A study on protein secondary structure phase transition based on an effective hydrogen bonding potential

Leong, Hon Wai

2013

Leong, H. W. (2013). A study on protein secondary structure phase transition based on an effective hydrogen bonding potential. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/54751>

<https://doi.org/10.32657/10356/54751>

**A STUDY ON PROTEIN SECONDARY STRUCTURE
PHASE TRANSITION BASED ON AN EFFECTIVE
HYDROGEN BONDING POTENTIAL**

LEONG HON WAI

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCES

2013

**A STUDY ON PROTEIN SECONDARY STRUCTURE
PHASE TRANSITION BASED ON AN EFFECTIVE
HYDROGEN BONDING POTENTIAL**

LEONG HON WAI

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

2013

Acknowledgement

First of all, I would like to express my heartiest gratitude to Asst/Prof. Dr. Chew Lock Yue, for his dedicated advice, patience and forgiveness throughout my PhD candidature journey. I wouldn't have been able to produce this thesis without his continuous guidance. There was a time when I thought of giving up my studies, and many thanks to Dr. Chew for his sincere words of encouragement for pulling me back into what I have started six years ago.

I would like to express my appreciation to Dr. Kerson Huang, Professor Emeritus from Massachusetts Institute of Technology (MIT). Dr. Huang has constructed the foundation of this research. I am grateful to be under his guidance at the beginning of my research.

Next, I would also like to extend my acknowledgement to the staff at High Performance Computing Centre for providing the computing resources for my research. The important results which lay the significance of this thesis wouldn't exist without their HPC resources.

Finally, I would like to thank my family, especially my wife, Ms. Go Ling Hong. This thesis is dedicated to them as an appreciation for their continuous support during my PhD studies.

By *Hon Wai, LEONG*

16th June 2013

Table of Contents

Acknowledgement	i
Summary	iii
Chapter 1: Introduction	1
Chapter 2: The Modeling Concept of Polypeptide Cranks.....	12
2.1 Origin: Conditioned Self-Avoiding Walk (CSAW)	12
2.2 The Polypeptide Cranks Model	15
2.3 The Theory of Our Polypeptide Mathematical Model	19
2.4 Scopes of Work	23
Chapter 3: Hamiltonian Formalism and Normal Mode Analysis	25
3.1 Lagrangian and Hamiltonian Formulation of Polypeptide Secondary Structures	25
3.2 Effective Potential of the Polypeptide Conformation.....	29
3.3 Potential Energy of Hydrogen Bonding.....	31
3.4 Normal Modes	33
3.5 The α -Helix.....	35
3.6 The β -Sheet.....	37
3.7 Comparison with Other Research Work	39
3.8 α - β Phase Transition	40
Chapter 4: Phase Transition in α - β -Coil Protein Secondary Structure	45
4.1 Canonical Formalism of Protein Chain.....	45
4.2 Grand Partition Function of Protein Chain	46
4.3 The Statistical Physics of Protein Phase Transition.....	48
4.4 α -helix to Coil Phase Transition	48
4.5 α -Helix to β -Sheet to Coil Phase Transition	49
4.6 Computation of the Change of Heat Capacity in Polypeptide	50
4.7 Analysis of Results and Comparison with Literature	53
4.8 Revisiting Phase Transition with Respect to Helmholtz Free Energy.....	54
4.9 Further Analysis and Discussion on Protein Phase Transition	55
Chapter 5: Conclusion.....	65
Bibliography	70
Appendix A: Rotation Matrix	78
Appendix B: Analytical Solution of Partition Function for polypeptide cranks.....	80
Appendix C: Effective Potential of Protein Chain Configuration	81
Appendix D: MATLAB Programming Codes	107

