**This document is downloaded from DR-NTU (https://dr.ntu.edu.sg)
Nanyang Technological University, Singapore.**

Application of molecular dynamics simulation in the study of protein structure and function

Siti Raudah Mohamed Lazim

2015

Siti Raudah Mohamed Lazim. (2015). Application of molecular dynamics simulation in the study of protein structure and function. Doctoral thesis, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/62909

https://doi.org/10.32657/10356/62909

Downloaded on 13 Mar 2024 16:49:15 SGT

Application of MD simulation in the study of proteins LAZIM, R 2015

APPLICATION OF MOLECULAR DYNAMICS SIMULATION IN THE STUDY OF PROTEIN STRUCTURE AND FUNCTION

SITI RAUDAH BINTE MOHAMED LAZIM

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCES

2015

APPLICATION OF MOLECULAR DYNAMICS SIMULATION IN THE STUDY OF PROTEIN STRUCTURE AND FUNCTION

SITI RAUDAH BINTE MOHAMED LAZIM

School of Physical and Mathematical Sciences

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2015

ACKNOWLEDGEMENT

First and foremost, I would like to thank my supervisor, Asst Prof Zhang Dawei, for giving me the chance to pursue my interest in research in his group. His guidance and support has allowed me to grow as a computational chemist and his continuous advice on research has greatly helped me in completing the work reported in this thesis.

I would also like to thank all members of Dr Zhang's group for all the valuable discussions that we have had with regard to research work and also for their continued support and friendship during the course of my graduate study.

Last but never the least, I would like to express my deepest gratitude to my parents, siblings, cousins and nieces for their support, encouragement, love and understanding. Thank you so much for believing in me and supporting me during my graduate study.

TABLES OF CONTENT

Acknowledgement…………………………………………………………………....i

Abstract ………………………………………………………………………….vii

List of Figures..………………………………………………………………..…..ix

List of Tables……………………………………………………………………….xvi

Chapter 1 Introduction …………………………………………………………….1

1.1 Overview……………………………………………………...…………1

1.2 Classical molecular dynamics simulation…………………….....………2

1.3 Principles of molecular dynamics simulation…………………....……4

1.4 Current applications of molecular dynamics simulation in the study of protein dynamics……………………………………………………........…7

1.5 Outline of thesis ………………………………………….………….......10

Chapter 2 Current problems in MD simulation …………………………….....13

2.1 Current limitations in MD simulations.……………………....……13

2.1.1 Improvement in sampling efficiency……………………...…13

2.1.2 Improvement in force fields………………..……………….....16

2.2 Development of polarizable and polarized force fields…….……......…20

2.3 Development of adaptive hydrogen bonds-specific charge (AHBC) scheme……………………………………………………………….....…24

Chapter 3 All-atom molecular dynamics simulation of structure variation from α/4β-fold to 3α-fold protein…………………………….…………………29

3.1 Introduction………………………………………….…....…….29

3.1.1 Protein folding problem: Protein misfolding..…………...….29

3.1.2 All-atom simulation of the evolution in protein structure from α/ 4β-fold to 3α-fold pattern……………………...........……………32

3.2 Methodology…………………….…………………….....…...…….33

3.3 Results: Protein conformational changes from α/4β-fold to 3α-fold……………………….………………………….……...........34

3.4 Conclusion………………………………………….......………..44

Chapter 4 Importance of electrostatic polarization in the *ab initio* folding of helical structures…………........……………………………………47

4.1 Introduction….………………………………...….………47

4.2 Methodology……………………………………...……...…49

4.3 Results: Comparative study involving the folding of 2khk using polarized and non-polarized force field….………………………….............53

4.4 Conclusion……..……………………………………….......67

Chapter 5 Importance of electrostatic environment in the modeling of the reduction process of metalloproteins……..……………...…………………..69

5.1 Introduction………………………………………...………...69

5.2 Methodology……………………………………...………74

5.3 Results: Improving the prediction of the reduction potential of mutated rubredoxin…….………………………………………...…….79

5.4 Conclusion..………………………………………….….....….88

Chapter 6 Utilization of MD simulation to predict the effect of mutation on the stability of proteins…….…………………..…………………………..91

6.1 Introduction…………………………………………….......91

6.2 Methodology..…………………………………………….......94

6.3 Results: Comparing the stability of apomyoglobin upon mutation…………………………………………………………...…..97

6.4 Conclusion….……………………………………………………......117

Chapter 7 Summary…………………………………………………………119

Appendix……………………………………………………………………123

References…………………………………………………………………125

List of Publications………………………………………………………161

ABSTRACT

Molecular dynamic (MD) simulation is a powerful theoretical tool which equips users with the ability to study the structures and functions of proteins by monitoring the intricate dynamics of proteins at the atomic level. With the utilization of MD simulation, dynamic processes occurring in the biological system such as protein folding and unfolding can be examined in great detail. In this thesis, the use of MD simulation in the dynamical study of protein folding and changes in protein conformations will be discussed. In addition, the predictive power of MD simulation will also be highlighted through two studies involving the determination of the stability of apomyoglobin variants and the reduction potential of rubredoxin. Besides its application, this thesis will also bring attention to a limitation of MD simulation which is the lack of polarization energy terms in classical force fields that effectively describe the inhomogeneous electrostatic properties of proteins. The important role of polarization effect in effectively modeling protein folding will be demonstrated through a study concerning the folding of a helical peptide, 2KHK. The folding of 2KHK close to its NMR structure was observed when polarization effects of hydrogen bonds were considered through an on-the-fly charge scheme termed adaptive hydrogen bond-specific charge (AHBC) scheme while a non-native structure was attained when non-polarized force field was used. The AHBC scheme periodically updates the charges of amino acids involve in hydrogen bonding hence incorporating into the force field the variation in charge distribution between hydrogen bond pairs upon formation/disruption of hydrogen bonds. This effectively represents the changes in the protein environment arising from the rapid configurational changes of proteins during folding. The significance of electrostatic environment in accurately determining the reduction potential of rubredoxin will also be illustrated in this thesis through the use of two

charge derivation schemes which differ in terms of the number of amino acid residues included during quantum mechanical calculations. From this study, a greater consideration of the electrostatic environment surrounding the iron atom of rubredoxin led to predictions which approached the experimentally determined reduction potentials. Also, the ability of the basic force field to model proteins will be portrayed through simple simulations that have been conducted to examine the conformational change of a protein from α/4β-fold to 3α-fold and to determine the stability of apomyoglobin upon mutation. Combined with appropriate tools for data analysis, MD simulations conducted were able to provide insights on the dynamics involved during the conformational change from α/4β-fold to 3α-fold and predict the stability of apomyoglobin variants relative to a wild type protein with reasonable agreement to experiment.

List of Figures

Figure 1.1: Visual illustration of terms listed in equation 1.2 which contribute to the empirical potential energy function of the force field. (adapted from Scientific American blog [39])**..7**

Figure 2.1: Schematic representation of replica exchange happening during REMD simulation. (adapted from http://www.weizhang.us/replica-exchange-simulation/)**............15**

Figure 2.2: Cartoon representation of the prion protein with the region coloured yellow being subjected to the on-the-fly AHBC scheme. (*above*) Schematic illustration of possible hydrogen bonds formed at the β-sheet upon mutation of the prion protein with the hydrogen bonds formed in the native protein enclosed in the orange dotted box. (*middle*) The sequences of the amino acids that are treated using AHBC scheme for rat, bovine and human are provided. (*below*) [21] **...17**

Figure 2.3: ***(left)*** Changes in distance between 127@O and 165@N (magenta), 131@O and 161@N (blue) and 133@N and 159@O (green) from MD simulations conducted using the on-the fly AHBC scheme for PrP^{Sc} of (A) rat, (B) bovine and (C) human at acidic pH. The plots for the distance between 133@N and 159@O (green) in (B) and (C) are shifted down by 2.0 Å to prevent overlaps between plots. ***(right)*** Changes in distances between 127@O and 165@N (magenta), 131@O and 161@N (blue) and 133@N and 159@O (green) from simulations conducted using AMBER force field for PrP^{Sc} of (A) rat, (B) bovine and (C) human. [21]**..18**

Figure 2.4: Changes in the secondary structures of the β-sheets of the prion proteins acquired through DSSP assignment of simulations performed using AHBC scheme and AMBER ff03 force field at acidic pH. β-structures which are originally not present in the starting structure are coloured green. [21]**...20**

Figure 2.5: MFCC method. (A) Amino acid fragments generated by cutting the peptide bond are capped at the ends using the conjugate cap, CH_2R_1CO-$NHCH_2R_2$. (B) Additional conjugate caps, $HCONH_2$-$HCONH_2$ are used to mimic hydrogen bonds in the AHBC scheme. The dotted red line represents the hydrogen bond. (adapted from Ref. 42)…………………………………………………………………………**26**

Figure 3.1: Schematic illustration of GA88 (PDB code: 2JWS) and GB88 (PDB code: 2JWU) obtained from experiment. Individual domains of the two proteins are labeled accordingly…………………………………………………………………………**33**

Figure 3.2: (A) Plots of the variation of the C_α-RMSD with time and the distribution of C_α-RMSD of simulated structures calculated over residue 9 to 53 at 270 K and 304 K. (B) Overlap between the experimental structure of GA88 (purple) (PDB code: 2JWS) and the folded structure with lowest C_α-RMSD (magenta) obtained at 270 K. (C) Overlap between experimental structure of GA88 (purple) (PDB code: 2JWS) and the folded structure with lowest C_α-RMSD (orange) obtained at 304 K……………………………**35**

Figure 3.3: Cluster analysis. Schematic illustrations of representative structures of the 5 clusters with the percentage occurrences of each cluster stated accordingly……………**36**

Figure 3.4: Cartoon representation of the experimental structure of GA88 and the representative structures of Cluster 4 and 5, derived from the cluster analysis conducted, with charged amino acids namely Lys13, Lys28, Lys31, Lys46 and Glu48 presented using licorice representation…………………………………………………………………**39**

Figure 3.5: (A) Variation of the C_α-RMSD of H1, H2 and H3 with time relative to the experimental structure of GA88 (PDB code: 2JWS) at 270 K. (B) Free-energy contour maps of H1, H2 and H3 of 3α-GA88 acquired at 270 K and 304 K……………………**40**

Figure 3.6: Secondary structure assignment (DSSP) of structures sampled at 270 K and 304 K during the REMD simulation. (Green: Parallel β-sheet, Magenta: Antiparallel β-sheet, Yellow: Mixed β sheet, Cyan: 3_{10}-helix, Blue: α-helix, Orange: π-helix)**………..41**

Figure 3.7: Time-dependent variation of the motion mode of the protein projected from PC1 to PC5 at 270 K and 304 K acquired from the first 15 ns of the simulation**……….43**

Figure 3.8: Time-dependent variation of the activity of the protein projected from PC1 and PC2 at 270 K during the first 15 ns of the simulation and schematic illustration of the motion modes of PC1 and PC2**……………………………………………………..44**

Figure 4.1: Flow chart summarizing the protocol for charge derivation using the AHBC scheme**……………………………………………………………………………52**

Figure 4.2: Schematic illustration of the NMR structure of 2khk with the glutamic acid residues at position 10 and 42 marked by two red circles**……………………………..…53**

Figure 4.3: Variation of the RMSD of the proteins folded through simulations performed using (A) AHBC scheme and (B) Amber charge**…………..……………………………54**

Figure 4.4: Changes in the helical content of the folded 2khk with time acquired from simulations conducted using the AHBC scheme (black) and AMBER force field (red)**...55**

Figure 4.5: Variation of the distance between the CA atoms of Glu10 and Glu42 for simulations conducted using (A) the AHBC scheme and (B) AMBER ff03 force field. Experimental distance between the two residues in 2khk are marked by the red line**………..57**

Figure 4.6: Variation of the dipole of the main-chain carbonyl group of folded 2khk, from residues 9 to 44, obtained through simulations conducted using (A) the AHBC scheme and (B) the Amber force field. Protein structures given in (A) and (B) represent the population with the highest percentage of occurrence in the trajectory and their respective macro-dipole indicated by red arrows**……………………………………………….…..59**

Figure 4.7: Variation of the hydrogen bond length formed between Lys29—Thr33, Ala30—Asp34, and Ser31—Gln35. (red: AHBC scheme, black: AMBER ff03 charge) ……………………………………………………………………………………………**61**

Figure 4.8: Schematic illustration of the representative structures of 5 clusters acquired through the cluster analysis performed for the last 20 ns of the simulation which were carried out using AMBER ff03 charge. The structures are ordered according to the sequence of appearance during the simulation with percentage occurrence stated. Amino acids that may form salt bridges are represented using licorice representation……………………**63**

Figure 4.9: Contour maps for the folding of 2khk carried out using (A) the AHBC scheme and (B) the Amber ff03 force field. (C) The free energy profile of 2khk……..**64**

Figure 4.10: (A) Variation of the RMSD of C34 compared to the X-ray crystal structure coded 1AIK over residues 2 to 35. Schematic illustration of the overlap between the X-ray structure (orange) and the best RMSD structure of the folded C34 (yellow) included. (B) Variation of the RMSD of N36 compared to the X-ray crystal structure coded 1AIK over residues 2 to 36. Schematic illustration of the overlap between the X-ray structure (blue) and the best RMSD structure of the folded N36 (grey) included..………………..**66**

Figure 5.1: (A) Cartoon representation of rubredoxin with the redox center illustrated using licorice representation. (B) Schematic diagram representing the six hydrogen bonds (HN---γS) formed between coordinated γS-Cys atoms and the backbone amide hydrogen atoms of Val8, Cys9, Tyr11, Leu41, Cys42 and Val44. Hydrogen bonds are represented as red lines and residues Leu41 and Val44 which are mutated to obtain L41A, V44A and V44G rubredoxin mutants are highlighted in green………………………………………..**73**

Figure 5.2: Licorice representation of iron atom and residues in the first and second coordination sphere. Schematic representation of amino acid residues isolated for DFT charge calculation in Scheme I and II……………………………………………………**75**

Figure 5.3: Electrostatic potential map of wild type rubredoxin in oxidized state with charges acquired using Scheme I and II. Only iron, Cys6, Val8, Cys9, Cys39, Cys42 and Val44 are depicted in this figure for clarity. The red arrows point to the backbone amide hydrogen atoms of Cys9 and Cys42 while the black arrows point to backbone amide hydrogen atoms of Val8 and Val44. The potential energies of all electrostatic potential maps range from -0.1 to 0.1 kcal/mol and the colors go from negative to positive in the following order: red < orange < yellow < green < cyan < blue < magenta……………………..**82**

Figure 5.4: Plots of ⟨∂V/∂λ⟩ versus simulation time for wild type rubredoxin with partial charges derived using Scheme II……………………………………………………**85**

Figure 5.5: Plots of ⟨∂V/∂λ⟩ versus λ for rubredoxin variants namely wild type, L41A, V44A and V44G. Samplings of ∂V/∂λ are conducted by employing charges derived using Scheme I (green) and Scheme II (red)…………………………………………………**86**

Figure 6.1: Cartoon representation of myoglobin and apomyoglobin with the individual helices alphabetically labeled A to H…………………………………...……………..**93**

Figure 6.2: Flowchart summarizing the modeling of apoMb variants in explicit 2 M urea solution…………………………………………………………………………….…..**96**

Figure 6.3: (A) Variation of RMSD with time and (B) the distribution of the RMSD of wild type apoMb (black), E109A (red), E109G (green) and G65A/G73A (blue) at Helix F to H…………………………………………………………………………..………..**98**

Figure 6.4: Variation in native contacts with time for wild type apoMb (black), E109A (red), E109G (green) and G65A/G73A (green)………………………………………**101**

Figure 6.5: Schematic representation of the solvent accessible surface area (SASA) of apoMb with the helical domains and two residues found within the hydrophobic core namely Trp7 and Trp14 labeled………………………………………………………**102**

Figure 6.6: Variation of the solvent accessible surface area (SASA) of wild type apoMb (black), E109A (red), E109G (green) and G65A/G73A (blue) with time……………..…**104**

Figure 6.7: (A) Changes in the solvent accessible surface area (SASA) of Trp7 with time and (B) the distribution of the SASA of Trp7 in wild type (black), E109A (red), E109G (green) and G65A/G73A (blue)……………………………………………………..**105**

Figure 6.8: (A) Correlation maps of wild type apoMb, E109A, E109G and G65A/G73A with the eight helical domains represented by color-coded boxes going from Helix A (blue) to Helix H (purple). Schematic representations of apoMb with the helices and loops involved in (B) Region 1, (C) Region 2 and 3 and (D) Region 4 highlighted…………**107**

Figure 6.9: (A) Cartoon representations of fluctuations observed for WT, E109A, E109G and G65A/G73A through PCA. The color transition from first to last frame goes from red to white to blue. The RMSF of the CA atoms of (B) Loop EF and (C) Helix F………..**112**

Figure 6.10: (A) Cartoon representation of apoMb with four hydrogen bonds formed between Glu6 and Lys133, between His12 and Asp122, between His116 and Gln128 and between Asp27 and Arg118 displayed using licorice representation. (B) Cartoon representation of apoMb with two hydrogen bonds formed between Glu4 and Lys79 and between His82 and Asp141 displayed using licorice representation. The hydrophobic core of apoMb is represented using surface representation colored magenta…………………..**114**

Figure 6.11: Plot of distance between six hydrogen bond pairs namely Glu6 and Lys133, Asp27 and Arg118, His12 and Asp122, Gln128 and His116, Glu4 and Lys79, and Asp141 and His82 against time…………..……………………………………….……………**115**

Figure 6.12: (A) Schematic representation of urea-water clusters in the vicinity of hydrogen bonds formed between (A) Glu4 and Lys79 and His12 and Asp122 (B) Glu4 and Lys79, Glu6 and Lys133, His82 and Asp141 and His116 and Gln128. Ball-and-stick representation was used for urea (carbon atoms in cyan) and water molecules. This figure was generated using Schrodinger Suite 2011. [55]**……………………………………..117**

List of Tables

Table 5.1: Partial charges of iron and iron-coordinated sulphur atoms of cysteine residues of wild type rubredoxin obtained using charge schemes I and II and that obtained by Gámiz-Hernández et al. for comparison. [28]**……………………………………………80**

Table 5.2: Partial charges of iron and atoms in the first and second coordination spheres of wild type rubredoxin**……….……………………………………………………….81**

Table 5.3: $\Delta\Delta E_{\mathrm{cal}}$ of rubredoxin mutants *viz.* L41A, V44A and V44G based on charges acquired through Scheme I and II**…………………………………………………..87**

Chapter 1 Introduction

1.1 Overview

Interest in the study of proteins by using computers began circa 1975 when classical molecular dynamics (MD) simulation was first implemented in the folding study of a bovine pancreatic trypsin inhibitor, a simple mini-protein involved in the inhibition of the trypsin protein. [1] This protein, even though lacking in dynamical interest, served as the genesis of MD simulation contributing towards the better understanding of the underlying theory of MD simulation in the study of biomolecules such as proteins and nucleic acids. [1-3] Other than the use of classical mechanics in MD simulation, works done by Martin Karplus, Michael Levitt and Arieh Warshel in the dawn of molecular modeling also extended the implementation of computer-based molecular modeling from exclusively protein-centric systems to systems comprised of organic/inorganic moieties such as protein-ligand complexes through the combination of classical and quantum physics. [4-6]

The coupling of classical and quantum mechanics has permitted the cyber exploration of chemical processes such as enzymatic reactions occurring in living organisms, and with the progress made in MD simulation, investigation related to the vital processes occurring in living organisms could be done using computers with comparable precision to that of experiment. [4-6] With the introduction of quantum mechanics into classical MD simulation, the application of MD simulation has broadened to include a wide spectrum of applications, other than the exploration of protein folding. [7-11] These applications include computer-aided drug discovery, free energy calculation of ion/ligand migration in biological systems and determination of the reduction potential of electron transfer proteins. [7-11] These achievements in multi-scale molecular modeling were recognized worldwide through the conferment of the Novel Prize in Chemistry for the year 2013 to Martin Karplus, Michael Levitt and Arieh Warshel. [12, 13] Molecular modeling, in recent years, has progressively improved, catching up to its traditional experimental counterparts in the field of protein folding. [12, 14-17] This advancement in molecular modeling was aptly described in the press release for the Nobel Prize in Chemistry 2013 [12] which stated that,

> *"Today the computer is just as important a tool for chemists as the test tube. Simulations are so realistic that they predict the outcome of traditional experiments."*

Alongside the positive growth of MD simulation, the development of robust computational analytical tools also played a significant role in establishing a bridge between simulations and experiments. In this thesis, MD simulation will be used to study the structures, dynamics and functions of proteins. Protein folding will be examined in Chapter 3 and 4 using MD simulations with Chapter 4 emphasizing the significance of including the polarization effect of the protein environment in the study of protein folding. The incorporation of quantum mechanics in classical MD simulations was also put to good use in Chapter 5 for the study of a metalloprotein, specifically rubredoxin, and lastly, the practicality of MD simulation in predicting the stability of a protein upon mutation will be discussed in Chapter 6.

1.2 Classical molecular dynamics simulation

Classical MD simulation is a powerful computational device which allows the atomistic modeling of the structure and dynamics of biological macromolecules such as proteins, carbohydrates and nucleic acids. [15-20] On a side note, other than biological structures listed, the use of MD simulation in the study of non-biological components such as liquids, polymers and nano-clusters has also been reported. [20-23] The effective modeling of complex systems such as proteins at the atomic level is possible through the implementation of the classical equation of motion which provides microscopic information related to the time-dependent variation in position, velocity and acceleration of particles in a many-body system. [16] The combination of the numerical solution of the equation of motion and the force field models that are often empirical approximations, equipped MD simulation with the means to evaluate the potential energy and interatomic interactions between particles. [7, 14, 17] This permits the investigation of the macroscopic properties of the system of interest in isolation as well as the complex phenomena of a myriad of reactions occurring simultaneously in the natural domain of the protein. [7, 14, 17, 24, 25]

In experiments, macroscopic properties such as binding affinity and protein stability, are not based on observations made on a single protein structure, instead these observations are averages acquired from a pool of configurations within a sample that is made up of millions to billions of atoms/molecules. [7, 14-17] Therefore, to convincingly model experiments using theoretical methods, sampling is a key aspect in MD simulations as better sampling will ensure that ample generation of microscopic states or representative conformations of the system at equilibrium is attained during the simulation. [16, 17] These microscopic states are linked to the macroscopic properties of the system through the implementation of statistical mechanics whereby the time average of possible systems having different microscopic states but similar macroscopic or thermodynamic states are evaluated to achieve observations that are of significant agreement to experiment. [16, 17] Hence, the use of MD simulations empower researchers to conduct experiments through the use of computers to explore thermodynamics and kinetics events occurring in proteins such as the binding of ligands to enzymes, configurational changes of proteins induced by the binding of a ligand, aggregation of proteins and enzymatic reactions. [5, 7, 8, 14, 15-17, 27-29]

The utilization of the classical MD simulation in the study of proteins is one of the common applications of MD simulation and will be the main subject of this thesis. This theoretical tool is a valuable gadget in the study of proteins due to the capacity of MD simulation to capture the rapidity and complexity of protein folding which are beyond the limits of time- and space-dependent resolution of experimental procedures such as X-ray crystallography, nuclear magnetic resonance (NMR) and nuclear Overhauser effect (NOE) spectroscopy. [14-18, 28, 29] With the establishment of principal force fields such as AMBER, CHARMM and GROMOS, intricate dynamics of proteins can be monitored at the atomic level using classical MD simulation, to comprehend the structure and function of proteins in greater detail. [15-17, 28-33] These force field models, being amino-acid specific in nature, have also enabled the universal application of these conventional force fields to a wide variety of protein types hence permitting the theoretical study of a diverse range of complicated, dynamic processes occurring in our biological systems such as protein folding, protein misfolding, enzymatic reactions, enzyme inhibitions, molecular recognitions and configurational changes of proteins. [7, 14-18, 27-29, 34-36]

1.3 Principles of molecular dynamics simulation

In this thesis, all MD simulations conducted were performed using the Amber 10 molecular dynamics simulation package which consists of a collection of programs that assist users in the modeling and simulation of biomolecules which in our case involves proteins. [37] *Sander*, a module in the Amber simulation package, enables users to relax the structure of biomolecules using minimization methods such as the steepest descent and conjugate gradient methods and facilitates the execution of MD simulations by implementing Newton's second law of motion [16, 17, 37]:

$$\mathbf{F}_i = m_i \ddot{\mathbf{r}}_i \qquad \mathbf{F}_i = -\frac{\partial U(\mathbf{r}_1,\ldots,\mathbf{r}_N)}{\partial \mathbf{r}_i} \tag{1.1}$$

where $\mathbf{F}_i$ is the accumulated force acting on the ith atom in a many-body system with the potential energy, U, of N interacting atoms defined as a function of the Cartesian coordinates of the atoms, $\mathbf{r}_i = (x_i, y_i, z_i)$. m_i represents the mass of the ith atom.

The determination of an appropriate potential energy which imitates realistically the potential energy surface of a system is one of the most significant obstacles in the MD simulation. The *sander* module in Amber 10 simulation package rectifies this problem to a certain extent by supporting a variety of force field models that permits the simulation of biomolecules such as proteins and nucleic acids and at the same time enables the modeling of non-biological elements such as water molecules and organic solvents. [37] The force field model adapted in the *sander* module captures the most fundamental description of the key features of atoms/molecules in the condensed phase and is represented in the following form [17, 37]:

$$\begin{aligned} U(\mathbf{r}_1,\ldots.,\mathbf{r}_N) &= \sum_{bonds} K_b\,(b-b_0)^2 + \sum_{angles} K_\theta\,(\theta-\theta_o)^2 \\ &+ \sum_{dihedrals}\left(\frac{V_n}{2}\right)(1+\cos[n\phi-\delta]) \\ &+ \sum_{nonbij}\left(\frac{A_{ij}}{r_{ij}^{12}}\right)-\left(\frac{B_{ij}}{r_{ij}^{6}}\right)+\left(\frac{q_i q_j}{\varepsilon r_{ij}}\right) \end{aligned} \tag{1.2}$$

where the summation of the first three terms represents the interatomic interactions between atoms bound together through the formation of covalent bonds while the last term represents pairwise interactions between atoms that are not chemically bonded to each other and are separated by a distance of $r_{ij} = r_i - r_j$. [17, 37]

The first two terms in equation 1.2 define the energies associated with the deviation of the bond length, b, and the bond angle, θ, from its equilibrium values, b_o and θ_o respectively, which are acquired from empirical data or *ab initio* calculations. [17] K_b and K_θ are the designated force constants for the energy terms related to a bond length and a bond angle respectively. The third term presented in the equation 1.2 represents restrictions imposed on rotations occurring around covalent bonds formed between atoms and are described by periodic energy terms such as the potential energy barrier, V_n, the periodicity which is determined by the value n, and the equilibrium dihedral angle, δ. [17]

Finally, the last term of equation 1.2 accounts for the non-bonded interaction between atom pairs. The Lennard-Jones 6-12 potential model and the Coulomb equation are subsets of the last term and are routinely used in various force fields to depict van der Waals (VDWs) and electrostatic interactions, respectively. The Lennard-Jones 6-12 potential model is a simple mathematical approximation of the interactions occurring between pairs of neutral atoms/molecules and the most typical expression of this model is [38]:

$$U_{LJ} = \sum_{i<j}^{atoms} \left(\frac{A_{ij}}{r_{ij}^{12}}\right) - \left(\frac{B_{ij}}{r_{ij}^{6}}\right) = \sum_{i<j}^{atoms} 4\varepsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right] \qquad (1.3)$$

where ε is the depth of the potential well and σ represents the distance between atoms at which the potential of the interatomic interaction is zero. [38] While the Lennard-Jones potential model settles the interactions between neutral particles, the Coulomb equation given in equation 1.2 deals with the interactions between atoms arising from the partial charges attached to the atoms which are predefined by the force field of choice. The parameters q_i and q_j in equation 1.2 represent the charges of the ith and jth atoms. All five energy terms discussed above are visually illustrated in Figure 1.1.

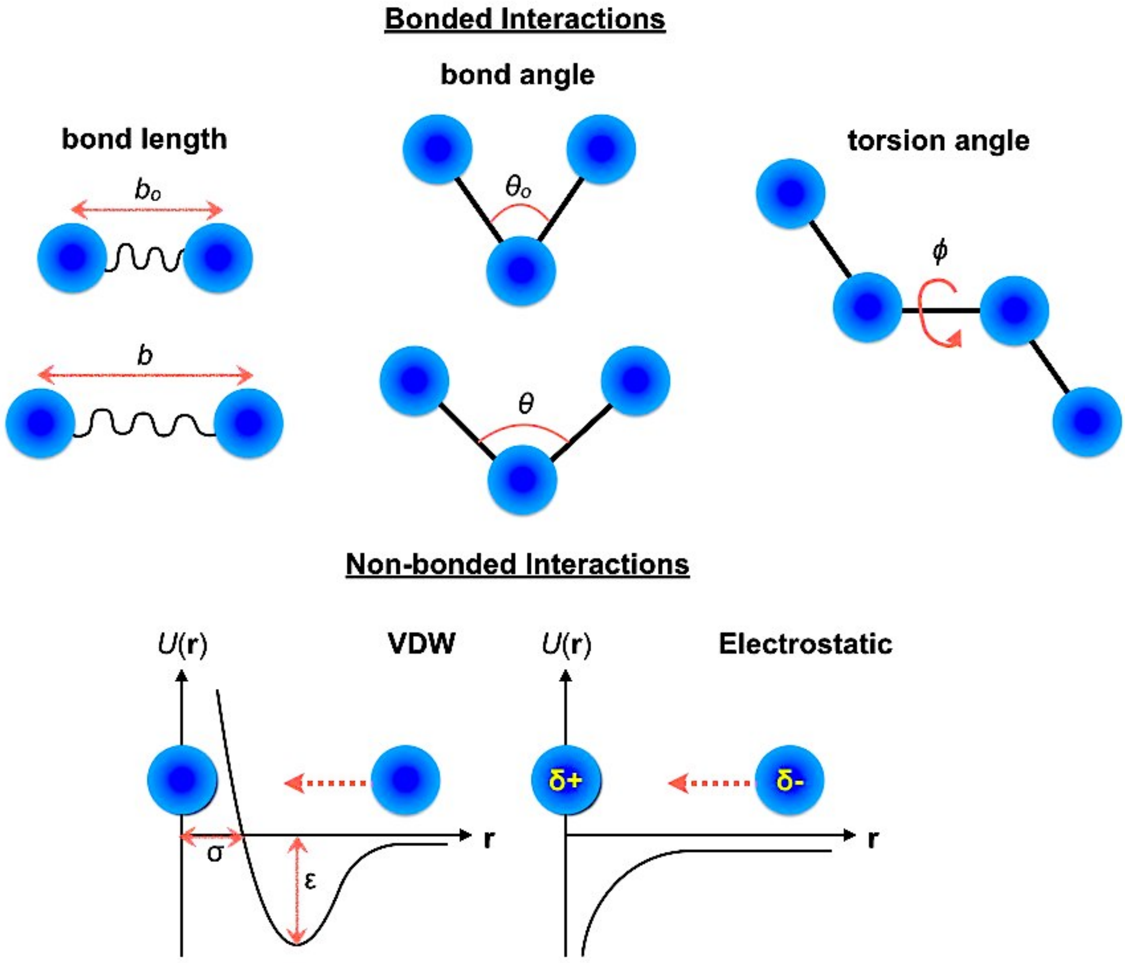

Bonded Interactions
bond angle
bond length
torsion angle
b_o
θ_o
ϕ
b
θ
Non-bonded Interactions
U(r)
VDW
U(r)
Electrostatic
σ
ε
r
δ+
δ-
r

Figure 1.1: Visual illustration of terms listed in equation 1.2 which contribute to the empirical potential energy function of the force field. (adapted from Scientific American Blog [39])

1.4 Current applications of molecular dynamics simulation in the study of protein dynamics

The advancement in MD simulations has seen the application of this useful technology to a wide variety of studies exploring underlying biological phenomena such as protein folding, protein aggregation, substrate channeling in proteins, protein-ligand interactions and protein-DNA interactions. [10, 27, 40] Compared to the first folding simulation of bovine pancreatic trypsin inhibitor which consists of around 500 atoms, current advancement in MD simulations has permitted the simulation of larger systems (~10^4 to 10^5 atoms) owing to the significant development in computational methods and the accessibility to powerful computing resources. [1, 41, 42] The use of MD simulations in protein researches is gaining significant recognition and some of the reasons which aptly summarize the growing popularity of this technology includes, (1) the ability to establish a relation between the structure and function of proteins, (2) the ability to provide thermodynamics information, (3) the ability to provide useful information related to intra- and intermolecular interactions which is an important factor in drug design and discovery and (4) the accessibility of affordable computational resources. [40] In order to celebrate the progress in MD simulation over recent years, this section will highlight some interesting applications of MD simulations which use modern MD methodologies and techniques to improve the quality of protein research in the virtual world.

The use of MD simulation in the study of protein folding is becoming commonplace in the theoretical field of protein research owing to the ability of MD simulations to offer key perspectives of the folding mechanism of proteins in nature with details at the atomic level. [40-42] The development of MD simulations has led to the introduction of complementary computational method such as replica-exchange molecular dynamics (REMD) which enhances the sampling efficiency of MD simulations as well as enables researchers to acquire thermodynamics information related to the folding process. [43, 44] For example, Lei et al. carried out the *ab initio* folding of a 35-residue villin headpiece using two sets of REMD simulations of 200 ns each and from the simulations conducted, the free energy landscape acquired from the REMD simulation provided details pertaining to the free-energy difference among different states of the protein, hence aiding in the determination of a feasible folding pathway for the villin headpiece. [43] Additionally, the use of REMD simulation had also empowered Lei et al. to fold the protein with accuracy which has not been achieved before by others. [43] The lowest C_α-RMSD of the folded villin headpiece compared to the experimental structure was 0.46 Å while previous *ab initio* folding simulations of the villin headpiece had only managed to achieve best C_α-RMSD of around 3.0 Å. [44-48]

Other than REMD simulation, steered molecular dynamics (SMD) simulation is another efficient sampling method which involves the application of an external harmonic force onto a protein to influence a change in its structure during MD simulations. [49, 50] Through the use of SMD simulations, changes that can only be observed by performing long simulations can be accelerated within a practical time scale and the free-energy changes of the studied process can be calculated from the SMD simulation conducted. [49, 50] Some examples of biological occurrences that can be explored using SMD simulations include, substrate channeling in proteins, unbinding of ligands from proteins and protein unfolding. [9, 10, 49-51] An example of the use of SMD simulation in the modeling of substrate channeling is the study conducted by Amaro et al. who simulated the channeling of ammonia from one active site to the next along a $(\beta/\alpha)_8$ barrel protein of HisH-HisF, a multi-domain globular protein complex. [9, 10] From this study, the potential of mean force for the channeling of ammonia through a gated channel in the known closed and proposed open configurations were reconstructed hence enabling the understanding of the mechanism governing the channeling of ammonia along the $(\beta/\alpha)_8$ barrel protein. [9, 10] On top of that, SMD simulation also allowed for the scrutinization of interactions formed between ammonia and conserved hydrophilic amino acids and water molecules within the predominantly hydrophobic channel hence providing insights on interactions which are crucial for the conductance of ammonia along the protein channel. [9, 10]

Other than MD methodologies that promotes efficient sampling, coarse-grained molecular dynamics (CGMD) simulation is another complementary tool of MD simulation which bridges experimental methods and molecular modeling. [52-54] Often times, the modeling of biological system using all-atom MD simulation is constrained by the simulation time and the size of the system of interest. [52-54] However, with the introduction of CGMD, the modeling of large systems such as membrane protein-lipid bilayer and protein aggregates are possible. [52-54] In CGMD simulation, a single coarse-grained (CG) bead represents a group of atoms, that is to say, the atoms of one amino acid can be represented by one CG bead or a monomer of a macromolecule can be represented by one CG bead. [52-54] Hence through the implementation of CGMD, the modeling of crucial biological processes such as protein aggregation from monomer to fibril can be examined thus enabling the better understanding of protein-related diseases such as Alzheimer's disease and diabetes mellitus. [53, 54]

While the methods discussed above are only some of the MD methodologies currently used, it aptly highlights the advancement in MD simulations over the years. With the continuous development of MD simulations, the variety of biological processes that can be explored using computers with atomistic precision will continue to broaden.

1.5 Outline of thesis

MD simulations, as highlighted (*vide supra*), enable theorists to study numerous complex dynamic processes taking place in biological systems [7, 11, 14-18] and in this thesis, the benefits of MD simulation will be exploited to investigate some of the dynamic processes of the biological system *viz.* protein folding and unfolding, variation in protein conformations and stability of protein. Additionally, the limitations of available force fields will also be explored through two separate studies whereby the importance of the protein environment in the modeling of protein folding and the reduction process of metalloproteins will be elaborated. The content of this thesis is organized into seven chapters:

Chapter 1 provides a brief overview of classical MD simulation with the basic theory related to the MD algorithm concisely described. The current applications of MD simulation in the study of proteins were also discussed.

Chapter 2 focuses on the current limitations of MD simulation which include sampling efficiency and force field. The importance of incorporating polarization effect into available force fields is also highlighted in this chapter through a brief description of a comparative study in which the stabilization of the β-structure of a prion protein was investigated by conducting simulations using both polarized (AHBC) and non-polarized (AMBER ff03) force field. Descriptions of AHBC, which is the short form for adaptive hydrogen bond specific charge scheme, are also furnished in this chapter. A brief explanation of REMD simulation is also provided in this chapter.

Chapter 3 explores the use of REMD simulation to investigate the mechanism dictating the structural variation of a protein from an α/4β-fold to a 3α-fold. The proteins involved in this study are two engineered proteins named GA88 and GB88, which have a "sequence identity" of 88%. In this chapter, the "protein folding problem" will also be highlighted and is the main motivation for this structural variation study.

Chapter 4 scrutinizes the importance of polarization effect in the folding of an extended helical structure termed 2khk. The *ab initio* folding of 2khk using both polarized (AHBC) and non-polarized (AMBER ff03) force field will be discussed in this chapter. The influence of polarization effect in precisely folding long helical domains will be further corroborated through the folding simulations of two additional helical peptides namely N36 and C34.

Chapter 5 further emphasizes the importance of electrostatic interactions in the theoretical study of proteins. In this chapter, the role of the electrostatic environment in accurately determining the reduction potential of rubredoxin, an iron-containing metalloprotein, will be showcased through the use of two charge derivation schemes which differ by the number of residues considered for quantum mechanical (QM) calculations. An increase in the number of residues subjected to QM calculation concomitantly leads to a greater consideration of the electrostatic environment around the iron metal center. The charges derived through the two charge schemes are incorporated into the force field used and conventional MD simulation with the implementation of thermodynamic integration will be conducted to derive the reduction potential of three rubredoxin mutants, namely L41A, V44A and V44G, relative to a wild type protein.

Chapter 6 describes a simple study conducted using MD simulation to determine the stability of a protein relative to wild type upon mutation. In this study, conventional MD simulations of native apomyoglobin (apoMb) and mutated apoMb are conducted in explicit 2 M urea solution to effectuate the denaturation of the proteins. Analyses related to the configurational fluctuations of the proteins will be used to determine the stability of apoMb upon mutation. Significant agreement between theoretical observation and experimental data is attained through this study hence showcasing the practicality of using MD simulation for the determination of protein stability upon mutagenesis.

Chapter 7 provides a summary of all the studies that have been discussed in this thesis.

Chapter 2 Current problems in MD simulation

2.1 Current limitations in MD simulation

While the previous chapter highlights the benefits of MD simulation, this chapter will cover two limitations of MD simulation that has concerned the MD community. These two highly concerned drawbacks of MD simulation include [1-3]:

(i) the need for more computational resources to generate sufficient sampling to study protein folding.

