAOOO3OOOOPPOS
1st NMR	-616.35	1.38	0.69	OAAAAAA3OOOAOO3SPPOO
1	-693.98	4.03	3.25	OOAAOA3OSOO3AOAOCCOO
2	-693.65	4.04	3.26	OOAAOA3OSOO33OAOCCOO
3	-692.49	3.97	3.28	OOAAOA3OOOO3AOAOCCOO
4	-692.07	2.09	1.40	OAAAAAAAAOOA3OOOPCOO
5	-691.51	4.49	3.69	OOAAOA3OSOO3AOAOCCOO
6	-691.41	3.91	3.14	OOAAAAAAAOA3OOOPCOOO
7	-690.65	4.05	3.38	OOAAAAOASOOOOOAOCPOA
8	-690.46	2.28	1.36	OOAAOAAAAOOA3OOOPCOO
9	-689.81	3.89	3.09	O3AAAAAAAOOOOAOPOOOS
10	-689.38	4.13	3.44	OPA3OA3OSOOA3OAOPPOO

Six representative results are shown in Fig. 9.3.1.2. The lowest energy conformation of Trp-cage sampled was a structure similar to the αββ motif (Fig. 9.3.1.2a), with a potential energy of -694 kcal/mol. The αββ motif represents the conformation of an α-helix at N-terminus, followed by two β-strands. This structure has an apolar hydrophobic interface (side chains of Pro-18, Pro-17, Pro-12, Leu-7, Trp-6, and

Tyr-3) on one side and a polar hydrophilic interface (side chains of Gln-5, Lys-8, Asp-9, Ser-13, Ser-14, and Arg-16) on the other side. There were myriad local potential energy minima between -694 and -636 kcal/mol. Their conformations are diverse but are all well ordered, with each having more than 10 H-bonds (data not shown). The heavy atom and backbone atom RMSDs from the averaged NMR structure of different conformations differ by as much as 9.06 Å and 8.07 Å, respectively, so the different potential energy minima represent distinctly different structural types.



(a) E = - 694.0 kcal/mol **(b)** E = -692.1 kcal/mol **(c)** E = -678.5 kcal/mol

(d) E = -675.0 kcal/mol **(e)** E = -669.2 kcal/mol **(f)**: E = -666.4 kcal/mol

Figure 9.3.1.2. Representative conformations and potential energies (kcal/mol) of potential energy minima sampled in a six-trajectory DIVE simulation of Trp-cage in GB/SA implicit solvent, starting from a randomly coiled structure. The conformations are shown in ribbon diagram for backbone and stick model for heavy atoms. All hydrogen atoms are removed. **(a)** The sampled lowest potential energy minimum conformation has a structure similar to the αββ conformation (backbone RMSD = 3.25 Å, heavy RMSD = 4.04 Å). **(b)** A native-like conformation does not have the 3_{10} helical segment (backbone RMSD = 1.40 Å, heavy RMSD = 2.09 Å). **(c)** A conformation has a long 3_{10} helical segment at the α-helical position of these NMR structures (backbone RMSD = 4.14 Å, heavy RMSD = 5.17 Å). **(d)** A conformation similar to NMR structures, noticeably containing a second α-helix instead of the original 3_{10} –helical section (backbone RMSD = 2.80 Å, heavy RMSD = 3.70 Å). **(e)** A conformation has two correct helices but the wrong Trp-cage packing structure (backbone RMSD = 4.97 Å, heavy RMSD = 5.91 Å). **(f)** A conformation has two correct helices but shows extended-like structure (backbone

RMSD = 5.78 Å, heavy RMSD = 6.41 Å). These images were generated using PyMOL (http://pymol.sourceforge.net/).

Clustering analysis was conducted to evaluate the diverse conformations sampled. A total of 325 clusters were identified for the 72,000 structures. The most-populated cluster included 2459 structures, corresponding to the experimentally determined native structure. The second largest cluster population was a native-like structure, which included 2189 structures. Eleven clusters contained 1000 to 2000 structures. An additional 187 clusters contained more than 100 structures. The last 125 clusters were sparsely populated, with less than 100 structures each. The largest of the pair-wise all heavy atom RMSDs between the members of a cluster was approximately 2.5 Å. Each minimum energy conformation shown in Fig. 9.3.1.2 belongs to a different cluster. Increasing the cluster radius to 4 Å caused the structures shown in Fig. 9.3.1.2b and Fig. 9.3.1.2d to be sorted into the same cluster. The number of clusters observed was close to the number of minimum regions obtained in the simulations. The small size and the large number of clusters observed are due to the efficient diverse conformational space sampling of the DIVE simulations.

In the DIVE simulation protocol, a threshold temperature of 2 K corresponds to a potential energy minima mapped within approximately 1 - 2 kcal/mol. However, at threshold temperatures near 0 K, many cooling steps will be required below 10 K. Furthermore, at these low kinetic energies, the conformations will vary little and may remain trapped in the same energy wells. Therefore, longer simulation times are required to achieve the same number of cooling and heating cycles. In order to maximize the efficiency of conformational space sampling in a limited simulation time, a threshold temperature of 10 K is generally used. At this temperature, a good balance is achieved between the conformational searches and potential energy minimization. To illustrate how accurately the minimum energies are located in our molecular dynamics simulations, we extended the minimization processes of these sampled potential energy minima. Table 9.3.1.3 displays the sampled and minimized potential energies of the six energy minima and the closest native structure in the six-trajectory DIVE simulation. Also shown are the various energy differences in the Table.

Table 9.3.1.3. Sampled potential energies and minimized potential energies (in kcal/mol) of six energy minima and the closet native structure of Trp-cage in a six-trajectory DIVE simulation. Also shown are the energy differences of the sampled potential energies, energy differences of the minimized potential energies, and energy offsets of the sampled energy differences from the minimized energy differences between the neighboring conformations. The first six conformations were sampled between 9 K and 10 K. The seventh conformation was sampled at the temperature of 161.2 K. The bold numbers indicate the large difference of the potential energies (or energy differences) between the molecular dynamics simulation and molecular mechanics minimization.