Summary

In this thesis, we define the single unit of a polypeptide as crank. Using this convention, we formulate the Hamiltonian of polypeptide. In our canonical formalism, the dihedral angles of polypeptide chain are treated as generalized coordinates, and we take the assumption that the effective potential of a polypeptide chain is contributed by hydrogen bonding alone. We have employed such formalism to analyze the normal modes distribution of polypeptide secondary structures, such as the α -helix and the β -sheet. Interestingly, our results are in agreement with literature. Following this success, assuming that the dihedral angles of each crank only folds into five distinct states, we derive the grand partition function of polypeptide chain. We study the phase transition of polypeptide secondary structures and its statistical properties. Our theory has successfully demonstrated the α - β -coil phase transition of polypeptide secondary structure in a single plot. Our results emphasize the role of hydrogen bonding network in polypeptide secondary structure phase transition. At the conclusion of this thesis, we discuss the potential of our research and the obstacles to be overcome in order to refine the results obtained so far. The major results presented in this thesis have been reviewed and published in the following journal articles:

1. H. W. Leong, L. Y. Chew and K. Huang, *Physical Review E*, **82**, 011915 (2010).
2. H. W. Leong and L. Y. Chew, *Physical Review E*, **86**, 031902 (2012).

We have presented the results of our articles in this thesis with permission from American Physical Society.

Chapter 1: Introduction

Various aspects of protein have been studied since many decades ago. This includes analysis of biological functions, computation of normal modes, study of protein-protein interaction etc., and most importantly the folding mechanism of protein. Due to its large scale and complex molecular structures, most studies of protein folding are limited to small chain⁽¹⁾. But yet, the study of small protein chain is still considered complex, albeit they are amenable to studies through computational method.

In the review on "The Protein Folding Problem" published in the Annual Review of Biophysics 2008⁽²⁾, the authors defined three closely related puzzles of the protein folding problem:

- (a) What is the folding code?
- (b) What is the folding mechanism?
- (c) Can we predict the native structure of a protein from its amino acid sequence?

It was explained that protein solves its large global optimization problem as a series of smaller local optimization problems, growing and assembling the native structure from peptide fragments, i.e. local structures first. The major goal of computational research on protein folding is to predict a protein's three-dimensional native structure from its amino acid sequence in order to accelerate drug discovery and annotate protein function from genome sequences. Computational biology can reduce experimental cost of slower prediction from structural biology experiments⁽²⁾.

Protein is a large organic complex formed by multiple atoms. The basic unit of a protein is known as an amino acid residue. An amino acid residue is a molecule which contains an amine group, a carboxylic acid group, and a side chain. Figure 1.1 illustrates the generic representation of an amino acid residue.

Figure 1.1: An amino acid residue⁽³⁾.

There are altogether 20 types of standard amino acid residues of different side chains (Figure 1.2). Each of these residues possesses different characteristics from each others. Figure 1.3 shows how a peptide bond is formed by joining two residues into chain. In fact, these residues join together to form a very long chain. Different combinations of these amino acid residues constitute a long sequence, known as primary structure.

Figure 1.2: Different type of amino acid residues.

Figure 1.3: Formation of a peptide bond between two amino acid residues through condensation. ⁽⁴⁾

The non-covalent interactions between different residues in a primary structure drive the formation of secondary structure such as α -helices and β -sheets. These secondary structures interact with each other to form the tertiary structure, a three-dimensional subunit complex, a.k.a. native structure of the protein chain. The specific biological function of a protein is known to be determined by its native structure. Multiple folded protein chains will then form the quaternary structure. (See Figure 1.4)

Figure 1.4: Evolution of protein from its primary structure into native quaternary structure ⁽⁵⁾.

Proteins are known to perform various biological functions ⁽⁶⁾: they are enzymes that catalyze chemical reactions occurring in biological systems and increase rates of the reactions at least by a factor of 10^6 ; they transport chemical compounds ranging from electrons to macromolecules, for example haemoglobin that transports oxygen and iron in the circulatory system; they are hormones acting as messengers and are responsible for coordinating cells in tissues; they are immunoglobulin or antibodies that bind to specific foreign invaders such as bacteria or viruses to prevent them to attack the living cells; they serve as structure for cells and tissues, such as actin filaments and myosin filaments which are the major component of muscle tissues that determine the shape of cells; they are motor proteins which convert chemical energy into mechanical energy; they perform self regulation of body fluids, such as forming ion pump to keep the concentration balance of ions in cells, such as Na⁺-K⁺-ATPase; and provide storage for energy.