(ii) the need to improve available force fields so as to accurately model proteins in nature.

2.1.1 Improvement in sampling efficiency

The majority of MD simulations conducted hitherto are limited to the microsecond time scale which may not suffice in the study of protein folding where rigorous sampling of configurational states is required. [1-2] However, with the continuous search for computer algorithms that permit longer simulation time scales, limitations arising from the lack of computer resources to carry out sufficient conformational sampling can be rectified in the near future. [1-4] Recently, millisecond simulations have been reported by Shaw et al. who used a self-built supercomputer named Anton which was built with the intention of conducting one microsecond of simulation per day. [3] The development of Anton is seen as a stepping stone in the field of MD simulations whereby simulations in the millisecond time scale will permit better sampling of conformational states that enables the reliable observation of protein dynamics crucial for the investigation of protein folding and unfolding and the binding of drugs to receptors for drug discovery purposes. [1-3]

As mentioned in the previous chapter, sampling efficiency has been continuously improved through the implementation of MD methods such as replica-exchange molecular dynamics (REMD) simulation. In examining the dynamics involved during protein folding, one significant obstacle faced by theorists is the difficulty in achieving authoritative distributions of structures from MD simulation conducted at low temperatures. [5] This is due to the tendency of these simulations to be entrapped in the local energy minimum wells of the potential energy surface of the system. [5] The introduction of the REMD algorithm proposed by Sugita and Okamoto rectified this problem by incorporating the parallel tempering method, an algorithm commonly used in Monte Carlo simulations, into the MD algorithm. [5-11] The use of REMD simulations in the folding study of proteins is fairly common due to the ability of REMD algorithm to promote the system to perform replica exchanges through energy space. [5] This helps the system to escape local energy minima states while boosting the search for conformations conforming to the global energy minimum state by broadening the conformations space sampled by the simulation. [5, 12-15]

REMD simulations, in general, begin with the selection of *N* independent replicas of a system which is simultaneously subjected to isothermal MD simulations at different temperatures. [5] Subsequently, exchanges between two neighboring replicas of the system attained at different temperatures are attempted after a predefined number of simulation steps and the success of this exchange is governed by the Metropolis criterion [5]:

$$\begin{aligned} w(X \rightarrow X') &\equiv w(x_m^i | x_{m+1}^j) \\ &= \min\{1, \exp(-\Delta) \\ where\ \Delta &= \left(\frac{1}{kT_{m+1}} - \frac{1}{kT_m}\right)(E_i - E_j) \end{aligned} \quad (1.4)$$

Based on the equation above, the proposed exchange of replicas i and j which have neighboring temperatures of T_m and T_{m+1} is accepted if Δ is less than or equal to 0. [5] Meanwhile, if Δ is greater than 0, the acceptance ratio for the replica exchange between i and j will be governed by exp(-Δ) in equation 1.4. [5] Successful exchanges occurring between replicas during REMD simulation promote the accessibility of high energy configurations to simulations conducted at low temperatures and reciprocally. [5] This equips researchers with the ability to sample both low and high energy conformations during REMD simulations hence contributing to the widening of the conformational space sampled. [5] A schematic explanation of replica-exchange and the effective sampling of configurations at both low and high temperatures during REMD simulation is illustrated in Figure 2.1.

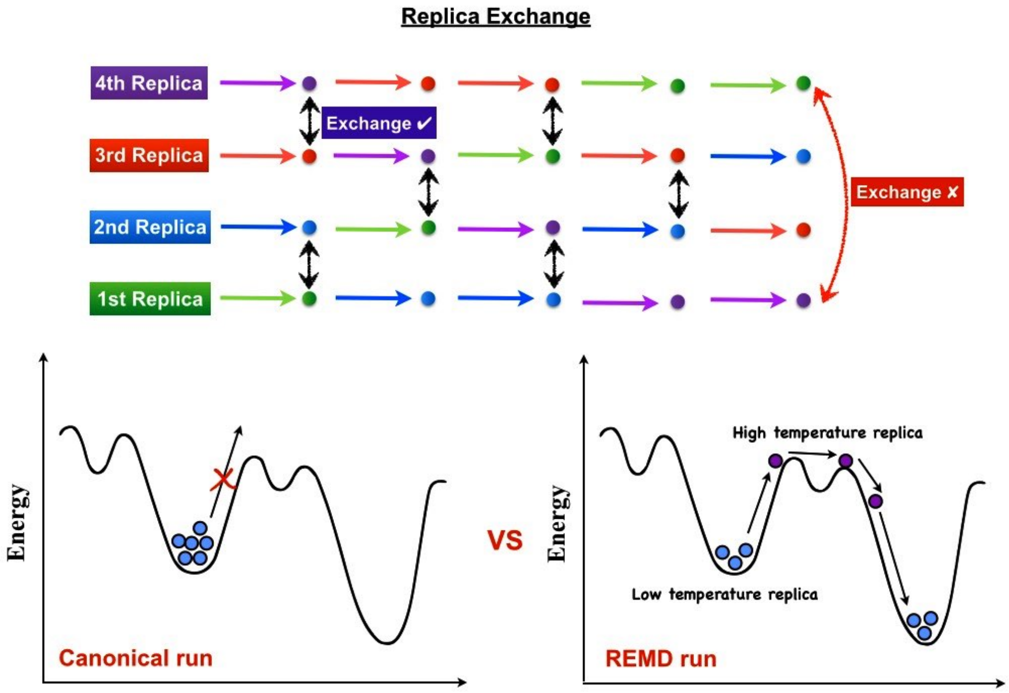

Replica Exchange
4th Replica
3rd Replica
2nd Replica
1st Replica
Exchange ✓
Exchange ✗
Energy
Canonical run
VS
Energy
High temperature replica
Low temperature replica
REMD run

Figure 2.1: Schematic representation of replica exchange happening during REMD simulation. (adapted from http://www.weizhang.us/replica-exchange-simulation/)

2.1.2 Improvement in force fields

Another problem that researchers face in the use of MD simulation to model protein dynamics is the availability of force fields that can precisely model protein dynamics in nature. Traditional force fields such as AMBER, CHARMM and GROMOS, are often based on semi-empirical parameters which approximate the quantum mechanical properties of the atoms of the protein. [1, 16-19] Albeit the successful application of these traditional force fields in the investigation of protein dynamics, these force fields are not suitable for simulations where non-protein atoms are involved such as small drug molecules and transition metals. [1] Another widely known shortcoming these principal force fields have, is the absence or lack of electronic polarization effect considered for the protein. [16-19] The charges utilized in traditional force fields are acquired through the charge fitting of the electrostatic potential (ESP) of individual amino acids, without considering the polarization effect exerted by the protein frame. [20] That is to say, the charges of an amino acid used in the aforementioned force fields are static and are not influenced by changes in the environment of the amino acid. While such an approximation is practical in some cases, most of the time, this approximation may lead to the erroneous observation of protein dynamics such as protein folding whereby the consideration of polarization effect is greatly needed to account for the rapid changes in conformation that lead to the constant changes in the electrostatic environment of the amino acids.

Here, we will highlight a study conducted by us to emphasize the importance of polarization effects in effectively modeling proteins. [21] We conducted a comparative study involving the modeling of the elongation of the β-sheet of prion protein during the initial transition stage of PrP^{C} to PrP^{Sc} using two force fields namely the AMBER ff03 force field and another force field that incorporates charges derived using a newly developed on-the-fly charge update scheme termed adaptive hydrogen bond-specific charge (AHBC) scheme which will be elaborated in another section of this chapter. [21] The AHBC scheme took into account the changes in the electrostatic environment of the protein caused by the formation and/or disruption of hydrogen bonds and this action led to the better description of the secondary structures of the protein. [21] To observe the elongation and/or disruption of the β-sheet in prion during the simulation, the distances between 127O and 165N (magenta), 131O and 161N (blue) and 133N and 159O (green), illustrated in Figure 2.2, were calculated to monitor the formation of hydrogen bonds between the aforementioned hydrogen bond pairs. [21] These hydrogen bonds may contribute to the elongation of the β-sheet of the prion protein of rat, human and bovine. [21]

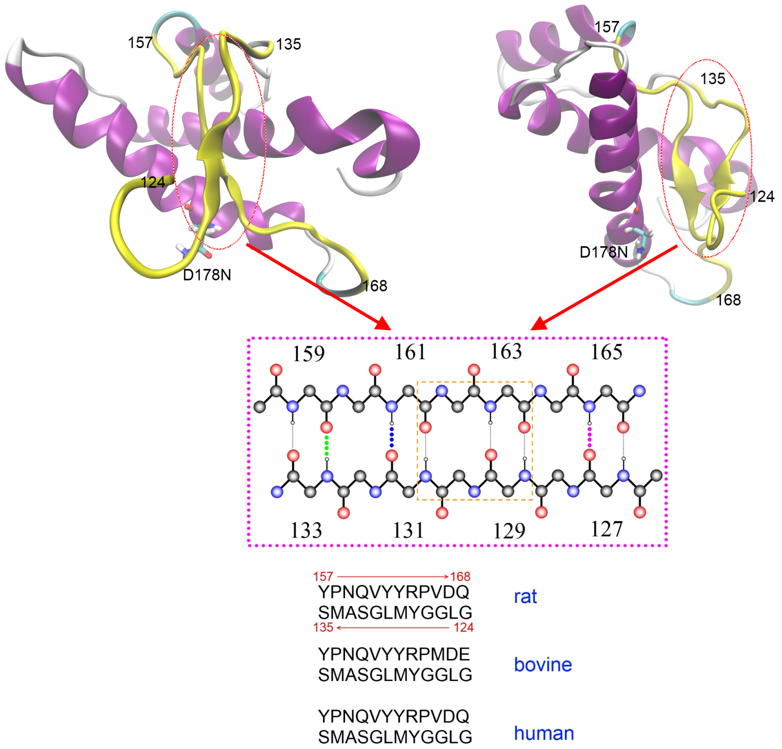

157
135
124
D178N
168
157
135
124
D178N
168
159
161
163
165
133
131
129
127
157
168
YPNQVYYRPVDQ
SMASGLMYGGLG
135
124
rat
YPNQVYYRPMDE
SMASGLMYGGLG
bovine
YPNQVYYRPVDQ
SMASGLMYGGLG
human

Figure 2.2: Cartoon representation of the prion protein with the region coloured yellow being subjected to the on-the-fly AHBC scheme. (*above*) Schematic illustration of possible hydrogen bonds formed at the β-sheet upon mutation of the prion protein with the hydrogen bonds formed in the native protein enclosed in the orange dotted box. (*middle*)

The sequences of the amino acids that are treated using AHBC scheme for rat, bovine and human are provided. (*below*) [21]

From Figure 2.3, simulations conducted using the AHBC scheme showed the formation of hydrogen bond between 131O and 161N (blue) for all three species hence signifying the elongation of the β-sheets across all species which are in agreement with experiment. [21] Additionally, hydrogen bond between 133N and 159O (green) was also established for the prion proteins in bovine and human but not in rats. On the contrary, the simulations conducted using AMBER ff03 showed no elongation of the β-sheet of the prion proteins in rat and bovine, evident from the lack of hydrogen bonds formed between 127O and 165N (magenta) and 131O and 161N (blue). However, the latter is present in human prion protein. [21]

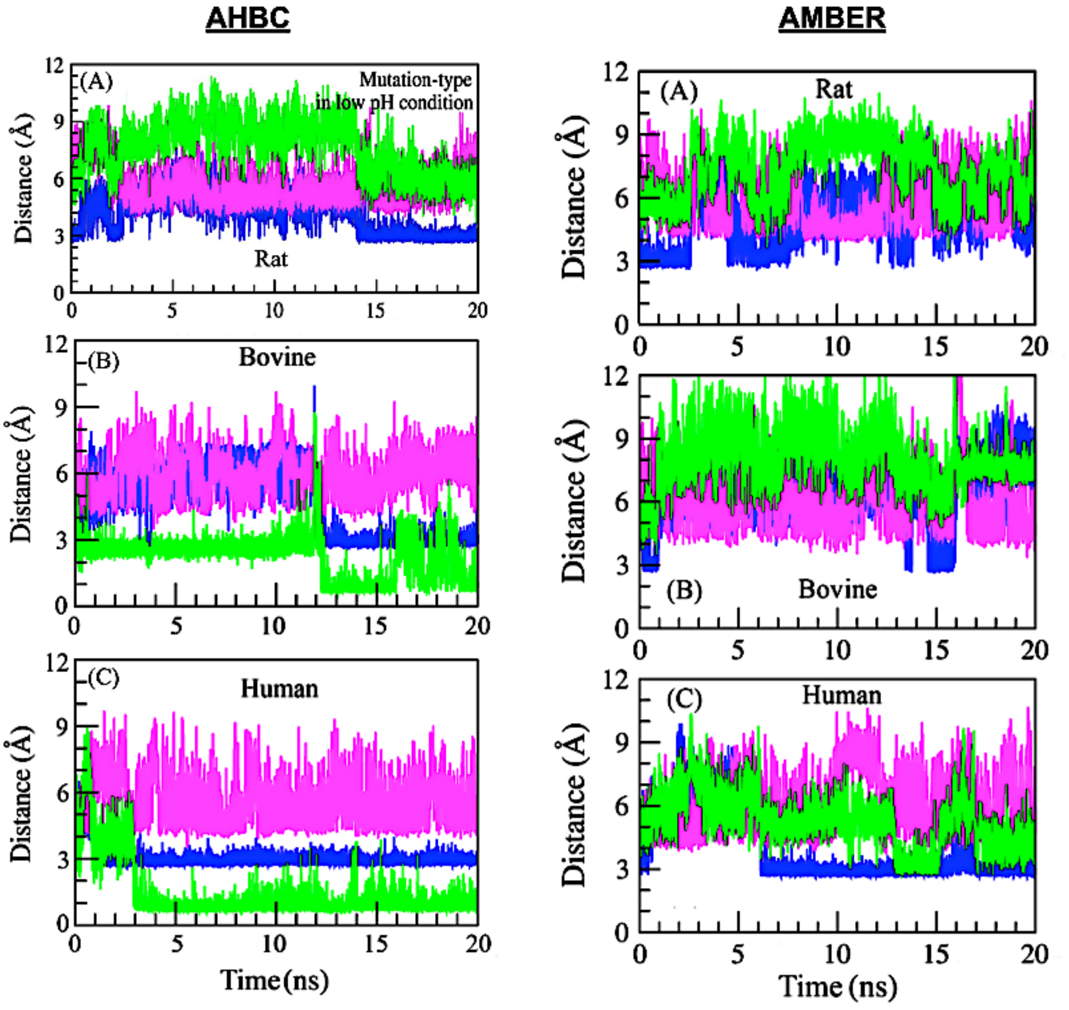


Figure 2.3: **(*left*)** Changes in distance between 127@O and 165@N (magenta), 131@O and 161@N (blue) and 133@N and 159@O (green) from MD simulations conducted using the on-the fly AHBC scheme for PrP^{Sc} of (A) rat, (B) bovine and (C) human at acidic pH. The plots for the distance between 133@N and 159@O (green) in (B) and (C) are shifted down by 2.0 Å to prevent overlaps between plots. **(*right*)** Changes in distances between 127@O and 165@N (magenta), 131@O and 161@N (blue) and 133@N and 159@O (green) from simulations conducted using AMBER force field for PrP^{Sc} of (A) rat, (B) bovine and (C) human. [21]

The elongation of the β-sheet was also monitored using DSSP analysis presented in Figure 2.4. DSSP (**D**efine **S**econdary **S**tructure of **P**roteins) is a conventional procedure employed to assign secondary structures to the amino acids of a protein. [22] As seen in Figure 2.4 below, the β-sheets of all three prion proteins were stabilized when AHBC scheme was implemented. However, the simulation of the prion proteins using mean field showed the destabilization of the β-sheet of the prion proteins especially for rats and bovine. [21] The disparity in observations noted for simulation conducted using AHBC and mean-field, with the former providing better agreement with studies proposing the elongation of the β-sheet during the transition of PrP^{C} to PrP^{Sc}, highlights the importance of incorporating polarization effect into force fields to accurately model the dynamic behavior of protein parallel to nature. In another section of this chapter, details with regard to the AHBC scheme will be provided to furnish readers with the basic information necessary to understand the charge update scheme.

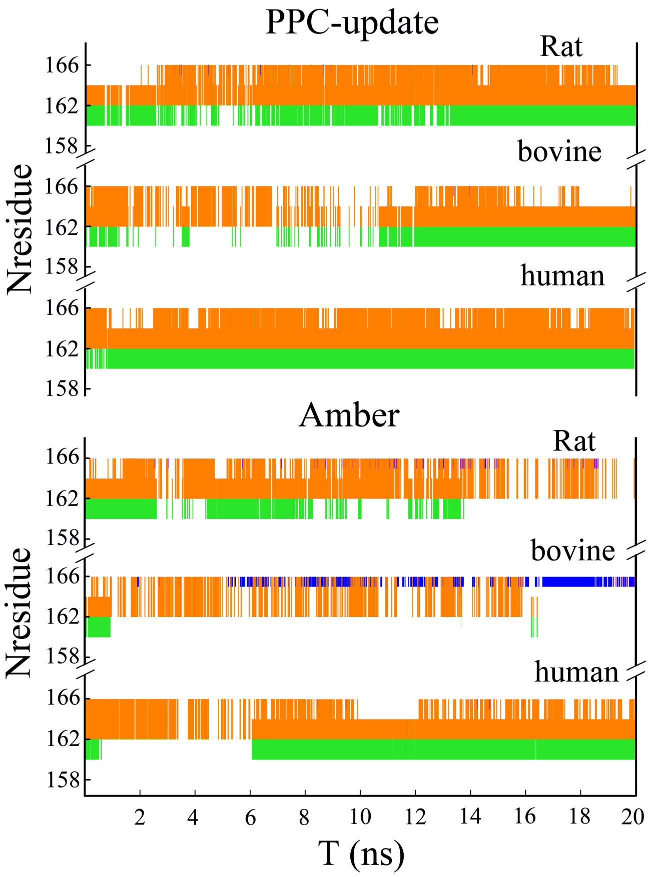


Figure 2.4: Changes in the secondary structures of the β-sheets of the prion proteins acquired through DSSP assignment of simulations performed using AHBC scheme and AMBER ff03 force field at acidic pH. β-structures which are originally not present in the starting structure are coloured green. [21]

2.2 Development of polarizable and polarized force fields

Availability of traditional force fields such as AMBER, GROMOS and CHARMM has empowered researchers to explore the structure and function of a wide range of proteins at the atomic level. [16-18] However, the amino acid-specific nature of these popular force fields gives rise to a key flaw that rendered these force fields incapable of providing a precise electrostatic description of amino acids under different environmental conditions. [23-28] Regardless of changes in the dielectric environment, atoms modeled using the aforementioned pairwise additive force fields are designated as fixed charges, ignoring any changes in the polarizability of the atoms. [23-28] This shortcoming can be rectified through the introduction of polarizable/polarized force field that considers the non-uniform polarization of amino acids in different physical domains hence furnishing researchers with a reasonably accurate model of protein in nature. [23-28]

The implementation of polarizable force fields permitted the variation of the charge distribution of the protein atoms as a response to changes in the dielectric environment. [23] Traditional force fields such as AMBER, CHARMM and OPLS have introduced complementary polarizable force fields by adding an additional "polarization" energy term into the total potential energy. [23-30] There are three common methods in which polarization is accounted and these include induced dipole model, fluctuating charge model and Drude oscillator model. [23, 27-30] The induced dipole model is one of the more popular models used to account for polarization in protein systems and is currently used in the polarizable counterparts of AMBER and AMOEBA force fields. [23, 27, 29] In the induced dipole model, each atom is assigned an inducible point dipole based on its fixed charges and the induced dipole moment of the atom is equivalent to the polarizability of the atom and the electric field exerted on the atom by other induced dipoles. [23, 27] In the fluctuating charge model, the atomic charge is allowed to vary in response to changes in the environment following the principles of electronegativity equalization. [23, 28] This principle states that charge transfer is permitted between atoms until the instant at which similar electronegativity is reached for both atoms. [23, 28, 29] The Drude oscillator model, on the other hand, describes polarization effect by including a supplementary atom to each polarizable centers and these two charged particles are connected by a harmonic spring. [29, 30] In this model, atomic polarizability is mirrored by the changes in the displacement between the two charge points effectuated by the variation in the local electric field of the atom. [29, 30] Even though these models have different protocols, there is one factor in common which is to initiate a molecular response towards the variation in the surrounding environment. [23, 24, 26-30]

While continuous developments are being made for polarizable force fields, other methods aiming to incorporate polarization effect into force fields have been developed. In recent years, Zhang et al. had designed a charge fitting procedure termed polarized protein-specific charge (PPC) scheme which had enabled the building of the polarization effect of the protein's native state into the atomic charges of the protein atoms. [25, 31] In the PPC scheme, quantum mechanical (QM) calculations of proteins in continuum solvation model are performed by combining molecular fractionation with conjugate caps (MFCC) approach and Poisson-Boltzmann solvation model. [25, 31-34] Through the implementation of MFCC, full quantum calculations can be conducted for large proteins in its native or selected states by fragmenting the protein into individually capped amino acids and performing separate QM calculations on each of the capped fragments. [25, 31-33] In this way, the computational expenses associated with the full QM calculations of proteins will be significantly reduced as the computational expenditure of MFCC method scales linearly with the number of fragments used, making it practical for application to large proteins. [25, 31-33, 35] There are two advantages associated to the use of the PPC scheme in accounting for the polarization of the protein system:

(i) The PPC scheme embodies the simplicity of conventional force fields as polarization effect is incorporated into the atomic partial charges. [25, 31-33] This ensures the straightforward application of PPC in MD simulations as one will only need to replace the static AMBER charges with that of PPC while keeping all other parameters constant. [25]

(ii) The determination of the partial charges using *ab initio* methods ensures that additional error caused by the addition of empirical information is prevented. [25] This is contrary to the polarizable force fields discussed (*vide supra*) which involved the inclusion of additional polarization energy terms into the total potential energy of the system. [23-28]

At present time, the PPC scheme has been implemented in numerous studies pertaining to the structures and dynamics of proteins. Ji et al. had successfully predicted the shift in pKa for buried Asp close to the experimental value through the employment of the PPC scheme in free energy perturbation simulations. [25] Similar calculations using the standard charges of AMBER and CHARMM gave rise to a value approximately twice that of experiment. [25] Besides this study, Ji et al. also utilized the PPC scheme in MD simulation to examine the dynamics involved in the binding of rosiglitazone to PPAR-γ. [36] The utilization of polarized force field in this study saw the stabilization of the protein-ligand complex. [36] Furthermore, the use of the PPC scheme has led to the stabilization of structural features crucial for the binding stability of ligand-PPAR-γ complex through the conservation of a critical hydrogen bond which was broken when non-polarized force field was used. [36] The observations achieved using PPC scheme concurred well with experiment. Through another study carried out by Tong et al., the successful application of the PPC scheme in the MD simulations of NMR order parameters of five proteins was proven through the consistency achieved between theoretically determined order parameters and that of experiment. [37] Similar to the study conducted by Ji et al., the ability to reproduce experimental observations was attributed to the ability of the PPC scheme to stabilize hydrogen bonds through the consideration of electronic polarization. [37] This prevents the excessively large fluctuations of structural domains such as coils and loops as was seen in MD simulations conducted using mean-field. [37] In a recent publication by Wei et al., the use of PPC scheme has also enabled the accurate determination of the reduction potential of mutated azurin proteins relative to that of the wild type protein. [38] The calculated relative reduction potential values of the mutated proteins acquired through free-energy perturbation simulations came close to the experimental values with the consideration of polarization effect using the PPC scheme. [38] These highlighted studies are some of the published researches detailing the successful implementation of the PPC scheme in the study of the structure and dynamics of proteins. The PPC scheme has also been applied to studies pertaining to binding affinities and structural refinement. [39-41]

2.3 Development of adaptive hydrogen bond-specific charge (AHBC) scheme

While the use of PPC scheme has seen far-reaching advantages towards studies focusing on the structural, thermodynamics and kinetic properties of proteins, the use of PPC in the *ab initio* folding of protein may not be feasible as protein folding routinely entails significant structural diversification. [31, 42] Frequent variations in the protein structure during folding leads to the fluctuations of the atomic partial charges due to the continuous changes in the physical environment experienced by the atoms in the protein. [31, 42] As such, by using PPCs acquired based on the native state or selected structures, we will be instilling an inherent bias for the native fold into the force field which may consequently obscure the actual folding pathway of the protein. [25, 31] Therefore, to consider the changing polarization effect of the protein during folding, charge update of all atoms for every time step will be necessary, but this action may give rise to an expensive computational expenditure which will not be feasible for large proteins. [31] On that account, it is crucial for us to discern sets of atoms that are crucial guides for protein folding as limiting the charge updating scheme to these atoms may likely lower the computational cost without compromising accuracy. Adaptive hydrogen bond-specific charge (AHBC) scheme adapts the rationale outlined above, giving rise to an on-the-fly charge update scheme which provide a precise description of the electrostatic polarization effect of hydrogen bonds in its instantaneous physical environment by incorporating the polarization effect of the hydrogen bonds into the atomic charge of the hydrogen bond donor and acceptor. [31]

Hydrogen bond is one of the main interaction that governs the development and the stabilization of the secondary structures of proteins such as α-helix and β-sheet. [31, 42] During the formation of hydrogen bonds, the electron clouds of the hydrogen bond pair are distorted and the fixed charges used in traditional force fields such as AMBER are not able to provide the energetic description for this phenomenon. [42] The AHBC scheme aims to provide an accurate description of the interaction energies between atoms of hydrogen bonds by exploiting the MFCC approach to perform QM calculations on capped amino acids involved in hydrogen bonding with the consideration of solvation effect by solving the linear Poisson-Boltzmann equation. [32-34, 42] In the PPC scheme, a pair of conjugate caps (CH_2R_1CO-$NHCH_2R_2$) is use to cap the fragments at the two ends where the peptide bonds between two amino acids are cut to form the capped amino acid fragments. [32, 42] The conjugate caps replace the atomic position previously occupied by the cutoff protein to ensure minimal introduction of artifacts. [32, 42] The same protocol is also practiced in the AHBC scheme with the addition of another conjugate cap ($HCONH_2$-$HCONH_2$) that mimics the hydrogen bond that has been cut off. [32, 42] The MFCC treatment to fragments generated by the PPC and the AHBC scheme is illustrated in Figure 2.5 for better understanding of the MFCC method. Besides the MFCC treatment to protein fragment shown in Figure 2.5, there are other types of conjugate caps used in the AHBC scheme to seal the amino acid fragments generated. [42]

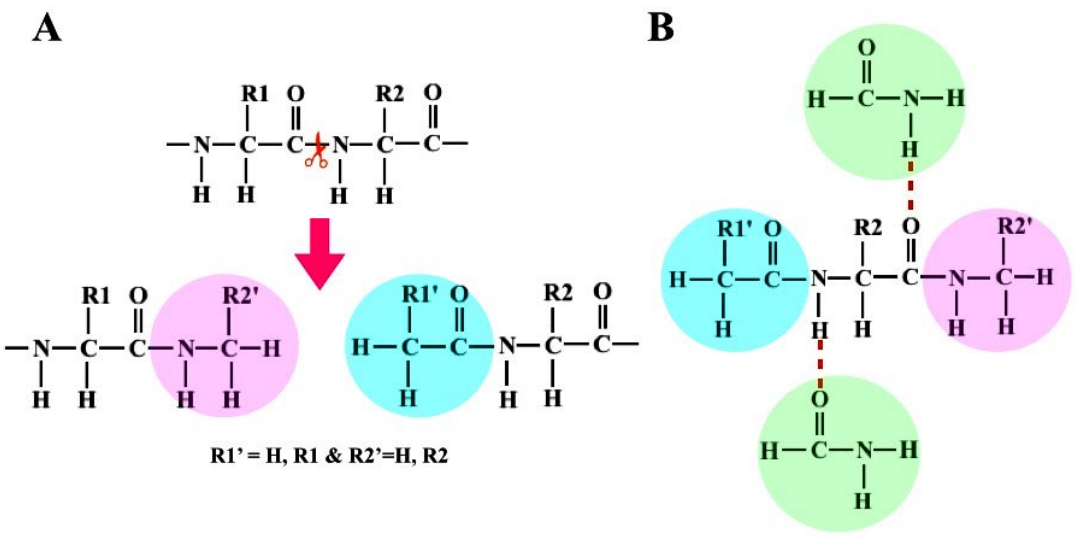


Figure 2.5: MFCC method. (A) Amino acid fragments generated by cutting the peptide bond are capped at the ends using the conjugate cap, $CH_2R_1CO\text{-}NHCH_2R_2$. (B) Additional conjugate caps, $HCONH_2\text{-}HCONH_2$ are used to mimic hydrogen bonds in the AHBC scheme. The dotted red line represents the hydrogen bond. (adapted from Ref. 42)

Numerous studies highlighting the benefits of inculcating inhomogeneous polarization effect into MD simulations have been conducted and all these studies have shown encouraging results such as the enhanced stabilization of folded protein close to the native state and the stabilization of crucial intra-protein hydrogen bonds. [20, 31, 36, 37, 41, 43, 44] These studies highlight the significance of polarization effect in the stabilization of the secondary structure of proteins and hence justify the choice of updating the atomic charges of amino acids involved in hydrogen bonding in the AHBC scheme. [20, 31, 36, 37, 41, 43, 44] The application of AHBC scheme in MD simulations investigating protein folding is also highly practical given the intrinsic nature of protein folding which requires the continuous forming and breaking of hydrogen bonds. Charge update of the amino acids participating in development and/or disruption of hydrogen bonds at fixed interval ensures that the changes in the physical environment of the hydrogen bond donor and acceptor are constantly updated as this will augment the electrostatic interaction between the hydrogen bond pairs resulting in the stabilization of hydrogen bonds that are vitally crucial in maintaining the structural integrity of α-helices and β-sheets of proteins.

The AHBC scheme has been utilized in several studies involving protein folding. Duan et al. successfully fold a 17-residue helical peptide using the AHBC scheme with lowest backbone RMSD of 0.5 Å, relative to the NMR structure, achieved. [31] A parallel MD simulation using AMBER charges, on the other hand, did not succeed in achieving a folded structure close to the native state. [31] The AHBC scheme has also been utilized by Wei et al. to differentiate the helical propensity of three polyalanine variants. [45] In this study, the folding of polyalanine variants using the charges derived from the AHBC scheme saw the folding of these helices to achieve helical content in agreement to that of experiment. [45] However, when the default partial charges of the AMBER force field were used, the helical contents of the three helices folded was similar to each other hence emphasizing the importance of considering explicitly the polarization along a hydrogen bond to attain folding which are sequence-dependent. [45] Another published research that has utilized the AHBC scheme is a study conducted by Xu et al. who examined the effect of solvent on the folding of E6aP. [46] Through this study, the difference in the propensity of E6aP to fold into a helical structure in TFE solvent and water was reproduced. [46] The lowest backbone RMSD acquired for E6aP folded in TFE solvent is 0.9 Å relative to the NMR structure while the helical content of E6aP folded in water was calculated to be 23% which is close to the experimental value acquired through CD spectroscopy (20%). [46] In this thesis, using similar simulation methods as Duan et al, the

AHBC scheme was successfully used to fold three helical peptides and this study will be elaborated in Chapter 4.

Chapter 3 All-atom molecular dynamics simulation of structure variation from α/4β-fold to 3α-fold protein

3.1 Introduction

3.1.1 Protein folding problem: Protein misfolding

The “protein folding problem” remains as one of the main highlights of researches concerning protein folding over the last fifty years, both experimentally and theoretically. [1-5] Through the studies conducted to understand this problem, insights pertaining to protein misfolding, which is the main causing factor of protein-related diseases, has also been acquired, facilitating the comprehension of folding diseases such as diabetes mellitus, Alzheimer’s disease and Parkinson’s disease. [1-5] The questions posed through “the protein folding problem” cover three areas in protein research which are highlighted by Dill et al. in Ref. 1. The questions raised include [1]:

(i) How does the physicochemical data contained in the primary sequence of a protein governs its three-dimensional structure?

(ii) How does a protein fold rapidly in the microsecond time scale with multitudes of probable conformations possibly assumed by the protein during protein folding?

(iii) Is it possible to innovate a computer algorithm that predicts the three-dimensional structure of a protein on the basis of its amino acid sequence?

Proteins are complex biological molecules that fold effectively into their native conformations using information stored in their primary sequence, i.e. the primary sequence of a protein determines its tertiary conformation. [6-13] The folding of proteins into their native structures are generally governed by non-covalent interactions such as hydrogen bonds, electrostatics, van der Waals and hydrophobic interactions. [6-8] The stability of the overall three-dimensional configuration of a protein will ensure the effective functioning of the protein in its native domain. [6-8] Almost all known functions of proteins have crucial influences on biological processes responsible for life sustenance. [6-8] Some of the critical functions of proteins include the involvement of proteins in cell signaling and biochemical reactions as well as being a part of the structural and mechanical features of cells. [6-8] These functions reflect the importance of proteins as an indispensable component in all living organisms hence highlighting the importance of understanding the structure and function of proteins in greater detail. [6-18]

Efforts focusing on the decoding of the amino acid sequence to predict the tertiary structure of proteins are rampantly carried out to answer the “protein folding problem.” [4, 5] Rose et al. offered an interesting viewpoint on this question by suggesting an alternative tactic which involves the understanding of the relationship between amino acid sequence and its propensity to favor one fold over the other. [10] This proposal has led to various attempts to take on the challenge of engineering a protein pair with primary sequences that have “high sequence identity” but vary in terms of native folds. [3, 5, 10, 16] These attempts are made with the aim of deducing the minimum variation that can be made on the primary sequence of a protein to change its inclination for one configuration over another. [3, 5, 10, 16] In a study conducted by Alexander et al., the connection between a single mutation and a switch in three-dimensional conformation of a protein was established with the engineering of a protein pair that has a 95% sequence identity. [3]

This type of study is especially useful in harnessing information that is beneficial towards the understanding of protein misfolding which is one of the main causes of numerous diseases in humans such as diabetes, Alzheimer's disease and Creutzfeldt-Jakob disease. [17-21] Protein misfolding often occurs through point mutations and these studies have provided us with some insights on how a single change in the amino acid sequence of a protein may result in the alteration of its tertiary structure from the native conformation.

While the above paragraph highlights the experimental effort put into understanding the translation of the primary sequence to the tertiary structure of a protein, theoretical investigations have also been launched to predict the three-dimensional conformation of a protein from its amino acid sequence through the development of computer algorithms. [1] In the year 1994, an event named CASP (**C**ritical **A**ssessment of Protein **S**tructure **P**rediction) was inaugurated and has since been organized every two years to challenge research groups to develop computer algorithms to predict the three-dimensional structures of primary sequences which structures are known but not released publicly as yet. [1] From these CASP events, it was realized that most successful predictions were based on the hypothesis that similar primary sequences result in conformations which are essentially the same. [1] The innovation of computer algorithms capable of accurately determining the structure of a protein from its primary sequence helps to ease the laborious experimental work related to protein structure prediction and hence expedites experiment-based researches associated with drug discovery and determination of protein structures. [1]

In the next section, REMD simulation, a commonly used computational tool in protein folding studies, was used to explore the mechanism governing the transition of a protein from an $\alpha/4\beta$-fold to a 3α-fold conformation. Through this study, some insights pertaining to protein misfolding could be attained by monitoring the conformational transition of the β-sheet to α-helix. The limitations of MD simulation that arise from this study were also briefly discussed to emphasize the importance of polarization effect in the modeling of proteins as discussed in the previous chapter.

3.1.2 All-atom simulation of the evolution in protein structure from α/4β-fold to 3α-fold pattern

In this study, the structural evolution of a protein from α/4β-fold to 3α-fold was scrutinized through the application of REMD simulation. GA88 and GB88 are two G proteins with a “sequence identity” of 88%, designed by He et al. through the mutation of 24 residues and 17 residues of their respective wild type proteins respectively. [5] As illustrated in Figure 3.1 below, the GA88 protein consists of three α-helical domains labeled H1 (residues 9 to 23), H2 (residues 27 to 34) and H3 (residues 39 to 51) while its partner GB88, contains one α-helical domain labeled HB (residues 23 to 36) and four β strands labeled B1 (residues 1 to 8), B2 (residues 13 to 20), B3 (residues 42 to 46) and B4 (residues 51 to 55). [5]

GA88, which has a native 3α-fold conformation, will be the main focus of this study as GB88 was merely used as a blueprint for the modeling of the starting structure which possessed the primary sequence of GA88 encompassed in the native α/4β conformation of GB88. Using REMD simulation, the mechanism directing the structural evolution of GA88 from the non-native α/4β-fold to its original 3α-fold configuration was explored. The use of a starting structure that has the primary sequence of GA88 and a three-dimensional configuration homologous to GB88 ensures the structural evolution of the protein from α/4β-fold to 3α-fold during the simulation. Since the structure used at the start of REMD simulation conducted is not the actual GB88 protein, the protein will be termed α/4β-GA88 in this study. On the other hand, the resulting protein from the REMD simulation, which possesses the wild type conformation of GA88, will be termed 3α-GA88.

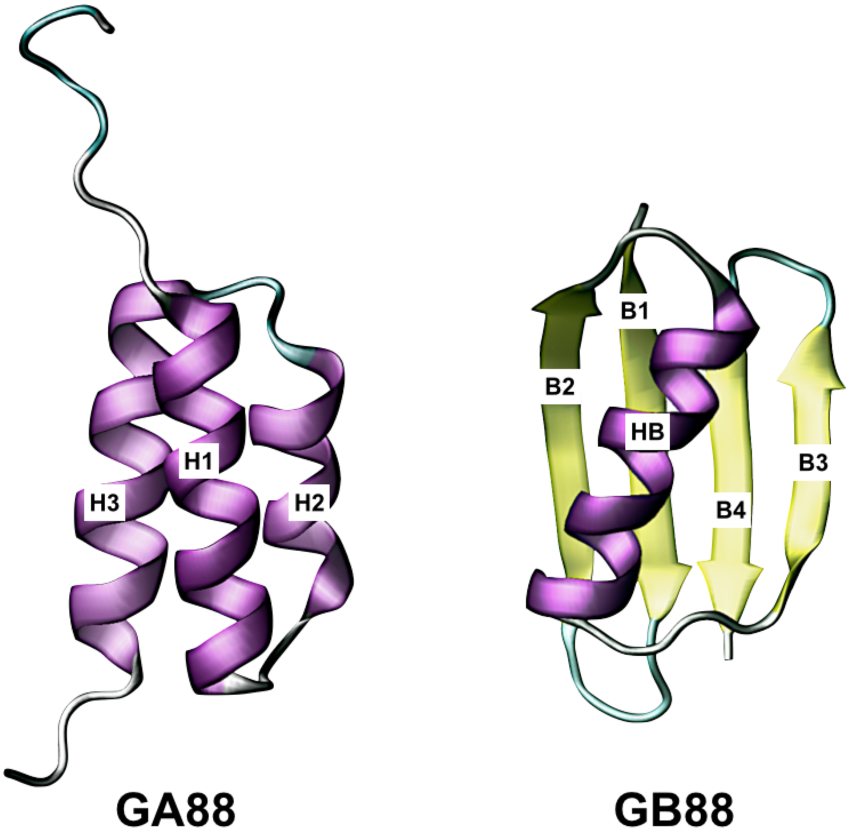


Figure 3.1: Schematic illustration of GA88 (PDB code: 2JWS) and GB88 (PDB code: 2JWU) obtained from experiment. Individual domains of the two proteins are labeled accordingly.