Energy Structure	Potential energy (kcal/mol)	Energy difference (kcal/mol)	Minimized potential energy (kcal/mol)	Minimized energy difference (kcal/mol)	Energy difference offset
Fig. 9.3.1.2a	-694.0		-697.8		
		1.9		1.8	0.1
Fig. 9.3.1.2b	-692.1		-696.0		
		13.6		14.3	-0.7
Fig. 9.3.1.2c	-678.5		-681.7		
		3.5		3.4	0.1
Fig. 9.3.1.2d	-675.0		-678.3		
		5.9		6.0	-0.1
Fig. 9.3.1.2e	-669.1		-672.3		
		2.6		1.9	0.7
Fig. 9.3.1.2f	-666.5		-670.4		
		135.3		**27.3**	**108**
Fig. 9.3.1.1b	**-531.2**		**-643.1**		

Three important conclusions can be drawn from these data. First, the potential energies of the energy minima sampled in the simulations are not absolute minima, but are energy states at kinetic energies corresponding to 9 K – 10 K. In fact, a difference of 3 – 4 kcal/mol from the true minima is typically observed for the Trp-cage. In some cases, differences as large as 6 – 8 kcal/mol may exist. Second, though the potential energies of the minima sampled during molecular dynamics simulations differ by several kcal/mol from the true minima, the relative energies between different conformations are still accurate to ±1 kcal/mol, which are accurate enough to calculate energy differences. Third, the seventh conformation was sampled at the temperature of 161.2 K and its potential energy is far above the minimum energy, so the difference is very large (111.9 kcal/mol). At such high temperatures, the ensemble temperatures can vary by more than 10 K. As a result, the errors for the relative energies can reach at least several kcal/mol. These errors are large enough to mis-identify the global energy minimum or miscalculate energy differences.

9.3.2. Folding pathways

Despite decades of intensive research, the identification of the rate-limiting step, transition states and intermediate states in protein folding still remains a major challenge. Molecular dynamics simulations are currently better than experimental methods in tracking the time evolution of structural changes at high temporal resolutions [1]. This advantage is also present in DIVE simulations of fast folding proteins. In the following analysis, VMD [37] and Moil-view [39] were used to visualize folding and unfolding trajectories. The distances of hydrogen bonds and salt-bridges, etc. were calculated and tracked during the simulations.

Before exploring the folding pathways for Trp-cage, the folded and unfolded structures must be identified from the many possible conformations. For these studies, folded structures were defined as those with all heavy atoms RMSDs less than 2.5 Å from the averaged NMR structure. Since 2.5 Å is near the upper bound of the pair-wise RMSDs of the 38 models in the NMR ensemble, this was a reasonable choice. In fact, many simulated conformations with all heavy atoms RMSDs between 2.5 Å and 3.0 Å exhibit very similar tertiary structures to the NMR structures. Conformations with all heavy atom RMSDs between 3.0 Å and 4.0 Å that conserve secondary structure will have locally distorted tertiary structure and those that conserve tertiary structure will have locally distorted secondary structure. Either of these situations we define as a "locally" misfolded structure. For example, a conformational cluster was identified with well conserved multiple secondary structures, but the two ends of the main chain (PPII helical and α-helical segments) were orientated incorrectly. In this study, unfolded structures are defined as those with heavy atoms RMSD greater than 4 Å. Using these criteria, folding and unfolding of the native structures were observed more than ten times in our simulations.

A series of stabilizing factors account for the many diverse conformations (including the native structures) with low potential energies that are observed for this mini-protein. These stabilizing factors include (1) hydrophobic packing among five or six member rings of Tyr-3,Trp-6, Pro-12, Pro-17, Pro-18 and Pro-19; (2) π-stacking of the aromatic rings of Tyr-3 and Trp-6; (3) steric packing of the Trp-6, Tyr-3, Pro-12, and even the backbone inside the groove of the PPII helix of three consecutive prolines; (4) interactions between the side chains of hydrophobic amino acids; (5) hydrogen bonds between backbones and side chains of the hydrophilic amino acids; and (6) salt bridges between Asn-1, Lys-8, Asp-9, Arg-16, and Ser-20.

Several major stabilizing factors, such as hydrophobic packing, π-stacking and the steric packing among the residues of Tyr-3, Trp-6, Pro-12 and the 3 consecutive prolines, play an important role in the initial transitions. Driven by these stabilizing interactions, the fully extended structure is quickly collapsed into several coiled or looped structures [1,12,13]. These initial transitions were considered by Chowdhury *et al.* [1]to be the first stage of folding, because the structures show some native-like backbone or tertiary topology after hydrophobic collapse. However, we consider these quickly relaxed structures to be unfolded structures in the natural environments, not a first stage of folding. This consideration is consistent with the point that the unfolded states are likely to have some residual elements of the native structure [40]. Furthermore, the unfolded states in the experiments do display some residual hydrophobic cluster formation [7]. The possibility of these initial transitions involving the early formation of some nonlocal contacts in the unfolded structures is not addressed here.

In DIVE simulations the mini-protein folds and unfolds quickly, because the simulations focus on structural transitions between diverse conformations rather than equilibrating or trapping in a local region of conformational space. Therefore, a large number of diverse folding pathways is observed in a very short simulation time (6 ns for each trajectory, 36 ns altogether). The pathways start from various high energy conformations (coiled, looped etc.), cross different energy barriers and transition states, and finally reach the native structures. The same stabilizing factors take on different significance in the different folding processes.

No completely non-reversible pathway from the extended structure to the NMR structures was observed in the simulations. Therefore, qualitative aspects of the molecular mechanism for protein folding, such as the sequence of major events, transition states, and the rate-limiting step, were analyzed. The transition states were identified as those transient states that appear commonly between the unfolded and folded structures and their energies are higher than those of the unfolded, misfolded and folded conformations. The rate-limiting step was primarily decided by the simulation time required to make a transition. Minor events are more likely to involve reversible folding and unfolding of local structures, and thus are either omitted or described imprecisely during the exploration of a major folding pathway. Tens of folding and unfolding trajectories were enough to sketch clearly the common process, major stabilizing factors and pathways for the folding of the Trp-cage motif.