Proteins function by binding to other molecules in a specific manner. The ability of protein to bind specifically to its partners induces formation of aggregates that perform discrete sets of function. Study of protein-protein interaction is the key to understand important aspects of cellular function. The functional characteristics of proteins can be studied through *in vitro* or *in vivo* experiments, and by means of numerical simulations, analysis and computation. In this thesis, we shall focus our studies based on the approach of numerical analysis and computation.

Though experimental values of physical quantities of such many-particle system as a protein can be found via ensemble average, experimental systems are so large that it is impossible to determine its ensemble average by summing over all accessible states in a computer. Two popular computational methods for determining these physical quantities as statistical averages over a restricted set of states are the molecular dynamics ⁽⁷⁾⁽⁸⁾ and Monte Carlo ⁽⁹⁾ methods. These two methods have been used widely in solving protein folding problem.

Molecular dynamics (MD) ⁽¹⁰⁾ is a widely used method for studying protein folding. It essentially integrates the equations of motion of the system numerically. It can therefore be viewed as a simulation of the system as it develops over a period of time. The great advantage of the MD method is that it not only provides a way to evaluate expectation values of static physical quantities; dynamical phenomena, such as transport of heat or charge, or relaxation of systems far from equilibrium can also be studied.

In Monte Carlo (MC) method⁽¹¹⁾, the classical many-particle system is simulated by introducing artificial dynamics based on ‘random’ numbers. There are three types of Monte Carlo simulations: direct Monte Carlo, Monte Carlo integration and Metropolis Monte Carlo. In most cases, protein folding problem is studied using Metropolis Monte Carlo method, in which a sequence of distributions of the protein system is generated in a so-called Markov chain.

In general, a generic computational protein model possesses degrees of freedom which are described by a set of generalized coordinates, with the force field being given semi-empirically, consisting of energy terms due to stretching, bending, torsion, as well as van der Waals and electrostatic interaction. The all-atom full potential field as described by Karplus *et. al.*⁽⁸⁾ can be written as

$$V = \sum_i \frac{K}{2} (b_i - b_i^0)^2 + \sum_i \frac{C}{2} (\theta_i - \theta_i^0)^2 + \sum_i D [1 + \cos(n\phi_i - \delta)] + \sum_{i,j} \left(\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^6} \right) + \sum_{i,j} \frac{q_i q_j}{\chi r_{ij}}$$

Equation 1.1

In the right side of Equation 1.1, the first term is the stretching potential; second term is the bending potential; third term is the torsion potential; fourth term is the van der Waals potential; and the fifth term is the electrostatic potential.

Ongoing protein research involves the studies of different terms based on Equation 1.1, inclusive of adaption of different values into the parameters as well as modification of the empirical formula for each term. The objective is to formulate an unified potential to model any amino acid sequence into its native structure. There have been different force-fields used to simulate protein folding problems. Particularly for molecular dynamics, the more popular conventional force fields computational package are Assisted Model Building with Energy Refinement (AMBER)⁽¹²⁾, Chemistry at HARvard Macromolecular Mechanics (CHARMM)^{(13) (14) (15)}, Merck Molecular Force Field (MMFF)⁽¹⁶⁾, Not Another Molecular Dynamics (NAMD)⁽¹⁷⁾, GRONingen MACHine for Chemical Simulations (GROMACS)⁽¹⁸⁾ etc. These force field packages were developed by different research institutes based on different assumptions on the force fields parameters. Among different algorithms developed for protein folding prediction, the more advanced algorithms include replica exchange molecular dynamics (REMD)⁽¹⁹⁾ and Reduced Protein Folding Model (RPFM)⁽⁶⁾ which incorporate the Monte Carlo simulation method.