3.2 Methodology

REMD simulation involving replicas of α/4β-GA88 at 42 distinct temperatures was conducted using AMBER 10 simulation package. [22] The distribution of the temperature that ranges from 270 K to 710 K was determined to ensure a targeted acceptance ratio of 0.2. The starting structure used in the REMD simulation was prepared using the NMR structure of GB88 which was acquired from the Protein Data Bank (PDB) with

PDB code of 2JWU and transformed into α/4β-GA88 by performing mutations namely A24G, T25I, F30I, Y33I, Y45L, T49I and K50L, using the LEaP module. [5, 22, 23] The simulations are conducted using AMBER ff03 force field and generalized Born (GB) model. [24, 25] In this study, to reduce the likelihood of overstabilizing the β-sheets of α/4β-GA88, the non-polar solvation term commonly calculated using the surface area, was excluded. [26-27]

Prior to the REMD run, the starting structure was minimized for 500 steps using the steepest descent method followed by minimization using the conjugate gradient method till the energy gradient of the system converged to 0.01 kcal/mol/Å. This was accompanied by the heating of each replica to their designated temperature for 100 ps using Langevin thermostat with a collision frequency of 4 ps^{-1}. [28] During the REMD simulation, replica exchange was attempted every 10000 steps with the simulations conducted for each replica lasting for 75 ns. A time step of 2 fs was used for all simulation conducted. All covalent bonds involving hydrogen atoms were constrained using the SHAKE algorithm and non-bonded interactions were curbed at 12 Å. [29] An acceptance ratio of at least 0.2 was observed for all the replicas during the REMD simulation suggesting that the system is not limited by local free energy minimum trapping. Each replica was also observed to have explored all temperatures several times during the simulation.

3.3 Results: Protein conformational changes from α/4β-fold to 3α-fold

To monitor the structural transition of α/4β-GA88 to 3α-GA88, the C_α-RMSD (root-mean-square deviation) of the folded protein relative to the NMR structure of GA88 (PDB code: 2JWS) over residues 9 to 53 was calculated and plotted against time in Figure

3.2. The first eight amino acids at the N-terminus and the three amino acids at the C-terminus end of the protein were excluded when computing the C_α-RMSD as these residues are part of a disordered random coil as suggested in the NMR structure of GA88. [5] From Figure 3.2, the C_α-RMSD calculated for the protein structures acquired at 270 K, showed an overall decrease in value as the simulation progresses signifying the global folding of the protein to a structure close to that of experiment, with the lowest C_α-RMSD recorded to be 4.34 Å. This is within expectation as the folding of GA88 was reported to fold completely at 298 K. [30] The C_α-RMSD of the folded protein obtained at 304 K was also computed to observe the behavior of the protein at a temperature slightly above 298 K where complete folding of GA88 occurs. Based on Figure 3.2, the protein simulated at 304 K showed similar albeit less apparent decrease in C_α-RMSD relative to that observed at 270 K, with lowest C_α-RMSD recorded to be 4.75 Å.

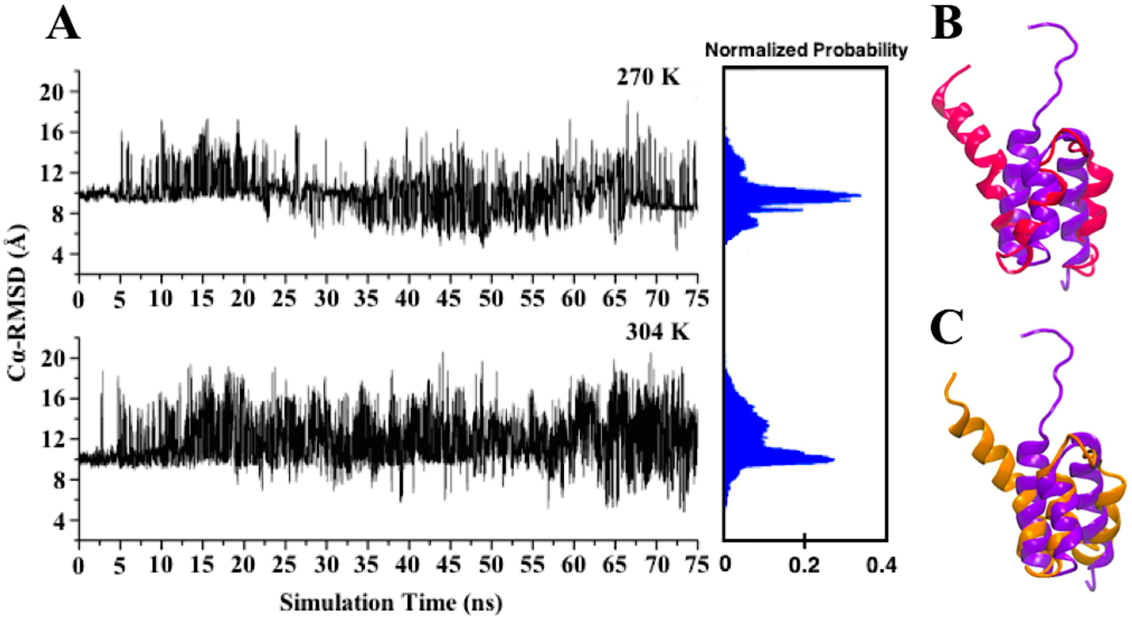


Figure 3.2: (A) Plots of the variation of the C_α-RMSD with time and the distribution of C_α-RMSD of simulated structures calculated over residue 9 to 53 at 270 K and 304 K. (B) Overlap between the experimental structure of GA88 (purple) (PDB code: 2JWS) and the

folded structure with lowest C_α-RMSD (magenta) obtained at 270 K. (C) Overlap between experimental structure of GA88 (purple) (PDB code: 2JWS) and the folded structure with lowest C_α-RMSD (orange) obtained at 304 K.

Even though an overall decreasing trend was observed for the C_α-RMSD computed for structures acquired at 270 K and 304 K relative to the NMR structure of GA88, large fluctuations in C_α-RMSDs and the greater distributions of folded structures with large C_α-RMSD values was apparent in Figure 3.2. Hence, to verify that the structural switch from α/4β-GA88 to 3α-GA88 has indeed occurred, cluster analysis was conducted for the trajectory acquired at 270 K. Out of the five clusters acquired and presented in Figure 3.3, at least 50% of the trajectory consists of structures with the individual helical domains of 3α-GA88 folded. This observation substantiated the structural transformation of the simulated protein from α/4β-fold to 3α-fold conformation.

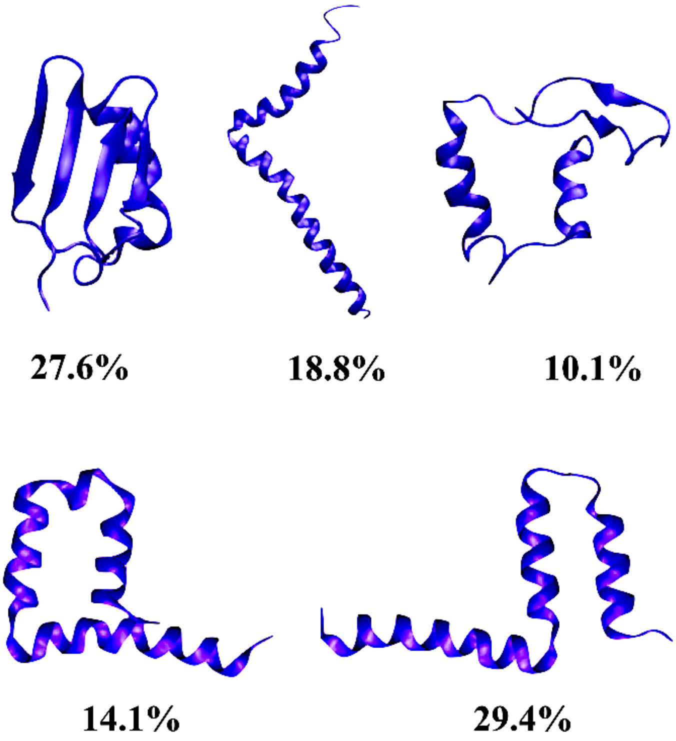


Figure 3.3: Cluster analysis. Schematic illustrations of representative structures of the 5 clusters with the percentage occurrences of each cluster stated accordingly.

Further scrutinization of the folded structures at 270 K and 304 K in Figure 3.2 B and 3.2 C respectively, showed additional helices being formed at the N-terminus end of GA88 which was supposed to be a random coil according to experiment. [5] Judging this observation on the basis of helix propensity, the observation highlighted (*vide supra*) is well-founded since the average helical propensity of the eight residues (TTYKLILN) at the N-terminus end of GA88 is 0.45 kcal/mol and this is comparable to the average helical propensity of part of the H3 domain (residues 43 to 51) which properly folds into an α-helix (average helical propensity: 0.44 kcal/mol). [5, 31] The helical propensities calculated here are based on the helical propensities scale derived by Pace et al. for each amino acid through experimental studies of proteins and peptides. [31] While the helical propensities explain the folding of the supposedly random coil structure to an α-helix during the REMD simulation, this observation still goes against the experimental observations reported by He et al. [5] He et al. observed a net reduction in the helical propensity of residues 6 to 8 upon mutation (A6I, N7L and S8N), hence the preference for a random coil structure instead of an α-helical conformation as seen in the parent protein. [5] This conflicting observation between experiment and theory may indicate the likely presence of an inherent bias in the force field model towards helical structures. [32, 33] This is corroborated by studies conducted by Wang and Wade who emphasized over preference of AMBER ff03 force field for α-helix upon observing the replacement of the unfolded β-sheet with an α-helix during a simulation. [32] Contrariwise, the first four residues of the H3 domain (VEGV) fold into a random coil instead of an α-helix and this is understandable as the helical propensity of this section of the H3 domain that fails to fold into an α-helix (residues 39 to 42) has an average helical propensity of 0.65 kcal/mol hence evincing the preference of the short sequence, VEGV, to fold into a random coil rather than an α-helix. [31]

From the folded structures presented in Figure 3.2 B and 3.2 C, the evident inability of the helix bundle of 3α-GA88 to assemble precisely was also pointed out. This observation may be due to the inefficacy of the implicit solvation model to precisely account for the entropic gain of the system during the expulsion of water molecules from the hydrophobic core during the folding process. [34] This caused the folded protein to adopt a lower energy conformation which is unlike the native structure of GA88. [34] The inability of the helix bundle to aggregate accurately may also be due to the preference of an implicit solvation model for salt bridges rather than hydrophobic interactions. [35-37] This was observed in Figure 3.4 which showed the preference of charged amino acid residues specifically Lys13 and Lys46 to be fully exposed to the external environment as opposed to the orderly packing of these residues in the hydrophobic core as seen in the experimental structure of GA88. Additionally, the representative structure of Cluster 5 also showed the development of two salt bridges between Lys28 and Glu48 and between Lys31 and Glu48 which were previously absent from the NMR structure of GA88. (Figure 3.4) These happenings may potentially hinder the aggregation process of the helix bundle as the overestimation of salt bridges in the implicit solvation model may consequently undermine hydrophobic interactions hence resulting in a final ensemble which is different from the native configuration of GA88. [5, 35-37]

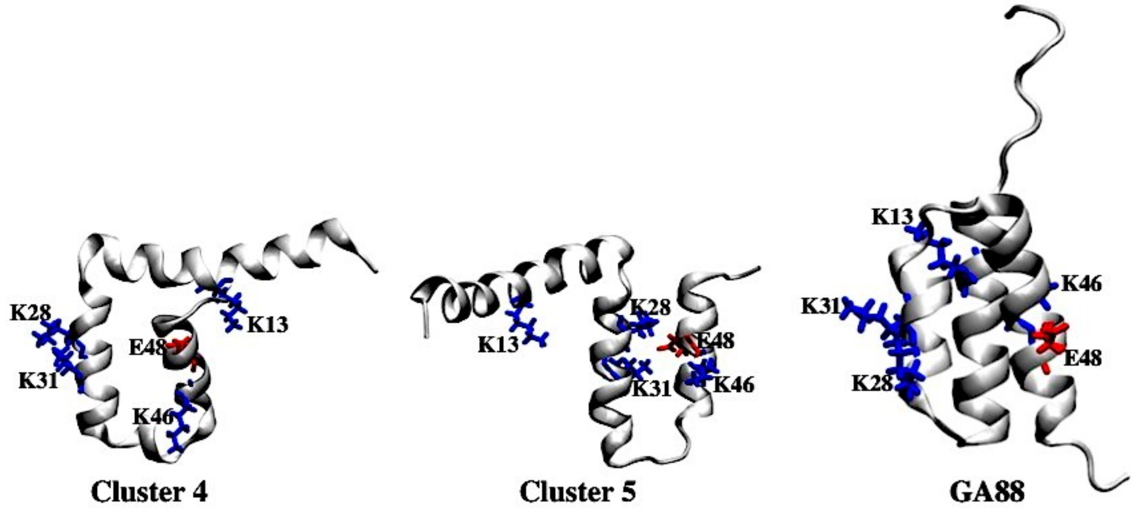


Figure 3.4: Cartoon representation of the experimental structure of GA88 and the representative structures of Cluster 4 and 5, derived from the cluster analysis conducted, with charged amino acids namely Lys13, Lys28, Lys31, Lys46 and Glu48 presented using licorice representation.

Besides examining the folded structure of 3α-GA88 which were acquired through the REMD simulation, the folding mechanism of 3α-GA88 from the non-native α/4β-fold were also explored. To do so, we separately monitored the folding of each helical domain of 3α-GA88, The C_α-RMSDs of the three helical domains namely H1, H2 and H3 were computed relative to the NMR structure of GA88 and plotted in Figure 3.5 A. Since the C_α-RMSD plots of structures acquired at 270 K and 304 K are relatively similar, we have only included the C_α-RMSD plots of structures acquired at 270 K. Based on the plots presented in Figure 3.5 A, the folding of both H1 and H3 were evident from the downward slope seen in the C_α-RMSD plots of these helical domains. However, the large fluctuations exhibited by the C_α-RMSD plots of H1 and H3 may suggest the recurrence of the folding and unfolding process of H1 and H3 during the REMD simulation and this opinion is corroborated by the DSSP plot [38, 39] presented in Figure 3.6 which showed the interchange between α-helix and β-sheet at residues 9 to 23 (H1) and residues 39 to 51

(H3) during the 75 ns simulation. On the other hand, the C_α-RMSD of H2 remained approximately uniform during the simulation hence implying the preservation of the helical domain of α/4β-GA88 throughout the 75 ns simulation. This observation was supported by the DSSP plot [38, 39] in Figure 3.6 which showed the conservation of the α-helical domain at residues 27 to 34 during the simulation. In addition, the DSSP plot in Figure 3.6 also demonstrated the transformation of B1-loop-B2 to H1 and B3-loop-B4 to H3 and this proves the variation of the structure of the simulated protein from α/4β-fold to 3α-fold conformation.

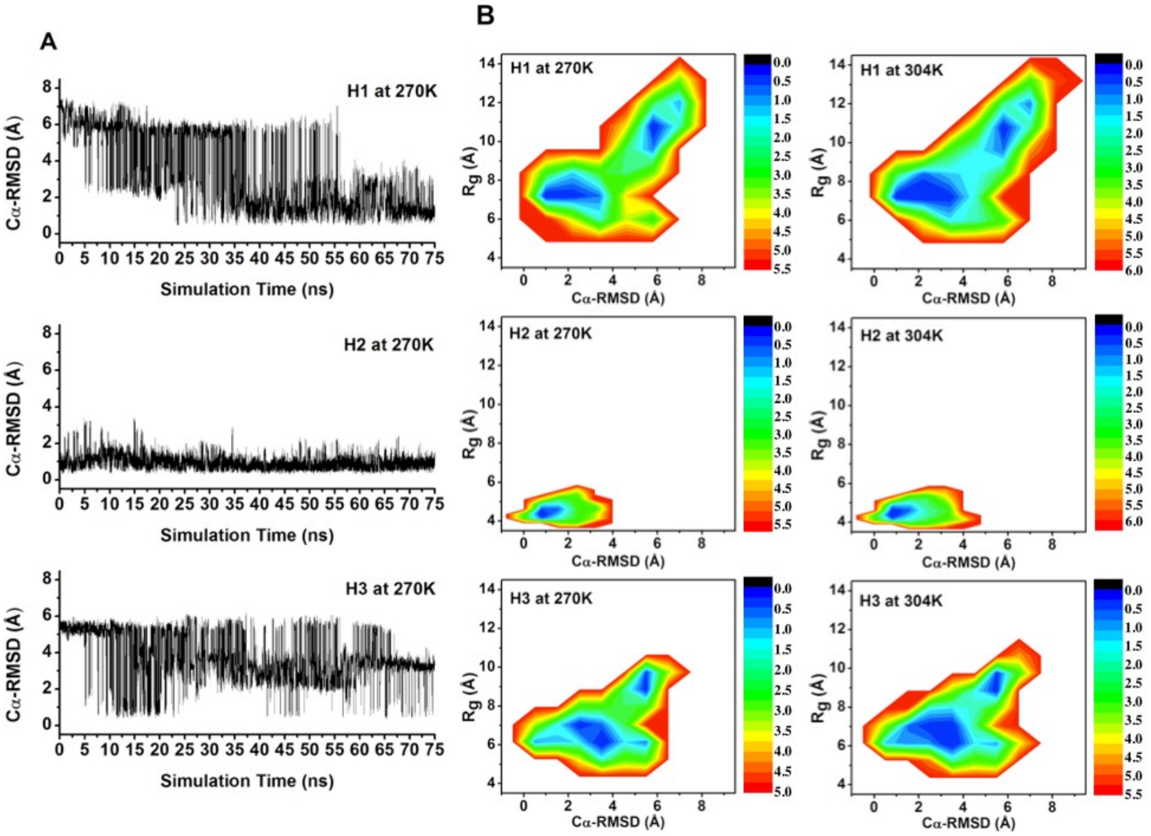


Figure 3.5: (A) Variation of the C_α-RMSD of H1, H2 and H3 with time relative to the experimental structure of GA88 (PDB code: 2JWS) at 270 K. (B) Free-energy contour maps of H1, H2 and H3 of 3α-GA88 acquired at 270 K and 304 K.

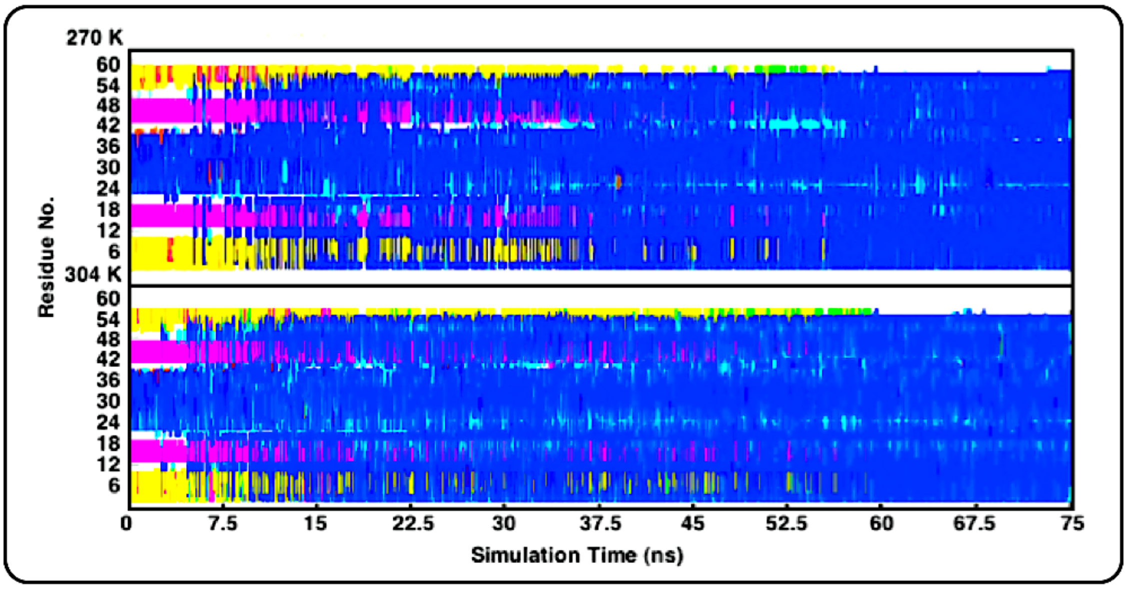


Figure 3.6: Secondary structure assignment (DSSP) of structures sampled at 270 K and 304 K during the REMD simulation. (Green: Parallel β-sheet, Magenta: Antiparallel β-sheet, Yellow: Mixed β sheet, Cyan: 3_{10}-helix, Blue: α-helix, Orange: π-helix)

In order to observe the major structural populations found in the trajectories obtained from the REMD simulation, the free energy landscapes of H1, H2 and H3 were plotted in Figure 3.5 B above, with C_α-RMSD and radius of gyration (R_g) selected as the reaction coordinates. Through plots of free energy landscapes, the scores of dimensionality accommodated within the MD trajectories were simplified into two-dimensional information which is easier to analyze. The free energy landscape acquired for both H1 and H3 at 270 K and 304 K in Figure 3.5 B showed two significant populations in the lower left and the upper right regions of the plots which correspond to the folded and unfolded states of the helical domains respectively. The two populations may account for the continuous switch from α-helix to β-sheet and vice versa during the simulation which were corroborated by substantial fluctuations of the C_α-RMSDs of H1 and H3 as showed in Figure 3.5. On the other hand, the sole population observed for H2 at both temperatures validated the observation made earlier pertaining to the conservation of HB of α/4β-GA88 to form H2 of 3α/4β-GA88.

Other than the folding of 3α-GA88, the unfolding pathway of α/4β-GA88 was also explored in this study using principal component analysis (PCA). PCA is a well-known analytical method that is commonly used to study protein folding and unfolding. [4, 40-44] Through the application of PCA, the complexity of the MD trajectories are minimized by restricting the $3N$ degrees of freedom of a protein system to essential degrees of freedom that represents functionally crucial motions of the protein. [4, 40-44] PCA was conducted on the first 15 ns of the trajectories to acquire motion modes that principally describe the unfolding pathway of α/4β-GA88. Five principal components (PCs) were obtained through PCA and out of the five PCs, only the first two PCs namely PC1 and PC2, showed notable contributions to the unfolding of α/4β-GA88 with PC1 having a greater influence on the unfolding process compared to PC2 (Figure 3.7). From Figure 3.7, greater magnitude in the modal activity of PC1 to PC5 at higher temperatures was observed suggesting the ability of the protein to explore more conformations at higher temperatures.

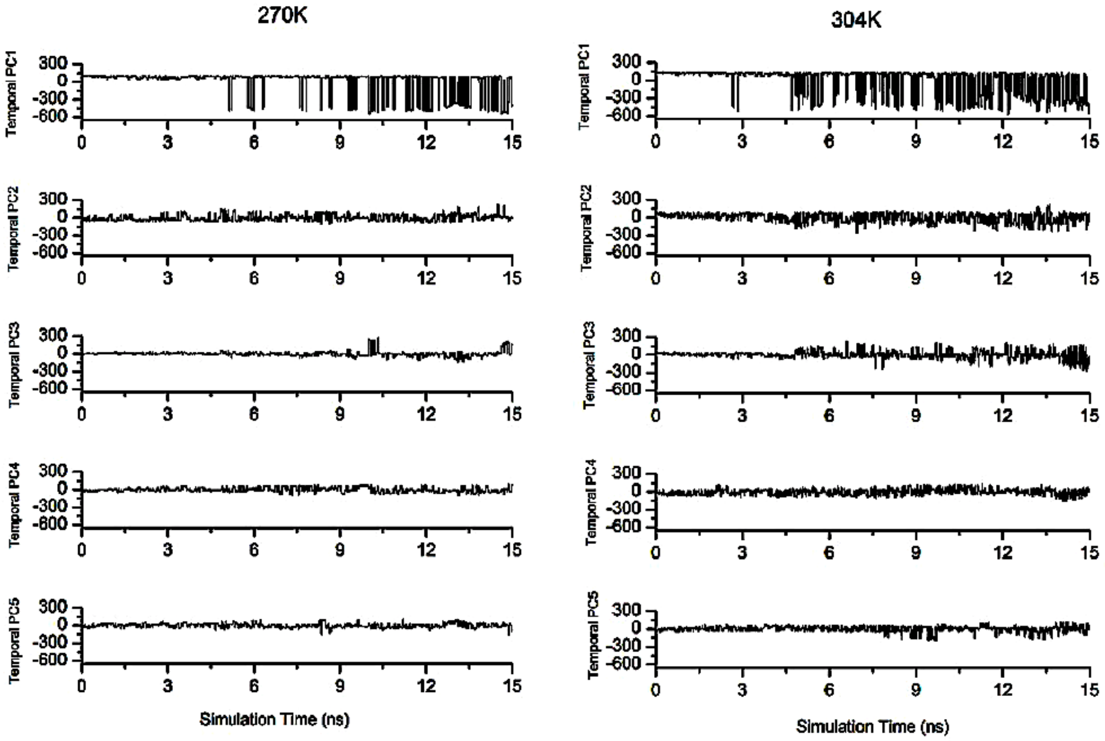


Figure 3.7: Time-dependent variation of the motion mode of the protein projected from PC1 to PC5 at 270 K and 304 K acquired from the first 15 ns of the simulation.

The crucial motion modes identified from PCA were also analyzed visually using interactive essential dynamics (IED) program with visual molecular dynamics (VMD) serving as a display interface to observe the movement of the protein which governs the unfolding of α/4β-GA88. [45, 46] PC1 and PC2 contributed to the unfolding of α/4β-GA88 through the separation of the two β-sheets to form the H1 and H3 domains in 3α-GA88. PC1 describes the pulling motion separating B1 and B4 while PC2 represents the bending of the HB domain in α/4β-GA88. (Figure 3.8) These motions resulted in the unpacking of the hydrophobic core of α/4β-GA88 which constitute of the following residues; Tyr3, Leu5 and Leu7 in B1, Ala26, Ile30 and Ala34 in HB, Trp43 and Leu45 in B3 and Phe52 and Val54 in B4. The unpacking of the hydrophobic core may be a crucial step in the unfolding of the β-sheets of α/4β-GA88 to form the H1 and H3 domains of 3α-GA88.

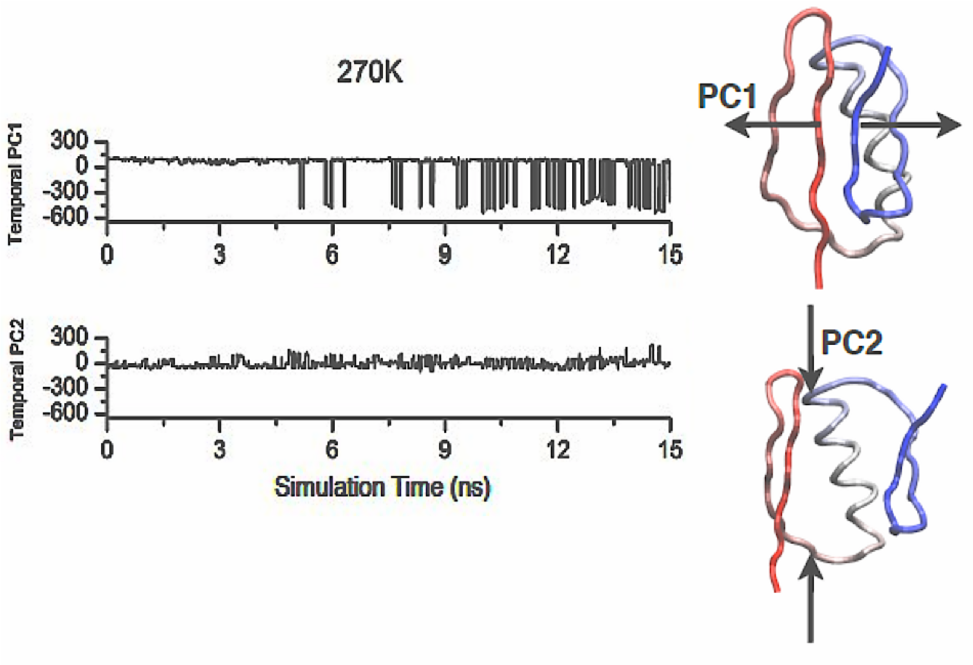


Figure 3.8: Time-dependent variation of the activity of the protein projected from PC1 and PC2 at 270 K during the first 15 ns of the simulation and schematic illustration of the motion modes of PC1 and PC2.

3.4 Conclusion

In this chapter, the mechanism governing the structural variation of a protein from α/4β-fold to 3α-fold using a protein that has the primary structure of GA88 but the tertiary structure of GB88 was investigated using REMD simulation. The transition from α/4β-GA88 to 3α-GA88 was noted despite the large C_α-RMSD computed for the folded structures at 270 K (4.34 Å) and 304 K (4.75Å). Through the analyses conducted, the process of conformation change from β-sheet to α-helix was understood and insights pertaining to protein misfolding were attained. Based on the analyses conducted, the main event that controls the change in conformation from β-sheet to α-helix is the unpacking of the hydrophobic core. From this observation, we deduce that hydrophobic interaction might be one of the crucial interactions that when maintained will prevent the occurrence of protein misfolding.

The underlying bias of the force field model used in this study was further discussed in this chapter. The preference of AMBER ff03 force field for α-helix was highlighted through the inclination of the N-terminus of the folded 3α-GA88 to fold into a helix rather than its native random coil. The inability of 3α-GA88 to aggregate accurately into the native helix bundle conformation of GA88 was also attributed to the undermining of the hydrophobic interactions brought about by the overestimation of salt bridges by the implicit solvation model used. While conducting the REMD simulation in explicit water may be favorable, the increase in the size of the system with the addition of water molecules will be quite expensive as the number of replicas needed increases proportionally to the square root of the degree of freedom of the system being simulated. [47, 48]

While implicit solvation model may be one reason for causing the improper aggregation of the helix bundle, another reason could be the lack of polarization effect considered during the simulation. The use of standard charges in the AMBER force field is not enough to describe the changes in the dielectric polarization effect of the amino acids caused by changes in the dielectric environment during the folding process. In the next chapter, the role of electrostatic polarization effect in the folding of proteins in implicit water will be elaborated to showcase the importance of polarization effect in the modeling of proteins.

Chapter 4 Importance of electrostatic polarization in the *ab initio* folding of helical structures

4.1 Introduction

Protein folding is a complex event in the biological system that many wish to apprehend because of its significant contribution towards the understanding of the dynamics of proteins in nature rather than the static experimental protein configuration that can be obtained from the Protein Data Bank (PDB). [1-3] Through the implementation of MD simulation, protein dynamics could be explored at the atomic level and enable us to monitor the physical folding pathway of proteins, something which is hard to achieve using experimental techniques. [4, 5] The continuous advancement in computer resources and the accessibility to the state-of-the-art computational devices have equipped theorists with computational tools capable of precisely predicting structures of fast folding proteins and occasionally large proteins, and enabling folding simulations on the milliseconds time scale. [6-10]

AMBER, CHARMM and GROMOS are just some of the force field models commonly used to conduct MD simulations. [11-14] The amino acid-specific character of these principal force fields contributes towards the straightforward implementation of these force fields to a variety of protein types. [11-14] However, the use of static charges for each amino acid in these principal force fields has led to an inherent limitation which prevents the force fields from effectively providing accurate electrostatic portrayal of amino acids encompassed within different environments. [15-17] As such, polarizable/ polarized force fields which consider the non-uniform polarization of amino acid residues under different environment have been developed hence equipping researchers with a more accurate modeling of proteins in nature. [15-17]

In nature, protein folding is steered by the formation of intermolecular interactions such as VDWs, electrostatic interactions, hydrogen bonds and hydrophobic interactions. [18, 19] Of particular significance is the formation of backbone hydrogen bonds which are crucial for the stabilization of secondary structures such as α-helix and β-sheet. [18-20] In this study, adaptive hydrogen bond specific charge (AHBC) scheme which was introduced in Chapter 2, will be used to fold three α-helical peptides namely b30-82 domain of F_1F_0 adenosine triphosphate synthase from Escherichia Coli (E.coli) (2khk, PDB id: 2KHK) and N36 and C34 domain of gp41 from HIV-1 envelope glycoprotein (PDB id: 1AIK). [20-23] The AHBC scheme utilizes on-the-fly charge fitting whereby the charges of amino acids participating in the forming and/or breaking of hydrogen bonds during protein folding are updated at fixed interval during the simulation by subjecting these amino acids to fragment quantum mechanical calculation (MFCC). [18, 20, 21, 24-27] By applying the AHBC scheme to the folding of the three peptides, the persistent forming and breaking of hydrogen bonds that essentially occur during the protein folding will be accounted for hence providing us with a more practical mechanism pertaining to the folding of helical peptides in nature.

The main objective of this study is to confirm the significance of electrostatic polarization effect of hydrogen bonds in accurately folding the three helical peptides listed above. To do so, a comparative study involving the folding of 2khk using both polarized (AHBC) and non-polarized (AMBER ff03) force field was conducted and the trajectories obtained were analyzed to observe the difference in the folding accuracy of the two force fields. To verify the reliability of the AHBC scheme in assisting the accurate folding of α-helices, *ab initio* folding of N36 and C34 were also conducted using the AHBC scheme.

4.2 Methodology

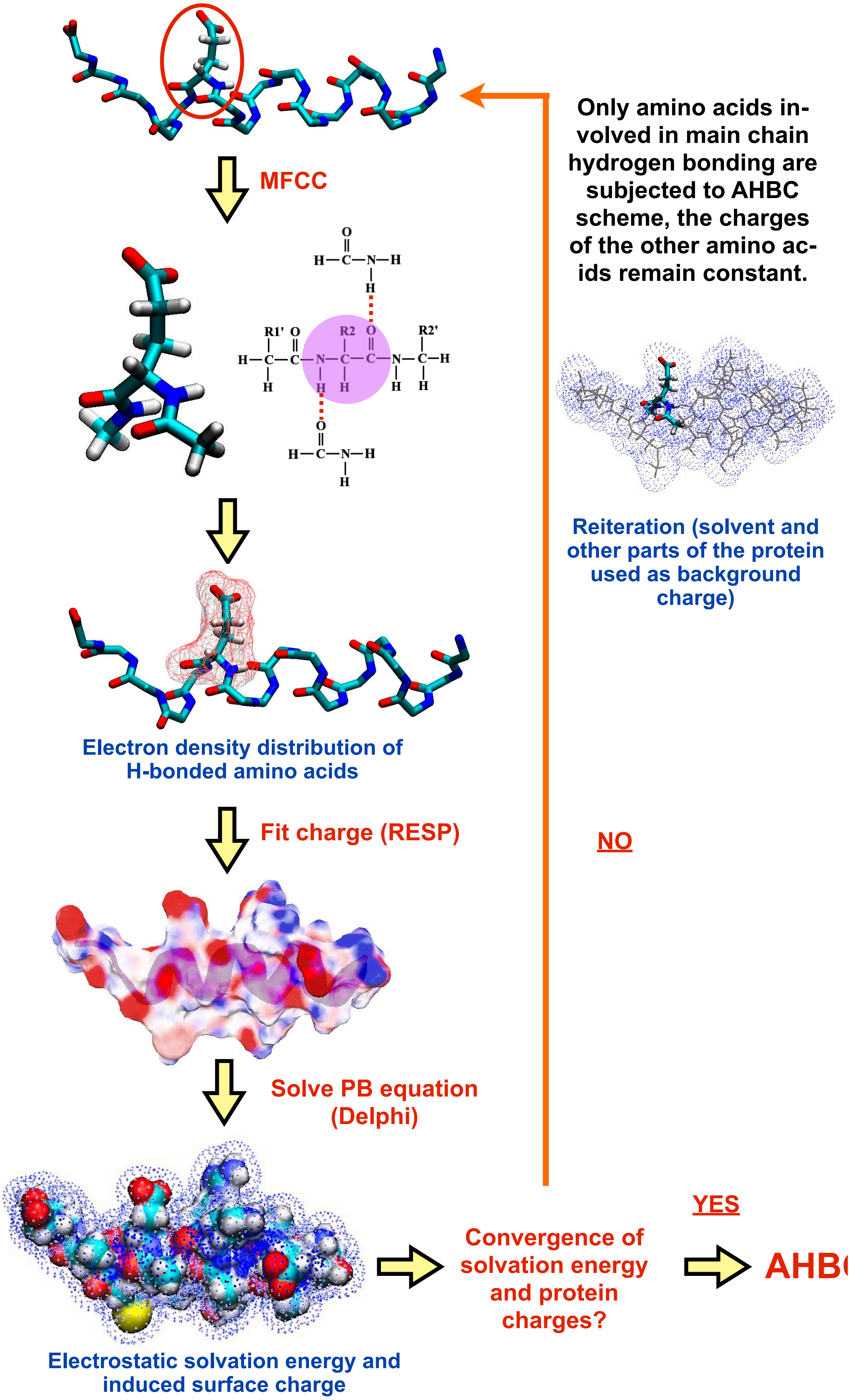
MD simulations conducted in this study commenced from the linear conformation of the three proteins, namely 2khk, N36 and C34, which were constructed using the LEaP module in AmberTools. [22, 23, 28] AMBER ff03 and implicit generalized Born solvation model of Onufriev et al. were utilized to represent the protein and solvation effect respectively. [13, 28, 29] Each peptide was minimized for 10000 steps using the steepest descent method and followed up with minimization by the conjugate gradient method until an energy gradient of 0.01 kcal/mol/Å was reached. The minimized peptides were then heated in 100 ps to their respective experimental temperatures (288 K for 2khk and 300 K for both N36 and C34) using Langevin thermostat with collision frequency of 4 ps^{-1}. [22, 23, 30] After heating, MD simulations using AHBC scheme were conducted for 2khk, N36 and C34 for 14 ns , 19 ns and 50 ns respectively. The difference in simulation time is due to the difference in time taken for the protein to fold into structures close to experiment with C_α-RMSD of less than 3.0 Å. To conduct a comparative study, a simulation using AMBER ff03 charge was also performed for 2khk and this simulation lasted for 70 ns. All simulations were conducted using a time step of 2 fs. The SHAKE algorithm was applied to all bonds involving hydrogen atoms and the cut-off for non-bonded interactions was set to 12 Å. [31] For simulation conducted using the AHBC scheme, all parameters of the AMBER ff03 force field were kept the same except for the charges of amino acids involved in backbone hydrogen bondings which were periodically updated every 20 ps. Amino acids involved in backbone hydrogen bonds were identified using the HBplus program. [32]

The modus operandi of the AHBC scheme used in this study is similar to the steps outlined in Refs 17, 33 and 34. Firstly, the electron densities of the amino acid fragments participating in the forming or breaking of hydrogen bonds were acquired by treating these amino acids using the gas phase fragment quantum mechanics calculation (MFCC). [17, 18, 20, 21, 24-27] The strategy of the MFCC method basically involved the decomposition of the peptide into amino acid fragments which were capped at the terminal using conjugate caps shown in Figure 2.5 in Chapter 2. The total electron densities of the capped fragments were subsequently calculated using the following expression: [20]

$$\rho = \sum_{i=1}^{N_f} \rho_i - \sum_{j=1}^{N_{cc*}} \rho_j^{cc*} - \sum_{j=1}^{N_{hc}} \rho_j^{hc} + \Delta p \qquad (4.1)$$

where the first term represents the sum of the electron densities of capped amino acid fragments involved in either forming or breaking of hydrogen bonds. The second and third terms represent the total electron densities of conjugate caps used to seal the fragments where the peptide bonds and the hydrogen bonds were cut during the molecular fragmentation process respectively. N_f, N_{cc*} and N_{hc} represent the total number of capped fragments, conjugate caps for peptide bonds and conjugate caps for hydrogen bonds used during the MFCC method respectively. Δp is the error.