Two different major folding pathways were observed. In each pathway, multiple secondary structures of the main chain were folded first, followed by the formation of the tertiary structure. While the folding of the longer α-helix (rather than the shorter 3_{10}-helix) plays a significant role in the first pathway, the correct packing of the Trp-6 indole ring is crucial in the second pathway. Each is the rate-limiting step for its pathway.

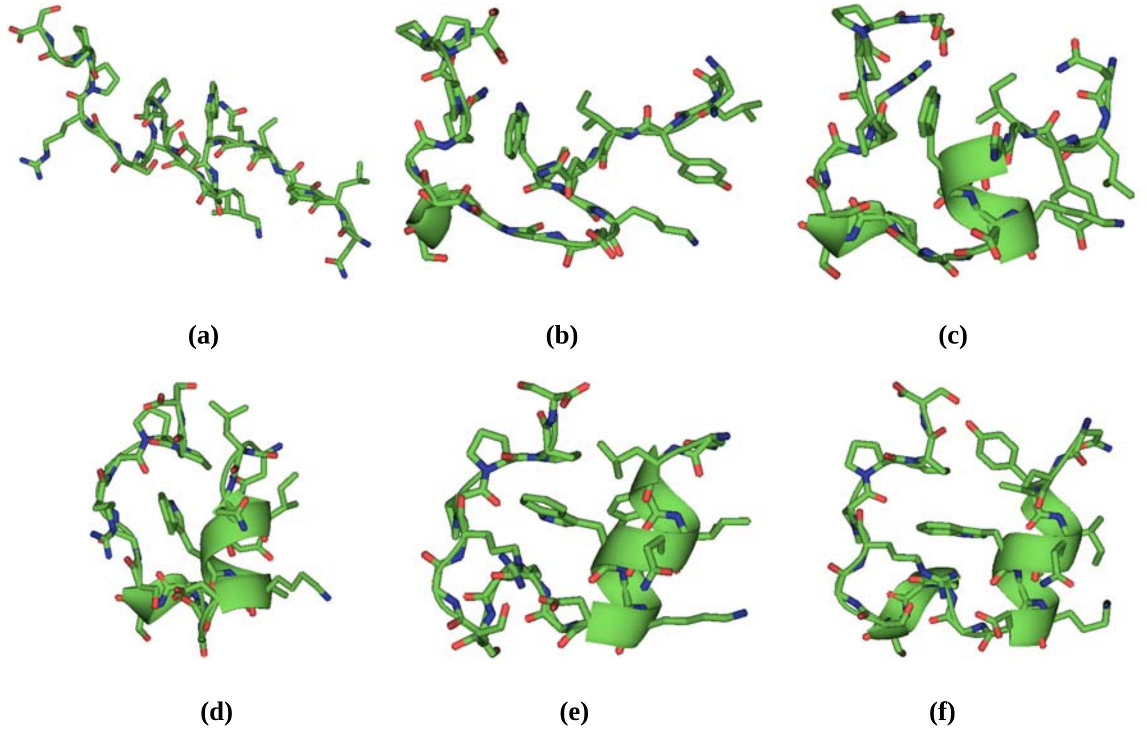
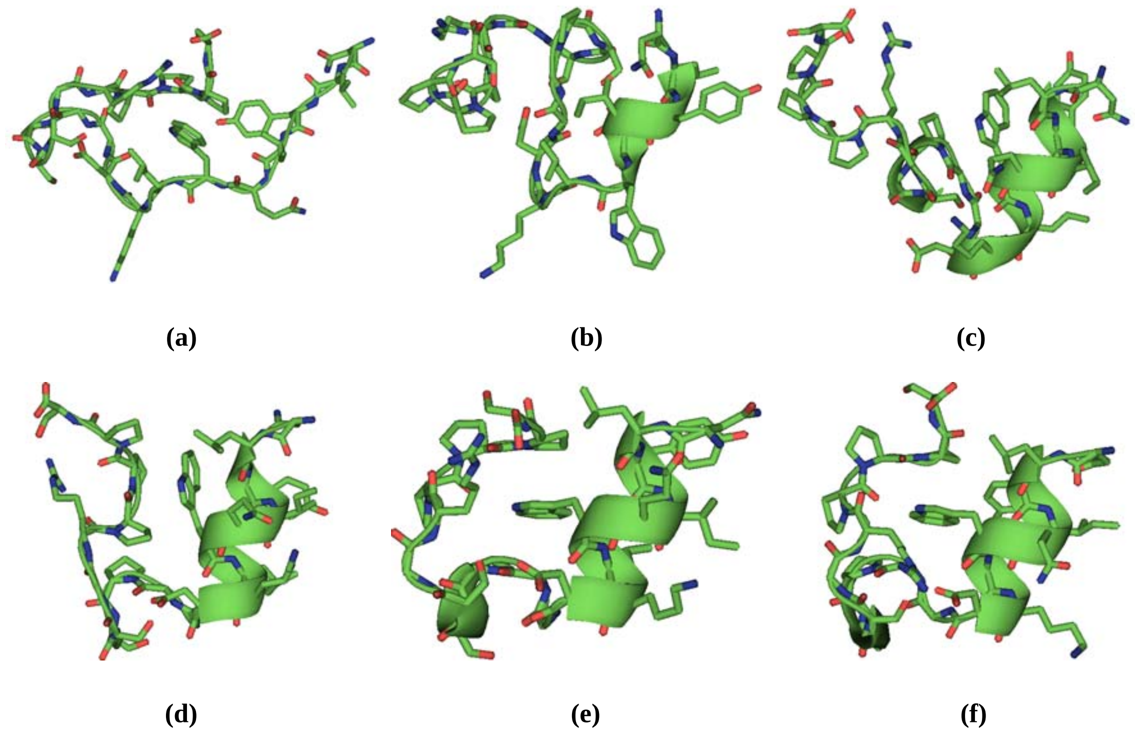


Figure 9.3.2.1. Snapshots of the folding structural change for the Trp-cage motif in the first mechanism. The conformations are shown in ribbon diagram for backbone and stick model for heavy atoms. All hydrogen atoms are removed. **(a)** An unfolded structure in which a large loop exists between Pro-12 and Trp-6. **(b)** 3_{10}-helix and the hydrophobic packing between the side chains of Trp-6 and Pro-17 appear. **(c)** α-helix extends. **(d)** The hydrogen bond between the $N^{\varepsilon 1}H^{\varepsilon 1}$ of Trp-6 and the O of Pro-17 forms. **(e)** Packing among the side chains of Trp-6, Pro-12, and Pro-19 occurs. **(f)** A native-like folded structure. The images were generated by using PyMOL (http://pymol.sourceforge.net/).