The research conducted in this thesis is greatly inspired by Chen's RPFM theory in formulating the effective potential of protein folding. The RPFM⁽⁶⁾ is a reduced off-lattice model in which peptide chains are represented by explicit backbone structures with simplified side chain units. The degrees of freedom for backbones are based on the Ramachandran angles ϕ and ψ . In order to reduce the complexity and the time for computation, there is no internal degree of freedom for side-chains. Since water molecules are not included explicitly in this model, their effects are incorporated into several effective potentials. The first important interaction is the hydrogen bond interaction. This potential has been considered as the major interaction responsible for stabilizing secondary structures. The second one is the hydrophobic interaction which is a mesoscopic potential induced by collective motion of water molecules, and is considered responsible for the compact globule formation of peptides. In addition, two new interactions were introduced in their works: dipole-dipole interactions and local hydrophobic interactions. The forms for these two potentials are proposed explicitly considering detailed structures of proteins.

Despite large amount of research being spent in computational protein science, protein folding remains an open problem. Though certain well developed force-fields are able to stabilize certain proteins, there is as yet a unified force-field being discovered that can fold any amino acid sequence. Based on Equation 1.1, certain terms can bias the folding towards a particular structure. Our main observations through literature studies have found that protein folding is unlikely to be dominated by electrostatic interactions between different side chains as most side chains are neutral, with only five of them considered to be partially charged. The stability of protein is somehow not affected by pH or salt concentration of the solvent. Small charges on side chains typically do not affect the stability of the structure. However, hydrogen bonding seems important, since hydrogen bonds are key to the formation of secondary structure, which stabilizes the native structure. Similarly, the tight packing observed in proteins implies that the van der Waals interactions are important. Table 1.1 shows diverse characteristics of different protein side chains. Depending on the assumptions, different values of hydrophobic index have been calculated, which are shown in Table 1.2.

Table 1.1: Properties of 20 different types of amino acid residues. (Reproduced from Ref. (20))

Amino Acid	3-Letter ⁽²¹⁾	1-Letter ⁽²¹⁾	Side Chain Polarity ⁽²¹⁾	Side Chain Charge ⁽²¹⁾	Hydrophobic Index ⁽²²⁾
Alanine	Ala	A	Non-Polar	Neutral	1.8
Arginine	Arg	R	Polar	Basic (strong)	-4.5
Asparagine	Asn	N	Polar	Neutral	-3.5
Aspartic Acid	Asp	D	Polar	Acidic	-3.5
Cysteine	Cys	C	Polar	Neutral	2.5
Glutamic Acid	Glu	E	Polar	Acidic	-3.5
Glutamine	Gln	Q	Polar	Neutral	-3.5
Glycine	Gly	G	Non-Polar	Neutral	-0.4
Histidine	His	H	Polar	Basic (weak)	-3.2
Isoleucine	Ile	I	Non-polar	Neutral	4.5
Leucine	Leu	L	Non-polar	Neutral	3.8
Lysine	Lys	K	Polar	Basic	-3.9
Methionine	Met	M	Non-Polar	Neutral	1.9
Phenylalanine	Phe	F	Non-Polar	Neutral	2.8
Proline	Pro	P	Non-Polar	Neutral	-1.6
Serine	Ser	S	Polar	Neutral	-0.8
Threonine	Thr	T	Polar	Neutral	-0.7
Tryptophan	Trp	W	Non-Polar	Neutral	-1.3
Tyrosine	Tyr	Y	Polar	Neutral	-1.3
Valine	Val	V	Non-Polar	Neutral	4.2

Table 1.2: Different values of hydrophobic index. KD - Kyte and Doolittle ⁽²²⁾; RW - Radzicka and Wolfenden ⁽²³⁾; WW - Wimley and White ⁽²⁴⁾; HH - Hessa, Kim, Bihlmaier, Lundin, Boekel, Andersson, Nilsson, White and von Heijne ⁽²⁵⁾; AK, Andrey Karshikoff ⁽²⁶⁾.