The atomic charges of these amino acid fragments were then fitted using the conventional RESP (**R**estrained **E**lectro**s**tatic **P**otential) procedure. In order to imitate the polarization effect of solvent on the proteins, discrete charges on the surface of the proteins were obtained from the reaction field generated by solving the Poisson-Boltzmann (PB) equation using the Delphi program. [35] To solve the PB equation, a probe radius of 1.4 Å and grid size of 4.0 grids/Å were selected to compute the solvent accessible surface area. The external dielectric constant was set to 80 (water) while the internal dielectric constant was set to unity as the molecular polarizability of the protein atoms were incorporated explicitly through the quantum mechanical (QM) calculations conducted. The parameters used here are similar to the parameters used by Ji et al. [17] The charges of the amino acid fragments were calculated again using QM calculation but this time round, discrete surface charges were treated as background charges. The new atomic charges acquired for the amino acid fragments were subsequently used to calculate new surface charges and these steps outlined will repeat until convergences of both the protein dipole and the surface charges were attained. All QM calculations were conducted at B3LYP/6-31G* level of theory. The steps taken to derive AHBC are illustrated by means of a flow chart in Figure 4.1.


MFCC
O
H—C—N—H
H
R1' O
R2 O
R2'
H—C—C—N—C—C—N—C—H
H H H H H
O
H—C—N—H
H
Fit charge (RESP)
Solve PB equation
(Delphi)
NO
YES
AHB

Only amino acids involved in main chain hydrogen bonding are subjected to AHBC scheme, the charges of the other amino acids remain constant.

Reiteration (solvent and other parts of the protein used as background charge)

Electron density distribution of H-bonded amino acids

Convergence of solvation energy and protein charges?

Electrostatic solvation energy and induced surface charge

Figure 4.1: Flow chart summarizing the protocol for charge derivation using the AHBC scheme.

4.3 Results: Comparative study involving the folding of 2khk using polarized and non-polarized force field

The main objective of this section is to compare the accuracy of the two force fields namely AMBER ff03 and AHBC, in directing the folding of a helical peptide, in this case 2khk, to a final structure close to experiment (PDB code: 2KHK). [22] The preciseness of the two force fields in driving the folding of 2khk were monitored through the calculation of the time-dependent changes in backbone RMSD (root-mean-square deviation) of the folded peptide relative to the experimental structure of 2khk using the ptraj module of AmberTools 1.2. [28] The backbone RMSDs were calculated over the helical domain which spans from residue 10 to 42 as illustrated in Figure 4.2.

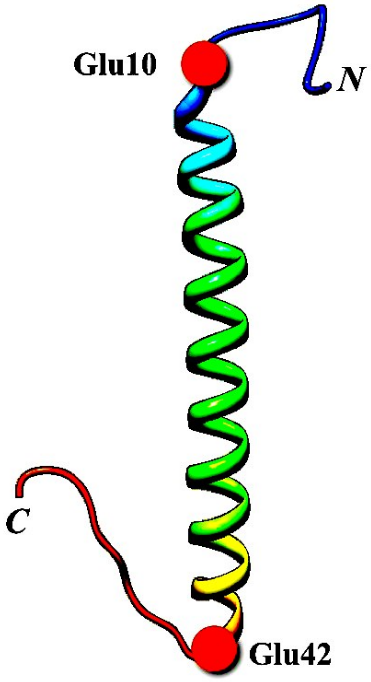


Figure 4.2: Schematic illustration of the NMR structure of 2khk (PDB code: 2KHK) with the glutamic acid residues at position 10 and 42 marked by two red circles.

Based on the plots of the backbone RMSD of the proteins folded using AHBC and mean-field in Figure 4.3, the protein folded using the AHBC scheme showed better folding of the helical region. This observation was indicated by the decrease in the backbone RMSD of the protein as the simulation progresses, with lowest backbone rmsd of 1.3 Å achieved. However, when 2khk was folded using mean field, the backbone rmsd of 2khk showed large fluctuations with a jump in backbone RMSD to around 11 Å after 50 ns suggesting the folding of the 2khk to a structure different from that of experiment. These observations suggest the feasibility of the AHBC scheme to improve the folding of helical structures by considering the polarization effect of backbone hydrogen bonds formed or broken during the protein folding.

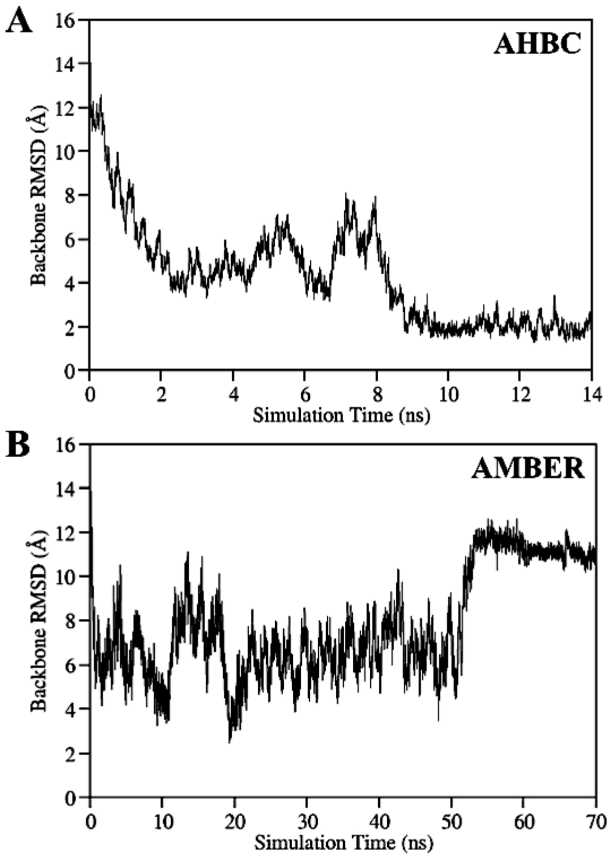


Figure 4.3: Variations in the RMSD of the proteins folded through simulations performed using (A) AHBC scheme and (B) Amber charge.

Besides the calculation of backbone RMSD, the folding of 2khk was also verified through the calculation of the changes in the helical content of 2khk during the course of the simulation which is presented in Figure 4.4. In computing the helical content of the folded structures, only amino acids with dihedral angles of (φ, ψ) = (-60 ± 20, -40 ± 20) were considered as helical. [36] From the helical content plotted in Figure 4.4, both force fields demonstrated the instantaneous folding of the peptide into an α-helix as showed by the sharp rise in the helical content of 2khk to around 45% in under 2 ns. The continuous increase in the helical content of the protein folded using the AHBC scheme was observed and its helical content fluctuated around 75 to 80% after 10 ns. This concurred with experimental observation as the helical content of 2khk was reported to be 71%. [22] The 2khk folded using AMBER ff03 charge exhibited a helical content of not more than 60%. However, to conclude the inability of mean-field to fold extended helices based on these observations may be too crude as the difference in the helical content between the folded 2khk and the experimental structure is only 10% which is equivalent to five residues.

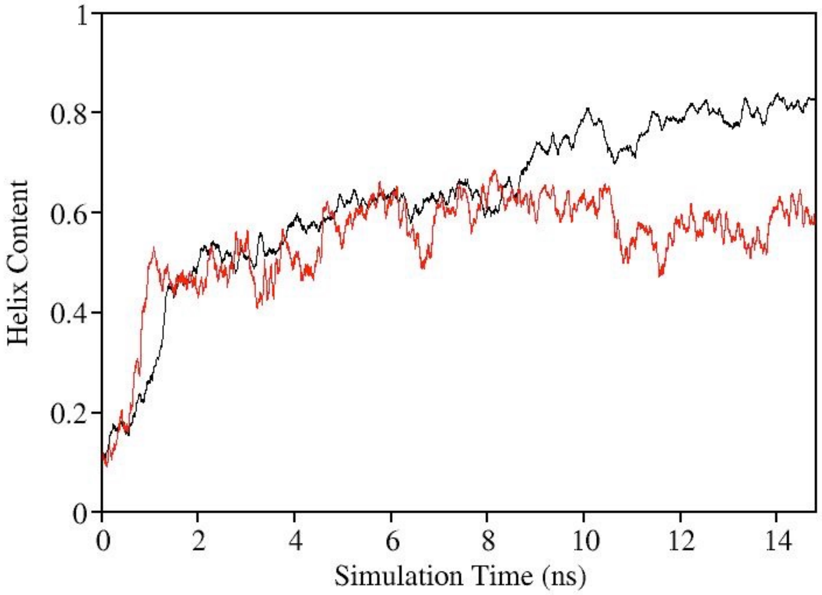


Figure 4.4: Changes in the helical content of the folded 2khk with time acquired from simulations conducted using the AHBC scheme (black) and AMBER force field (red).

Therefore, to obtain more concrete evidence on the importance of polarization effect in the folding of helix, the end-to-end distance between two amino acids at the two tips of the helical region of 2khk, specifically Glu10 and Glu42, were measured and plotted in Figure 4.5. During the folding of 2khk, the shortening of the peptide length is expected to occur since each helix turn consist of 3.6 amino acids. This was observed in Figure 4.5 A for the simulation conducted using AHBC whereby a drop in the end-to-end distance between Glu10 and Glu42 was noted as the simulation progresses till it attains a compactness almost identical to that of the NMR structure (48.07 Å) [22] Contrariwise, the protein folded using AMBER ff03 charge showed an unusual shortening in the distance between Glu10 and Glu42 to around 20 Å after 50 ns and this was followed by another drop in the end-to-end distance to around 7.5 Å at the 60 ns mark. (Figure 4.5 B) This observation may suggest that the two terminal ends of 2khk are approaching each other. Before the two termini move to a position close to each other, fluctuation in the backbone RMSD around 6 Å was discerned and further scrutinization of the MD trajectory revealed the folding of 2khk to a conformation close to the experimental structure with best backbone RMSD of 2.5 Å. These observations evinced the significance of considering the polarization effect contributed by the ever changing hydrogen bonds during folding simulations. The inefficacy of the AMBER force field to stabilize the backbone hydrogen bonds essential for the folding of the extended helix was also highlighted by the curving of the α-helix of the 2khk folded using the AMBER charge.

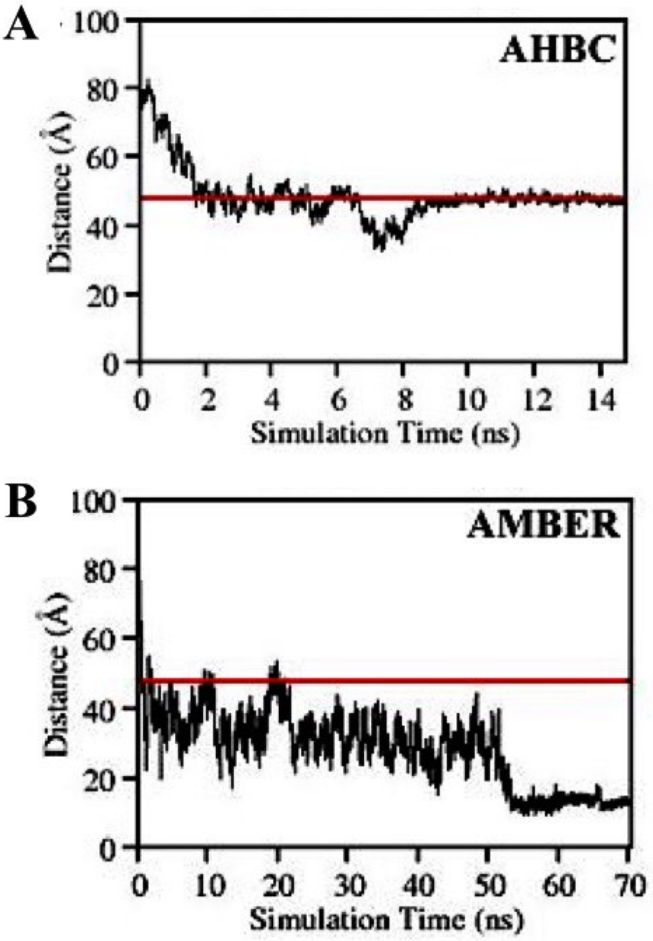


Figure 4.5: Variation of the distance between the CA atoms of Glu10 and Glu42 for simulations conducted using (A) the AHBC scheme and (B) AMBER ff03 force field. Experimental distance between the two residues in 2khk are marked by the red line.

Another method that we have exploited to observe the capacity of the two force fields to maintain the extended helical conformation of 2khk was through the calculation of the variation of the dipole moment of the helix during the simulation. The net dipole of the 2khk protein was measured by calculating the net dipole moment of the backbone carbonyl group (C=O) from residues 9 to 44 since the dipole vector of the α-helix is in the same direction as the dipole vector of backbone C=O, both of which are directed towards the N-terminal of 2khk. Since the macro-dipole of a helix is the summation of all the dipole moments contributed by the peptide bond, an increase in the helix dipole will be expected as the helical structure develops. From Figure 4.6 A, helix growth was evident from the increase in the net dipole moment of the backbone C=O of the protein folded using the AHBC scheme. On top of that, the stabilization of the helix dipole was observed after 10 ns in Figure 4.6 A corroborating the successful folding of 2khk as shown by the RMSD data in Figure 4.3 A. The precise folding of 2khk by the AHBC scheme was also confirmed through cluster analysis conducted for the whole trajectory whereby the most populated cluster revealed a representative structure similar to the NMR structure of 2khk and this is illustrated in Figure 4.6 A (inset).

From the plot of the net dipole moment acquired for 2khk simulated under AMBER ff03 force field in Figure 4.6 B, the development of the helical domain of 2khk followed by the stabilization of this domain was observed during the first 50 ns of the simulation. However, the net dipole of the backbone C=O of the 2khk peptide (AMBER) began to drop to a minimum value of 0.66 Debyes after 50 ns. Combined with the earlier analyses conducted for the simulation using the AMBER force field, the drop in the net dipole to a value close to zero may arise because of the bending of the extended conformation of the α-helix. In addition, cluster analysis conducted revealed a populated cluster with a representative structure as showed in Figure 4.6 B (inset) which corresponds to a helix-turn-helix (HTH) configuration.

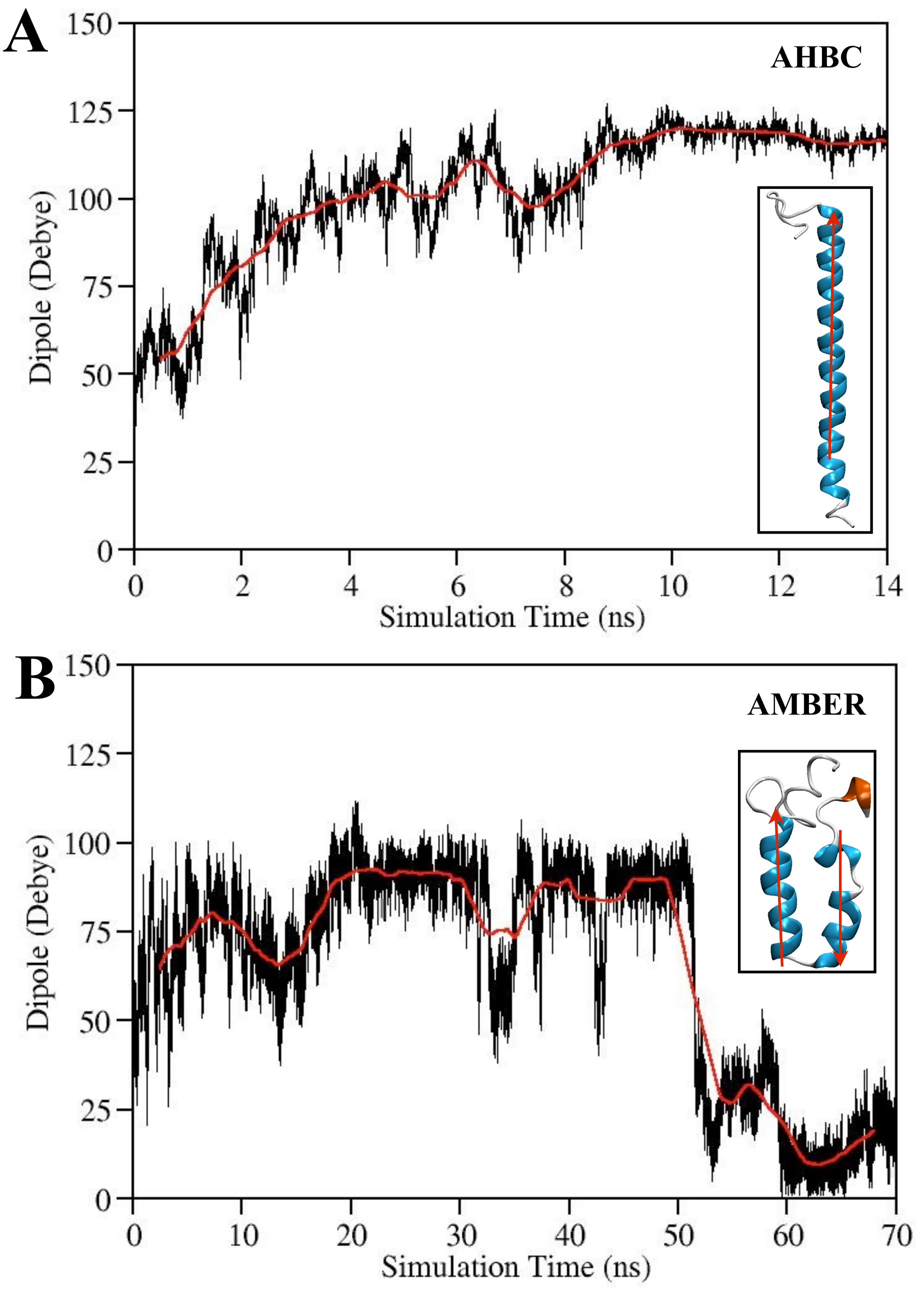


Figure 4.6: Variation of the dipole of the main-chain carbonyl group of folded 2khk, from residues 9 to 44, obtained through simulations conducted using (A) the AHBC scheme and (B) the Amber force field. Protein structures given in (A) and (B) represent the population with the highest percentage of occurrence in the trajectory and their respective macro-dipole indicated by red arrows.

The curving of the extended helix to form the HTH conformation in the simulation performed using the AMBER charge may suggest the inability of the mean-field to stabilize the backbone hydrogen bonds. Based on the HTH structure presented in Figure 4.6 B (inset), the breaking of the three backbone hydrogen bonds between Lys29 and Thr33, Ala30 and Asp34, and Ser31 and Gln35 may caused the unfolding of the helix at residues 30 to 34 and resulted in the formation of a random coil. These events promoted the bending of the folded helix. The variations of the hydrogen bond lengths of the aforementioned hydrogen bonds were plotted in Figure 4.7 and from these plots, the breaking of the hydrogen bonds listed above were noted after 50 ns for the simulation conducted using AMBER ff03 charge and this corroborated the dramatic jump in backbone RMSD observed in Figure 4.3 B. While the three hydrogen bonds mentioned above were disrupted during the simulation conducted using AMBER charge, these hydrogen bonds exhibited significant stabilization upon formation when AHBC scheme was used to fold the 2khk peptide. This denotes the importance of considering the electronic charge rearrangement between hydrogen bond pairs upon establishment and/or disruption of hydrogen bonds during the folding process, as practiced by the AHBC scheme. This, in turn, will augment the electrostatic interactions between hydrogen bond donors and acceptors leading to the stabilization of established hydrogen bonds.

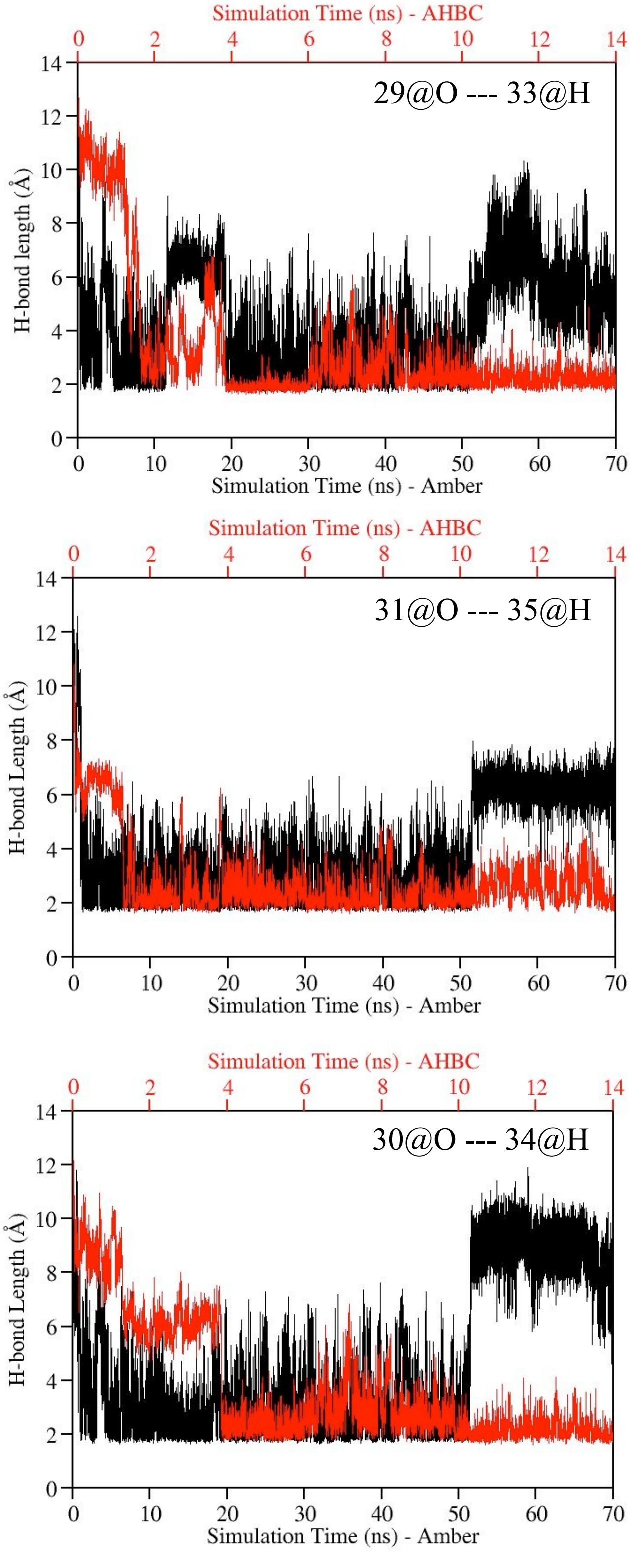


Figure 4.7: Variation of the hydrogen bond length formed between Lys29—Thr33, Ala30—Asp34, and Ser31—Gln35. (red: AHBC scheme, black: AMBER ff03 charge)

Based on Figure 4.3 B, the 2khk peptide folded using the AMBER force field exhibited a large increase in the backbone RMSD after 50 ns, and this was followed by the stabilization of the backbone RMSD around 11 Å in the next 20 ns. This trend is evident in all analyses conducted hitherto for the folding of 2khk under AMBER ff03 force field. Therefore, to provide answers to this interesting observation, cluster analysis was conducted over the last 20000 frames of the AMBER simulation and the representative structures of the five clusters are presented in Figure 4.8 with the structures ordered according to the sequence it showed up during the folding process. 86.2% of the last 20000 frames of the AMBER simulation were made up of HTH conformation mentioned (*vide supra*) hence implying the preference of 2khk to remain in the HTH conformation upon formation as demonstrated by the stabilization of the backbone RMSD around 11 Å after 50 ns (Figure 4.3B). The folding of 2khk to a structure close to the experimental one was also observed at the beginning of the last 20 ns of the simulation hence corroborating the initial deduction made that the HTH conformation was formed through the bending of the extended α-helix.

Another question that came up with the observation of the HTH conformation was why the HTH conformation was preferred in the simulation performed using the AMBER ff03 charge and not in the simulation that followed the AHBC scheme. Through the scrutinization of the HTH structure showcased in Figure 4.8, two salt bridges between Glu11 and Arg54 and Arg21 and Asp43 were discerned and the formation of these interactions may cause the stabilization of these HTH structure. The curving of the extended helix to form HTH, as portrayed in Figure 4.8, may indicate the preference of 2khk, modeled under the AMBER ff03 force field, to establish interactions involving charged amino acids at the cost of hydrogen bond destabilization.

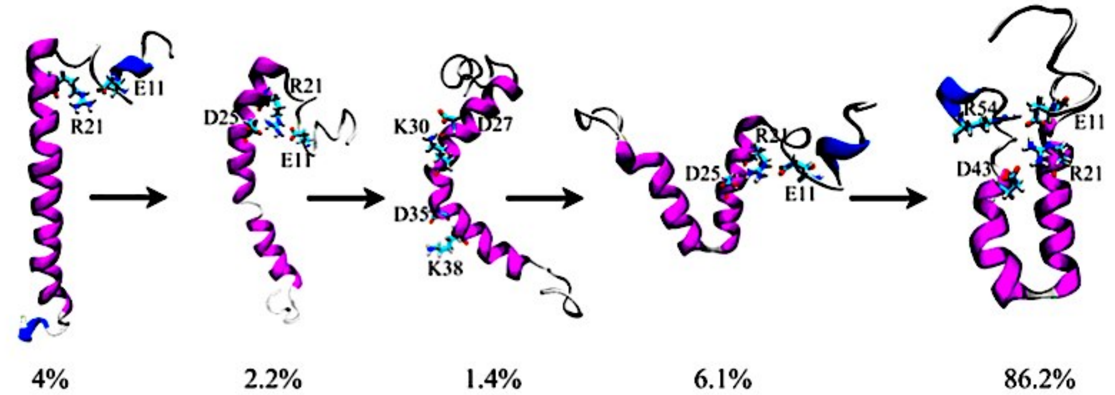


Figure 4.8: Schematic illustration of the representative structures of 5 clusters acquired through the cluster analysis performed for the last 20 ns of the simulation which were carried out using AMBER ff03 charge. The structures are ordered according to the sequence of appearance during the simulation with percentage occurrence stated. Amino acids that may form salt bridges are represented using licorice representation.

To conclude the list of analyses conducted in this study, we have also generated the free energy contour maps for the folding simulations performed using the AHBC and AMBER ff03 charges. The weighted histogram method (WHAM) was used to generate the free energy contour maps which were plotted with reaction coordinates corresponding to the backbone RMSD and the radius of gyration (R_g) and displayed in Figure 4.9 A and B. [37-39] Based on Figures 4.9 A and 4.9 B, two major populations were discerned for each contour maps and are labeled R1 and R2 accordingly. R2 population corresponds to the folded states of 2khk under the AHBC and AMBER force field. However, while R2 corresponds to the lowest free energy configuration for MD simulation conducted using the AHBC scheme, R1 is the population that represents the lowest free energy state for simulations conducted using AMBER charge. This showcased the incapacity of the AM-

BER force field to preserve the elongated helical conformation of 2khk upon development, resulting in the collapse of the extended structure to the HTH conformation.

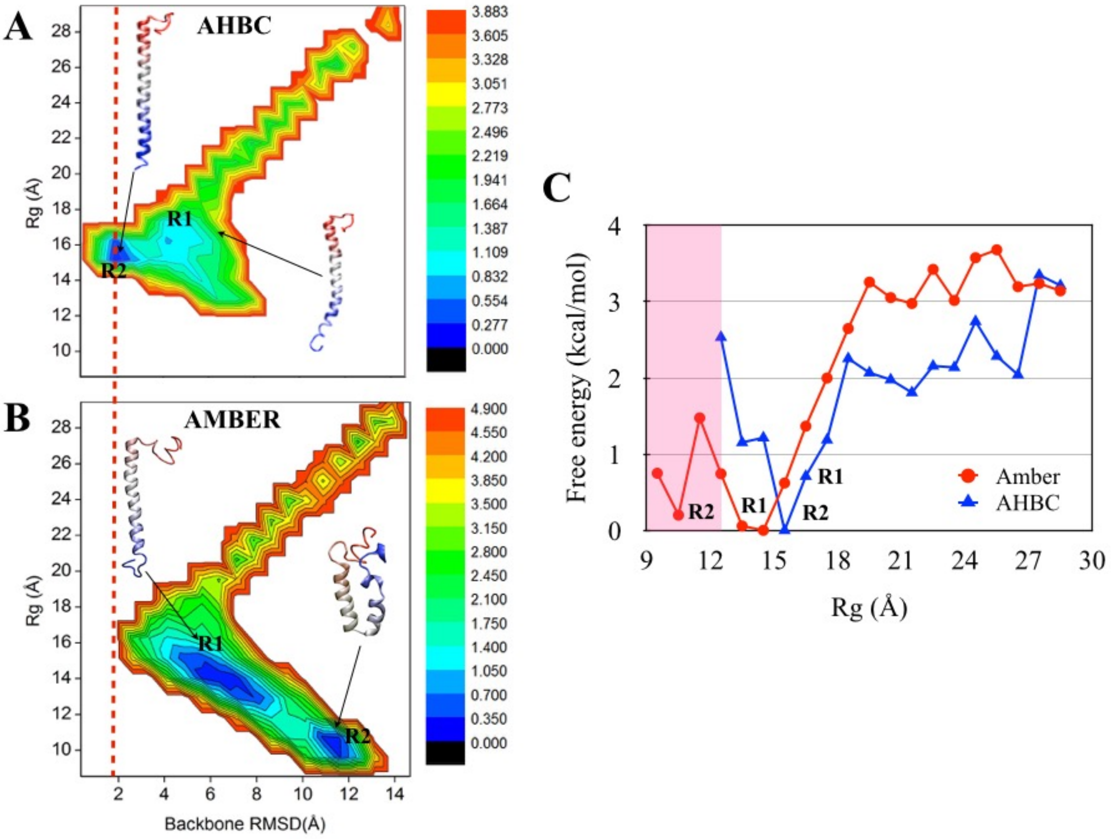


Figure 4.9: Contour maps for the folding of 2khk carried out using (A) the AHBC scheme and (B) the Amber ff03 force field. (C) The free energy profile of 2khk.

The free energy profiles plotted based on the R_g of the folded 2khk for both AHBC and AMBER, substantiated the observations made through the free energy contour maps. Based on Figure 4.9 C, the folded 2khk, obtained using the AHBC scheme, was required to overcome an energy barrier of at least 2.5 kcal/mol to undergo a conformational change. However, this energy barrier was lowered to around 1.5 kcal/mol for the simulation carried out using the AMBER ff03 charge thus the preference for the HTH conformation. These observations pointed out the importance of electronic polarization between hydrogen bond donors and acceptors in augmenting the electrostatic interactions between hydrogen bond pairs resulting in the stabilization of hydrogen bonds vital for the folding of 2khk.

Based on all the analyses conducted (*vide supra*), the AHBC scheme showed better accuracy in the folding of 2khk compared to its non-polarized counterpart. However, in order to strengthen our notion that polarization is indeed important for the precise folding of α-helices, we conduct two additional folding simulations of two distinct helical proteins namely C34 and N36 using the AHBC scheme. [23] As illustrated in Figure 4.10, the successful folding of the two peptides into structures similar to the their respective experimental structures were evident with best backbone RMSD of 0.73 Å and 0.72 Å achieved for C34 and N36 respectively. The folding of the two helices in Figure 4.10 was comparable to the folding of 2khk in that the extended helix was stabilized upon formation, evident from the constant fluctuation of the backbone RMSD of the two protein around 2 Å during the last 10 ns of the simulations. These observations supplement our notion that the polarization effect between hydrogen bond pairs is important for the stabilization of the backbone hydrogen bonds of the helix hence eliminating the chances of the extended helix to collapse into a structure different from native.

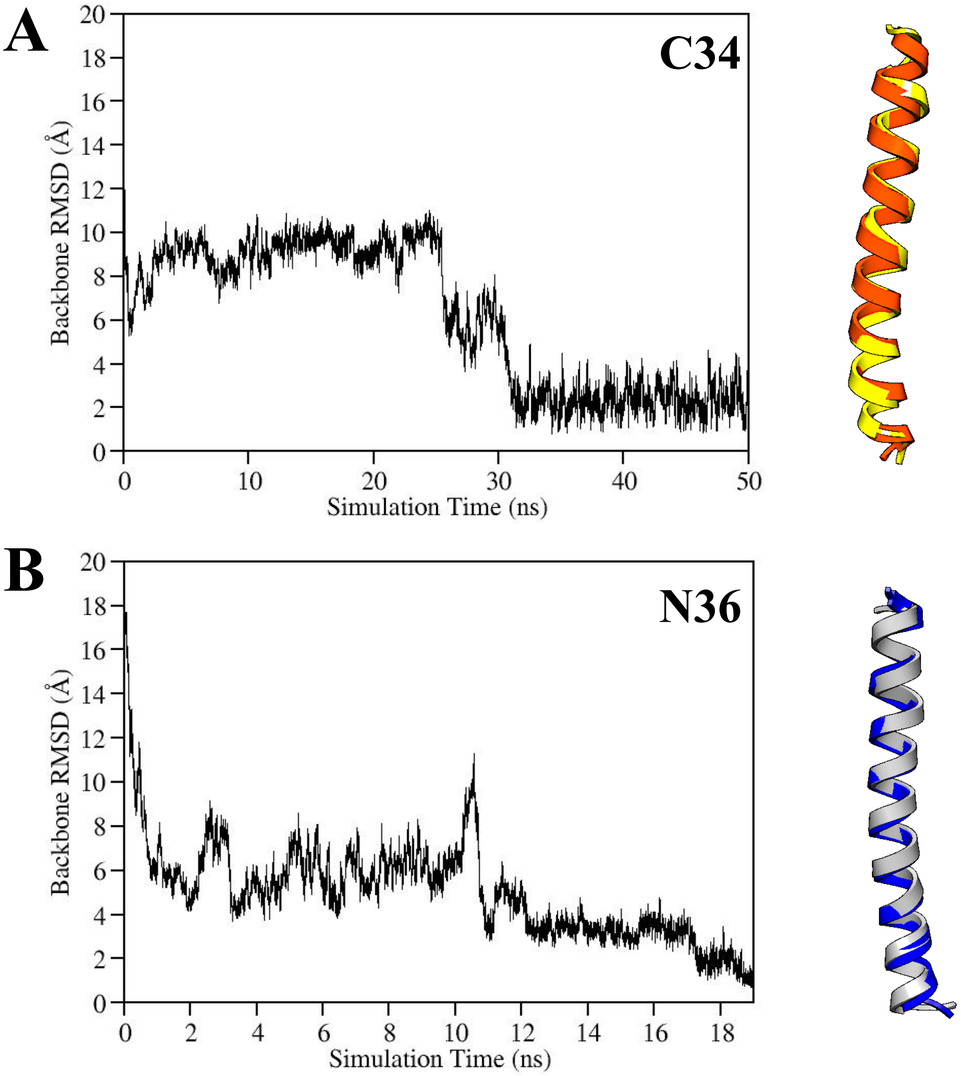


Figure 4.10: (A) Variation of the RMSD of C34 compared to the X-ray crystal structure coded 1AIK over residues 2 to 35. Schematic illustration of the overlap between the X-ray structure (orange) and the best RMSD structure of the folded C34 (yellow) included. [40] (B) Variation of the RMSD of N36 compared to the X-ray crystal structure coded 1AIK over residues 2 to 36. Schematic illustration of the overlap between the X-ray structure (blue) and the best RMSD structure of the folded N36 (grey) included. [40]

4.4 Conclusion

In this chapter, the *ab initio* folding of 2khk, an exclusively helical protein, was carried out using a newly developed on-the-fly charge fitting scheme called AHBC. The AHBC scheme took into account the electrostatic polarization of backbone hydrogen bonds that are constantly formed and broken during the folding process of the protein. A parallel simulation of 2khk using mean field, in this case AMBER ff03 force field, was also performed to compare the accuracy of the two force fields to precisely fold 2khk from its linear structure to a conformation close to the experimental structure of 2khk (PDB code: 2KHK). [22] Analyses such as the changes in the backbone RMSD, end-to-end distance, net dipole moment and hydrogen bond length were conducted to monitor the folding of 2khk using both polarized and non-polarized force field. Based on these analyses, the AHBC scheme performed better than the AMBER ff03 force field in accurately folding the 2khk protein. This finding emphasizes the vital role played by the polarization effect of hydrogen bonds in the folding of helices.

The results obtained from the folding of 2khk using the AHBC scheme was further validated through the folding of two more helical peptides namely C34 and N36. The best backbone RMSDs for all three peptides simulated using the AHBC scheme are below 2 Å and this reflects the capacity of the AHBC scheme to fold helical structures precisely. The study elaborated in this chapter stresses the importance of polarization effect in the folding of elongated helices as the consideration of the changes in the electrostatic environment of the amino acid during protein folding, using the AHBC scheme, aid in the stabilization of hydrogen bonds crucial in maintaining the elongated form of the helix. This subsequently prevents the formation of the HTH conformation that was acquired through AMBER simulations without charge update.

Chapter 5 Importance of electrostatic environment in the modeling of the reduction process of metalloproteins

5.1 Introduction

In the previous chapter, the role of polarization effect in accounting for the changes in the electrostatic environment caused by the formation and disruption of hydrogen bonds during protein folding was explored. Here, we will look into the role of electrostatic environment in the modeling of the reduction process occurring in metalloproteins. Metalloproteins are metal-containing proteins that constitute the majority of proteins that exist in nature. [1-3] Members of this protein group serve a diverse range of functions from structural to biological purpose. [1-3] In the metalloprotein family, electron transfer (eT) proteins which are known to drive various biological processes such as respiration, signal transduction and photosynthesis, are garnering attention from both theorists and experimentalists due to their ability to regulate the reduction potential of the redox center to accommodate to the needs of the biological system. [4, 5] The modulation of the reduction potentials of metalloproteins is often executed through the reorganization of amino acids and solvent molecules adjacent to the metal center. [4, 5] The capacity of eT protein to control the route of electron transfer by controlling the reduction potential at their redox centers have piqued the interest of many due to its beneficial application in biotechnological inventions such as biosensors. [6] The increasing popularity of eT proteins in the research arena is evident from the surge in the number of studies conducted in a bid to comprehend the structural and dynamical aspects of metalloproteins. These studies are instrumental in acquiring knowledge pertaining to the redox processes occurring in metalloproteins which subsequently facilitates the design of redox active proteins that are of relevance to the field of biotechnology and nanotechnology. [6, 7-17]

The use of theoretical devices in the prediction of the reduction potentials of metalloproteins have been widely conducted thus far using various methods ranging from quantum mechanics (QM) to hybrid quantum mechanics/molecular mechanics (QM/MM). [18-28] QM is one of the most frequently utilized tool in the study of metalloproteins owing to the proficiency of the QM approach to precisely model the electronic structures of metalloproteins while providing useful information pertaining to energies related to the reactivity of the metalloproteins. [18-20, 27-29] However, QM studies of metalloproteins are often confined to the modeling of the active site as metalloproteins, as a whole, are normally too large for QM to accommodate. As a result, considerations such as sterics, electrostatics and entropy of the protein frame and solvent, enclosing the redox active site, are overlooked even though these factors may potentially exert pivotal influence on the redox activity of the metalloprotein. Other approaches resolving the issue associated with the incorporation of protein and solvent environment in studying the redox activity of metalloproteins have emerged in the form of molecular mechanics (MM) and hybrid QM/MM. [19, 22-27] Through these methods, factors such as sterics and electrostatic effects are considered during the modeling of metalloproteins, boosting the reliability of the reduction potential determined. [19, 22-27] Besides the consideration of steric and electrostatics influences, dynamical aspects of metalloproteins should also be accounted when examining the redox properties of metalloproteins. [4, 5, 15-17, 30] Fluctuations in the structural motifs adjacent to the redox center have been proposed to have a significant effect on the reduction potentials of metalloproteins. [4, 5, 15-17, 30]

The vital role played by protein dynamics and solvent effect in accurately embodying the electrostatic interactions between the redox active site of a metalloprotein and the environment enclosing it has been showcased in various free energy studies. [19, 22, 31-33] Of particular interest is the study conducted by Olsson et al. whereby both QM/MM and classical MD simulation were utilized to predict the reduction potential of two metalloproteins namely plastocyanin and rusticyanin, in a bid to compare the accuracy between these two computational methods. [19] Through this study, comprehensive sampling of protein conformations was highlighted as one of the main concerns that should be carefully addressed while determining the reduction potential of metalloproteins since proper sampling provided a better depiction of the protein and solvent environment encircling the redox center. [19] This observation attained by Olsson et al. emphasized the importance of considering protein dynamics in predicting the reduction potentials of metalloproteins. [19] With the use of MD simulation, an accurate portrayal of the redox center can be achieved as factors such as entropy are accounted through temperature scaling using thermostats such as Andersen temperature coupling scheme and Langevin dynamics.