In the first pathway, folding starts from an unfolded structure in which a large loop exists between Pro-12 and Trp-6. The distance between the side chains of Pro-12 and Trp-6 is short, and hydrophobic interactions exist (Fig. 9.3.2.1a). A hydrogen bond between the NH of Gly-11 and the backbone O of Trp-6 may appear. Hydrophobic packing between Pro-12 and Trp-6, along with the hydrogen bond between the NH of Gly-11 and the O of Trp-6, plays an important role in the early stages of the relaxed transition. The early association of Trp-6 and Pro-12 is well supported by the NMR data [7]. Two helices are initiated from Lys-8 and Ser-13, respectively. The folding appears highly cooperative, involving several stabilizing factors. $N^{\varepsilon 1}H^{\varepsilon 1}$ of Trp-6 and the backbone carbonyl oxygen of Tyr-3 can form a hydrogen bond, as can the NH of Gly-11 and the O of Arg-16. These two extra hydrogen bonds may help to break the large loop temporarily and build the helical structures on both sides of the loop. After the breaking of the hydrogen bond between the NH of Gly-11 and the O of Arg-16, a 3_{10}-helix can be formed from Gly-11 to Ser-13 (Fig. 9.3.2.1b). The hydrophobic packing between Trp-6 and Pro-17 appears as well. The short 3_{10}-helix is folded much earlier than the longer α-helix. However, it is not maintained and experiences conformational changes (folding and refolding) with concurrent changes in nearby secondary or tertiary structure. In fact, these changes are necessary for the proper folding of the Trp-cage tertiary structure later. With the break of the hydrogen bond between $NH^{\varepsilon 1}$ of Trp-6 and the backbone carbonyl oxygen of Tyr-3, the α-helix extends through the established backbone hydrogen bond between Lys-8 and Gln-5 (Fig. 9.3.2.1c). The α-helix from Asp-9 to Leu-2 is further formed when two ends of the backbone are well separated. In addition, the hydrogen bond between the NH of Gly-11 and the O of Trp-6 is reformed. Next, the ends of the PPII and α-helices move closer, and a hydrogen bond is formed between the $N^{\varepsilon 1}H^{\varepsilon 1}$ of Trp-6 and the O of Pro-17 (Fig. 9.3.2.1d). The indole ring of Trp-6 can temporarily pack into the groove of the PPII helix, which is flattened out a little. With the partial unfolding and refolding of the 3_{10}-helical segment, the packing of side chains between Pro-12 and Trp-6 occurs, followed by the packing of the Pro-19 ring (Fig. 9.3.2.1e). The salt bridge between Asp-9 and Arg-16 appears stable at this stage. Finally, the cage is closed by the packing of the Tyr-3 phenyl ring (Fig. 9.3.2.1f). In this mechanism, the $N^{\varepsilon 1}H^{\varepsilon 1}$ in the indole ring of Trp-6 from the calculated structure faces inside, opposite to the orientation of that from the native structure. Thus, the hydrogen bond between the N$^{\varepsilon1}$H$^{\varepsilon1}$ of Trp-6 and the O of Arg-16 was not found.



(a) **(b)** **(c)**

(d) **(e)** **(f)**

Figure 9.3.2.2. Snapshots of the folding structural change for the Trp cage motif in the second mechanism. The conformations are shown in ribbon diagram for backbone and stick model for heavy atoms. All hydrogen atoms are removed. **(a)** A globular coiled structure with an empty and ill-defined cage. **(b)** Two-turn α-helix initiated from the N-terminus. **(c)** Well-defined secondary structure forms. **(d)** Packing between the side chains of Pro-12 and Pro-17, and of Trp-6 and Pro-18 occur. **(e)** Packing between Trp-6, Pro-12, and Pro-19 appears. **(f)** Tertiary structure of Trp-cage. The images were generated by using PyMOL (http://pymol.sourceforge.net/).