Amino Acid			KD	RW	WW	HH	AK
Alanine	A	Ala	1.8	0.05	-0.17	0.11	0.4
Arginine	R	Arg	-4.5	-0.42	-0.81	2.58	-1.4
Asparagine	N	Asn	-3.5	-0.28	-0.42	2.05	-0.8
Aspartic Acid	D	Asp	-3.5	-0.35	-1.23	3.49	-1.1
Cysteine	C	Cys	2.5	0.01	0.24	-0.13	2.1
Glutamic Acid	E	Glu	-3.5	-0.21	-2.02	2.68	-0.9
Glutamine	Q	Gln	-3.5	-0.25	-0.58	2.36	-0.3
Glycine	G	Gly	-0.4	0.00	-0.01	0.74	0.0
Histidine	H	His	-3.2	-0.17	-0.96	2.06	0.2
Isoleucine	I	Ile	4.5	0.15	0.31	-0.60	2.5
Leucine	L	Leu	3.8	0.15	0.56	-0.55	2.3
Lysine	K	Lys	-3.9	-0.21	-0.99	2.71	-1.4
Methionine	M	Met	1.9	0.04	0.23	-0.10	1.7
Phenylalanine	F	Phe	2.8	0.06	1.13	-0.32	2.4
Proline	P	Pro	-1.6	0.00	-0.45	2.23	1.0
Serine	S	Ser	-0.8	-0.21	-0.13	0.84	-0.1
Threonine	T	Thr	-0.7	-0.15	-0.14	0.52	0.4
Tryptophan	W	Trp	-0.9	0.03	1.85	0.30	3.1
Tyrosine	Y	Tyr	-1.3	-0.03	0.94	0.68	1.3
Valine	V	Val	4.2	0.13	-0.07	-0.31	1.7

Though charges on side chain do not seem important, it is in fact a characteristic which differentiate the side chains from one another . Based on our literature studies, it is interesting to note that besides the size of the side chain, it is the charge carried by each atom in the side chain which

physically differentiate the side chains from each other. In fact, there are various studies that investigate into the empirical partial charge values of different protein atoms^{(27) (28) (29)}.

The bonds between different atoms in a protein vibrate at very high frequencies. Various studies have been carried out that analyze the normal modes of these high frequency vibrations^{(30) (31) (32) (33) (34) (35) (36)}. The structure of a protein can be identified from these high frequency vibrations through NMR techniques⁽³⁷⁾. Furthermore, extensive research has also been carried out to understand how a specific native structure can arise from a given sequence of amino acids. These native structures, which determine the functions of proteins, are known to exhibit a certain range of slow vibrational modes⁽³⁸⁾.

The slow vibration modes of protein play an important role in the catalytic function of protein. Studies on these slow modes have been performed both experimentally and theoretically⁽³⁹⁾. Among these, the normal-mode analysis (NMA) is a technique that explores the flexibility and the slow range of motion of diverse protein configurations^{(40) (41) (42) (43) (44) (45)}. In the classical NMA approach, the degrees of freedom are described by a set of generalized coordinates, with the force field being given semi-empirically, consisting of energy terms due to stretching, bending, torsion, as well as van der Waals and electrostatic interactions⁽⁴¹⁾. Although classical NMA has reduced the computational expense on the investigation of low-frequency collective modes of proteins in comparison to molecular dynamics, it is still limited to small proteins. This has motivated the formulation of coarse-grained models such as the Go model^{(46) (47)}, the elastic network model^{(48) (49) (50)}, and the Gaussian network model^{(51) (52) (53)}, which leads to greater computational efficiency. Interestingly, the low-frequency normal modes determined from these simplistic models correspond closely not only to those obtained from more sophisticated force fields, but they also correspond to the conformational transitions observed experimentally^{(54) (55)}. In particular, the application of NMA to these simplistic models has led to theoretical studies on model refinement of crystallographic or diffraction data^{(56) (57)}, cooperative and hinge-bending motion in enzyme⁽⁵⁸⁾, mechanism of allosteric communication⁽⁵⁹⁾, NMR order parameters^{(60) (61)}, and conformational changes in large complexes such as viruses^{(62) (63)}. Remarkably, the results from these studies are found to correlate well with those obtained experimentally.