In this study, rubredoxin from Clostridium pasteurianum (Cp) will be used to evaluate the accuracy of two charge schemes that will be elaborated later in this chapter, in determining the reduction potential of the rubredoxin variants relative to that of the wild type protein. [34, 35] Rubredoxin is an iron-containing metalloprotein composed of a redox center with an iron atom tetrahedrally coordinated to the thiolates of four cysteine residues namely Cys6, Cys9, Cys39 and Cys42. [4, 16, 17, 34, 35] These amino acids are encompassed within two separate CysXXCys loops as showed in Figure 5.1. [4, 16, 17, 34, 35] Within the two CysXXCys loops is a network of six hydrogen bonds (HN---γS) formed between the iron coordinated sulphur atoms of Cys and the backbone amide of residues in the first and second coordination spheres which include Val8, Cys9, Tyr11, Leu41, Cys42 and Val44 (Figure 5.1). [4, 16, 17, 34, 35] The overall strength of these hydrogen bonds has been proposed by Lin et al. to affect the reduction potential of rubredoxin. [16] However, in a recent theoretical study conducted by Gámiz-Hernández et al., the importance of hydrogen bonds in modulating the reduction potential of rubredoxin was disputed as this study successfully reproduced the reduction potential of 16 rubredoxin variants by keeping the hydrogen bond geometry of all the variants the same. [36] This study also suggested that the subtle shifts of the protein backbone and amino acid side chains which led to a slight variation in the charge distribution may be the main contributor towards the difference in the reduction potentials among rubredoxin variants. [36] Even though hydrogen bonds have been suggested to have a minor role in the determination of the reduction potential of rubredoxin, it may indirectly contribute to the reduction potential shifts of rubredoxin. For example, in a recent study conducted by Zheng et al., the stabilities of Fe(III)-thiolate bonds at the redox center were correlated to the strength of HN---γS hydrogen bonds. [17] Albeit the ongoing debate about the importance of hydrogen bonds in rubredoxin, all these findings highlighted the importance of understanding the electrostatic environment of the redox center of the protein in order to manipulate its reduction potential. [4, 16, 17, 34, 35]

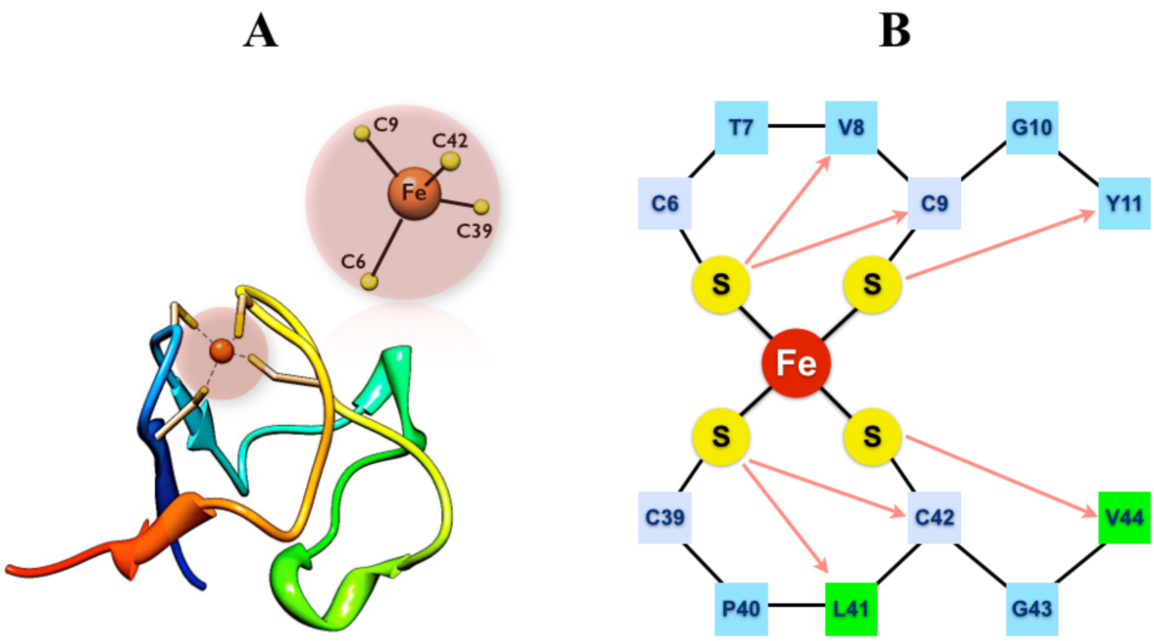


Figure 5.1: (A) Cartoon representation of rubredoxin with the redox center illustrated using licorice representation. (B) Schematic diagram representing the six hydrogen bonds (HN---γS) formed between coordinated γS-Cys atoms and the backbone amide hydrogen atoms of Val8, Cys9, Tyr11, Leu41, Cys42 and Val44. Hydrogen bonds are represented as red lines and residues Leu41 and Val44 which are mutated to obtain L41A, V44A and V44G rubredoxin mutants are highlighted in green.

Rubredoxin is an ideal system in experimental studies due to its small dimensions, low reorganization energy and high electron transfer rate. [4, 15-17, 35-40] This led to the utilization of rubredoxin in multitudes of studies pertaining to the reduction potential of iron-containing metalloproteins, hence providing ample access to experimental data required to validate the accuracy of the theoretically determined reduction potential of the three rubredoxin mutants namely L41A, V44G and V44A. Free energy calculations of the redox processes occurring in the rubredoxin variants will be performed by conducting classical MD simulations with the implementation of thermodynamic integration (TI) formalism using a combination of AMBER ff03 force field and charges derived through density functional theory (DFT) calculation. [41] From this free energy study, the reduction potential of the mutated rubredoxin with reference to that of wild type rubredoxin was acquired and compared with experimental data to provide some insights on the influence of the electrostatic environment in the modeling of metalloproteins with precision.

5.2 Methodology

Derivation of atomic charges

In this study, two different charge schemes namely Scheme I and II, were utilized to obtain the atomic charges of the iron center and selected residues in the first and second coordination shells of the metalloprotein. In Scheme I, the atomic charges of iron and amino acid residues in the first coordination shell namely Cys6, Cys9, Cys39 and Cys42 were obtained using DFT calculation in the gas phase. Similarly, these residues were also considered for DFT charge fitting in Scheme II with the inclusion of four other amino acids namely Thr7, Val8, Gly43 and Val44 (Ala44 and Gly44 for mutants), which were located in the second coordination shell and formed hydrogen bonds with cysteine residues that are directly coordinated to the metal center. Amino acids included for DFT calculation in Scheme I and II are showcased in Figure 5.2. The additional residues considered in Scheme II reflected the greater consideration of the electrostatic environment surrounding the iron atom compared to Scheme I. This way, we will be able to shed some light on the pivotal role of electrostatic environment in the modeling of metalloproteins and in the determination of their reduction potentials.

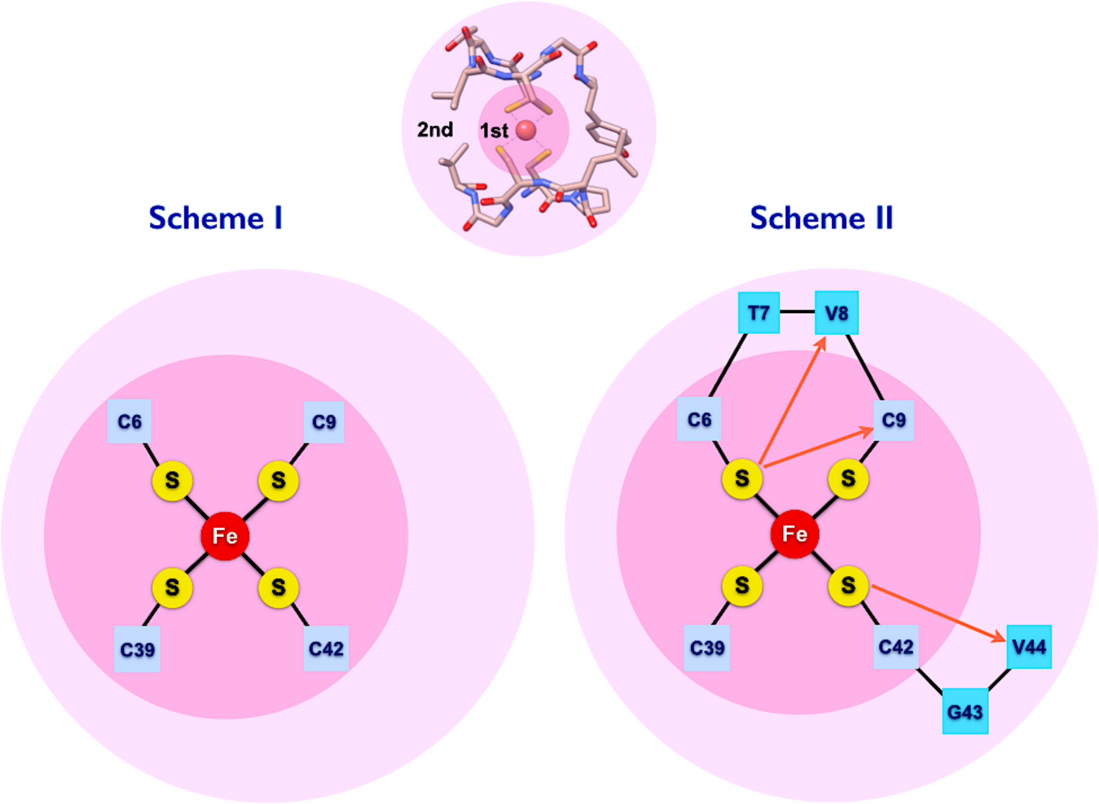


Figure 5.2: Licorice representation of iron atom and residues in the first and second co-ordination sphere. Schematic representation of amino acid residues isolated for DFT charge calculation in Scheme I and II.

In conducting the DFT calculations, the amino acids isolated from rubredoxin for charge derivation using Scheme I and II were capped using CH_3CO-$NHCH_3$. The coordinates of the CH_3CO-$NHCH_3$ caps replaced the coordinates of the cutoff protein in order to minimize the introduction of artifacts resulting from the terminating groups occupying atomic positions which were previously vacant. QM calculations in Scheme I and II were conducted using Gaussian09 at B3LYP/6-31G* level. [42-44] After QM calculations, the charge fitting of the residues highlighted in Figure 5.2 were conducted using the RESP (**R**estrained **E**lectro**s**tatic **P**otential) method while the remaining atoms of the proteins were assigned AMBER ff03 charges. [41, 45]

System Preparation

Modeling of the wild type rubredoxin was conducted by acquiring the X-ray crystal structure of the protein from the Protein Data Bank (PDB) with PDB id of 5RXN. [34, 35, 46] Hydrogen atoms were added to the structure of rubredoxin and the protein immersed in an octahedral TIP3P water box with minimal distance between the protein and the box boundary set to 10 Å using the LEaP module in AmberTools 1.2. [47] The rubredoxin-water system was subsequently subjected to energy minimization and the resulting protein structure was used as a template for the manual mutation of the wild type protein using LEaP module to obtain the rubredoxin mutants namely L41A, V44A and V44G. [47] The prepared mutants were then relaxed to eliminate improper bond lengths which may impede the convergence during QM calculations. All parameters used were consistent with AMBER ff03 force field with the exception of the partial charges of the residues featured in Figure 5.2, which were acquired using charge schemes I and II outlined above. [41] The redox center of the protein was represented using a non-bonded model with the necessary van der Waals parameters for iron obtained from Giammano et al. ($R=1.20$ Å, $\varepsilon=0.05$ kcal/mol) which correspond to the VDWs parameters of heme. [48]

MD Free Energy Simulation

To attain the free energy expended during the reduction process ($Fe^{3+} + e^{-} \rightarrow Fe^{2+}$) of rubredoxin at its redox center, MD simulations with the implementation of thermodynamic integration (TI), also known as the free energy perturbation method, were conducted. In applying the TI formalism for the measurement of free energy afforded during the reduction process, the rearrangement of the metalloprotein from its initial to the final state were accounted by a linear potential energy function given below:

$$U(\lambda) = (1-\lambda)U_{Ox} + \lambda U_{Rd} \quad (5.1)$$

where U_{Ox} and U_{Rd} correspond to the potential energies of the oxidized and reduced states of the metalloprotein respectively. λ represents the coupling parameter which range from 0 to 1 and the choice of λ is vital for the precise modeling of the reduction process. Another equation of interest in the application of TI in MD simulation is the Gibbs free energy difference between the oxidized and reduced states of the metalloprotein of interest:

$$\Delta G = U_{Rd} - U_{Ox} = \int_0^1 d\lambda \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda \quad (5.2)$$

where the brackets represent a cumulative ensemble average of the partial derivatives of the potential energies calculated at each λ value.

Using TI resources in the sander module of AMBER 10 simulation package, a total of fourteen MD simulations were conducted using coupling λ values ranging from 0 to 1, as performed by Satelle et al. [47, 49, 50] The free energy expended during the reduction process of the four rubredoxin variants namely wild type, L41A, V44A and V44G were calculated using equation 5.2 and the 14-point Gaussian quadrature. The fourteen λ values utilized in this study were 0, 0.00922, 0.04794, 0.11505, 0.20634, 0.31608, 0.43738, 0.56262, 0.68392, 0.79366, 0.88495, 0.95206, 0.99078 and 1, which were obtained from the AMBER 10 manual. [47] From the Gibbs free energy calculated, the difference in the reduction potential of the mutated protein and wild type protein was computed to attain the relative reduction potential of each mutant. The equation used in the computation of the relative reduction potential is as given below:

$$\Delta E^\circ_{Mut} - \Delta E^\circ_{WT} = -\frac{\Delta\Delta G}{n\mathrm{F}} \quad (5.3)$$

where $\Delta\Delta G$ stands for the difference between the free energy obtained during the reduction processes of the mutants and native rubredoxin while ΔE^0 represents the reduction potential of the mutants or its wild type. The number of electrons involved in the reduction process of rubredoxin is denoted as n, which in this case is 1, and F is the Faraday constant.

To attain the relative reduction potentials of the mutated proteins, a total of fourteen MD simulations with distinct λ values, as listed above, were conducted for each of the rubredoxin variants. All rubredoxin variants used in this study were solvated in TIP3P water confined in an octahedral box with the minimum distance between the protein and the edge of the water box set to 10 Å, and the protein-water system was neutralized by adding sodium ions using the LEaP module. [47] Other than the exclusive partial charges allocated to the reduced and oxidized states of the rubredoxin variants, the rest of the force field parameters and coordinates used for the two states are the same. This approach was used in this study as the geometry of the redox center of rubredoxin was assumed to undergo minimal change after electron transfer since the reorganization energy of rubredoxin is relatively low. [4, 15-17, 35-40]

Preceding the MD simulation is the relaxation of the solvent molecules and the protein-solvent system for 2000 steps and 50000 steps respectively using the steepest descent method. Following the minimization step, the whole system was heated from 10 K to 300 K using a Langevin thermostat with a collision frequency of 2 ps^{-1} in NVT (canonical) ensemble. During the heating process, weak restraint of 10 kcal/mol/Å^2 was imposed on the protein to prevent erroneous fluctuation of the protein. Equilibration and production MD were conducted after the heating step, which were carried out using NPT ensemble for 200 ps and 1 ns respectively. A time step of 2 fs was used in all simulations conducted. The calculation of long-range electrostatic interactions proceeded using the particle mesh Ewald method while all bonds involving hydrogen atoms were constrained using the SHAKE algorithm. [51, 52] MD simulations for all the rubredoxin variants were conducted using the AMBER 10 program. [47] Standard abscissas and weights for each λ value were taken from the AMBER 10 manual. [47]

5.3. Results: Improving the prediction of the reduction potential of mutated rubredoxin

Multitudes of studies working towards the understanding of the physical and dynamical aspects of eT proteins have been conducted to harness information pertaining to factors that influence the diverse reduction potential of eT proteins. [5-11, 27] From these studies, three key elements accounting for the difference in the reduction potential of eT proteins of the same type have been pointed out and these factors include: (i) the electrostatic environment of the redox center contributed by neighboring amino acids, (ii) the development of hydrogen bonds in the vicinity of the redox center and (iii) the accessibility of the redox center to solvent penetration. [6-11, 16] From the factors listed above, the first two factors are considered in Scheme I and II respectively and this allowed us to examine the significance of considering the electrostatic environment of the metal atom in theoretically determining the reduction potentials of metalloproteins.

Partial charges of rubredoxin variants

Partial charges of wild type rubredoxin calculated in this study using Gaussian09 at B3LYP/6-31G* level of theory were compared to charges calculated by Gámiz-Hernández et al. using Jaguar 5.5 at B3LYP/LACVP** level in Table 5.1. [28, 36, 42-44, 53] The partial charges obtained for iron and iron-coordinated sulphur atoms of the cysteine residues in wild type rubredoxin are comparable to the partial charges derived by Gámiz-Hernández et al. hence substantiating the reliability of the charges derived in this study. [28]

Atom	Oxidation state	S-I	S-II	Gámiz-Hernández et al.[28]
Fe	Oxidized	1.130	1.086	0.978
	Reduced	1.334	1.285	1.083
Sγ (Cys6)	Oxidized	-0.663	-0.656	-0.602
	Reduced	-0.840	-0.857	-0.797
Sγ (Cys9)	Oxidized	-0.506	-0.500	-0.580
	Reduced	-0.736	-0.728	-0.797
Sγ (Cys39)	Oxidized	-0.639	-0.643	-0.602
	Reduced	-0.824	-0.824	-0.797
Sγ (Cys42)	Oxidized	-0.592	-0.502	-0.580
	Reduced	-0.822	-0.700	-0.797

Table 5.1: Partial charges of iron and iron-coordinated sulphur atoms of cysteine residues of wild type rubredoxin obtained using charge schemes I and II and that obtained by Gámiz-Hernández et al. for comparison. [28]

Comparing the partial charges of the iron atom derived using Scheme I and II with the formal charges of the iron atom in the oxidized and reduced states, a drop in partial charge of the iron atom was apparent with the inclusion of more electrostatic environment during charge fitting. (Table 5.2) This is in agreement with the behavior of redox proteins in nature which tend to exhibit sizable redistribution of the charges in the redox center due to the strong interactions between the metal, specifically a transition metal, and the neighboring amino acids. [54] This emphasizes the importance of considering the electrostatic environment of the redox center when modeling metalloproteins theoretically.

Residue	Atom	Oxidation state	Formal Charge	S-I	S-II
Fe	Fe	Oxidized	3.000	1.130	1.086
		Reduced	2.000	1.334	1.285
Cys6	Cβ	Oxidized	-0.241	-0.034	-0.018
		Reduced	-0.241	-0.050	0.011
	Sγ	Oxidized	-0.884	-0.663	-0.656
		Reduced	-0.884	-0.840	-0.857
Val8	N	Oxidized	-0.450	-0.450	-0.153
		Reduced	-0.450	-0.450	-0.125
	H	Oxidized	0.440	0.440	0.213
		Reduced	0.440	0.440	0.212
Cys9	N	Oxidized	-0.416	-0.444	-0.431
		Reduced	-0.416	-0.576	-0.512
	H	Oxidized	0.272	0.163	0.080
		Reduced	0.272	0.222	0.117
	Cβ	Oxidized	-0.241	0.057	0.033
		Reduced	-0.241	0.074	0.056
	Sγ	Oxidized	-0.884	-0.506	-0.500
		Reduced	-0.884	-0.736	-0.728
Cys39	Cβ	Oxidized	-0.241	0.078	0.039
		Reduced	-0.241	0.073	-0.004
	Sγ	Oxidized	-0.884	-0.639	-0.643
		Reduced	-0.884	-0.824	-0.824
Cys42	N	Oxidized	-0.416	-0.323	-0.272
		Reduced	-0.416	-0.444	-0.365
	H	Oxidized	0.272	0.072	0.034
		Reduced	0.272	0.131	0.082
	Cβ	Oxidized	-0.241	0.123	0.022
		Reduced	-0.241	0.143	0.035
	Sγ	Oxidized	-0.884	-0.592	-0.502
		Reduced	-0.884	-0.822	-0.700
Val44	N	Oxidized	-0.450	-0.450	-0.028
		Reduced	-0.450	-0.450	-0.037
	H	Oxidized	0.440	0.440	0.042
		Reduced	0.440	0.440	0.045

Table 5.2: Partial charges of iron and atoms in the first and second coordination spheres of wild type rubredoxin.

Besides the partial charge of iron, the partial charges of atoms interacting with the iron center were also examined. Electrostatic potential maps showcasing the charge distribution at the redox center of wild type rubredoxin in the oxidized state are depicted in Figure 5.3 for charges derived using both Scheme I and II. The electrostatic potential map of oxidized rubredoxin obtained using the formal charge of iron and the AMBER ff03 charge of protein was also included in Figure 5.3 for comparison.

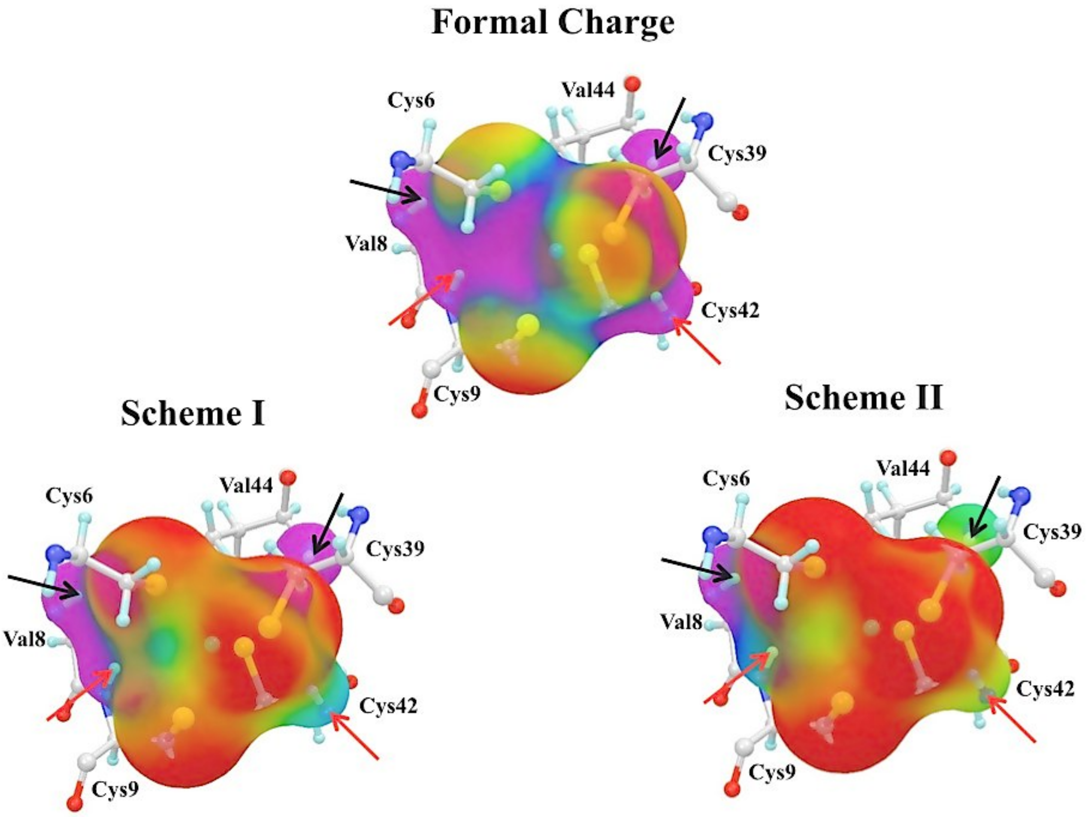

Formal Charge
Cys6
Val44
Cys39
Val8
Cys42
Cys9
Scheme I
Cys6
Val44
Cys39
Val8
Cys42
Cys9
Scheme II
Cys6
Val44
Cys39
Val8
Cys42
Cys9

Figure 5.3: Electrostatic potential map of wild type rubredoxin in oxidized state with charges acquired using Scheme I and II. Only iron, Cys6, Val8, Cys9, Cys39, Cys42 and Val44 are depicted in this figure for clarity. The red arrows point to the backbone amide hydrogen atoms of Cys9 and Cys42 while the black arrows point to backbone amide hydrogen atoms of Val8 and Val44. The potential energies of all electrostatic potential maps range from -0.1 to 0.1 kcal/mol and the colors go from negative to positive in the following order: red < orange < yellow < green < cyan < blue < magenta.

Rubredoxin in its oxidized state has a net negative charge of -1 at the redox center and this is illustrated in Figure 5.3 whereby the electrostatic potential around the redox center becomes more negative as more amino acids surrounding the metal center are factored into the QM calculations performed in Scheme I and II. At the same time, the more negative electrostatic potentials observed in Scheme I and II compared to the formal charge, around regions where the sulphur atoms of Cys6, Cys9, Cys39 and Cys42 coordinate to iron, suggest the instance of charge transfer from the coordinated Cys residues to the iron atom as seen in nature. [41] The backbone amide hydrogens (-NH) of Cys9 and Cys42 were also noted to have a more negative potential as we went from the use of mean-field charge to QM derived charges of Scheme I and II. These observations were aptly portrayed in Figure 5.3 (regions pointed by red arrows) and numerically supported in Table 5.2. The more negative potential observed for the amide hydrogen atoms of Cys9 and Cys42 in Scheme I and II concurred with the experimental observations reported by Xia et al. whereby the ^{15}N NMR study conducted revealed the delocalization of electrons along the Fe-γS---H-N pathway which resulted from the formation of hydrogen bonds between Cys6 and Cys9 and between Cys39 and Cys42. (Figure 5.1) [55]

The significance of long-range interactions between the iron atom and amino acids in the second coordination sphere was also considered through Scheme II with the inclusion of Thr7, Val8, Gly43 and Val44 during the DFT calculation. (Figure 5.2) The hydrogen bonds (HN---γS) between γS-Cys6 and HN-Val8 and between γS-Cys42 and HN-Val44, portrayed in Figure 5.2, have been documented in addition to the hydrogen bonds discussed in the previous paragraph. [4, 16, 34, 35] In Figure 5.3, the electrostatic potential of the backbone amide hydrogen atoms of Val8 and Val44, pinpointed by the black arrows, became more negative with the inclusion of residues in the second coordination shell in Scheme II. [41] This may be due to the delocalization of the electron density from the lone pair orbital of γS-Cys6 and γS-Cys42 to the backbone amide hydrogen atoms of Val8 and Val44 respectively, along the hydrogen bonds formed. [55-57]

Reduction potential of rubredoxin variants

The atomic charges derived using Scheme I and II were incorporated into the force field and free energy MD simulation was performed to calculate the free energy (ΔG) expended during the reduction process of rubredoxin. The partial derivatives of the potential energies, $\partial V/\partial\lambda$, attained through TI implementation were integrated over an interval of 0 to 1 to acquire ΔG (Equation 5.2). Hence, to ensure the accurate calculation of ΔG for the determination of the reduction potential of rubredoxin, ample simulation length is necessary to eliminate inaccuracies arising from the lack of convergence in the $\partial V/\partial\lambda$ attained through the TI calculation. As such, prior to the calculation of ΔG, it is essential for us to monitor the convergence of $\partial V/\partial\lambda$ at each λ value for all four rubredoxin proteins by plotting the graphs of cumulative $\partial V/\partial\lambda$ ($\langle\partial V/\partial\lambda\rangle$) against time as shown in Figure 5.4. Based on Figure 5.4, the convergence of the cumulative $\partial V/\partial\lambda$ was noted for wild type rubredoxin in Scheme II, particularly during the last 400 ps of the simulation from which the ΔG was calculated. Similar convergence pattern of $\partial V/\partial\lambda$ with time was observed for all the simulations conducted in this study hence were not be shown here.

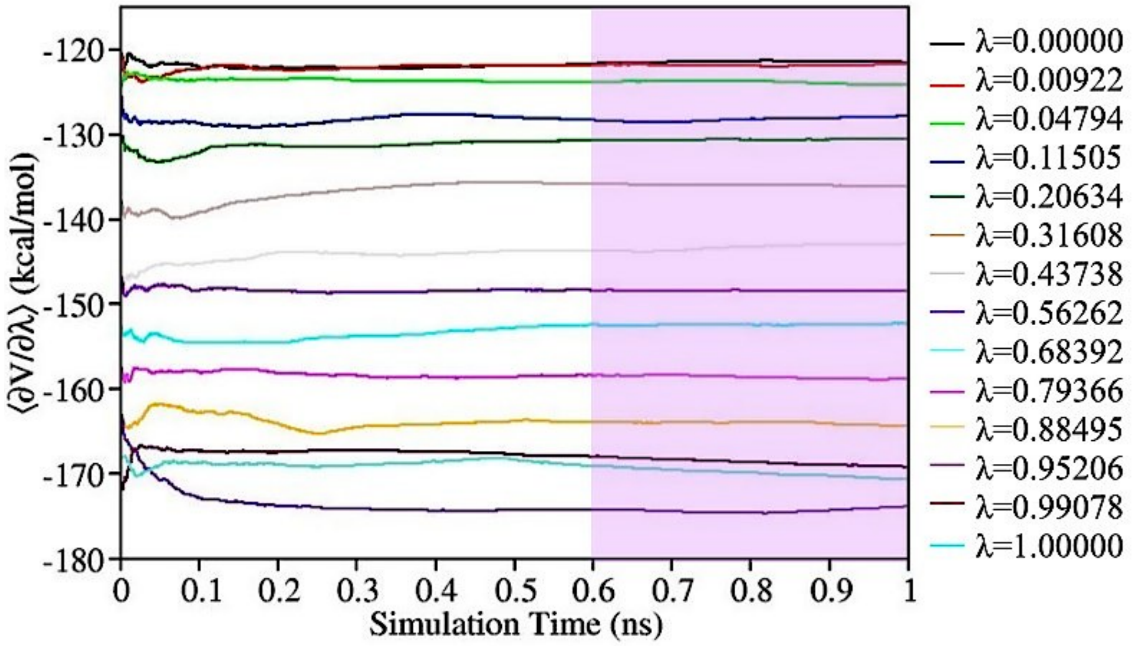


Figure 5.4: Plots of ⟨∂V/∂λ⟩ versus simulation time for wild type rubredoxin with partial charges derived using Scheme II.

Another step taken to ensure the accuracy of the reduction potential calculated was to examine the adequacy of the ∂V/∂λ sampling prior to the calculation of ΔG. This was carried out by plotting a graph of ⟨∂V/∂λ⟩ versus λ since a linear relationship between these two elements indicated the sufficient sampling of ∂V/∂λ through the MD simulations conducted. From Figure 5.5, a linear relationship between ⟨∂V/∂λ⟩ and λ was indeed observed suggesting the sufficient sampling of ∂V/∂λ for all rubredoxin variants. At the same time, the small standard deviation observed for all graphs exhibited in Figure 5.5, minimized the error inherited from the ∂V/∂λ calculated through TI for the calculation of ΔG using Equation 5.2, hence justifying the reliability of the relative reduction potential ($\Delta\Delta E_{\text{cal}}$) calculated in this study.

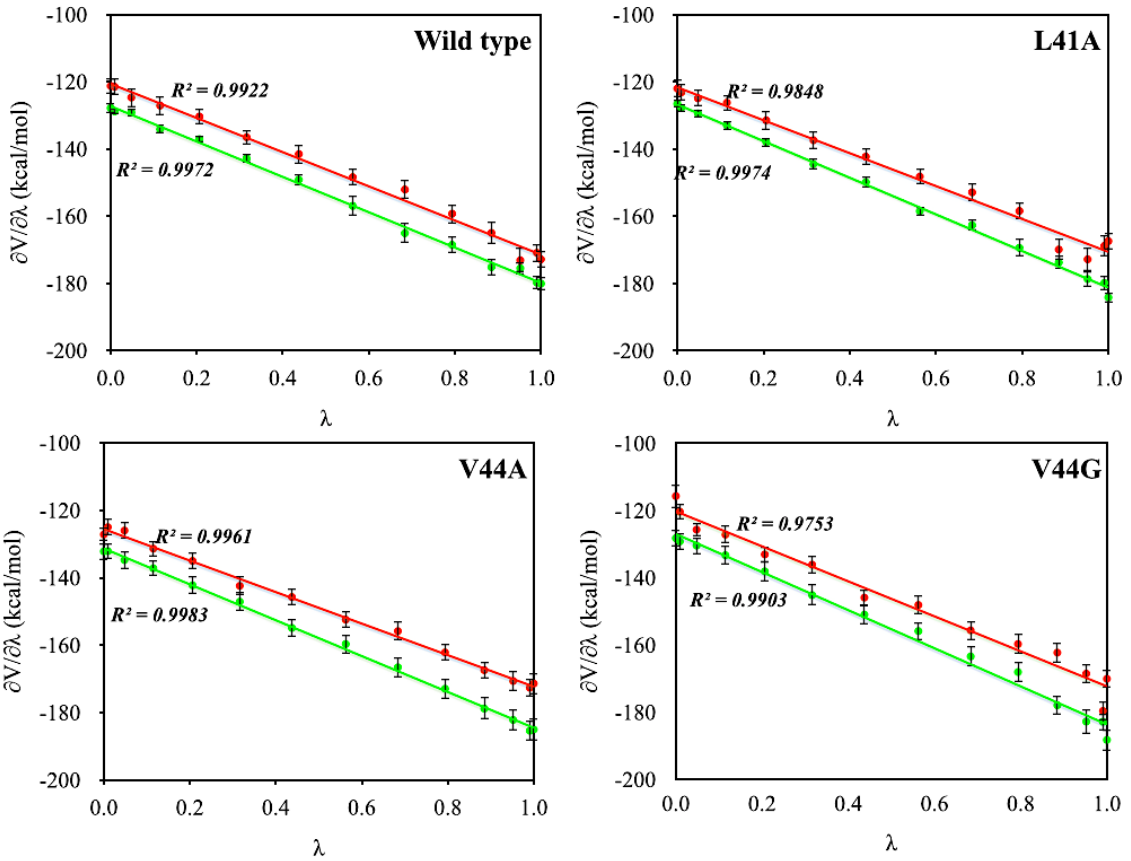
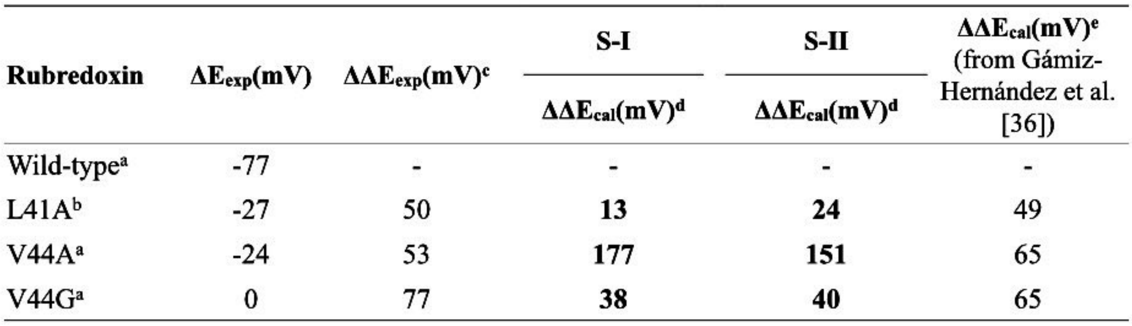


Figure 5.5: Plots of ⟨∂V/∂λ⟩ versus λ for rubredoxin variants namely wild type, L41A, V44A and V44G. Samplings of ∂V/∂λ are conducted by employing charges derived using Scheme I (green) and Scheme II (red).

After verifying the reliability of the simulations conducted, the relative reduction potentials ($\Delta\Delta E_{\mathrm{cal}}$), defined in Equation 5.3, of the mutated rubredoxin proteins were calculated. From the $\Delta\Delta E_{\mathrm{cal}}$ obtained, we observed a significant difference in the $\Delta\Delta E_{\mathrm{cal}}$ obtained using Scheme I and II albeit the slight variation in the partial charges derived through the two charge schemes. The $\Delta\Delta E_{\mathrm{cal}}$ obtained using Scheme II showed better agreement with experimental values compared to that derived using Scheme I. (Table 5.3) [36] This observation reflected the importance of evaluating the electrostatic interactions surrounding the redox center during DFT-based calculations as the consideration of additional residues in Scheme II led to a better prediction of $\Delta\Delta E_{\mathrm{cal}}$. This result is justifiable as the consideration of Val44/Ala44/Gly44 in Scheme II enabled us to consider the variation of the electrostatic potential at the redox center. Changes in the electrostatic potential of the redox center have been suggested by Ichiye et al. to be likely caused by the subtle shift in the protein backbone brought about by the difference in the size of the amino acid side chains of rubredoxin variants at position 44. [5, 16, 17, 37, 38, 58, 59] Hence, with the inclusion of four additional residues namely Thr7, Val8, Gly43 and Val44 in Scheme II, changes in the protein environment exerted by slight fluctuations in the protein backbone and side chains were accounted during the QM calculation.



Rubredoxin	**ΔE_{exp}(mV)**	**$\Delta\Delta E_{exp}$(mV)^c**	**S-I**	**S-II**	**$\Delta\Delta E_{cal}$(mV)^e** (from Gámiz-Hernández et al. [36])
			$\Delta\Delta E_{cal}$(mV)^d	**$\Delta\Delta E_{cal}$(mV)^d**	
Wild-type^a	-77	-	-	-	-
L41A^b	-27	50	**13**	**24**	49
V44A^a	-24	53	**177**	**151**	65
V44G^a	0	77	**38**	**40**	65

(a) Experimental data obtained from Xiao et al. [58]

(b) Experimental data obtained from Park et al. [59]

(c) $\Delta\Delta E_{exp}$ is the difference between the reduction potential of the mutant and wild type protein acquired experimentally.

(d) $\Delta\Delta E_{cal}$ is obtained by calculating the difference in ΔG between mutant and wild type protein which is then converted to $\Delta\Delta E$ using the Nerst equation. (Equation 5.3)

(e) $\Delta\Delta E_{cal}$ obtained from Gámiz-Hernández et al. [36] for comparative purposes with the calculated $\Delta\Delta E_{cal}$ computed in this study.

Table 5.3: $\Delta\Delta E_{cal}$ of rubredoxin mutants *viz.* L41A, V44A and V44G based on charges acquired through Scheme I and II.

Nevertheless, the inclusion of residues in the second coordination sphere of the iron atom was not enough to ensure a precise determination of $\Delta\Delta E$ and this was showed by the less accurate reduction potential determined in this study as compared to that computed by Gámiz-Hernández et al. in Table 5.3. [36] This observation is understandable as the latter used a combination of three methods namely; (i) molecular mechanics that was used to acquire the most probable protonation states of rubredoxin by conducting self-consistent geometry optimization of side chains that may influence the redox potential, (ii) the inclusion of electrostatics energies by solving the linearized Poisson-Boltzmann equation, and (iii) the implementation of Metropolis-Monte-Carlo algorithm to perform the statistical averaging of conformations with lowest electrostatic energies, all of which improved the accuracy of the redox potential calculated albeit being more computationally extensive. [36] The first two steps of the protocol above were also iterated until a constant protonation state at a defined pH was achieved.[36] Furthermore, Gámiz-Hernández et al. also considered several crystal structures with different protonation pattern and side chain conformations in both oxidized and reduced states were all included during Boltzmann-averaging hence increasing the accuracy of the redox potential determined. [36] Despite the less impressive agreement of $\Delta\Delta E_{\mathrm{cal}}$ obtained using Scheme I and II with that of experiment, this study was still able to showcase the importance of incorporating the effect of the electrostatic environment in the modeling of the reduction process of metalloproteins.