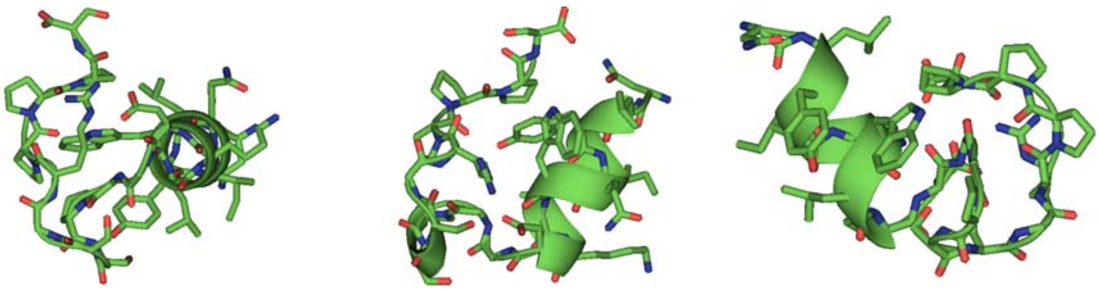
In the second pathway, folding starts from a loosely globular compact structure (Fig. 9.3.2.2a). Other than the stretched PPII-helix, the unfolded structure has no well ordered secondary structures. The phenyl ring of Tyr-3 can be buried in the groove of three prolines (steric packing) to close a poorly defined cage and a hydrogen bond between the O of Pro-17 and the O$^{\varepsilon1}$H$^{\varepsilon1}$ of Tyr-3 can be formed. Alternatively, the phenyl ring can be moved outside to open the cage and make hydrophobic interactions with the side chains of the three consecutive prolines. The opening and closing of the cage is determined by the movement of the two termini. The indole ring of Trp-6 was outside and removed from the five-member ring of Pro-12, which may have hydrophobic interactions with the side chain of Leu-7 or with the three prolines. Folding was initiated from the N-terminus (Fig. 9.3.2.2b). Beginning with Leu-2, the first contact was made in our simulations between Leu-2 and Gln-5, followed by the contact between Gln-5 and Lys-8, and completed by Asp-9. The hydrophobic packing between Tyr-3 and the three prolines (mostly Pro-18) may aid in the early formation of the α-helix at the N-terminus, but the extension of this important secondary structure requires the separation of the two termini of the backbone chain. If the separation of the two end of the backbone chain is not achieved, the polypeptide maintains the loose, overall compact structure of the empty cage. The formation of the C-terminal end of the α-helix is highly cooperative with the folding of the second helical structure (the short 3_{10}-helix) and the backbone structure between Gly-10 and Arg-16 (Fig. 9.3.2.2c). That is, the hydrogen bond of the 3_{10}-helical backbone and the hydrogen bond between the NH of Gly-11 and the O of Trp-6 are formed during this period and the correct backbone topology is almost folded. The 3_{10}-helix and the backbone structure between Gly-10 and Arg-16 are, however, not maintained and can be partly unfolded, with the different hydrophobic packing being observed among the side chains of Trp-6, Pro-12, and three consecutive prolines. Structural rearrangements at other places such as the hydrogen bond between $O^{\varepsilon1}H^{\varepsilon1}$ of Tyr-3 and O of Pro-17 can also affect the backbone structure between Gly-10 and Arg-16. In contrast, the α-helix maintains its structure throughout most of the simulations. Hydrophobic packing between the side chains of Pro-12 and Trp-6 occurs early in the following simulation. Tyr-3 interacts with the side chain of Trp-6. The packing between the aromatic rings of Trp-6 and Tyr-3 can co-exist also. In this situation, the indole ring of Trp-6 sits between the side chains of Pro-12 and Tyr-3. Furthermore, the two sides of the backbone chain move together. Parallel packing between the side chains of Pro-12 and Trp-6 temporarily disappears and packing between the side chains of Pro-12 and Pro-17, as well as the side chains of Pro-18 and Trp-6, occurs following the formation of the hydrogen bond between the NH of Gly-11 and the O of Trp-6 (Fig. 9.3.2.2d). The steric packing of the Trp-6 indole ring inside of the PPII helix is not significant because the groove of the PPII helix faces slightly outside of the cage. When the end of the PPII helix turns to the inside, parallel packing of Pro-19 and Trp-6 occurs, and the hydrogen bond between the O of Pro-17 and the $N^{\varepsilon1}H^{\varepsilon1}$ of Trp-6 forms. The 3_{10}-helix further re-forms, along with the salt bridge between Asp-9 and Arg-16 (Fig. 9.3.2.2e). The aromatic ring of Tyr-3 can be in the correct position, or can re-pack again to close up the cage if the packing between the side chains of Tyr-3 and Trp-6 is lost during the folding trajectory. In this mechanism, the indole ring of Trp-6 from the calculated structure has the same orientation as that from the native structure (Fig. 9.3.2.2f). The hydrogen bond between the $N^{\varepsilon1}H^{\varepsilon1}$ of Trp-6 and the O of Arg-16 could appear temporarily, but the O of Pro-17 dominates over the O of Arg-16 to form the hydrogen bond with the $N^{\varepsilon1}H^{\varepsilon1}$ of Trp-6 in the folded structures.

From the very beginning, five locations were observed to nucleate helical secondary structures: Leu-2, Lys-8, Ser-13, Gly-15, and Arg-16 (according to their ψ, φ torsion angles). From the different trajectories, the helical structure may form at any of these nucleation sites first, and any nucleation site can become any helix type, including a

3_{10}-helix or an α-helix. However, consistent with the native structure, the residues from Leu-2 to Asp-9 prefer an α-helix, and the residues from Gly-12 to Ser-14 prefer a 3_{10}-helix. Sometimes the α-helix formation is preceded by formation of a β-turn or a 3_{10}-helix. Throughout the simulations, transitions between 3_{10}- and α-helices occur frequently. Given the on-going debate on the relative significance of 3_{10}- versus α-helices [1,41,42], this observation is noteworthy, indicating the possibility of their coexistent. A 3_{10}-helix may or may not precede formation of an α-helix depending upon the particular nearby environment and folding pathways. During the simulations, a half- or even a one-turn helix folds rather easily, but can unfold quickly. A two-turn helical structure, however, is very stable, once folded. In particular, the folding of the two-turn helix from Leu-2 to Asp-9 is highly cooperative. This helix segment may be the most stable local structure for the mini-protein observed in our simulations. The transition between different helical types still occurs, but complete unfolding occurs less often than that of any other local structure except for the rigid PPII-helix.

The formation of the correct protein tertiary topology begins with the well-defined backbone topology, including at a minimum the stretched PPII helical and long α-helical secondary structures. The secondary structure from Gly-10 to Arg-16 is slightly flexible, and can be folded earlier or later than the globular packing of the Trp-6 indole ring, as described above. In the tertiary structure folding, the hydrophobic packing of the side chains of Tyr-3, Trp-6, Pro-12 and the three consecutive prolines plays a very important role in stabilizing diverse, low energy conformations. This is similar to the relaxation of the fully extended chain, but slightly different from the formation of the well-defined backbone topology from the unfolded, collapsed structures. Interestingly, the different packing or contacts between these hydrophobic residues build several conformational clusters of low potential energies, with all heavy atom RMSDs between ~ 2.5 Å and ~ 4 Å when compared to the native structure. On the other hand, for the folding process of the current Trp-cage motif, particular side chain packing or contacts between the Trp-6 and other hydrophobic residues, in a specific sequence, are required. The hydrogen bond between the O of Pro-17 and the $N^{\varepsilon1}H^{\varepsilon1}$ of Trp-6, not the hydrogen bond between the O of Arg-16 and the $N^{\varepsilon1}H^{\varepsilon1}$ of Trp-6 play the crucial role in the formation of the experimentally demonstrated native structures.