It is known that the very low-frequency modes of a protein can be associated with those of secondary structures, namely α -helices and β -sheets. A specific sequence folds into a specific structure⁽¹⁾. However, there have been reported cases of protein misfolding, in which a transition

causes the protein to fold into a different structure, resulting in a loss of normal biological functions^{(64) (65) (66)}.

Protein misfolding is known to be a source of debilitating diseases, which results from an aggregation of misfolded proteins, such as prion protein found in mad cow disease⁽⁶⁴⁾, bovine spongiform encephalopathy (BSE)⁽⁶⁵⁾, Creutzfeldt-Jakob disease (CJD)^{(67) (68)}, beta amyloid found in Alzheimer's disease⁽⁶⁹⁾. It is of particular interest to study possible transitions between these structures. A number of theories to study the transition of protein structure have been presented. Among them are Zimm-Bragg⁽⁷⁰⁾, Lifson-Roig⁽⁷¹⁾, Yakubovich *et. al.*^{(72) (73) (74)}, Ding *et. al.*⁽⁷⁵⁾, Hong-Lei⁽⁷⁶⁾, Yasar-Demir⁽⁷⁷⁾, Gibbs-DiMarzio⁽⁷⁸⁾, Peng *et. al.*⁽⁷⁹⁾ etc.

Current research suspects that an α - β transition occurs in prion when there is a change in the chemical properties of the solvent containing the protein. In experiments, addition of urea alone could unfold the recombinant human prion protein (huPrP), while an addition of both urea and NaCl induced a transition into the β -sheet⁽⁸⁰⁾. On the other hand, cellular prion protein (PrP^C) is also found to convert into fibrillar state in the presence of low concentrations of sodium dodecyl sulfate and NaCl, and the conversion is dependent on the concentrations of the solution⁽⁸¹⁾.

Besides prion, the amino acid sequence EKAYLRT is found to exist both in the α -helix and β -sheet form⁽⁸²⁾. In a Monte Carlo folding simulation of EKAYLRT, α -helix is found to occur when there is no interaction between EKAYLRT with other molecules, while in the vicinity of a β -strand the peptide forms a β -sheet⁽⁸³⁾. The EKAYLRT peptide could serve as a simple model to study the transition behavior in protein, as reported by Daidone *et. al.*, since proteins could exhibit similar dynamical behavior in a α - β transition. This was shown in molecular dynamics simulation of H1 peptide from prion protein and the A β (12-28) fragment from the A β (1-42) peptide, which are responsible for the Alzheimer's disease⁽⁸⁴⁾.

The many-body (multiple polypeptide chains, different residues and many atoms) nature of proteins gives rise to a complex dynamical system (from as simple as primary structures to multi-domain quaternary structures) with many areas left to be explored. In fact, these areas include the protein folding problem, protein normal mode analysis, protein misfolding, and different computational methodologies to tackle these problems. Our work in this area was initiated by Dr. Kerson Huang, Professor of Physics Emeritus from Massachusetts Institute of Technology (MIT). Prof. Huang had introduced an alternative approach to study the protein folding problem, known as the Conditioned

Self Avoiding Walk (CSAW)^{(85) (86)}. His objective is to formulate the protein folding problem into the most fundamental form to uncover the underlying physics of protein folding. His work was further developed by Dr. Weitao Sun and Dr. Jinzhi Lei^{(76) (87)}. Our research presented in this thesis originates from the same theoretical framework of the CSAW model. It is interesting to note that a change in the shape of the protein native structure would severely affect the functionality of a protein of the same amino acid sequence. We are interested to explore the mechanism and factors that contribute to such structural transition in protein.

In this thesis, we present a conceptual model to address the phase transition phenomenon in protein secondary structures. In our model, we have defined the basic unit of a polypeptide chain as a "crank" instead of a residue^{(86) (85)}. This leads to viewing the protein from a different geometrical point of view. Our model emphasizes on the importance of hydrogen bonding network in forming the protein secondary structure. Furthermore, we have studied the statistical properties of both α -helix and β -sheet as a function of the number of hydrogen bonds.