5.4 Conclusion

The role of electrostatic environment in the modeling of the reduction process of rubredoxin has been examined through the employment of free energy MD simulations with the implementation of TI formalism. In conducting the simulations, two charge schemes namely Scheme I and II (Figure 5.2), were utilized to represent the difference in the electrostatic environment considered. This resulted in the reorganization of the partial charges at the redox site which concurred with mainstream opinion that the electrostatic description formulated in the mean field could not effectively reflect the real electrostatic interactions present in proteins.

The implementation of Scheme I and II in the determination of the $\Delta\Delta E$ of the three rubredoxin variants highlights the importance of considering the electrostatic interactions contributed by neighboring residues of the redox center in the modeling of metalloproteins. The inclusion of residues in the second coordination sphere during the DFT calculation considers the influence that the protein backbone and amino acid side chain have towards the reduction potential of rubredoxin resulting in the gradual approach of the predicted $\Delta\Delta E$ closer to the experimental data.

Chapter 6 Utilization of MD simulation to predict the effect of mutation on the stability of proteins

6.1 Introduction

The last few decades have seen the remarkable evolution of protein research from one which encompasses studies pertaining to the relationship between the proteins' structure and function to one involving the alteration of the structure and/or function of proteins to achieve desirable traits. [1-6] Protein stability and adaptation to extreme conditions are two protein attributes that are highly advantageous to industries and research laboratories for the benefits that include faster chemical reactions, enhanced substrate solubility, better immunity towards microbial contamination and easier storage and handling of proteins. [4-6] All of these facilitate the improvements in the efficiency of industrial processes and laboratory work. With the introduction of protein engineering, designing proteins with superior functions and better stability compared to their native counterparts is no longer impossible. [6-8] The growing popularity of protein engineering has led to the development of sophisticated tools that enables researchers to conduct protein modification to attain desirable protein traits, some of which include augmented protein stability under non-physiological conditions, better selectivity and catalytic activity of enzymatic proteins and improved electron transfer rate of redox proteins. [6-9]

In altering the structure and/or function of proteins experimentally or theoretically, the work of nature in the form of mutagenesis and sequence variation is adopted. [1, 6-8] Thorough understanding of the three dimensional structure of a protein prior to mutation is essential in order to prevent erroneous mutations that may cause the protein's tertiary conformation to be destabilized while simultaneously triggering the crippling of the native biological functionality of the protein. [1-8] The exploration of the three dimensional structure of a protein at the atomic level is possible through the incorporation of modern computational tools into protein engineering. [7, 8] The combination of MD simulation and protein engineering is gaining much interest as an affordable and effective alternative for rational protein design due to its ability to empower comprehensive observations of the intricate dynamics of biological processes related to proteins such as protein folding and unfolding, conformational changes and protein stability. [7, 8, 10] These interesting applications of MD simulation are especially useful in knowledge-based protein design for it enables researchers to gain insights on the feasibility of the mutation performed in preserving the overall three dimensional conformation of the native protein. Simultaneously, useful information of critical interactions such as hydrophobic interactions, van der Waals and hydrogen bonds, that may govern the stability of a protein and play a role in promoting the proper functioning of a protein could be attained through rigorous analyses of MD trajectories. These aspects of MD simulation decrease the likelihood of disrupting the tertiary structure of the native protein when mutagenesis are conducted experimentally. Hence, MD simulation is an attractive technology that can be harnessed to improve the efficiency while minimizing the cost of experimental work.

In this chapter, MD simulation will be utilized as the main tool to predict the stability of apomyoglobin upon mutation. Unlike previous chapters which emphasize the importance of polarization effect and electrostatic environment in the modeling of proteins through the implementation of QM calculations, this chapter will simply concentrate on the proficiency of classical MD simulation to provide reliable insights on the effects of mutation on the structure and function of a protein prior to experiment. Apomyoglobin (apoMb) is a small protein derived from myoglobin through the removal of heme. [11-13]

At near neutral pH, native apoMb consists of eight helical domains labeled A to H (Figure 6.1) and its overall conformation resembles that of myoglobin with the exception of a few structural changes such as the partial unfolding of helix F, N-terminal end of Helix G and C-terminal end of Helix H. [13, 14] These slight dissimilarity in the tertiary structure of apoMb and holomyoglobin were addressed by Eliezer and Wright who observed through NMR studies the greater fluctuation of apoMb at loop EF, Helix F, loop FG and the N-terminus of Helix G compared to the fluctuation of these domains in holomyoglobin. [14] Being small and compact in structure, apoMb serves as an ideal model for the folding of globular, single-domain proteins in general due to the accessibility of the intermediate states of apoMb. [15-20] Furthermore, the popularity of apoMb ensures the accessibility of its experimental data which is important in computational studies where empirical information is very much welcomed.

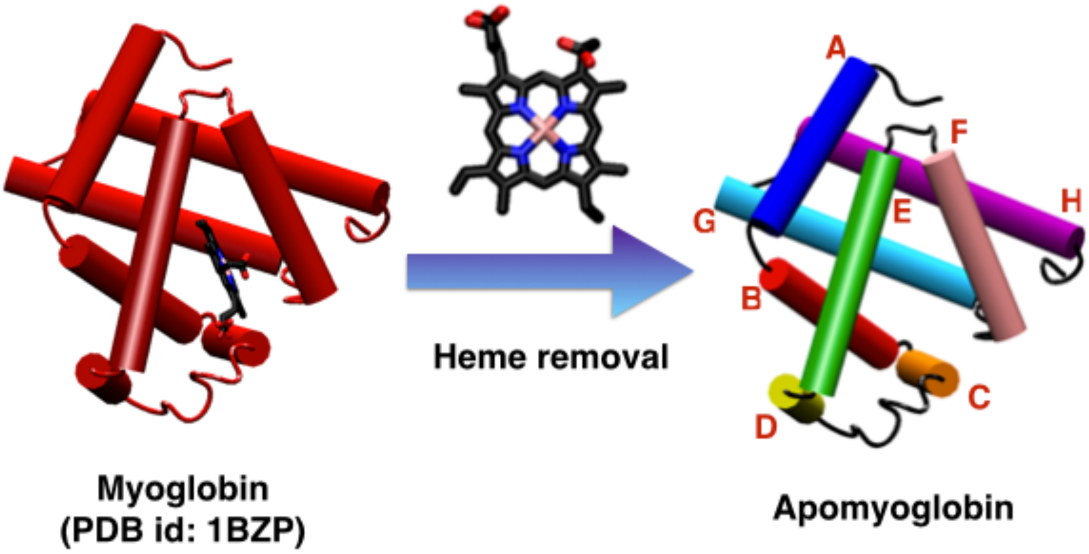


Figure 6.1: Cartoon representation of myoglobin and apomyoglobin with the individual helices alphabetically labeled A to H.

In this study, MD simulations for wild type apoMb and three of its mutants namely E109A (Helix G), E109G (Helix G) and G65A/G73A (Helix E) were conducted in explicit urea solution, which acts as the denaturant initiating the unfolding of apoMb. The trajectories obtained were analyzed to distinguish the difference in the stability of the three apoMb variants relative to the wild type protein. The stability of the apoMb variants determined in this theoretical work will be compared against the stability determined experimentally by Luo et al. [11]

6.2 Methodology

In this study, two MD simulations were performed for each apoMb variants using AMBER10 simulation package. [21] To model the wild type apoMb, X-ray crystal structure of wild type myoglobin from sperm whale was obtained from the Protein Data Bank (PDB) with PDB id of 1BZP. [22, 23] The heme moiety, crystallization waters and sulphate ions included in the PDB structure of myoglobin were deleted before any calculations were performed. Prior to simulations conducted for the apoMb variants in explicit urea solution, the prepared wild type apoMb was solvated in TIP3P water box and relaxed for 500 ps at neutral pH upon heme removal. [24] The details of this simple equilibration are provided in Appendix A located at the end of this thesis. From this relaxation MD, the structure from the last frame was used for the manual mutation of residues to obtain the mutated proteins of interest. After mutation, missing atoms were added and all histidine residues were doubly protonated to represent apoMb at pH 4.2. The pKa of histidine residues were calculated using PROPKA prior to the determination of their protonation state at pH 4.2. [25-28] These steps were simplified with the aid of the LEaP module in AmberTools 1.2. [21]

To model the apoMb variants in explicit urea solution, the urea-water system was constructed by diluting a pre-equilibrated 8 M urea box available in AMBER 10 simulation package using TIP3P water molecules to obtain a urea concentration of approximately 2 M. [21, 24] The resulting explicit 2 M urea-water system comprised of 330 urea molecules and 9118 water molecules enclosed within a rectangular box of dimensions 69.58 Å by 66.60 Å by 72.52 Å. The prepared urea-water solution was then minimized for 10000 steps using the steepest descent method and continued with another 10000 steps of minimization using the conjugate gradient method. This is subsequently followed by the heating of the prepared urea-water solution from 10 K to 277.15 K in 100 ps using Langevin thermostat with a collision frequency of 4 ps^{-1} in canonical (NVT) ensemble. [29, 30] Following the heating step, a simulation under isothermal-isobaric (NPT) ensemble was conducted for 3 ns to ensure proper mixing of the urea-water mixture under constant temperature (277.15K) and pressure. Generalized Amber force field (gaff) was employed to conduct the simulation elaborated above and parameters and charges used for urea molecule were obtained from Özpınar et al. [31, 32]

After preparing the equilibrated 2 M urea box, wild type apoMb and mutated apoMb proteins namely E109A, E109G and G65A/G73A were each solvated in an octahedral urea-water box with the minimum distance between the protein and the edge of the box kept to 15 Å. Using the LEaP module, Na^+ and Cl^- ions were added to neutralize the system. [21] Solvent molecules surrounding the protein were minimized using the steepest descent method followed by minimization using the conjugate gradient method for 10000 steps each. This was ensued by the minimization of the whole protein-solvent system for 5000 steps using the steepest descent method with further minimization of the system thereafter using the conjugate gradient method until an energy gradient of 0.01 kcal/mol/Å was reached. The entire system was then heated from 10 K to 300 K for 100 ps using the Langevin thermostat with a collision frequency of 4 ps^{-1} in the NVT ensemble. [29, 30] Weak restrain with harmonic potential of 5 kcal/mol/$Å^2$ was imposed on the protein during heating to counteract large fluctuations of the protein when the temperature varied. Upon equilibration of the protein-urea system at 300 K, two production runs of 20 ns each were conducted for each protein using the NPT ensemble. A time step of 2 fs was used for all the simulations conducted with the long-range electrostatic interactions calculated using the particle mesh Ewald method. [33] The SHAKE algorithm was used to constrain covalent bonds involving hydrogen atoms. [34] The force field used to carry out the simulations outlined for protein-urea system are AMBER ff99SB and gaff. [31, 35, 36] Figure 6.2 showed a flowchart summarizing the procedures highlighted in this section for a clearer portrayal of the steps taken to conduct the simulation of apoMb in urea solution.

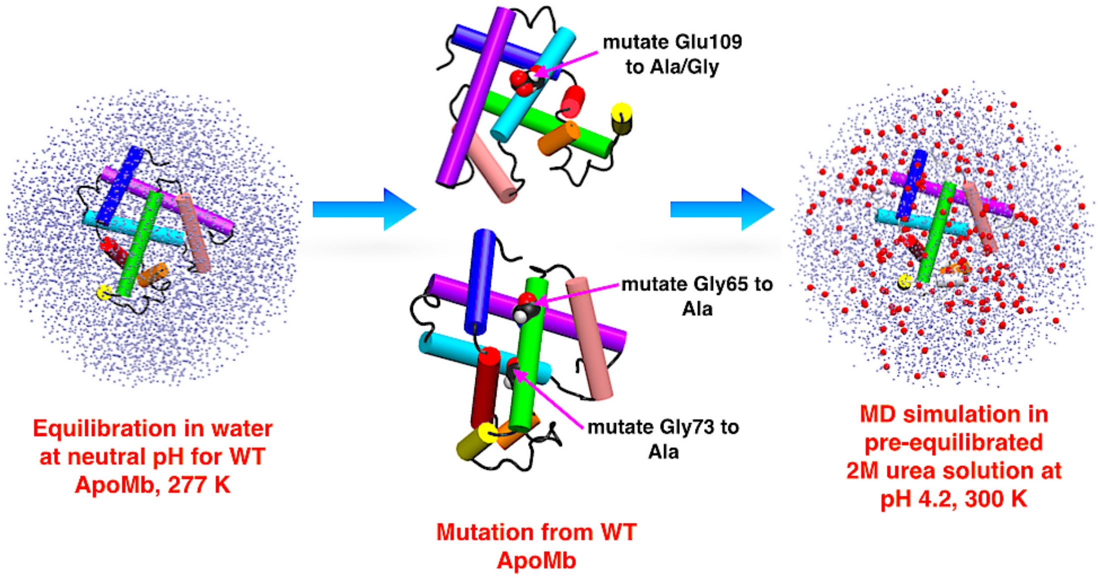

mutate Glu109 to Ala/Gly
mutate Gly65 to Ala
mutate Gly73 to Ala
Equilibration in water at neutral pH for WT ApoMb, 277 K
Mutation from WT ApoMb
MD simulation in pre-equilibrated 2M urea solution at pH 4.2, 300 K

Figure 6.2: Flowchart summarizing the modeling of apoMb variants in explicit 2 M urea solution.

6.3 Results: Comparing the stability of apomyoglobin upon mutation

To compare the stability of the four apoMb variants namely wild type, E109A, E109G and G65A/G73A theoretically, MD simulations of these proteins were conducted in the presence of explicit urea solution which was employed to effectuate accelerated protein denaturation. The theoretical comparison of the stability of the aforementioned apoMb variants will be compared to the stability order of these four proteins determined experimentally by Luo et al. who scrutinized the effect of Ala → Gly mutation on the stability of apoMb molten globule. [11] Luo et al. studied the effect of mutation on the stability of eight apoMb mutants namely wild type, E109A, E109G, Q8A, Q8G, K140A, G23A/G25A and G65A/G73A, by monitoring the reversible unfolding of these proteins in urea solution using circular dichroism and fluorescence spectroscopy. [11] From the difference in the free energy calculated by Luo et al. with respect to the wild type protein, E109A showed better stability compared to wild type while E109G showed the least stability among the four variants utilized in this theoretical study. [11] The doubly mutated protein variant (G65A/G73A), on the other hand, showed insignificant destabilization effect on apoMb. [11]

Trajectories acquired from the simulations performed were analyzed to validate the reliability of MD simulation in determining the stabilities of the four apoMb variants by comparing the results obtained theoretically against the experimental results attained by Luo et al. [11] The extent of denaturation of the protein during the 20 ns simulation conducted was observed by examining the protein dynamics of apoMb in urea solution and the conformational changes that took place during the simulation. Since two simulations were conducted for each apoMb variants, data analyzed from the two simulations were averaged and presented in this thesis.

Root-mean-square deviation (RMSD)

Calculating the RMSD of the folded protein relative to the experimental structure is one common option used in theoretical studies to quantify the degree of conformational changes that happen during the course of the simulation conducted. To compare the extent of variation in the protein conformation during the simulation, RMSDs of the apoMb variants relative to the crystal structure of sperm whale myoglobin (1BZP) were calculated. [23] Only the conformational changes of Helix F to H were considered during the calculation of the RMSD of apoMb variants relative to 1BZP since the most significant structural difference between myoglobin and wild type apoMb, as observed by Eliezer and Wright through NMR, lies mostly in Helix F and part of Helix G and H. [14, 23] Based on Figure 6.3 A, the RMSD of Helix F to H of all apoMb variants showed a concomitant increase with time. This concurred with the presence of urea molecules in the explicit solvent and the low pH condition which accelerated the unfolding of apoMb.

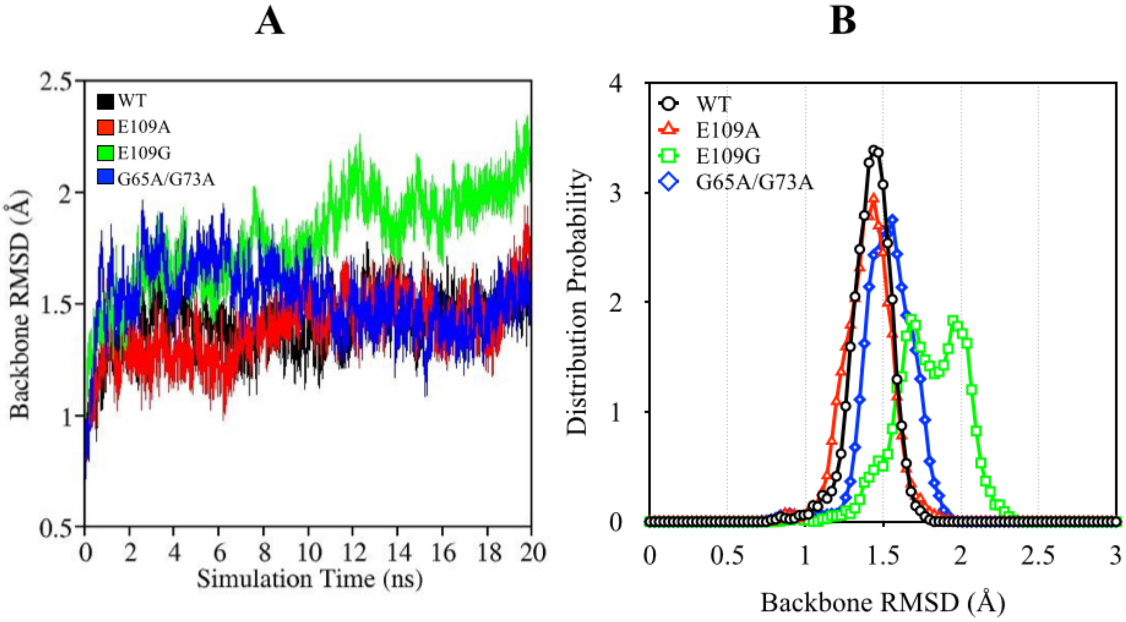


Figure 6.3: (A) Variation of RMSD with time and (B) the distribution of the RMSD of wild type apoMb (black), E109A (red), E109G (green) and G65A/G73A (blue) at Helix F to H.

Utilizing the RMSD plot for wild type apoMb as the reference, we compared the stability of the mutants by scrutinizing the fluctuations in the RMSDs of the three mutants relative to that of the wild type protein. A greater increase in the RMSD of a apoMb variant would signify a more significant destabilization of the protein upon mutation. Based on Figure 6.3 A presented above, the plot of RMSD with time for E109G showed the largest hike in RMSD among the protein variants during the last 10 ns of the simulation and this substantiated the destabilizing effect that E109G has on apoMb as accounted by Luo et al. in Ref 11.

Due to the proximity of the RMSD plots for wild type apoMb, E109A and G65A/G73A, the RMSD distributions for all apoMb variants were plotted, as showed in Figure 6.3 B, to clearly distinguish the degree of configurational changes that occurred during the simulation of each protein. In the study conducted by Luo et al., the free energy differences calculated for E109A, E109G and G65A/G73A with reference to wild type apoMb were 0.17, -0.89 and -0.11 kcal/mol respectively. [11] The variation in the stability of the apoMb variants portrayed numerically by Luo et al. is effectively reproduced in Figure 6.3 B which displayed the slight shifting of the RMSD distribution curve of E109A to lower RMSD values denoting better stability of apoMb with E109A mutation. On the other hand, the plots of the RMSD distribution for E109G and G65A/G73A were displaced to the right side of the graph towards larger RMSDs evincing the destabilization of apoMb when E109G mutation and G65A/G73A mutation are in place. In addition, the slight destabilization caused by G65A/G73A mutation evident from the numerical calculation performed by Luo et al. (*supra*) was also successfully reflected in Figure 6.3 B by the smaller shift in the RMSD distribution curve of the doubly mutated apoMb compared to that of E109G. These observations showed the sensitivity and reliability of MD simulation as a workable tool for predicting the stability of protein prior to experimental mutation of proteins.

Native Contacts

Even though the RMSD plots of the four apoMb variants concurred with the stability order obtained experimentally, determination of protein stability solely on the extent of conformational changes may not be feasible for some proteins. Therefore, it is imperative for us to explore intramolecular interactions within the protein as these interactions may provide us with a better gauge of the stability of the proteins following mutation(s) compared to the RMSD calculation of these proteins. One method that can be implemented to investigate the intramolecular interactions present in the apoMb variants is through the calculation of the fraction of native contacts conserved during the simulation. Calculating the native contacts numerically quantifies the amount of interactions between amino acids which are spatially next to each other but are not sequentially adjacent to each other in the primary sequence. ApoMb variants with destabilizing mutation tend to unfold at a much faster rate compared to their counterparts over the same period of simulation time leading to a greater loss in native contacts and this trend could be use as a guide in determining the stability of the mutants relative to wild type apoMb.

In Figure 6.4 presented below, the plots of native contacts conserved against time showed a descending trend that is in accordance with the presence of urea denaturants and low pH conditions. ApoMb with destabilizing E109G mutation exhibited the largest dive in native contacts conserved as the simulation progresses. Contrariwise, the apoMb variant with E109A mutation showed the highest preservation of native contacts showcasing the enhanced stability of apoMb when E109A mutation is in place. On the other hand, the percentage of native contacts preserved in G65A/G73A is comparatively similar to the wild type protein supporting the none to slight destabilizing effect of this mutation as documented by Luo et al. [11]

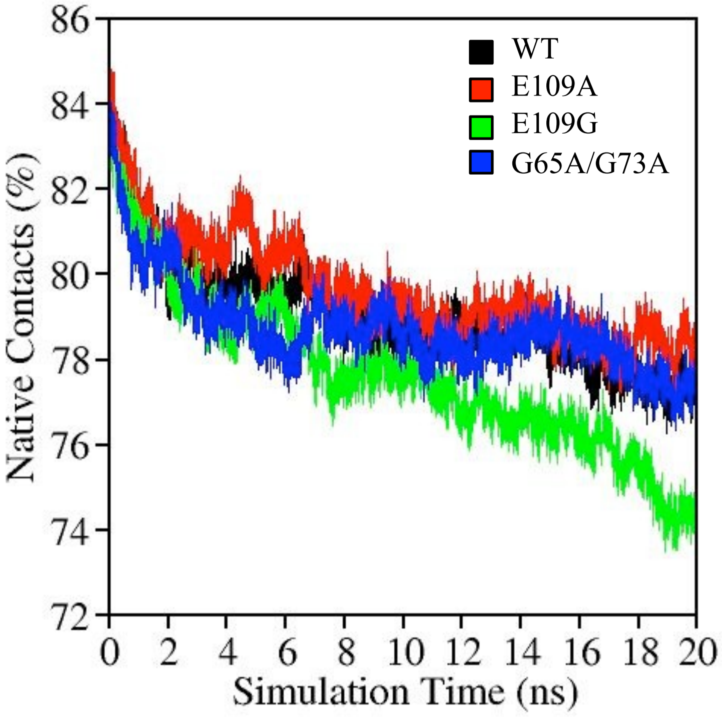


Figure 6.4: Variation in native contacts with time for wild type apoMb (black), E109A (red), E109G (green) and G65A/G73A (green).

Solvent Accessible Surface Area (SASA)

Besides the overall consideration of intramolecular interactions in terms of native contacts preserved, hydrophobic interaction established among non-polar amino acid residues is a key interaction that warrants attention in inquiring the stability of globular proteins. The formation of hydrophobic interactions assured the stability of globular proteins in solution by ensuring the protection of non-polar amino acids from the aqueous environment by encapsulating the hydrophobic residues within the hydrophobic core. [37] Experimental studies often exploit the intrinsic fluorescence property of proteins to provide responsive signals to the variation of the solvent accessibility of the hydrophobic core caused by changes in the tertiary structure of proteins during folding and unfolding. [11, 13, 38-43] The variation of the solvent accessibility of the hydrophobic core is routinely observed experimentally using UV fluorescence spectroscopy. [11, 13, 38-43] Protein unfolding is usually marked by the hydration of the hydrophobic core while protein folding is distinguished through the occurrence of hydrophobic collapse, an event that involves the sequestering of the hydrophobic residues away from the aqueous environment. [11, 13, 38-43] Unfolding of proteins during denaturation inevitably caused the exposure of the hydrophobic core to the aqueous environment leading to the loss of hydrophobic interactions within the non-polar amino acid clusters. Structural changes pertaining to the folding and unfolding of apoMb are commonly scrutinized by keeping track of the changes in the fluorescence emission of Trp7 and Trp14. [11, 13, 38-43] These amino acids are components of the hydrophobic core of apoMb which are confined by helices A, G and H. (Figure 6.5) [11, 13, 38-43]

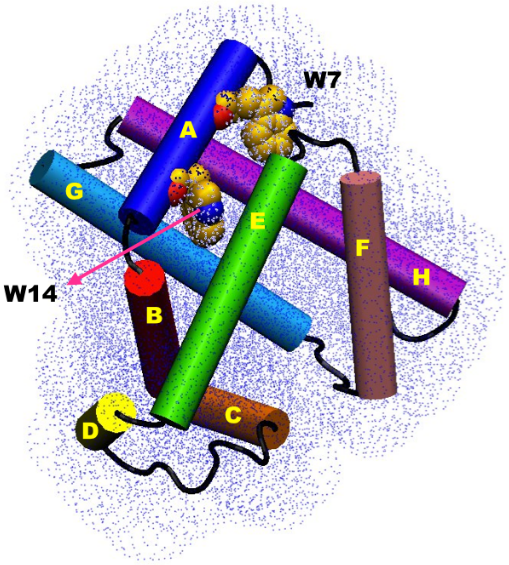


Figure 6.5: Schematic representation of the solvent accessible surface area (SASA) of apoMb with the helical domains and two residues found within the hydrophobic core namely Trp7 and Trp14 labeled.

In order to observe through theoretical means the changes in the accessibility of the four apoMb proteins to the surrounding solvent, the variation of the solvent accessible surface area (SASA) of the protein with time was computed. SASA, as the name suggests, measures the surface area of the protein that is unsheltered from the surrounding environment. During the process of denaturation, the SASA of proteins will naturally expand due to the hydration of the hydrophobic core caused by the disruption of hydrophobic interactions among non-polar clusters. The graph of SASA versus time in Figure 6.6, displayed an ascending trend in SASA of the four apoMb variants suggesting the solvation of the hydrophobic core of wild type apoMb, E109A, E109G and G65A/G73A. However, the extent of the hydration of the hydrophobic core differed across the four variants with E109G showing the steepest increase in SASA, denoting the decline in the stability of the apoMb protein upon E109G mutation. (Figure 6.6) Contrary to that of E109G, the SASA of the apoMb protein with E109A mutation showed a more gradual increase in SASA with time compared to wild type apoMb while the SASA of G65A/G73A apoMb showed a relatively similar increase in SASA with time to that of the wild type protein. (Figure 6.6) The stability order suggested through the calculation of SASA of the whole protein concurred with experimental observations whereby stabilization of apoMb was noted with E109A mutation while the destabilizing effect of E109G and G65A/G73A mutation was discerned, with the latter having none to slight destabilizing effect on apoMb compared to the former. [11]

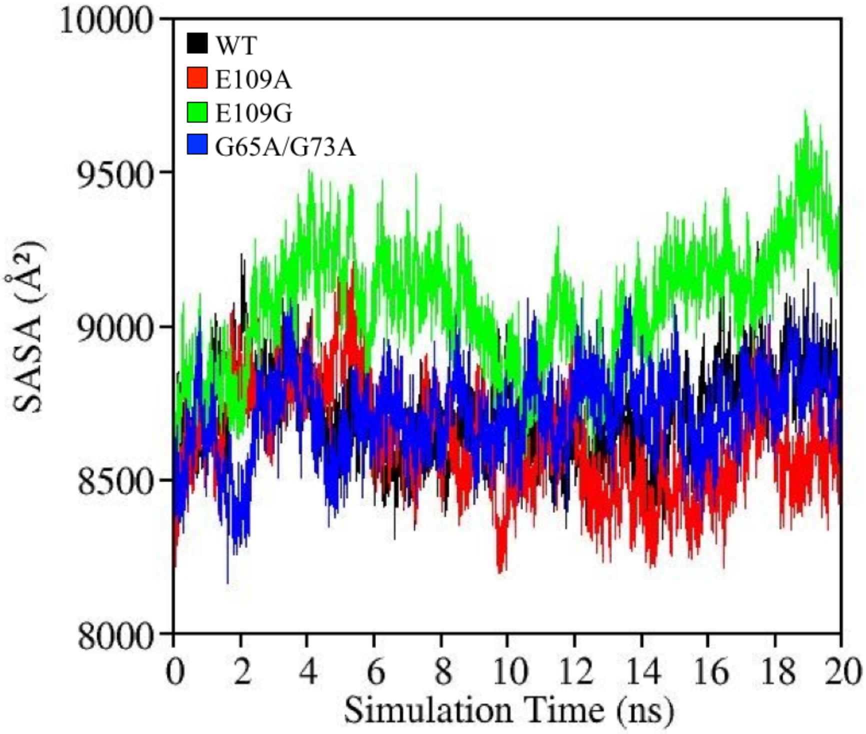


Figure 6.6: Variation of the solvent accessible surface area (SASA) of wild type apoMb (black), E109A (red), E109G (green) and G65A/G73A (blue) with time.

In addition to the SASA of the entire apoMb proteins, the stability order of the four apoMb variants was also acquired by calculating the variation of the SASA of Trp7 as the simulation progresses. Since the simulations were adequately performed to observe the early unfolding of apoMb, only the SASA of Trp7 was measured as this residue is adjacent to the surface of the protein and may provide a more substantial change in SASA compared to Trp14 which remain embedded in the hydrophobic core during the simulation. (Figure 6.5) [38-43] Computing the SASA of Trp7 enabled the direct inspection of the solvent accessibility of the hydrophobic core since Trp7 is situated in the hydrophobic core. [38-43] Plots of the SASA of Trp7 versus simulation time exhibited in Figure 6.7 A agreed with the stability order documented by Luo et al. for E109A, E109G and G65A/G73A mutants. [11] In order to illustrate the accessibility of Trp7 to the surrounding solvent molecules with better clarity, the distribution curves of the SASA calculated for Trp7 residue in the four apoMb variants were also plotted in Figure 6.7 B.

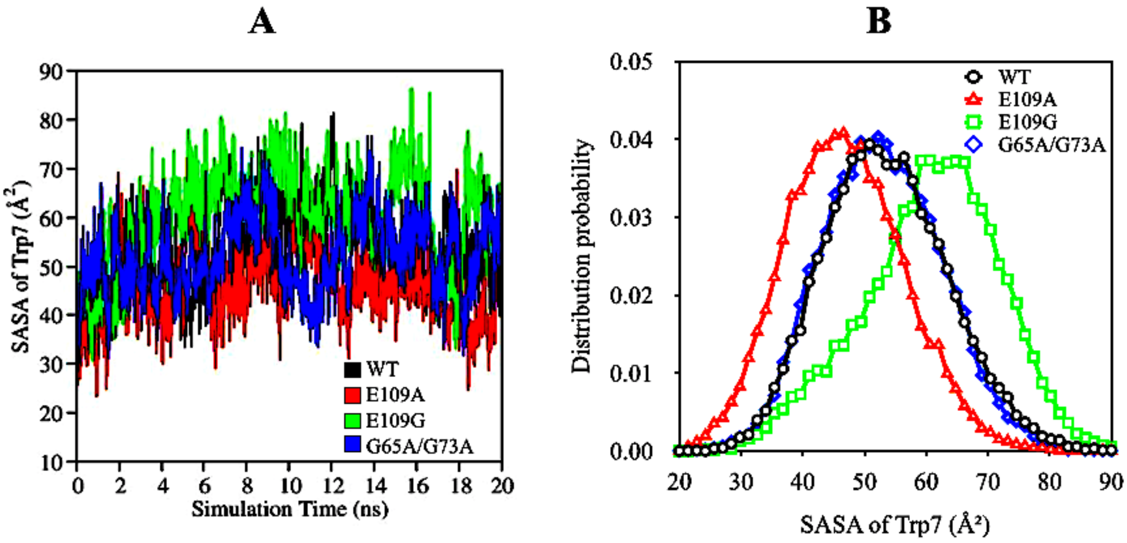


Figure 6.7: (A) Changes in the solvent accessible surface area (SASA) of Trp7 with time and (B) the distribution of the SASA of Trp7 in wild type (black), E109A (red), E109G (green) and G65A/G73A (blue).

The peaks of the distribution curves of both wild type apoMb and G65A/G73A overlap at approximately similar SASA values demonstrating the minimal effect that the double mutation has on the stability of apoMb in solution. Using wild type apoMb as the standard, enhanced protection of the hydrophobic core of apoMb with E109A mutation was observed as the distribution curve of E109A in Figure 6.7 B peaked at a smaller SASA value indicating the poor solvent accessibility of Trp7 compared to wild type apoMb. On the other hand, the hydrophobic core of apoMb became more accessible to the aqueous environment upon E109G mutation as the largest peak value was recorded for the distribution curve of E109G in Figure 6.7 B. These observations showcased the significance of hydrophobic core stabilization in securing the stability of apoMb as minimal exposure of the hydrophobic core to the surrounding solvent, as portrayed in E109A, led to the enhanced stability of the mutants relative to wild type apoMb.

Correlation Map

On top of comparing the stability of the apoMb variants based on their overall conformations (*vide supra*), the dynamics aspects of the unfolding of apoMb variants in urea solution at low pH were also explored to gain better understanding of the disparity in the dynamic activities of the proteins which led to the difference in stability. Correlation matrix is one useful method that is often utilized to depict the complex dynamical details of proteins in simple two-dimensional graphs. In this study, correlation matrices of four apoMb variants were calculated to observe the correlation in the dynamics of the helices and loops of apoMb during the last 5 ns of the simulation. Only one out of two trajectories acquired was used to obtain the correlation matrices of the four apoMb variants. The red regions on the correlation map portrayed in Figure 6.8 represent the concerted motion of residues in a similar direction while the blue regions represent the non-correlated movement of residue pairs. Correlated motions between the structural domains of apoMb may be the result of the formation of short-range or long-range interactions among helices and loops of apoMb. Inferences pertaining to the loss or gain in interactions between or among secondary structures of the apoMb variants could be made by scrutinizing the correlation maps plotted in Figure 6.8 and having a thorough understanding of the three-dimensional structure of apoMb.

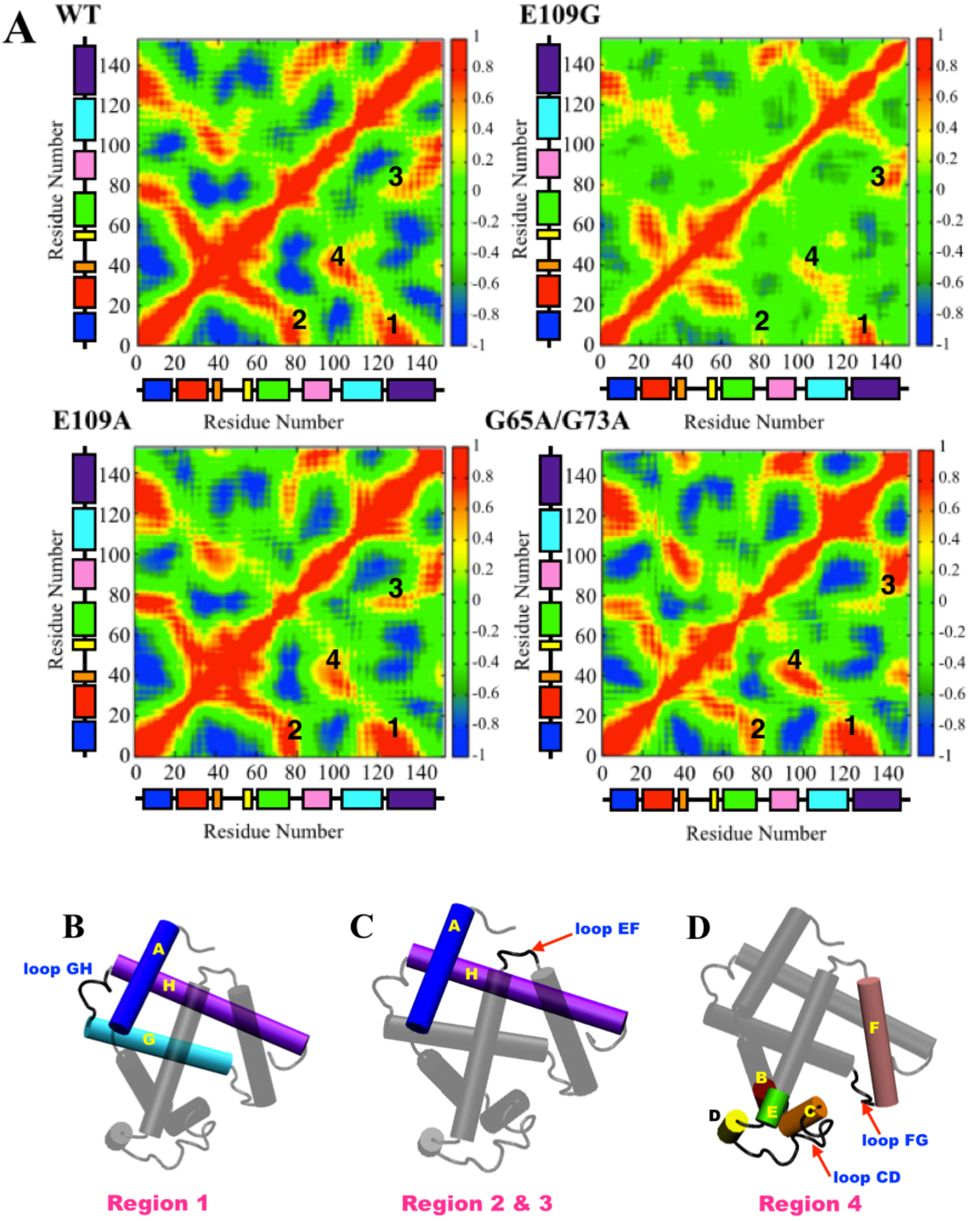


Figure 6.8: (A) Correlation maps of wild type apoMb, E109A, E109G and G65A/G73A with the eight helical domains represented by color-coded boxes going from Helix A (blue) to Helix H (purple). Schematic representations of apoMb with the helices and loops involved in (B) Region 1, (C) Region 2 and 3 and (D) Region 4 highlighted.

Four regions, with contrasting correlation intensities among the wild type apoMb, E109A, E109G and G65A/G73A have been pointed out and labeled accordingly in Figure

6.8 A above. Region 1 depicts the correlated movements of loop GH and helices A, G and H that encircled the hydrophobic core and this feature is commonly termed as the AGH-core. [13, 14, 41, 44-47] The assembly of the AGH-core is one of the early events that occur during the folding process of apoMb and this feature is highly stable under low pH and at high temperatures albeit the partial unfolding of helices A, G and H. [13, 14, 41, 44-47] The conservation of the AGH-core was evident by the positive correlation observed in Region 1 for all apoMb variants signifying the concerted movement of structural domains that made up the AGH-core. However, the difference in the intensities of the dynamical correlation of helices A, G and H across the four apoMb variants implied the difference in the stability of the AGH-core which may arise as a consequence of the disruption of hydrophobic interactions in the presence of urea and low pH conditions.