The different folding pathways proceed in different directions along the polypeptide chain show different timing of hydrophobic side chain packing and have different overall rate limiting steps. In the first pathway, the local secondary structure and even the tertiary structure of the residues between Gly-10 and Arg-16 forms first, followed by the folding of the globular Trp-cage. The hydrophobic packing initiates from the bottom of the cage (the backbone topology between the Gly-10 and Arg-16 in Fig. 9.3.2.1.d). Because the steric packing of the Trp-6 indole ring into the groove of the PPII-helix is highly cooperative with the folding of the flexible segment between Gly-10 and Arg-16 and occurs rapidly, the folding of globular Trp-cage is fast and the formation of a long helix may be the rate-limiting step. The side chain conformations of the N-terminal residues, in this situation, show a little more flexibility. Many folded structures within 1.5 Å backbone RMSD have all heavy atom RMSDs between 2 Å and 2.5 Å. A misfolded conformational cluster with heavy atom RMSDs between 3 Å and 4 Å has a nearly complete native-like backbone topology, but the ends of the long α-helix and the PPII helix are nearly vertical in Fig. 9.3.2.3a. This stable conformation is held together by a hydrogen bond between the O of Pro-17 and the $N^{\varepsilon1}H^{\varepsilon1}$ of Trp-6, tight steric packing of Trp-6 indole ring into the groove of the PPII-helix and hydrophobic packing of the side chains of Tyr-3 and Pro-12.



(a) E = - 686.4 kcal/mol **(b)** E= - 687.2 kcal/mol **(c)** E = - 681.8 kcal/mol

Figure 9.3.2.3. Three locally misfolded conformations and their minimum potential energies (kcal/mol). The conformations are shown in ribbon diagram for backbone and stick model for heavy atoms. All hydrogen atoms are removed. **(a)** A misfolded conformaiton with a nearly complete native-like backbone topology but the α-helix and the PPII helix are nearly vertical. **(b)** A misfolded conformation with a well-defined cage, but leaving the indole ring of Trp-6 outside. **(c)** A misfolded conformation with similar multiple secondary structures to the native structure but has two opposite hydrophilic and hydrophobic surfaces, respectively. The images were generated by using PyMOL (http://pymol.sourceforge.net/).

In the second pathway, the globular Trp cage forms first, followed by formation of the local secondary and tertiary structures between Gly-10 and Arg-16. In this case, the hydrophobic packing initiates from the door of the cage (two ends of the backbone in Fig. 9.3.2.2d). Therefore, the correct hydrophobic core packing of the indole ring of Trp-6 sandwiched between Pro-12 and other three proline residues appears difficult. In fact, several interesting misfolded conformational clusters were identified with incorrect packing of these side chains. For example, the aromatic ring of Tyr-3 can pack into the groove of PPII helix before Trp-6, with the formation of the hydrogen bond between the O of Pro-17 and the OH of Tyr-3. A well-defined cage is thus folded, but empty. The indole ring of Trp-6 remains outside (Fig. 9.3.2.3b). This conformational cluster has a very low backbone RMSD from the native structure (RMSD = ~1.2 Å) but the all heavy atom RMSD is about 2.8 Å. Another interesting conformational cluster, including the lowest global potential energy structure, is the hydrophobic-coiled structure (Fig. 9.3.2.3c). The cluster has similar multiple secondary structures to the native structure but has two opposite hydrophilic and hydrophobic surfaces, respectively. The hydrophobic surface arises from the hydrophobic interactions of the side chains of Pro-18, Pro-17, Pro-12, Leu-7, Trp-6, and Tyr-3. The hydrophilic surface includes the side chains of Gln-5, Lys-8, Asp-9, Ser-13, Ser-14, and Arg-16. The unfolding of these stable conformations involves moving the termini (the PPII-helix and α-helix) apart. During the overall folding pathway from the unfolded, collapsed structures to the Trp-cage motif, the correct packing of the Trp-6 indole ring is the rate-limited step. In fact, the correct packing of the Trp-6 indole ring to form the native tertiary structure was slower than the formation of the two-turn α-helix. Both mechanisms show residues between Gly-10 and Arg-16 to be flexible, a well-known observation in the literature [1,2,11,16,17].

Experiments imply that the Trp-cage folds with two-state kinetics [7,10]. That is, there is no stable intermediate between the unfolded and folded conformations. From these simulations, it is very difficult to evaluate this two-state folding kinetic model. On one hand, some stable misfolded conformations remain throughout the majority of folding and unfolding trajectories, which could be considered intermediate states. On the other hand, none of these stable structures are observed during all the diverse conformational transitions. Further, these misfolded conformations are usually sampled during the cooling cycle, and their low potential energies mean they are stable at low temperatures. However, these stable misfolded structures cannot directly convert into the native structure or vice versa. A common structural cluster at high temperatures always appears (similar to Fig. 9.3.2.1c and Fig. 9.3.2.2c). This structure has the well-defined α-helical secondary structure in which the end is moving away from the end of the PPII helix, and no hydrophobic packing is observed from the side chains between two terminal residues (especially between the side chains of Tyr-3 and three consecutive prolines). That is, tertiary contacts always exist between the two ends in the misfolded structures with the wrong packing or packing of the Trp-6 indole ring outside the cage. In order for these structures to transition to the Trp-cage motif, the ends of the chain need to break these tertiary contacts and separate, open the cage, and re-encapsulate the Trp-6 side chain into the sheath of Pro rings, until the correct packing of the Trp-cage side chain is obtained. Both folding pathways share this structural characterization. Therefore, members of this structural cluster are likely to be transition states for both folding pathways for the Trp-cage motif proposed here. However, the transition states of these two different folding pathways show some differences in the packing of Trp-6 and the three consecutive prolines. In the first mechanism, the indole ring of Trp-6 can pack into the groove of PPII-helix (Fig. 9.3.2.1d), while in the second mechanism the indole ring of Trp-6 directly packs with the five-member ring of Pro-18 (Fig. 9.3.2.2d).