Protein is a complex system consisting of a large number of atoms with a large number of degrees of freedom. From the point of view of the mechanics of protein folding, the independent variables are the dihedral angles between successive amino acid residues, which changes the shape of a protein. It is important to note that the degrees of freedom associated with the fast vibrating modes do not affect the general shape of the protein, and thus do not affect its function⁽⁸⁸⁾. Hence, we can ignore these fast degrees of freedom and reduce the number of variables of our model by considering dihedral angles as the only degrees of freedom.

In our study, we shall first derive the canonical formalism of the protein chain. Instead of using the cartesian coordinates to describe the configuration of the protein molecules, we use the set of the dihedral angles (ϕ and ψ) as generalized coordinates. Based on the Ramachandran plot⁽⁸⁹⁾, we have selected only five distinct sets of dihedral angles that a residue can adopt, such that given a chain with n residues, the combination of these sets of angles can fold into a total of 5^n conformations, among them include α -helix, β -sheet, β -turns and coil. We assume that hydrogen bonding is the dominant non-covalent interaction that stabilizes the formation of protein secondary structure. Such assumption simplifies the otherwise more complicated analysis that is required on the more detailed physical model involving many energy terms such as the stretching, bending, torsion, as well as the van der Waals and electrostatic interactions. Such simplifications have allowed us to derive the grand partition function of a protein chain in an analytical way.

We have plotted the relationship of α - β -coil structural transition against temperature change based on our analysis. We have investigated the phase transition of protein secondary structures and the heat capacity change as a function of temperature during these phase transitions. Although our model is much simplified compared to the sophisticated models, it has unexpectedly reveal certain fundamental properties of protein via the approach of statistical physics.

This thesis is organized as follow: In Chapter 2, we touch on the origin and milestones of our work, together with a detailed definition of the polypeptide model used in our computation. In Chapter 3, we define the force field, i.e. the hydrogen bonding network, and derive the formulae for the computation of normal modes and partition function. In Chapter 4, we extend our computation to plot the structural phase transition of protein secondary structure as a function of temperature based on the theory derived in Chapter 3. The findings presented in Chapter 3 and Chapter 4 have been reviewed and published in Physical Review E (Vol. 82, 011915 and Vol. 86, 031902). Finally in Chapter 5, we conclude our works with discussion of potential future research direction.

Chapter 2: The Modeling Concept of Polypeptide Cranks

2.1 Origin: Conditioned Self-Avoiding Walk (CSAW) ⁽⁸⁵⁾ ⁽⁸⁶⁾

Our works in this thesis are inspired by CSAW model developed by Prof. Kerson Huang. The conformation of the backbone of the protein in CSAW is completely specified by a set of dihedral angles $\{\phi_1, \psi_1; \phi_2, \psi_2; \dots\}$. CSAW only consider the dihedral angles as degrees of freedom, ignoring the small high frequency vibrations within the cranks.

CSAW is a modeling of protein folding simulation based on Monte Carlo and self-avoiding walk (SAW) algorithms. It is able to simulate the Brownian motion of a polypeptide model. The design of the model enables additional force fields to be inserted in a modular fashion. Such feature allows CSAW to be used as a theoretical laboratory to study the mechanism of protein folding and structure prediction. The first published CSAW method is proposed based on Brownian motion. The initial flat chain is folded through SAW algorithm. The SAW algorithm prohibit the molecules of the cranks from having overlapping. This very crude model is then enhanced by adding other important non-covalent interaction such as the hydrogen bonding and hydrophobic effect. The inclusion of these additional conditions are implemented via the Monte Carlo algorithm. Thus, the CSAW model is born. The four components of CSAW are: Brownian motion, Monte Carlo, SAW and the incorporated conditions.

In Brownian motion, a particle's position denoted by $x(t)$ is a stochastic function. It can be defined through Langevin equation:

$$m\ddot{x} = F(t) - \gamma\dot{x}$$

Equation 2.1

The resultant accelerations acted on the particle are given by a damping force $-\gamma\dot{x}$ and a random component $F(t)$. The statistical properties of $F(t)$ are as follows:

$$\begin{aligned}\langle F(t) \rangle &= 0 \\ \langle F(t)F(t') \rangle &= c_0\delta(t - t')\end{aligned}$$

Equation 2.2

The equation can be evaluated showing the presence of diffusion, with diffusion coefficient $D = c_0/(2\gamma^2)$.