AGH-core destabilization is the most discernible in E109G as seen by the weaker correlation in motion for residues in loop GH and helices A, G and H compared to the other apoMb variants. (Figure 6.8 A and Figure 6.8 B) In contrast, the residues in the AGH-core of E109A and G65A/G73A demonstrated a better correlation compared to wild type apoMb, connoting the augmentation of the stability of the AGH-core of apoMb by performing E109A and G65A/G73A mutation. Even though the results obtained for E109A and E109G are consistent with the experimental results reported by Luo et al., G65A/G73A mutation which was documented to cause a slight disruption in the stability of apoMb was shown to stabilize the AGH-core. [11] In this case, the improved stability of the AGH-core in G65A/G73A might suggest the likelihood of additional structural factors contributing to the destabilization of the overall conformation of apoMb.

Loop EF is a structural feature proximal to the AGH-core and may very likely aid in the stabilization of the hydrophobic center. (Figure 6.8 C) The regions labeled 2 and 3 in Figure 6.8 A reflected the presence of interactions between loop EF and helix A and between loop EF and helix H respectively. (Figure 6.8 C) Weaker correlations in the dynamics of loop EF and helices A and H in regions 2 and 3 were noted for E109G and G65A/G73A compared to that of wild type apoMb. On the other hand, the correlation map of E109A displayed similar dynamical correlations for the domains in regions 2 and 3 as that of wild type apoMb. The lack of interactions between loop EF and helix A/H in E109G and G65A/G73A may have ensued due to the disruption of hydrogen bonds previously formed between Glu4 (Helix A) and Lys79 (loop EF) and between Asp141 (Helix H) and His 82 (loop EF) which will be discussed later in this chapter. The breaking of the aforementioned hydrogen bonds could possibly cause loop EF to move away from Helix A/H making the hydrophobic core more susceptible to solvent penetration. Based on the variation of the correlation intensities portrayed by the four apoMb variants in Region 1 to 3 in Figure 6.8 A, the importance of hydrophobic core stabilization in either maintaining or enhancing the overall stability of apoMb in solution was aptly reflected by E109A and wild type apoMb. Both proteins showed significant correlation in motion of residues in Region 1 to 3. The significant correlation in these regions signified the presence of interactions among structural domains enclosing the hydrophobic core (Helices A, G and H, and loop EF and GH). These interactions most likely play a crucial role in protecting the hydrophobic center from hydration, hence stabilizing the three dimensional conformation of apoMb.

Another feature of the apoMb that is of relevance to this study is the heme pocket which was left unoccupied after the removal of heme from myoglobin to generate apoMb. (Figure 6.1) Haliloglu and Bahar conducted a coarse-grained dynamic Monte Carlo simulation which revealed the substantial fluctuations of structural motifs located in the vicinity of the heme pocket namely Helix F, loop FG and sections of apoMb starting from the

C-terminus of Helix B and ending at the N-terminus of Helix E. [48] (Figure 6.8 D) The intensified fluctuations of the aforementioned structural motifs were reported by Haliloglu and Bahar to be triggered by the collapsing of these motifs into the vacant heme pocket. [48] The tendency of Helix F, loop FG and the C-terminus of Helix B to the N-terminus of Helix E to move into the empty heme pocket was reproduced through this study. Positive correlations in motion displayed in the region labeled 4 in Figure 6.8 A for all apoMb variants indicated the movement of the residues in the aforementioned structural motifs along a similar direction. This observation implied the probable development of contacts among Helix F, loop FG and the C-terminus of Helix B to the N-terminus of Helix E upon collapsing into the vacant heme site.

The examination of the correlation map in Figure 6.8 A showed that the apoMb with E109G mutation exhibited the least correlated dynamics in region 4 while similar dynamical correlation was observed for the rest of the apoMb variants. This lack of correlation in the motion of Helix F, loop FG and the C-terminus of Helix B to the N-terminus of Helix E in E109G may be due to the loss of interactions among these structural motifs which subsequently increase the solvent accessibility of the heme pocket. This event may expedite the unfolding of E109G compared to its other counterparts as the solvation of the heme pocket may simultaneously cause the hydration of the hydrophobic center which is adjacent to the heme binding site.

Principal Component Analysis and RMSF

In the presence of factors promoting the denaturation of apoMb, several events related to the unfolding process could be observed through short MD simulations. Some of these events include the loosening of the compact hydrophobic core, which had been explored earlier in this chapter, and the increasing fluctuations of loops and helices of apoMb, in particular loop EF and helix F, which will be explored by performing principal component analysis (PCA) and calculating the root-mean-square fluctuations (RMSF) of apoMb residues. [14, 49] PCA is a well-established technique that simplifies the study of protein dynamics by limiting the 3*N* (*N* being the number of atoms in the protein) degrees of freedom of the protein to essential degrees of freedom that describe the functionally critical motions of the protein. [50, 51] PCA was performed on one of the two trajectories acquired for the apoMb variants and key motion modes representing the unfolding of apoMb were identified and visualized using Interactive Essential Dynamics (IED) program with Visual Molecular Dynamics (VMD) as a display interface to observe and compare motion modes associated to the fluctuations of loop EF and Helix F. [52, 53]

ApoMb proteins with destabilizing mutations are expected to experience greater fluctuations in loop EF and Helix F compared to native apoMb due to the more rapid unfolding of the mutated protein under denaturing conditions. E109G apoMb, which possess a destabilizing mutation, showed the most fluctuations at loop EF and helix F in Figure 6.9 A and this observation was further corroborated by the higher RMSF values discerned for residues in loop EF and helix F of E109G compared to the other apoMb proteins in Figure 6.9 B. The larger fluctuations of loop EF and Helix F of E109G monitored through PCA and RMSF plots in Figure 6.9 supports the weak correlation of the dynamics of Helix F, loop FG and the C-terminus of Helix B to the N-terminus of Helix E portrayed in the correlation map of E109G in Figure 6.8 A. The augmented fluctuations of loop EF and Helix F of E109G compared to wild type apoMb may have caused helix F and loop FG to move in the opposite direction resulting in the loss of interactions among the domains represented by region 4 in the correlation map plotted in Figure 6.8 A for E109G. Moreover, the outward shift of loop EF and Helix F in E109G away from the empty heme pocket supported the notion that the loss of interactions among Helix F, loop FG and the

C-terminus of Helix B to the N-terminus of Helix E, exposed the heme pocket to the surrounding solution. This occurrence concurrently promoted the hydration of the hydrophobic core which subsequently led to the dramatic increase in the SASA of E109G showed in Figure 6.6.

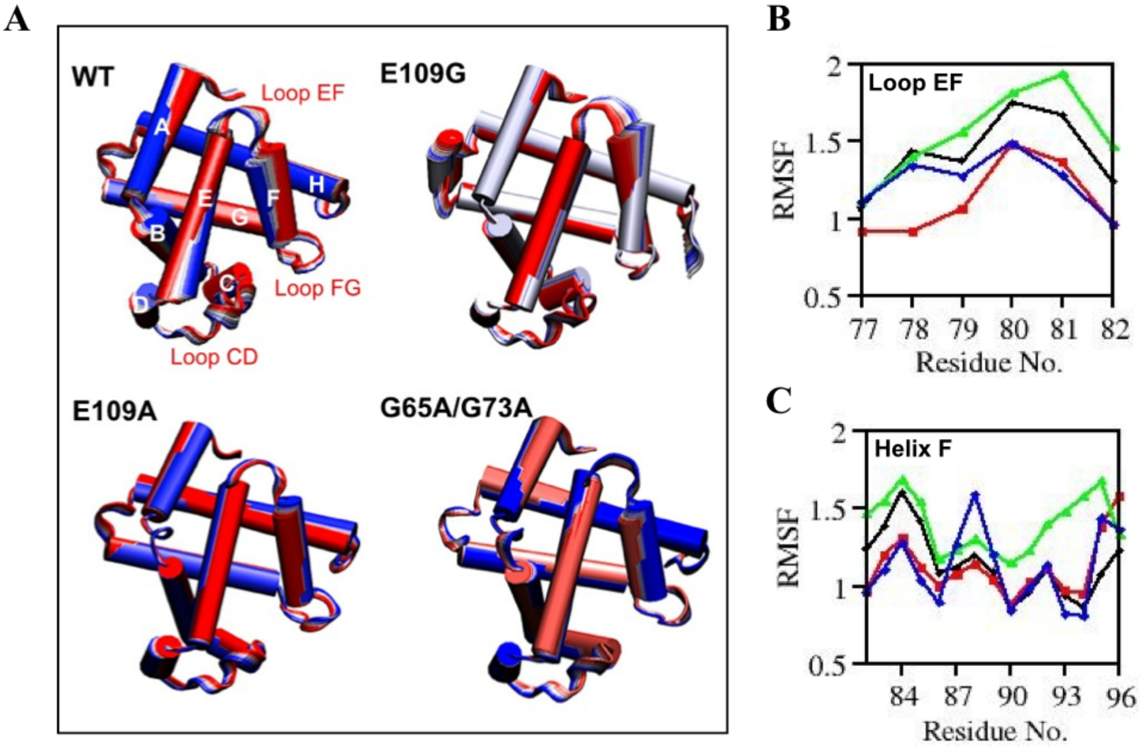


Figure 6.9: (A) Cartoon representations of fluctuations observed for WT, E109A, E109G and G65A/G73A through PCA. The color transition from first to last frame goes from red to white to blue. The RMSF of the CA atoms of (B) Loop EF and (C) Helix F.

Hydrophobic Core and Hydrogen Bonding

A tightly packed hydrophobic core is one of the most pivotal physical characteristics of globular proteins that govern the stability, native structure and properties of proteins in general. [54] This fact substantiates the analyses that have been conducted hitherto, which attributed the stability of apoMb to the conservation of the hydrophobic core, and this is best illustrated by E109A, the apoMb with stabilizing mutation. Globular proteins fold into its native state in a manner which ensures that the non-polar residues are buried in the protein interior while the polar residues will line the protein surface preventing water from interacting with the non-polar residues. In this part of the chapter, we explored the role of hydrogen bonds on the protein surface in protecting the hydrophobic center of apoMb from hydration. Hydrogen bonds that are established on the surface of E109G and may potentially serve as gates for solvent penetration when disrupted during unfolding were identified using ptraj module in AmberTools. [21] Hydrogen bond analysis was only conducted for E109G as this mutant showed the most increase in SASA during the simulation compared to the other apoMb variants denoting the greater accessibility of the hydrophobic core of E109G to the external environment. (Figure 6.6) This will provide us with a clearer correlation between the breaking of hydrogen bonds and the swelling of the hydrophobic core in apoMb.

From the hydrogen bond analysis conducted for E109G, six hydrogen bonds are identified and exhibited in Figure 6.10 A and B. The six hydrogen bond pairs distinguished include:

(i) Glu6 (Helix A) and Lys133 (Helix H)

(ii) Asp27 (Helix B) and Arg118 (Helix G)

(iii) His12 (Helix A) and Asp122 (loop GH)

(iv) His116 (Helix G) and Gln128 (Helix H)

(v) Glu4 (Helix A) and Lys79 (loop EF)

(vi) His82 (loop EF) and Asp141 (Helix H)

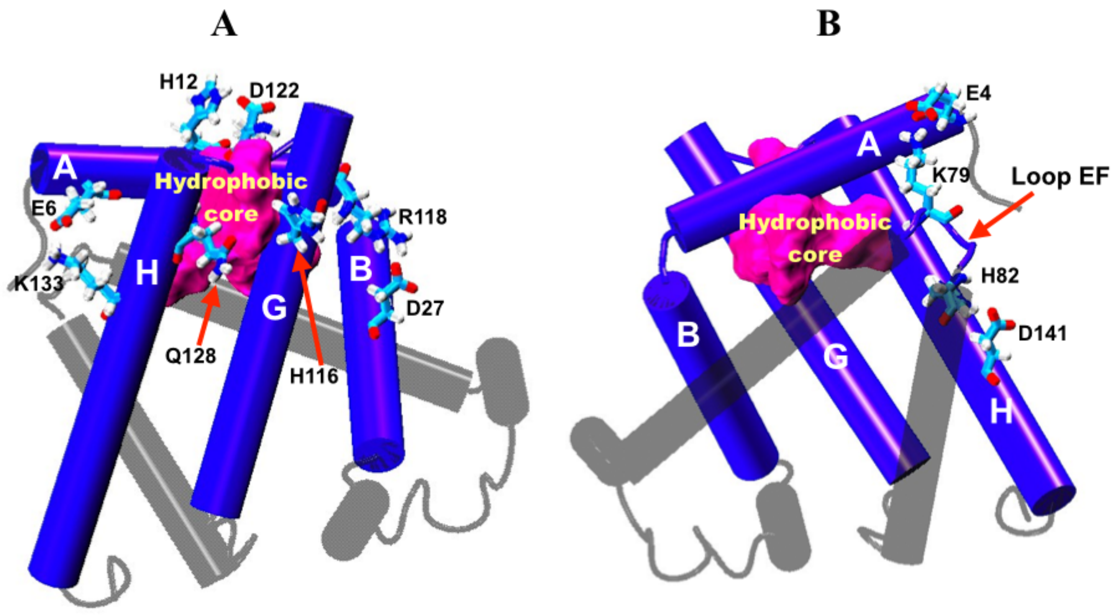


Figure 6.10: (A) Cartoon representation of apoMb with four hydrogen bonds formed between Glu6 and Lys133, between His12 and Asp122, between His116 and Gln128 and between Asp27 and Arg118 displayed using licorice representation. (B) Cartoon representation of apoMb with two hydrogen bonds formed between Glu4 and Lys79 and between His82 and Asp141 displayed using licorice representation. The hydrophobic core of apoMb is represented using surface representation colored magenta.

As portrayed in Figure 6.10, the six hydrogen bonds are well-distributed around the circumference of the AGH-hydrophobic core. To study the effect of hydrogen bond disruption on the stability of the hydrophobic core, changes in the distance between the hydrogen bond pairs listed above were calculated and showed in Figure 6.11. With the exclusion of the newly formed hydrogen bond between Asp27 and Arg118 observed in Fig 6.11, the rest of the hydrogen bond pairs identified above showed notable interruption in hydrogen bonding as evidenced by the increase in the hydrogen bond length with time. Based on the plot presented in Figure 6.11, hydrogen bonds formed between Glu4 and Lys79, between His12 and Asp122, and between His82 and Asp141, possibly play a vital role in guarding the hydrophobic core from the initial entry of solvent molecules as disruption of these hydrogen bonds were noted after 2 ns into simulation. This is consistent with the rise in the SASA of E109G around similar simulation time as showed in Figure 6.6. After 2 ns, the rise in the SASA of E109G persisted and this may stem from the breaking of two more hydrogen bonds between Glu6 and Lys133 and between His116 and Gln128.

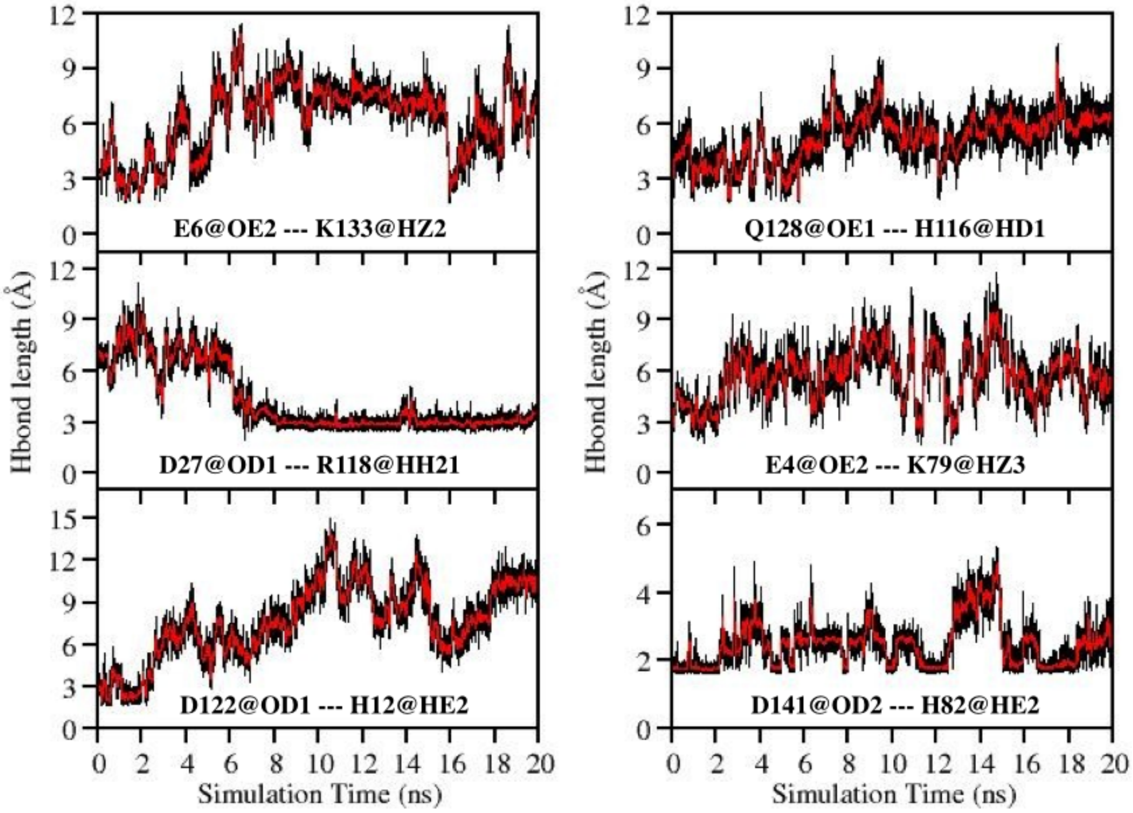


Figure 6.11: Plot of distance between six hydrogen bond pairs namely Glu6 and Lys133, Asp27 and Arg118, His12 and Asp122, Gln128 and His116, Glu4 and Lys79, and Asp141 and His82 against time.

The breaking of the hydrogen bond between His116 and Gln128 occurs 5 ns into the simulation and this event consequently resulted in the development of a new hydrogen bond between Asp27 and Arg118 at around the 6 ns mark, causing helices G and H to separate. This event sequentially led towards the solvation of the hydrophobic core which was corroborated by the continuous increase in SASA of E109G during the simulation.

(Figure 6.6) Through a computational study that involved the force-induced unfolding of apoMb, Choi et al. highlighted the loss of interactions between helices G and H as one of the main driving force prompting the unfolding of apoMb by promoting the disruption of the hydrophobic core. [13] Combining the distance plots in Figure 6.11 and the observations acquired from the correlation maps in Figure 6.8 A, the breaking of the hydrogen bonds between His82 and Asp141 and between His116 and Gln128 may contribute to the almost non-existent correlation in motion between loop EF and helix A/H of E109G (Region 2 and 3 in Figure 6.8 A). Additionally, the disruption of these two hydrogen bonds may also contribute to the increase in the fluctuations of loop EF and helix F that were depicted in Figure 6.9 (PCA and RMSF plots), which subsequently cause helix F, loop FG and the C-terminus of helix B to the N-terminus of helix E to be poorly correlated in motion compared to the other protein variants. (Figure 6.8 A, Region 4 and Figure 6.8 D) These sequence of occurrences resulted in the solvation of the protein interior which in turn loosened the compact hydrophobic core, validating the small correlation among residues in the AGH-core observed in the correlation map of E109G in Figure 6.8 A (Region 1).

The breaking of the five hydrogen bonds (Figure 6.11) may also be induced by interactions established between the urea-water solution and amino acids that were initially involved in the formation of the five hydrogen bonds namely Glu4, Glu6, His12, Lys79, His82, His116, Asp122, Gln128, Lys133 and Asp141. Hydrogen bond analyses conducted for the last 5 ns of the simulations performed for E109G showed the formation of hydrogen bonds between the urea-water solution and charged amino acids namely Glu4, Glu6, Asp122, Gln128 and Asp141. Since the hydrogen bonds formed between the urea-water solution and the charged amino acids were mostly transient, hydrogen bond distances were not calculated. Instead, Figure 6.12 below showcased the clustering of water and urea molecules around the hydrogen bond pairs which subsequently caused the hydrogen bonds to be disrupted hence providing an access for the urea-water solution to enter the hydrophobic core.

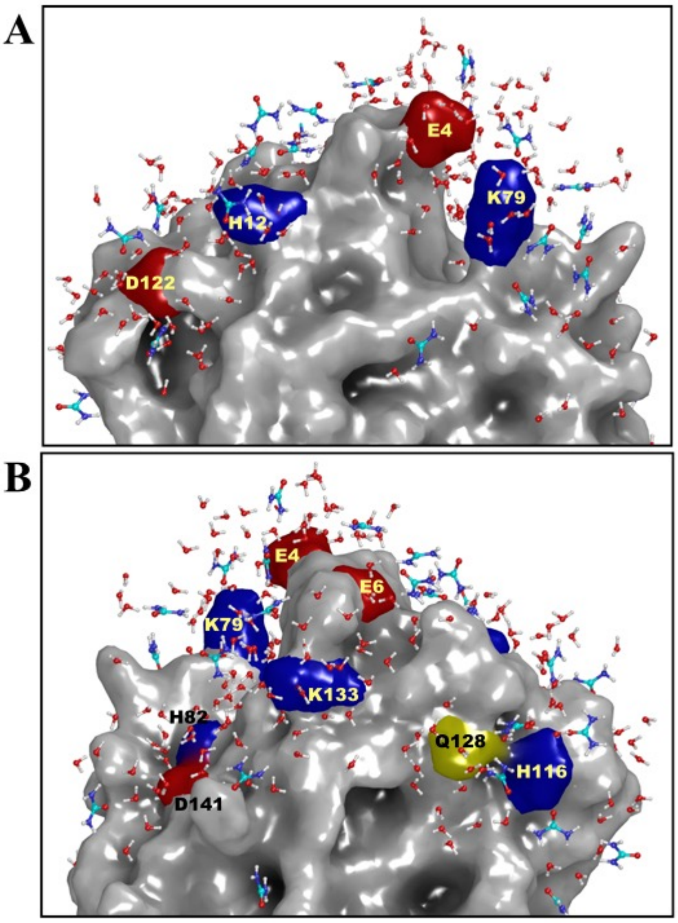


Figure 6.12: (A) Schematic representation of urea-water clusters in the vicinity of hydrogen bonds formed between (A) Glu4 and Lys79 and His12 and Asp122 (B) Glu4 and Lys79, Glu6 and Lys133, His82 and Asp141 and His116 and Gln128. Ball-and-stick representation was used for urea (carbon atoms in cyan) and water molecules. This figure was generated using Schrodinger Suite 2011. [55]

6.4 Conclusion

The analyses conducted thus far had justified the capacity of MD simulation in determining the stability of apoMb upon mutation, evident from the consistency in results attained theoretically here with the experimental observations documented in Ref 11. With the combination of suitable analytical tools, factors influencing the stability of the protein could be harnessed from the trajectories acquired hence eliminating the likelihood of committing erroneous mutation that may alter the tertiary structure and function of the protein upon mutagenesis. Despite the small benchmark study conducted here, the feasibility of applying MD simulation in predicting the stability of larger proteins upon mutation is possible with the ever-growing computational resources and the accessibility of state-of-art theoretical tools to researchers. Through the implementation of MD simulation as a test to predict the stability of proteins upon mutagenesis, the workload and expenditure of experimental researchers are lightened and the efficiency of research work can be potentially improved.

Chapter 7 Summary

MD simulation is one of the most frequently used computational tools in the theoretical study of biological macromolecules especially proteins. In the writing of this thesis, the use of MD simulation in the investigation of the fluctuation, dynamics and thermodynamics of proteins has been explored and this includes studies related to structural variation (Chapter 3), protein folding (Chapter 4), free energy study of metalloprotein (Chapter 5) and determination of protein stability (Chapter 6). Additionally, the importance of incorporating the effects of the electrostatic environment of the protein into the force field used was also evaluated through the study reported in Chapter 4 and 5 which showed results that have better consistency with experimental data upon explicit consideration of environment effects.

Protein misfolding is one of the leading causes of protein-related diseases such as diabetes, Alzheimer's disease and Parkinson's disease which prevalence is expected to increase due to the aging population. For this reason, numerous studies related to protein misfolding have been conducted to better understand the mechanism of these protein-related diseases. In Chapter 3 of this thesis, the mechanism causing the variation in the secondary structure of an engineered protein from β-sheet to α-helix was scrutinized using REMD simulation. The transition from β-sheet to α-helix was revealed to be prompted by the unpacking of the hydrophobic core of α/4β-GA88 which induces the unfolding of the β-sheet to form α-helix. This study underlines the capacity of MD simulation to model conformational changes which is especially useful in studies concerning protein misfolding which is crucial for the understanding of folding diseases. The MD simulation conducted in Chapter 3 also portrayed the underlying bias in the force field used which may likely be due to the inability of the conventional force field to provide effective descriptions for the changes in the environment of the protein atoms during folding and unfolding.

Conventional force fields popularly used in MD studies often lack the polarization energy term. While the use of basic force field is sufficient for a significant number of MD studies, most of the time, the lack of polarization effect may lead to the extraction of flawed structural and dynamical details from the MD simulation performed. In the case of the study pertaining to the determination of protein stability discussed in Chapter 6, the use of basic force field gave rise reasonable results which agreed well with experimental data. However, the lack of electrostatic environment effect in the *ab initio* folding of 2khk in Chapter 4 and the determination of the reduction potential of rubredoxin in Chapter 5 led to results which divert from experimental observations. For example, the use of AMBER ff03 force field in the folding of 2khk led to a structure which deviates from the NMR structure of 2khk. However, when polarization effects of hydrogen bonds were considered through the AHBC scheme, the folding of 2khk proceeded with ease to assume a structure close to experiment with best backbone RMSD of 1.3 Å. This study showcases the importance of polarization effect in the *ab initio* folding of proteins as the on-the-fly charge update offered by AHBC scheme will ensure that the changes in the environment of the protein's atoms induced by changes in conformation are considered during the simulation. Similarly, the importance of protein environment was emphasized through the study discussed in Chapter 5 whereby the gradual increase in the consideration of the electrostatic environment around the iron atom led to a better prediction of the reduction potential of the three rubredoxin mutants namely L41A, V44A and V44G.

Through this thesis, the predictive power of MD simulations to anticipate the outcome of experimental studies was also showcased in Chapter 4, 5 and 6 whereby the results achieved theoretically are comparable to that of experiment. The ease of the usage of MD simulation also pushes this computational tool as an essential gadget in experimental laboratories since useful informations such as intra- and intermolecular interactions, thermodynamics, kinetics, dynamics and binding affinities can be obtained at a lower cost within a shorter period of time compared to traditional experiments. Even though theoretical studies to date still require empirical information, the continuous advancement in *ab initio* calculations may propel MD simulation to be a powerful tool that can predict the dynamics and properties of proteins and other non-protein materials (organic and inorganic compounds, metal alloys and nano-materials) with the use of minimal experimental data in the future.

Appendix

Simulation details for the equilibration of wild type apomyoglobin in water

Prior to the mutation of apoMb to obtain mutants of interest namely E109A, E109G and G65A/G73A, wild type apoMb was obtained by removing the heme moiety from the sperm whale myoglobin structure (PDB id: 1BZP) and relaxed in TIP3P water for 500 ps. [1, 2] Before solvating the protein in water box, hydrogen atoms were added and histidine residues at position 36, 81 and 116 were doubly protonated to resemble apoMb at neutral pH. [3] The wild type apoMb was solvated in an octahedral TIP3P water box with a minimum distance of 10 Å set between the protein and edge of the water box. [2] Chloride ions were added to neutralize the system. These steps were simplified with the aid of the LEaP module in AmberTools 1.2. [4] AMBER ff99SB force field was employed to conduct this simulation. [5, 6] Particle mesh Ewald (PME) method was used to calculate long-range electrostatic interaction and SHAKE algorithm was implemented to constrain all covalent bonds involving hydrogen atoms. [7, 8]

Before proceeding with the equilibration step, the solvent molecules were minimized for 5000 steps with steepest descent minimization conducted for the first 1000 steps followed by minimization using conjugate gradient method thereafter. This was followed by the minimization of the entire system for 10000 steps using steepest descent method and another 10000 steps of minimization using conjugate gradient method. After minimization, the entire system was heated from 10 K to 277.15 K for 100 ps using Langevin thermostat in canonical ensemble with a collision frequency of 4 ps^{-1} and the protein was weakly restrained using a harmonic constant of 5 kcal/mol/Å^2 during heating. [9, 10] Subsequently, the protein was equilibrated in the water box for 500 ps at 277.15 K using a time step of 2 fs.

References

(References are arranged according to individual chapters)

Chapter 1

1. Levitt M, Warshel A. Computer simulation of protein folding. Nature 1975;253:694-698.
2. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature 1977;267:585-590.
3. Karplus M. Molecular dynamics simulations of proteins. Phys Today 1987;40:68–72.
4. Warshel A, Levitt M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J Mol Biol 1976;103:227-249.
5. Warshel A. Computer Modeling of Chemical Reactions in Enzymes and Solutions. Wiley, New York, 1992.
6. Field MJ, Bash PA, Karplus M. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. J Comput Chem 1990;11:700-733.
7. Durrant JD, McCammon AJ. Molecular dynamics simulation and drug discovery. BMC Biol 2011;9:71-79.
8. Fatmi MQ, Chang CA. The role of oligomerization and cooperative regulation in protein function: The case of tryptophan synthase. PLoS Comput Biol 2010;6:e1000994.
9. Amaro R, Tajkhorshid E, Luthey-Schulten Z. Developing an energy landscape for the novel function of a $(\beta/\alpha)_8$ barrel: Ammonia conduction through HisF. Proc Natl Acad Sci 2003;100:7599-7604.

10. Amaro R, Luthey-Schulten Z. Molecular dynamics simulations of substrate channeling through an α-β barrel protein. Chem Phys 2004;307:147-155.
11. Sattelle BM, Sutcliffe MJ. Calculating chemically accurate redox potentials for engineered flavoproteins from classical molecular dynamics free energy simulations. J Phys Chem A 2008;112:13053– 13057.
12. "The Nobel Prize in Chemistry 2013 - Press Release". *Nobelprize.org.* Nobel Media AB 2014. Web. 23 Jul 2014. http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/press.html
13. Nussinov R. The significance of the 2013 Nobel Prize in Chemistry and the challenges ahead. PLoS Comput Biol 2014;10:e1003423.
14. Tuckerman ME, Martyna GJ. Understanding modern molecular dynamics: Techniques and applications. J Phys Chem B 2000;104:159-178.
15. Lane TJ, Shukla D, Beauchamp KA, Pande VS. To milliseconds and beyond: challenges in the simulation of protein folding. Curr Op Struc Biol 2013;23:58-65.
16. Lindahl, ER, 2008, Molecular Dynamics Simulation, In: Kukol, A., ed., Molecular Modeling of Proteins, Humana Press, Totowa, NJ, p. 3-23.
17. Petrenko, Roman, and Meller, Jarosław(Mar 2010) Molecular Dynamics. In: eLS. John Wiley & Sons Ltd, Chichester. http://www.els.net [doi:10.1002/9780470015902.a0003048.pub2]
18. Bahar I, Lezon TR, Yang LW, Eyal E. Global dynamics of proteins: Bridging between structure and function. Annu Rev Biophys 2010;39:23-42.
19. Sapay N, Nurisso A, Imberty A. Simulation of carbohydrates, from molecular docking to dynamics in water. Methods Mol Biol 2013;924:469-483.
20. Cheatham III TE, Kollman PA. Molecular dynamics simulation of nucleic acids. Annu Rev Phys Chem 2000;51:435-471.

21. Ashurst WT, Hoover WG. Dense-fluid shear viscosity via nonequilibrium molecular dynamics simulation. Phys Rev A 1975;11:658-678.

22. Haughney M, Ferrario M, McDonald IR. Molecular-dynamics simulation of liquid methanol. J Chem Phys 1987;91:4934-4940.

23. Singh UC, Kollman PA. A combined *ab initio* quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the $CH_3Cl + Cl^-$ exchange reaction and gas phase protonation of polyethers. J Comput Chem 1986;7:718-730.

24. Tal AA, Müger EP, Abrikosov IA, Pilch I, Helmersson U. Molecular dynamics simulation of the growth of Cu nanoclusters from Cu ions in a plasma. Phys Rev B 2014;90:165421.

25. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. J Am Chem Soc 1995;117:5179-5197.

26. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J Comp Chem 2004;25:1157-1174.

27. Wei G, Mousseau N, Derreumaux P. Computational Simulations of the Early Steps of Protein Aggregation. Prion 2007;1:3-8.

28. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M. Simultaneous determination of protein structure and dynamics. Nature 2005;433:128-132.

29. Karplus M, Kuriyan J. Molecular dynamics and protein function. Proc Natl Acad Sci USA 2005;102:6679–6685.

30. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman PA. A point-charge force field for molecular

mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 2003;24:1999–2012.

31. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. J Comput Chem 2005;26:1668–1688.
32. MacKerell AD, Jr, Feig M, Brooks CL, III. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem 2004;25:1400–1415.
33. Scott WRP, Hünenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Krüger P, vanGunsteren WF. The GROMOS biomolecular simulation program package. J Phys Chem A 1999;103:3596–3607.
34. Lei H, Wu C, Liu H, Duan Y. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. Proc Natl Acad Sci 2007;104:4925–4930.
35. Kim E, Jang S, Pak Y. All-atom *ab initio* native structure prediction of a mixed fold (1FME): A comparison of structural and folding characteristics of various *ββα* miniproteins. J Chem Phys 2009;131:195102.
36. Lei HX, Wang ZX, Wu C, Duan Y. Dual folding pathways of an *α/β* protein from all-atom *ab initio* folding simulations. J Chem Phys 2009;131:165105.
37. Case DA, Darden TA, Cheatham TE, III, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossvary I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG,

Sagui C, Babin V, Kollman PA. Amber 10. San Francisco: University of California; 2008.

38. Lennard-Jones, J. E. On the Determination of Molecular Fields. Proc. R. Soc. Lond. A 1924;106:463-477.
39. Jogalekar, A. 2012, "Synthesis, enzymes and force fields: defining chemical elegance.” Scientific American Blog. http://blogs.scientificamerican.com/the-curious-wavefunction/2012/08/28/synthesis-enzymes-and-force-fields-defining-chemical-elegance/, Feb 2015.
40. Nowak, W, 2014, Applications of Computational Methods to Simulations of Protein Dynamics, In: Jerzy, L, ed., Handbook of Computational Chemistry, Springer Netherlands, p. 1127-1153.
41. Piana S, Klepeis JL, Shaw DE. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. Curr Op Struc Biol 2014;24:98-105.
42. Adcock SA, McCammon JA. Molecular dynamics: Survey of methods for simulating the activity of proteins. Chem Rev 2008;106:1589-1615.
43. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics for protein folding. Chem Phys Lett 1999;314:141-151.
44. Lei H, Wu C, Liu H, Duan Y. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. Proc Natl Acad Sci USA 2007;104:4925-4930.
45. Shen MY, Freed KF. All-atom fast protein folding simulations: the villin headpiece. Proteins 2002;49:439-445.

46. Zagrovic B, Snow CD, Shirts MR, Pande VS. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. J Mol Biol. 2002;323:927–937.
47. Jang S, Kim E, Shin S, Pak Y. Ab initio folding of helix bundle proteins using molecular dynamics simulations. J Am Chem Soc. 2003;125:14841–14846.
48. Ripoll DR, Vila JA, Scheraga HA. Folding of the villin headpiece subdomain from random structures. Analysis of the charge distribution as a function of pH. J Mol Biol. 2004;339:915–925.
49. Park S, Khalili-Araghi F, Tajkhorshid E, Schulten K. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. J Chem Phys 2003;119:3559-3566.
50. Park S, Schulten K. Calculating potentials of mean force from steered molecular dynamics simulations. J Chem Phys 2004;120:5946-5961.
51. Patel JS, Berteotti A, Ronsisvalle S, Rocchia W, Cavalli A. Steered molecular dynamics simulations for studying protein-ligand interaction in cyclin-dependent kinase 5. J Chem Inf Model 2014;54:470-480.
52. Saunders MG, Voth GA. Coarse-graining methods for computational biology. Annu Rev Biophys 2013;42:73-93.
53. Scott KA, Bond PJ, Ivetac A, Chetwynd AP, Khalid S, Sansom MSP. Coarse-grained MD simulations of membrane protein-bilayer self-assembly. Structure 2008;16:621-630.
54. Derreumaux, P, 2013, Coarse-Grained Models for Protein Folding and Aggregation, In: Monticelli, L., Salonen, E., ed., Biomolecular Simulations, Humana Press, Springer Science+Business Media New York, p. 585-600.

55. Wu C, Shea JE. Coarse-grained models for protein aggregation. Curr Op Struc Biol 2011;21:209-220.

Chapter 2

1. Durrant JD, McCammon AJ. Molecular dynamics simulation and drug discovery. BMC Biol 2011;9:71-79.
2. Lane TJ, Shukla D, Beauchamp KA, Pande VS. To milliseconds and beyond: challenges in the simulation of protein folding. Curr Op Struc Biol 2013;23:58-65.
3. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W.: Atomic-level characterization of the structural dynamics of proteins. *Science* 2010, 330:341-346.
4. Pierce LCT, Salomon-Ferrer R, de Oliveira CAF, McCammon JA, Walker RC. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. J Chem Theory Comput 2012; 8: 2997–3002.
5. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics for protein folding. Chem Phys Lett 1999;314:141-151.
6. Hansmann UHE, Okamoto JJ. Numerical comparisons of three recently proposed algorithms in the protein folding problem Comp Chem 1997;18:920-933.
7. Hukushima K, Nemoto K. Exchange Monte Carlo Method and Application to Spin Glass Simulations J Phys Soc Jpn 1996;65:1604-1608.
8. Swendsen RH, Wang JS. Replica Monte Carlo Simulation of Spin-Glasses. Phys Rev Lett 1986;57:2607-2609.
9. Geyer CJ, in: E.M. Keramidas (Ed.), Computing Science and Statistics: Proc. 23rd Symp. on the Interface, Interface Foundation, Fairfax Station, 1991, p. 156.

10. Tesi MC, van Rensburg EJJ, Orlandini E, Whittington SG. Monte carlo study of the interacting self-avoiding walk model in three dimensions. J Stat Phys 1996;82:155-181.
11. Marinari E, Parisi G, Ruiz-Lorenzo JJ, in: A.P. Young (Ed.), Spin Glasses and Random Fields, World Scientific, Singapore, 1998, p.59.
12. Lei H, Wu C, Liu H, Duan Y. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. Proc Natl Acad Sci USA 2007;104:4925-4930.
13. Lei H, Wang ZX, Wu C, Duan Y. Dual folding pathways of an α/β protein from all-atom *ab initio* folding simulations. J Chem Phys 2009;131:165105.
14. Kim E, Jang S, Pak Y. All-atom *ab initio* folding of a mixed fold (1FME): A comparison of structural and folding characterization of various ββα miniproteins. J Chem Phys 2009;131:195102.
15. Beck DA, White GW, Daggett V. Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations. J Struc Biol 2007;157:514–523.
16. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P. A point- charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 2003;24:1999–2012.
17. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. J Comput Chem 2005;26:1668–1688.
18. MacKerell Jr AD, Feig M, Brooks CL, III. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in repro-

ducing protein conformational distributions in molecular dynamics simulations. J Comput Chem 2004;25:1400–1415.