Neidigh et al. [7] reported the partial mechanism for the Trp cage folding in the case of a 39-residue exendin-4 peptide EX4. In that case, the Trp cage formation corresponded to the docking of the three consecutive prolines near the C-terminus onto an exposed Trp indole ring of a preformed helix [7]. This agrees with the first folding mechanism described above. We believe that the folding of the truncated 20-residue mini-protein into a Trp cage in their experiments also adopts this mechanism. Though direct experimental data were not observed, two pieces of evidence point in this direction. First, the early association of Trp-6 and Pro-12 is well supported by the NMR data. Second, the correct hydrophobic core packing is obtained much more easily in this mechanism than the second mechanism. In fact, in the first mechanism, the formation of the longer helix seems highly cooperative with the correct packing of the Trp-6 aromatic ring into the domain of the three prolines. This cooperative and super fast folding of tertiary structure along with secondary structure may block the experimental observation of the preformed helix in the 20-residue Trp-cage, which we consider as the benchmark transition state in the first folding mechanism for formation of the Trp-cage motif.

Finally, a comparison of folding mechanisms in our simulations to those in simulations reported by Nikiforovich *et al.* [16,17] is warranted. They suggested a possible locally driven folding pathway in which the formation of the backbone starts from an α-helical fragment 4-9, extends to fragment 4-12, and ends with the nearly complete segment 4 – 18. The folding of the globular tertiary structure is driven by key stabilizing, sequential interactions of Trp-6 and Pro-12, then Trp-6 and Pro-18, and finally Trp-6 and Tyr-3. This folding pathway is very similar to the first folding mechanism. However, the folding in our simulations started from the large loop between Trp-6 and Pro-12 rather than the α-helical fragment. Moreover, the folding of the secondary and tertiary topology in the first mechanism is highly cooperative, which cannot be observed in their simulations using stepwise elongation of the peptide chain. In the replica exchange simulations of Trp-cage with explicit water starting from a NMR structure, Zhou [11] suggests a folding mechanism which includes an intermediate state, as opposed to the two-state folding properties of this mini-protein. However, the intermediate states with two partially prepacked hydrophobic cores are likely to be the unfolding states in our first mechanism. As stated previously, we consider them unfolded states, not intermediate states, because they are part of the hydrophobic collapsed structures at room temperature. Furthermore, the highest temperature in Zhou's simulation was approximately 600 K, 100 K higher than our simulations. At such high temperatures, the unfolded structures are more extended. Zhou did not give any folding kinetics to describe exactly how the unfolded structures of two partially prepacked hydrophobic cores fold into the final native structure. Recently, Chowdhury and coworkers [1] proposed a possible folding pathway in which the folding was initialized from the N-terminal residues to form the α-helix and the correct packing of the Trp-6 indole ring, rather than the secondary structure formation, was considered as the rate-limiting step [1,2]. This folding mechanism is very similar to the second mechanism in our simulations, but the rate-limiting step is determined differently. In their simulations, it was determined by the different equilibrium times of the conformations before and after the transition in a single conventional NVT trajectory [1]. The rate-limiting step was determined in our simulations by the simulation time required for the transition. In fact, the long transition time of the rate-limiting step indicated the possibility of diverse misfolded conformations with the incorrect packing of the Trp-6 side chain when the transition states folded into the native structures. The possibility of the misfolded vs. folded conformations apparently indicate the important role of the conformational entropy loss during correct folding (because of restrictions placed on both the backbone and side chains in the native structure [8]).

Though different force fields and computational protocols were used, the molecular mechanisms of protein folding from different research groups essentially agree, especially in the key mechanism. Due to the complexity of the energy landscape however, different folding processes are likely to be observed. The wide sampling of conformational space in our DIVE simulations ensures a diversity of potential folding pathways being sampled, even in short aggregate simulation times. In our simulations, many different folding and unfolding trajectories between diverse unfolded conformations and the experimentally demonstrated native structures provide a good basis for analyzing protein folding mechanisms. In fact, the simulations not only reproduce the conformations and folding pathways explored by the experiments and other computational simulations in more detail, but also suggest additional features, consistent with experiments, that are not evident in the published simulations. These new features include the possible transition state and some misfolded structures.

9.4. Conclusions

The newly developed disrupted velocity (DIVE) search simulations in implicit solvent have been conducted to study the folded conformations and folding mechanisms of a 20-reside mini-protein Trp-cage. The simulations in this study used the AMBER99 force field along with the generalized Born/solvent-accessible surface area implicit solvent model. Starting from a fully extended conformation, the simulations produced many structures within 2.5 Å all heavy atom RMSD of the NMR structure in a very short simulation time (6 ns). Among these, the structure closest to the native NMR structure has a 1.6 Å all heavy atom RMSD and 0.9 Å all backbone atom RMSD. Besides these native-like structures, our simulations also explored many other well-ordered potential energy minima, including the lowest minimum energy conformation, which has an all heavy atom RMSD of 4.0 Å compared to the native structure. This conformation has a well-conserved α-helical segment similar to the native structure. However, the side chains of hydrophobic residues coil together to form an apolar hydrophobic interface on one side, whereas the side chains of the hydrophilic residues gathered to form a polar hydrophilic interface on the other side. The lowest potential energy sampled for this conformation may be an artifact of the force field, and this conformation possibly corresponds to a misfolded structure, arising from incorrect packing of the Trp-6 indole ring. The potential energies of myriad local minima range from -694 kcal/mol and -600 kcal/mol and the energy gap between any two neighboring minima is small (less than 2 kcal/mol, with a few exceptions for the minima of highest potential energies), indicating a very rugged multi-dimensional potential energy surface. Indeed, the potential energy surface exhibits no single deep well.

In addition to the search of diverse conformations, two major folding pathways for this mini-protein were also characterized. The first folding pathway begins with a large loop between Pro-12 and Trp-6. The hydrophobic contact between the side chains of these two residues is formed early. Two helical segments nucleate at either side of this large loop and the formation of 3_{10}-helix precedes the formation of the entire α-helical segment. The folding of the globular Trp-cage motif starts with the local tertiary structure of the residues from Gly-10 to Arg-16. The key Trp-cage-stabilizing contacts emerge from early association between Trp-6 and Pro-12, then Trp-6 and three prolines, and then Trp-6 and Tyr-3. The formation of the stable long α-helix is the rate-limiting step based on the simulation time required for the transition. The second folding pathway starts at the N-terminus with the formation of the α-helical segment when two ends of the backbone chain move apart. The entire α-helical segment is folded first, then the 3_{10}-helix, and then the overall Trp cage forms, followed by fine tuning of the tertiary fold of the residues between Gly-10 and Arg-16. The key Trp-cage stabilizing contacts emerge from the early association between Pro-12 and three prolines, then Trp-6 and Pro-12, Trp-6 and Try-3, and Trp-6 and three prolines. The correct Trp-cage packing is the rate-limiting step for this second pathway, based on the simulation time required for the transition. The salt bridge between Asp-9 and Arg-16 is formed at a later stage. Both folding pathways may have similar transition states in which the entire α-helical segment is folded, but the two ends of the backbone chain involving the PPII-helix and the α-helix are separated, allowing the cage to open. Nevertheless, the transition states in different folding pathways show some differences, in that in the first folding pathway, the indole ring of Trp-6 can packs into the groove of PPII-helix, while in the second folding pathway, the indole ring of Trp-6 directly packs with the five-member ring of three consecutive prolines, especially Pro-18.