19. Scott WRP, Hünenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Krüger P, vanGunsteren WF. The GROMOS biomolecular simulation program package. J Phys Chem A 1999;103:3596–3607.
20. Duan LL, Mei Y, Zhang QG, Zhang JZH. Intra-protein hydrogen bonding is dynamically stabilized by electronic polarization. J Chem Phys 2009;130:115102.
21. Xu Z, Lazim R, Mei Y, Zhang D. Stability of the β-structure in prion protein: A molecular dynamics study based on polarized force field. Chem Phys Lett 2012;539-540:239-244.
22. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577-2637.
23. Halgren TA, Damm W. Polarizable force fields. Curr Opin Struc Biol 2001;11:236–242.
24. Wang W, Skeel RD. Fast evaluation of polarizable forces. J Chem Phys 2005;123:164107.
25. Ji CG, Ye M, Zhang JZH. Developing polarized protein-specific charges for protein dynamics: MD free energy calculation of pKa shifts for Asp26/Asp20 in thioredoxin. Biophys J 2008;95:1080– 1088.
26. Warshel A, Kato M, Pisliakov AV. Polarizable force fields: History, test cases, and prospects. J Chem Theory Comput 2007;3:2034-2045.
27. Wang J, Cieplak P, Li J, Wang J, Cai Q, Hsieh MJ, Lei H, Luo R, Duan Y. Development of polarizable modes for molecular mchanical calculations II: Induced dipole models significantly improve accuracy of intermolecular interaction energies. J Chem Phys B 2011;115:3100-3111.

28. Patel S, MacKerell Jr AD, Brooks III CL. CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. J Comput Chem 2004;25:1504-1514.
29. Schmollngruber M, Lesch V, Schröder C, Heuer A, Steinhauser O. Comparing induced point-dipoles and Drude oscillators. Phys Chem Chem Phys 2015, Advance Article, doi: 10.1039/C4CP04512B.
30. Lopes PEM, Huang J, Shim J, Luo Y, Li H, Roux B, MacKerell Jr, AD. Polarizable force field for peptides and proteins based on the classical Drude oscillator. J Chem Theory Comput 2013;9:5430-5449.
31. Duan LL, Mei Y, Zhang D, Zhang QG, Zhang JZH. Folding of a helix at room temperature is critically aided by electrostatic polarization of intraprotein hydrogen bonds. J Am Chem Soc 2010;132: 11159–11164.
32. Zhang DW, Zhang JZH. Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein–molecule interaction energy. J Chem Phys 2003;119:3599-3605.
33. Gao AM, Zhang DW, Zhang JZH, Zhang Y. An efficient linear scaling method for ab initio calculation of electron density of proteins. Chem Phys Lett 2004;394:293-297.
34. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. J Comput Chem 2002;23:128–137.
35. Gordon MS, Fedorov DG, Pruitt SR, Slipchenko LV. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. Chem Rev 2012;112:632-672.
36. Ji CG, Zhang JZH. Protein polarization is critical to stabilizing AF-2 and helix-2’ domains in ligand binding to PPAR-gamma. J Am Chem Soc 2008;130:17129–17133.

37. Tong Y, Ji CG, Mei Y, Zhang JZH. Simulation of NMR data reveals that proteins' local structures are stabilized by electronic polarization. J Am Chem Soc 2009;131:8636–8641.
38. Wei C, Lazim R, Zhang D. Importance of polarization effect in the study of metalloproteins: Application of polarized protein specific charge scheme in predicting the reduction potential of azurin. Prot Struc Func Bioinform 2014;82:2209-2219.
39. Wei C, Mei Y, Zhang D. Theoretical study on the HIV-1 integrase–5CITEP complex based on polarized force fields. Chem Phys Lett 2010;495:121-124.
40. Mei Y, Yong LL, Zeng J, Zhang JZH. Electrostatic polarization is critical for the strong binding in streptavidin-biotin system. J Comput Chem 2012;33:1374-1382.
41. Xu Z, Mei Y, Duan L, Zhang D. Hydrogen bonds rebuilt by polarized protein-specific charges. Chem Phys Lett 2010;495:151–154.
42. Mei Y, Wu EL, Han KL, Zhang JZH. Treating hydrogen bonding in ab initio calculations in biopolymers. J Quant Chem 2006;106:1267-1276.
43. Dill KA. Dominant forces in protein folding. Biochem 1990;29: 7133–7155.
44. Myers JK, Pace CN. Hydrogen bonding stabilizes globular proteins. Biophys J 1996;71:2033–2039.
45. Wei C, Tung D, Yip YM, Mei Y, Zhang DW. Communication: The electrostatic polarization is essential to differentiate the helical propensity in polyalanine mutants. J Chem Phys 2011;134:171101.
46. Xu Z, Lazim R, Sun T, Mei Y, Zhang DW. Solvent effect on the folding dynamics and structure of E6-associated protein characterized from ab initio protein folding simulations. J Chem Phys 2012; 136:135102.

Chapter 3

1. Dill KA, MacCallum JL. The protein folding problem, 50 years on. Sci 2012;338:1042-1046.
2. Onuchic JN, Luthey-Schulten Z, Wolynes PG (1997) Theory of protein folding: The energy landscape perspective. Annu Rev Phys Chem 48:545-600
3. Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2009) A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci 106:21149-21154
4. Maaß A, Tekin ED, Schüller A, Palazoglu A, Reith D, Faller R (2010) Folding and unfolding characteristics of short beta strand peptides under different environmental conditions and starting configurations. Biochim Biophys Acta 1804:2003-2015
5. He Y, Chen Y, Alexander P, Bryan PN, Orban J (2008) NMR structures of two designed proteins with high sequence identity but different fold and function. Proc Natl Acad Sci 105:14412-14417
6. Lodish H., Berk A., Zipursky S. L., Matsudaira P., Baltimore D., Darnell J. Molecular Cell Biology. 4th edition. New York: W. H. Freeman; 2000. Section 3.1, Hierarchical Structure of Proteins. Available from:http://www.ncbi.nlm.nih.gov/books/NBK21581/
7. Ringe D, Petsko GA. Mapping protein dynamics by X-ray diffraction. Prog Biophys Mol Biol 1985;45:197–235.
8. Saven, JG. Computational protein design: engineering molecular diversity, nonnatural enzymes, nonbiological cofactor complexes, and membrane proteins. Curr Op Chem Biol 2011;15:452-457.
9. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223-230.
10. Rose GD, Creamer TP. Protein folding: predicting predicting. Proteins 1994;19:1–3.

11. Martin ACR, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JBO, Taroni C, Thornton JM. Protein folds and functions. Structure 1998;6:875-884.
12. Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Analyzing Protein Structure and Function. Available from: http://www.ncbi.nlm.nih.gov/books/NBK26820/
13. Hartl FU, Hayer-Hartl M. Converging concepts of protein folding *in vitro* and *in vivo*. Nature Struc Mol Biol 2009;16:574-581.
14. Chiti F, Dobson CM. Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem 2006;75:333–366.
15. Selkoe DJ. Cell biology of protein misfolding: The examples of Alzheimer's and Parkinson's diseases. Nat Cell Biol 2004;6:1054–1061.
16. Dalal S, Balasubramanian S, Regan L. Protein alchemy: Changing β-sheet into α- helix. Nat Struc Biol 1997;4:548-552.
17. Yang W-Z, Ko T-P, Corselli L, Johnson RC, Yuan HS. Conversion of a β-strand to an α-helix induced by a single-site mutation observed in the crystal structure of Fis mutant Pro^{26}Ala. Prot Sci 1998;7:1875-1883.
18. Dobson CM. Protein folding and misfolding. Nature 2003;426:884-890.
19. Zhou R. Trp-cage: Folding free energy landscape in explicit water. Proc Natl Acad Sci 2003;100:13280-13285
20. Prusiner SB. Prions. Proc Natl Acad Sci 1998;95:13363-13383
21. DeToma AS, Salamekh S, Ramamoorthy A, Lim MH. Misfolded Proteins in Alzheimer's Disease and Type II Diabetes. Chem Soc Rev 2012;41:608-621.
22. Case DA, Darden TA, Cheatham TE, III, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra

G, Kolossvary I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, Sagui C, Babin V, Kollman PA. Amber 10. San Francisco: University of California; 2008.

23. www.pdb.org Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H,Shindyalov IN, Bourne PE. The protein data bank. Nucl Acids Res 2000;28:235-242.
24. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations J Comput Chem 2003;24:1999-2012.
25. Tsui V, Case DA. Theory and applications of the generalized Born solvation model in macromolecular simulations. Biopoly (Nucleic Acid Sci) 2001;56:275-291.
26. Levy RM, Zhang LY, Gallicchio E, Felts AK. On the nonpolar hydration free energy of proteins: Surface area and continuum solvent models for the solute-solvent interaction energy. J Am Chem Soc 2003;125:9523-9530.
27. Lwin TZ, Zhou R, Luo R. Is Poisson-Boltzmann theory insufficient for proteinfolding simulations? J Chem Phys 2006;124:34902-34907.
28. Uberuaga BP, Anghel M, Voter AF. Synchronization of trajectories in canonical molecular-dynamics simulations: Observation, explanation, and exploitation. J Chem Phys 2004;120:6363-6374.
29. Ryckaert J-P, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. J Comput Phys 1977;23:327-341.

30. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. The design and characterization of two proteins with 88% sequence identity but different structure and function. Proc Natl Acad Sci 2007;104:11963-11968.

31. Pace NC, Scholtz MJ. A helix propensity scale based on experimental studies of peptides and proteins. Biophys J 1998;75:422-427.

32. Wang T, Wade RC. Force field effects on a β-sheet protein domain structure in thermal unfolding simulations. J Chem Theory Comput 2006;2:140-148.

33. Mittal J, Best RB. Tackling force-field bias in protein folding simulations: Folding of villin HP35 and pin WW domains in explicit water. Biophys J 2010;99:L26-L28.

34. Cheung MS, García AE, Onuchic JN. Protein folding mediated by solvation: Water expulsion and formation of the hydrophobic core occur after the structural collapse. Proc Natl Acad Sci 2002;99:685-690.

35. Zhou R, Berne BJ. Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water? Proc Natl Acad Sci 2002;99:12777-12782.

36. Geney R, Layten M, Gomperts R, Hornak V, Simmerling C. Investigation of salt bridge stability in a generalized Born solvent model. J Chem Theory Comput 2006;2:115-127.

37. Jang S, Kim E, Pak Y. Direct Folding Simulation of α-Helices and β-hairpins based on a single all-atom force field with an implicit solvation model. Proteins Struc Func Bioinfo 2007;66:53-60.

38. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577-2637.

39. Mezei M. Simulaid: A simulation facilitator and analysis program. J Comput Chem 2010;31: 2658-2668.

40. Lei H, Wu C, Liu H, Duan Y. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. Proc Natl Acad Sci 2007;104:4925-4930.
41. Lou H, Cukier RI. Molecular dynamics of apo-adenylate kinase: A principal component analysis. J. Phys. Chem. B 2006;110:12796-12808.
42. Das A, Mukhopadhyay C. Application of principal component analysis in protein unfolding: An all-atom molecular dynamics simulation study. J Chem Phys 2007;127:165103-(1-8).
43. Maisuradze GG, Liwo A, Scheraga HA. Principal component analysis for protein folding dynamics. J Mol Biol 2009;385:312-329.
44. Palazoglu A, Gursoy A, Arkun Y, Erman B. Folding dynamics of proteins from denatured to native state: Principal component analysis. J Comput Biol 2004;11:1149-1168.
45. Mongan J. Interactive essential dynamics J Comp Aided Molec Design 2004;18:433-436.
46. Humphrey W, Dalke A, Schulten K VMD: Visual molecular dynamics. J Molec Graphics 1996;14:33-38.
47. Fukunishi H, Watanabe O, Takada S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. J Chem Phys 2002;116:9058-9067.
48. Li X, Latour RA, Stuart SJ. TIGER2: An improved algorithm for temperature intervals with global exchange of replicas. J Chem Phys 2009;130:174106.

Chapter 4

1. Dill KA, MacCallum JL. The protein folding problem, 50 years on. Sci 2012;338:1042-1046.

2. Ringe D, Petsko GA. Mapping protein dynamics by X-ray diffrac- tion. Prog Biophys Mol Biol 1985;45:197–235.
3. www.pdb.org Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H,Shindyalov IN, Bourne PE (2000) The protein data bank. Nucl Acids Res 28:235-242
4. Karplus M. Molecular dynamics simulations of proteins. Phys Today 1987;40:68–72.
5. Karplus M, Kuriyan J. Molecular dynamics and protein function. Proc Natl Acad Sci USA 2005;102:6679–6685.
6. Shirts M, Pande VS. Screen savers of the world unite! Science 2000; 290:1903–1904.
7. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. Science 2011;334:517–520.
8. Liwo A, Khalili M, Scheraga HA. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. Proc Natl Acad Sci USA 2005;102: 2362–2367.
9. Voelz VA, Jager M, Yao S, Chen Y, Zhu L, Waldauer SA, Bowman GR, Friedrichs M, Bakajin O, Lapidus LJ et al.: Slow unfolded-state structuring in acyl-CoA binding protein folding revealed by simulation and experiment. J Am Chem Soc 2012;134:125645-77.
10. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W. Atomic-level characterization of the structural dynamics of proteins. Science 2010, 330:341-346.
11. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 2003;24:1999–2012.

12. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. J Comput Chem 2005;26:1668–1688.

13. MacKerell AD, Jr, Feig M, Brooks CL, III. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem 2004;25:1400–1415.

14. Scott WRP, Hünenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Krüger P, van Gunsteren WF.The GROMOS biomolecular simulation program package. J Phys Chem A 1999;103:3596–3607.

15. Halgren TA, Damm W. Polarizable force fields. Curr Opin Struc Biol 2001;11:236–242.

16. Wang W, Skeel RD. Fast evaluation of polarizable forces. J Chem Phys 2005;123:164107.

17. Ji CG, Ye M, Zhang JZH. Developing polarized protein-specific charges for protein dynamics: MD free energy calculation of pKa shifts for Asp26/Asp20 in thioredoxin. Biophys J 2008;95:1080– 1088.

18. Duan LL, Mei Y, Zhang D, Zhang QG, Zhang JZH. Folding of a helix at room temperature is critically aided by electrostatic polarization of intraprotein hydrogen bonds. J Am Chem Soc 2010;132: 11159–11164.

19. Davis ME, McCammon JA. Electrostatics in biomolecular structure and dynamics. Chem Rev 1990;90:509–521.

20. Mei Y, Wu EL, Han KL, Zhang JZH. Treating hydrogen bonding in ab initio calculations in biopolymers. J Quant Chem 2006;106:1267-1276.

21. Duan LL, Mei Y, Zhang QG, Zhang JZH. Intra-protein hydrogen bonding is dynamically stabilized by electronic polarization. J Chem Phys 2009;130:115102.
22. Priya R, Biukovic G, Gayen S, Vivekanandan S, Gruber G. Solution structure, determined by nuclear magnetic resonance, of the b30-82 domain of subunit b of Escherichia coli F1Fo ATP synthase. J Bacteriol 2009;191:7538–7544.
23. Chan DC, Fass D, Berger JM, Kim PS. Core structure of gp41 from the HIV envelope glycoprotein. Cell 1997;89:263–273.
24. Zhang DW, Zhang JZH. Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein-molecule interaction energy. J Chem Phys 2003;119:3599–3605.
25. Gao AM, Zhang DW, Zhang JZH, Zhang Y. An efficient linear scaling method for ab initio calculation of electron density of proteins. Chem Phys Lett 2004;394:293–297.
26. Hardin C, Eastwood MP, Luthey-Schulten Z, Wolynes PG. Associative memory Hamiltonians for structure prediction without homology:Alpha-helical proteins. Proc Natl Acad Sci USA 2000;97:14235– 14240.
27. Hardin C, Eastwood MP, Prentiss MC, Luthey-Schulten Z, Wolynes PG. Associative memory Hamiltonians for structure prediction without homology: α/β proteins. Proc Natl Acad Sci USA 2003; 100:1679–1684.
28. Case DA, Darden TA, Cheatham TE, III, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossvary I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, Sagui C, Babin V, Kollman PA. Amber 10. San Francisco: University of California; 2008.

29. Onufriev A, Bashford D, Case DA. Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins 2004;55:383.

30. Uberuaga BP, Anghel M, Voter AF. Synchronization of trajectories in canonical molecular-dynamics simulations: Observation, explanation, and exploitation. J Chem Phys 2004;120:6363–6374.

31. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of nalkanes. J Comput Phys 1977;23:327–341.

32. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;238:777–793.

33. Wei C, Tung D, Yip YM, Mei Y, Zhang DW. Communication: The electrostatic polarization is essential to differentiate the helical propensity in polyalanine mutants. J Chem Phys 2011;134:171101.

34. Xu Z, Lazim R, Sun T, Mei Y, Zhang DW. Solvent effect on the folding dynamics and structure of E6-associated protein characterized from ab initio protein folding simulations. J Chem Phys 2012; 136:135102.

35. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. J Comput Chem 2002;23:128–137.

36. Vila J, Williams RL, Grants JA, Wojcik J, Scheraga HA. The intrinsic helix-forming tendency of L-alanine. Proc Natl Acad Sci USA 1992; 89:7821–7825.

37. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J Comput Chem 1992;13:1011– 1021.
38. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. Multidimensional free-energy calculations using the weighted histo- gram analysis method. J Comput Chem 1995;16:1339–1350.
39. Roux B. The calculation of the potential of mean force using com- puter simulations. Comput Phys Commun 1995;91:275–282.

Chapter 5

1. Banci L. Molecular dynamics simulations of metalloproteins. Curr Opin Chem Biol 2003;7:143–149.
2. Waldron KJ, Robinson NJ. How do bacterial cells ensure that metalloproteins get the correct metal? Nat Rev Microbiol 2009;7:25–35.
3. Tipmanee V, Blumberger J. Kinetics of the terminal electron transfer step in cytochrome c oxidase. J Phys Chem B 2012;116:1876–1883.
4. Min T, Ergenekan CE, Eidsness MK, Ichiye T, Kang CH. Leucine 41 is a gate for water entry in the reduction of *Clostridium pasteurianum* rubredoxin Prot Sci 2001;10:613-621
5. Park IY, Youn B, Harley JL, Eidsness MK, Smith E, Ichiye T, Kang CH. The unique hydrogen bonded water in the reduced form of *Clostridium pasteurianum* rubredoxin and its possible role in electron transfer. J Biol Inorg Chem 2004;9:423-428.
6. Gilardi G, Fantuzzi A. Manipulating redox systems: application to nanotechnology. Trends Biotechnol 2001;19:468–476.
7. Rodgers KK, Sligar SG. Surface electrostatics, reduction potentials, and the internal dielectric constant of proteins. J Am Chem Soc 1991;113:9419–9421.

8. Bertrand P, Mbarki O, Asso M, Blanchard L, Guerlesquin F, Tegoni M. Control of the redox potential in c-type cytochromes: importance of the entropic contribution. Biochem 1995;34:11071– 11079.

9. Carter CW, Jr. New stereochemical analogies between iron-sulfur electron transport proteins. J Biol Chem 1977;252:7802–7811.

10. Adman E, Watenpaugh KD, Jensen LH. NH---S hydrogen bonds in Peptococcus aerogenes ferredoxin, Clostridium pasteurianum rubredoxin, and Chromatium high potential iron protein. Proc Natl Acad Sci USA 1975;72:4854–4858.

11. Kassner RJ. Effects of nonpolar environments on the redox potentials of heme complexes. Proc Natl Acad Sci USA 1972;69:2263– 2267.

12. Yanagisawa S, Banfield MJ, Dennison C. The role of hydrogen bonding at the active site of a cupredoxin: The Phe114Pro azurin variant. Biochemistry 2006;45:8812–8822.

13. Garner DK, Vaughan MD, Hwang HJ, Savelieff MG, Berry SM, Honek JF, Lu Y. Reduction potential tuning of the blue copper cen- ter in Pseudomonas aeruginosa Azurin by the axial methionine as probed by unnatural amino acids. J Am Chem Soc 2006;128:15608– 15617.

14. Berry SM, Baker MH, Reardon NJ. Reduction potential variations in azurin through secondary coordination sphere phenylalanine incorporations. J Inorg Biochem 2010;104:1071–1078.

15. Yelle RB, Park NS, Ichiye T. Molecular dynamics simulations of rubredoxin from Clostridium pasteurianum: Changes in structure and electrostatic potential during redox reactions. Proteins Struct Funct Genet 1995;22:154–167.

16. Lin I-J, Gebel EB, Machonkin TE, Westler WM, Markley JL. Changes in hydrogen-bond strengths explain reduction potentials in 10 rubredoxin variants. Proc Natl Acad Sci USA 2005;102:14581–14586.

17. Zheng P, Takayama SJ, Mauk AG, Li H. Hydrogen bond strength modulates the mechanical strength of ferric-thiolate bonds in rubredoxin. J Am Chem Soc 2012;134:4124–4131.

18. Olsson MHM, Ryde U. Geometry, reduction potential, and reorganization energy of the binuclear CuA site, studied by density functional theory. J Am Chem Soc 2001;123:7866–7876.

19. Olsson MHM, Hong GY, Warshel A. Frozen density functional free energy simulations of redox proteins: Computational studies of the reduction potential of plastocyanin and rusticyanin. J Am Chem Soc 2003;125:5025–5039.

20. Li H, Webb SP, Ivanic J, Jensen JH. Determinants of the relative reduction potentials of Type-1 copper sites in proteins. J Am Chem Soc 2004;126:8010–8019.

21. Si DJ, Li H. Quantum chemical calculation of Type-1 Cu reduction potential: Ligand interaction and solvation effect. J Phys Chem A 2009;113:12979–12987.

22. Cascella M, Magistrato A, Tavernelli I, Carloni P, Rothlisberger U. Role of protein frame and solvent for the redox properties of azurin from Pseudomonas aeruginosa. Proc Natl Acad Sci USA 2006;103: 19641-19646.

23. Barone V, De Rienzo F, Langella E, Menziani MC, Rega N, Sola M. A computational protocol to probe the role of solvation effects on the reduction potential of azurin mutants. Proteins Struct Funct Bioinf 2006;62:262–269.

24. Van den Bosch M, Swart M, Snijders JG, Berendsen HJC, Mark AE, Oostenbrink C, van Gunsteren WF, Canters GW. Calculation of the redox potential of the protein azurin and some mutants. ChemBio- Chem 2005;6:738–746.

25. Datta SN, Sudhamsu J, Pandey A. Theoretical determination of the standard reduction potential of plastocyanin in vitro. J Phys Chem B 2004;108:8007–8016.
26. Sattelle BM, Sutcliffe MJ. Calculating Chemically Accurate Redox Potentials for Engineered Flavoproteins from Classical Molecular Dynamics Free Energy Simulations J Phys Chem A 2008;112:13053-13057.
27. Sundararajan M, Hillier IH, Burton NA. Structure and redox properties of the protein, rubredoxin, and its ligand and metal mutants studied by electronic structure calculation. J Phys Chem A 2006; 110:785–790.
28. Gámiz-Hernández AP, Galstyan AS, Knapp EW. Understanding Rubredoxin Redox Potentials: Role of H-Bonds on Model Complexes J Chem Theory Comput 2009;5:2898-2908.
29. Bertini L, Bruschi M, Cosentino U, Greco C, Moro G, Zampella G, Gioia LD. Quantum mechanical methods for the investigation of metalloproteins and related bioinorganic compounds. Metalloproteins: Methods and Protocols, Methods in Molecular Biology 2014;1122:207-268.
30. Shenoy VS, Ichiye T. Influence of protein flexibility on the redox potential of rubredoxin: Energy minimization studies. Prot Struc Func Genet 1993;17:152-160.
31. Zeng XC, Hu H, Hu XQ, Yang WT. Calculating solution redox free energies with ab initio quantum mechanical/molecular mechanical minimum free energy path method. J Chem Phys 2009;130:164111.
32. Hu H, Yang WT. Free energies of chemical reactions in solution and in enzymes with ab initio quantum mechanics/molecular mechanics methods. Annu Rev Phys Chem 2008;59:573–601.

33. Formaneck MS, Li GH, Zhang XD, Cui Q. Calculating accurate redox potentials in enzymes with a combined QM/MM free energy perturbation approach. J Theor Comput Chem 2002;1:53– 67.
34. Watenpaugh KD, Sieker LC, Jensen LH. The structure of rubredoxin at 1.2 Å resolution. J Mol Biol 1979;131:509-522.
35. Watenpaugh KD, Sieker LC, Jensen LH. Crystallographic refinement of rubredoxin at 1·2 Å resolution. J Mol Biol 1980;138:615-633.
36. Gámiz-Hernández, A. P. L.; Knapp, E. W.; Understanding the redox behavior of transition metal complexes: from molecular models to proteins. Thesis. Berlin, August 2010.
37. Ergenekan CE, Thomas D, Fischer JT, Tan ML, Eidsness MK, Kang CH, Ichiye T. Prediction of Reduction Potential Changes in Rubredoxin: A Molecular Mechanics Approach. Biophys J 2003;85:2818-2829.
38. Luo Y, Ergenekan CE, Fischer JT, Tan ML, Ichiye T. The Molecular Determinants of the Increased Reduction Potential of the Rubredoxin Domain of Rubrerythrin Relative to Rubredoxin. Biophys J 2010;98:560-568.
39. Sigfridsson E, Olsson MHM, Ryde U. A comparison of the inner- sphere reorganization energies of cytochromes, iron-sulfur clusters, and blue copper proteins. J Phys Chem B 2001;105:5546–5552.
40. Dolan EA, Yelle RB, Beck BW, Fischer JT, Ichiye T. Protein Control of Electron Transfer Rates via Polarization: Molecular Dynamics Studies of Rubredoxin. Biophys J 2004;86:2030-2036.
41. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman PA. A point-charge force field for molecular

mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 2003;24:1999–2012.

42. Gaussian 09, Revision **A.1**, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. tNakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2009.
43. Becke AD. Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 1993;98:5648–5652.
44. Lee CT, Yang WT, Parr RG. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Phys Rev B 1988;37:785–789.
45. Bayly CI, Cieplak P, Cornell WD, Kollman PA. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. J Phys Chem 1993;97:10269– 10280.
46. www.pdb.org Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucl Acids Res 2000; 28:235-242.

47. Case DA, Darden TA, Cheatham TE, III, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossvary I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, Sagui C, Babin V, Kollman PA. Amber 10. San Francisco: University of California; 2008.
48. Giammano, D. A. Ph.D Thesis, University of California, Davis (1984)
49. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. J Am Chem Soc 1995;117:5179-5197.
50. Sattelle BM, Sutcliffe MJ. Calculating chemically accurate redox potentials for engineered flavoproteins from classical molecular dynamics free energy simulations. J Phys Chem A 2008;112:13053– 13057.
51. Darden T, York D, Pedersen L. Particle mesh Ewald: An *N*•log(*N*) method for Ewald sums in large systems. J Chem Phys 1993; 98:10089.
52. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 1977;23:327–341.
53. Jaguar, 5.5; Schroedinger, L.L.C.: Portland, OR, 1991-2003
54. Peraro MD, Spiegel K, Lamoureux G, De Vivo M, DeGrado WF, Klein ML Modeling the charge distribution at metal sites in proteins for molecular dynamics simulations. J Struc Biol 2007;157:444-453.
55. Xia B, Wilkens SJ, Westler WM, Markley JL. Amplification of One-Bond $^1H/^2H$ Isotope Effects on ^{15}N Chemical Shifts in *Clostridium pasteurianum* Rubredoxin by

Fermi-Contact Effects through Hydrogen Bonds. J Am Chem Soc 1998;120:4893-4894.

56. Sulpizi M, Raugei S, VandeVondele J, Carloni P, Sprik M. Calculation of Redox Properties: Understanding Short- and Long-Range Effects in Rubredoxin. J Phys Chem B 2007;111:3969-3976.
57. Wilkens SJ, Xia B, Weinhold F, Markley JL, Westler WM. NMR Investigations of Clostridium pasteurianum Rubredoxin. Origin of Hyperfine ^{1}H, ^{2}H, ^{13}C, and ^{15}N NMR Chemical Shifts in Iron−Sulfur Proteins As Determined by Comparison of Experimental Data with Hybrid Density Functional Calculations. J Am Chem Soc 1998;120:4806-4814.
58. Xiao S, Maher MJ, Cross M, Bond CS, Guss JM, Wedd AG. Mutation of the surface valine residues 8 and 44 in the rubredoxin from *Clostridium pasteurianum* : solvent access versus structural changes as determinants of reversible potential. J Biol Iorg Chem 2000;5:75-84.
59. Park IY, Eidsness MK, Lin I-J, Gebel EB, Youn B, Harley JL, Machonkin TE, Frederick RO, Markley JL, Smith ET, Ichiye T, Kang CH. Crystallographic studies of V44 mutants of *Clostridium pasteurianum* rubredoxin: Effects of side-chain size on reduction potential. Protein 2004;57:618-625.

Chapter 6

1. Saven, JG. Computational protein design: engineering molecular diversity, nonnatural enzymes, nonbiological cofactor complexes, and membrane proteins. Curr Op Chem Biol 2011;15:452-457.
2. Chaput JC, Woodbury NW, Stearns LA, Wiliams BAR. Creating protein biocatalysts as tools for future industrial applications. Expert Opin Biol Ther 2008;8:1087-1098.

3. Hellinga HW. Rational protein design: Combining theory and experiment. Proc Natl Acad Sci 1997;94:10015-10017.
4. Fu H, Grimsley G, Scholtz JM, Pace CN. Increasing protein stability: Importance of ΔC_p and the denatured state. Prot Sci 2010;19:1044-1052.
5. Motono C, Gromiha MM, Kumar S. Thermodynamic and kinetic determinants of Thermotoga maritime cold shock protein stability: A structural and dynamic analysis. Proteins 2008;71:655-669.
6. Bannen RM, Suresh V, Phillips Jr GN, Wright SJ, Mitchell JC. Optimal design of thermally stable proteins. Bioinformatics 2008;24:2339-2343.
7. Blundell TL, Elliot FRS, Gardner SP, Hubbard T, Islam S, Johnson M, Mantafounis D, Murray-Rust P, Overington J, Pitts JE, Sali A, Sibanda BL, Singh J, Stenberg MJE, Sutcliffe MJ, Thornton JM, Travers P. Protein Engineering and Design. Phil Trans R Soc Lond B 1989; 324:447-460.
8. Hellinga HW. Computational protein engineering. Nat Struc Biol 1998; 5:525-527.
9. Gilardi G, Fantuzzi A. Manipulating redox systems: application to nanotechnology. Trends Biotechnol 2001;19:468–476.
10. Karplus M. Molecular dynamics simulations of proteins. Physics Today 1987;40:68-72.
11. Luo YZ, Baldwin RL. How Ala → Gly mutations in different helices affect the stability of the apomyoglobin molten globule. Biochem 2001; 40:5283-5289.
12. Luo YZ, Kay MS, Baldwin RL. Cooperativity of folding of the apomyoglobin pH 4 intermediate studied by glycine and proline mutations. Nature 1997; 4:925-930.
13. Choi HS, Huh J, Jo WH. Similarity of Force-Induced Unfolding of Apomyoglobin to Its Chemical-Induced Unfolding: An Atomistic Molecular Dynamics Simulation Approach. Biophys J 2003; 85:1492-1502.

14. Eliezer D, Wright PE. Is apomyoglobin a molten globule? Structural characterization by NMR. J Mol Biol 1996; 263:531–538.

15. Gilmanshin R, Dyer RB, Callender RH. Structural heterogeneity of the various forms of apomyoglobin: Implications for protein folding. Prot Sci 1997; 6:2134-2142.

16. Kamei T, Oobatake M, Suzuki M. Hydration of apomyoglobin in native, molten globule, and unfolded state by using microwave dielectric spectroscopy. Biophys J 2002; 82:418-425.

17. Hughson FM, Wright PE, Baldwin RL. Structural characterization of a partly folded apomyoglobin intermediate. Science 1990; 249:1544–1548.

18. Hughson FM, Barrick D, Baldwin RL. Probing the stability of a partly folded apomyoglobin intermediate by site-directed mutagenesis. Biochemistry 1991; 30:4113–4118.

19. Barrick D, Baldwin RL. Three-state analysis of sperm whale apomyoglobin folding. Biochem 1993; 32:3790–3796.

20. Onufriev A, Case DA, Bashford D. Structural details, pathways, and energetics of unfolding apomyoglobin. J Mol Biol 2003; 325:555-567.

21. Case DA, Darden TA, Cheatham TE, III, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossvary I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, Sagui C, Babin V, Kollman PA. Amber 10. San Francisco: University of California; 2008.

22. www.pdb.org Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucl Acids Res 2000; 28:235-242.

23. Kachalova GS, Popov AN, Bartunik HD. A steric mechanism for inhibition of CO binding to heme proteins. Science 1999; 284:473-476.
24. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys 1983;79:926-935.
25. Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. Prot Struc Func Bioinfo 2005;61:704-721.
26. Bas DC, Rogers DM, Jensen JH. Very fast prediction and rationalization of pKa values for protein-ligand complexes. Prot Struc Func Bioinfo 2008;73:765-783.
27. Olsson MHM, Sondergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent treatment of internal and surface residues in empirical pKa predictions. J Chem Theory Comput 2011;7:525-537.
28. Sondergaard CR, Olsson MHM, Rostkowski M, Jensen JH. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pKa values. J Chem Theory Comput 2011;7:2284-2295.
29. Pastor RW, Brooks BR, Szabo A. An analysis of the accuracy of Langevin and molecular dynamics algorithm. Mol Phys 1988; 65:1409-1419.
30. Uberuaga BP, Anghel M, Voter AF. Synchronization of trajectories in canonical molecular-dynamics simulations: Observation, explanation, and exploitation. J Chem Phys 2004;120:6363–6374.
31. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general AMBER force field. J Comput Chem 2004;25:1157-1174.
32. Özpınar GA, Peukert W, Clark T. An improved generalized AMBER force field (GAFF) for urea. J Mol Model 2010;16:1427-1440.
33. Darden T, York D, Pedersen L. Particle mesh Ewald: An *N*•log(*N*) method for Ewald sums in large systems. J Chem Phys 1993; 98:10089.

34. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 1977;23:327–341.

35. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 2006;65:712-725.

36. Lauren W, Okur A, Simmerling C. Evaluating the performances of the ff99SB force field based on NMR scalar coupling data. Biophys J 2009;97:853-856.

37. Pace CN, Shirley BA, McNutt M, Gajiwala K. Forces contributing to the conformational stability of proteins. FASEB J 1996;10:75-83.

38. Tcherkasskaya O, Bychkova VE, Uversky VN, Gronenborn AM. Multisite fluorescence in proteins with multiple tryptophan residues. J Biol Chem 2000;275:36285-36294.

39. Glandières JM, Twist C, Haouz A, Zentz C, Alpert B. Resolved fluorescence of the two tryptophan residues in horse apomyoglobin. Photochem Photobiol 2000;71:382-386.

40. Twist C, Royer C, Alpert B. Effect of solvent diffusion on the apomyoglobin-water interface. Biochem 2002;41:10343-10350.

41. Xu M, Beresneva O, Rosario R, Roder H. Microsecond folding dynamics of apomyoglobin at acidic pH. J Phys Chem B 2012;116:7014-7025.

42. Lapidus LJ, Yao S, McGarrity KS, Hertzog DE, Tubman E, Bakajin O. Protein hydrophobic collapse and early folding steps observed in a microfluidic mixer. Biophys J 2007;93:218-224.

43. Brun L, Isom DG, Velu P, García-Moreno B, Royer CA. Hydration of the folding transition state ensemble of a protein. Biochem 2006;45:3473-3480.

44. Gulotta M, Rogatsky E, Callender RH, Dyer RB. Primary folding dynamics of sperm whale apomyoglobin: Core formation. Biophys J 2003;84:1909-1918.

45. Cocco MJ, Lecomte JTJ. Characterization of hydrophobic cores in apomyoglobin: A proton NMR spectroscopy study. Biochem 1990;29:11067-11072.

46. Eliezer D, Yao J, Dyson HJ, Wright PE. Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding. Nat Struct Biol 1998;5:148-155.

47. Nishimura C, Dyson HJ, Wright PE. The apomyoglobin folding pathway revisited: Structural heterogeneity in the kinetic burst phase intermediate. J Mol Biol 2002;322:483-489.

48. Haliloglu T, Bahar I. Coarse-grained simulations of conformational dynamics of proteins: Application to apomyoglobin. Proteins 1998;31:271-281.

49. Picotti P, Marabotti A, Negro A, Musi V, Spolaore B, Zambonin M, Fontana A. Modulation of the structural integrity of helix F in apomyoglobin by single amino acid replacements. Protein Sci 2004;13:1572-1585.

50. Lou H, Cukier RI. Molecular dynamics of apo-adenylate kinase: a principal component analysis. J Phys Chem *B* 2006; 110:12796-12808.

51. Maisuradze GG, Liwo A, Scheraga HA. Principal component analysis for protein folding dynamics. J Mol Biol 2009; 385:312-329.

52. Mongan J. Interactive essential dynamics. J. Comp. Aided Molec. Design 2004; 18:433-436.

53. Humphrey W, Dalke A, Schulten K. VMD - Visual Molecular Dynamics J Mol Graphics 1996; 14:33-38.

54. Munson M, Balasubramanian S, Fleming KG, Nagi AD, O'Brien R, Sturtevant JM, Regan L. What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties. Protein Sci 1996;5:1584-1593.

55. **Suite 2011**: Maestro, version 9.2, Schrödinger, LLC, New York, NY, 2011.

Appendix

1. Kachalova GS, Popov AN, Bartunik HD. A steric mechanism for inhibition of CO binding to heme proteins. Science 1999; 284:473-476.
2. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys 1983;79:926-935.
3. Onufriev A, Case DA, Bashford D. Structural details, pathways, and energetics of unfolding apomyoglobin. J Mol Biol 2003;325:555-567.
4. Case DA, Darden TA, Cheatham TE, III, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossvary I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, Sagui C, Babin V, Kollman PA. Amber 10. San Francisco: University of California; 2008.
5. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 2006;65:712-725.
6. Lauren W, Okur A, Simmerling C. Evaluating the performances of the ff99SB force field based on NMR scalar coupling data. Biophys J 2009;97:853-856.
7. Darden T, York D, Pedersen L. Particle mesh Ewald: An *N*•log(*N*) method for Ewald sums in large systems. J Chem Phys 1993; 98:10089.

8. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 1977;23:327–341.
9. Pastor RW, Brooks BR, Szabo A. An analysis of the accuracy of Langevin and molecular dynamics algorithm. Mol Phys 1988; 65:1409-1419.
10. Uberuaga BP, Anghel M, Voter AF. Synchronization of trajectories in canonical molecular-dynamics simulations: Observation, explanation, and exploitation. J Chem Phys 2004;120:6363–6374.

List of Publications

In writing this thesis, the following papers were produced:

Xu Z, **Lazim R**, Mei Y, Zhang D. Stability of the β-structure in prion protein: A molecular dynamics study based on polarized force field. Chem Phys Lett 2012; 539-540:239-244 **[Chapter 2]**

Lazim R, Mei Y, Zhang, D. Replica exchange molecular dynamics of structure variation from α/4β -fold to 3α-fold protein. J Mol Model 2012;18:1087-1095 **[Chapter 3]**

Lazim R, Wei C, Sun T, Zhang D. Ab initio folding of extended α-helix: A theoretical study about the role of electrostatic polarization in the folding of helical structures. Proteins 2013; 81:1610-1620 **[Chapter 4]**

Wei C[1], **Lazim R[1]**, Zhang D. Importance of polarization effect in the study of metalloproteins: Application of polarized protein specific charge scheme in predicting the reduction potential of azurin. Protein 2014. doi: 10.1002/prot.24584. ([1] Both authors contribute equally to the production of this paper) **[Chapter 5]**

Other publications:

Xu Z, **Lazim R**, Sun T, Mei Y, Zhang D. Solvent effect on the folding dynamics and structure of E6-associated protein characterized from ab initio simulations. J Chem Phys 2012;136:135102-135106

Hartono YD, **Lazim R**, Yip YM, Zhang D. Computational study of bindings of HK20 Fab and D5 Fab to HIV-1 gp41. Bioorg Med Chem Lett 2012;22:1695-1700