9.5. Bibliography

(1) Chowdhury, S.; Lee, M. C.; Xiong, G.; Duan, Y. *Journal of Molecular Biology* **2003**, *327*, 711.

(2) Chowdhury, S.; Lee, M. C.; Duan, Y. *Journal of Physical Chemistry B* **2004**, *108*, 13855.

(3) McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. *Nature structural biology* **1997**, *4*, 180.

(4) McKnight, C. J.; Doering, D. S.; Matsudaira, P. T.; Kim, P. S. *Journal of molecular biology* **1996**, *260*, 126.

(5) Dahiyat, B. I.; Sarisky, C. A.; Mayo, S. L. *Journal of Molecular Biology* **1997**, *273*, 789.

(6) Dahiyat, B. I.; Mayo, S. L. *Science* **1997**, *278*, 82.

(7) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. *Nature Structural Biology* **2002**, *9*, 425.

(8) Gellman, S. H.; Woolfson, D. N. *Nature Structural Biology* **2002**, *9*, 408.

(9) Neidigh, J. W.; Fesinmeyer, R. M.; Prickett, K. S.; Andersen, N. H. *Biochemistry* **2001**, *40*, 13188.

(10) Qiu, L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. *Journal of the American Chemical Society* **2002**, *124*, 12952.

(11) Zhou, R. *Proceedings of the National Academy of Sciences of the United States of America* **2003**, *100*, 13280.

(12) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *Journal of the American Chemical Society* **2002**, *124*, 11258.

(13) Snow, C. D.; Zagrovic, B.; Pande, V. S. *Journal of the American Chemical Society* **2002**, *124*, 14548.

(14) Pitera, J. W.; Swope, W. *Proceedings of the National Academy of Sciences of the United States of America* **2003**, *100*, 7587.

(15) Schug, A.; Herges, T.; Wenzel, W. *Physical Review Letters* **2003**, *91*, 158102/1.

(16) Nikiforovich, G. V.; Andersen, N. H.; Fesinmeyer, R. M. *Peptides 2002, Proceedings of the European Peptide Symposium, 27th, Sorrento, Italy, Aug. 31-Sept. 6, 2002* **2002**, 826.

(17) Nikiforovich, G. V.; Andersen, N. H.; Fesinmeyer, R. M.; Frieden, C. *Proteins: Structure, Function, and Genetics* **2003**, *52*, 292.

(18) Carnevali, P.; Toth, G.; Toubassi, G.; Meshkat, S. N. *Journal of the American Chemical Society* **2003**, *125*, 14244.

(19) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *Journal of the American Chemical Society* **1990**, *112*, 6127.

(20) Tsui, V.; Case, D. A. *Biopolymers* **2000**, *56*, 275.

(21) Tsui, V.; Case, D. A. *Journal of the American Chemical Society* **2000**, *122*, 2489.

(22) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Clarendon Press: Oxford, 1987.

(23) Wang, J.; Cieplak, P.; Kollman, P. A. *Journal of Computational Chemistry* **2000**, *21*, 1049.

(24) Jayaram, B.; Liu, Y.; Beveridge, D. L. *Journal of Chemical Physics* **1998**, *109*, 1465.

(25) Jayaram, B.; Sprous, D.; Beveridge, D. L. *Journal of Physical Chemistry B* **1998**, *102*, 9571.

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

(27) Hinsen, K. *Journal of Computational Chemistry* **2000**, *21*, 79.

(28) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *Journal of Chemical Physics* **1982**, *76*, 637.

(29) Tuckerman, M. E.; Martyna, G. J. *Journal of Physical Chemistry B* **2000**, *104*, 159.

(30) Palmer, B. J. *Journal of Computational Physics* **1993**, *104*, 470.

(31) Martyna, G. J.; Klein, M. L.; Tuckerman, M. E. *Journal of Chemical Physics* **1992**, *97*, 2635.

(32) Feig, M.; Karanicolas, J.; Brooks, C. L. *Journal of Molecular Graphics & Modelling* **2004**, *22*, 377.

(33) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Wang, J.; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Gohlke, H.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. AMBER 7; University of California: San Franciso, 2002.

(34) Carpenter, G. A.; Grossberg, S. *Applied Optics* **1987**, *26*, 4919.

(35) Karpen, M. E.; Tobias, D. J.; Brooks, C. L., III. *Biochemistry* **1993**, *32*, 412.

(36) Pao, Y. H. *Adaptive Pattern Recognition and Neural Networks*; Addison Wesley: New York, 1989.

(37) Humphrey, W.; Dalke, A.; Schulten, K. *Journal of molecular graphics* **1996**, *14*, 33.

(38) Voet, D.; Voet, J. G. Biochemistry; John Wiley & Sons, Inc.: New York, 1995; pp 144.

(39) Simmerling, C.; Elber, R.; Zhang, J. *Jerusalem Symposia on Quantum Chemistry and Biochemistry* **1995**, *27*, 241.

(40) Daggett, V. *Accounts of Chemical Research* **2002**, *35*, 422.

(41) Basu, G.; Kitao, A.; Hirata, F.; Go, N. *Journal of the American Chemical Society* **1994**, *116*, 6307.

(42) Tran Tran, T.; Zeng, J.; Treutlein, H.; Burgess Antony, W. *Journal of the American Chemical Society* **2002**, *124*, 5222.