By adding non-random external force $G(x)$, the Langevin equation may not be analytically solved, but it can be numerically evaluated via conditioned random walk computation. A trial step is generated at random, and the probability of acceptance is treated via Monte Carlo method. Using Monte Carlo method, the potential energy of the external force and the change of energy in the next Monte Carlo iteration, are denoted as E and ΔE , respectively. The Monte Carlo procedure is as follows:

- when ΔE is less than or equal to zero, the iteration step is accepted;
- if ΔE is greater than zero, the iteration step is accepted with a probability of $\exp(-\Delta E/k_B T)$.

The thermal fluctuation of the ensemble is simulated by the last condition above. After a sufficiently large number of iterations, generated conformation of the polypeptide chain would yield a Maxwell-Boltzmann distribution with potential energy E . In terms of Langevin equation, the conditioned random walk can be represented as follows:

$$m\ddot{x} = [F(t) - \gamma\dot{x}] + G(x)$$

Equation 2.3

The first term on the right, $[F(t) - \gamma\dot{x}]$ is simulated by self-avoiding random walk algorithm and the second term is simulated by Metropolis Monte Carlo algorithm.

A pivot algorithm is used to simulate SAW. At first, an initial chain configuration is chosen, and fixing one of the end. The algorithm follows a looping procedure:

- (1) An arbitrary pivot point is chosen in the chain. In CSAW, a C_α atom is randomly chosen as the pivot point.
- (2) The free end of the chain is rotated about the chosen pivot. This is performed by randomly rotating the dihedral angles about the pivot point.
- (3) The new rotated configuration is accepted as an update, if the rotation does not lead to any overlap with other atoms, otherwise the configuration is rejected. Overlapping of atoms are taken care by avoiding the distance between the centres of two atomic spheres to be larger than the summation of the atomic spheres' radii.

An uniform ensemble of SAW is generated by looping multiple times the above procedure. Such simulation forms a generalized Langevin equation :

$$m_k \ddot{x}_k = F_k(t) - \gamma_k \dot{x}_k + G_k(x_1, \dots, x_N), \quad k = 1, \dots, N.$$

Equation 2.4

$k = 1, \dots, N$ labels the crank unit of the chain; G_k represents the bonding forces that hold the successive cranks, which also prevent overlapping of atoms.

Combining Brownian motion, Monte Carlo and SAW, other non-covalent interactions such as hydrophobic effect, hydrogen bonding and electrostatic force can be included into the equation by introducing a force term, U_k :

$$m_k \ddot{x}_k = (F_k - \gamma_k \dot{x}_k + G_k) + U_k, \quad k = 1, \dots, N$$

Equation 2.5

CSAW is constructed based on a simplified model with constant bond lengths and bond angles, having dihedral angles as the only degrees of freedom. Its initial stage consists of only back bone atoms with no side chain. Such simple model has successfully folded α -helix, with just hydrophobic effect and hydrogen bonding interaction. The energy of CSAW model is given as:

$$E = -g_1 K_1 - g_2 K_2$$

$K_1 = \text{number of contacts of hydrophobic cranks}$
 $K_2 = \text{number of hydrogen bonds formed}$

Equation 2.6

The number of contacts of a crank is the number of nearby cranks in contact with it, excluding the neighboring cranks (next to the particular crank). Two cranks are considered to be in contact if their centre C_α atoms fall within a distance. Hydrogen bonds are considered to be formed when $N-H$ and $C=O$ groups of different cranks form an interaction, following a set of criteria defined by adjacent angles along $N - H \cdots O = C$ ($\angle NHO$ and $\angle HOC$), and the atomic distance between H and O atoms.

In an attempt to improve the CSAW model, the addition of electrostatic potential and hard-sphere side chain has successfully folded an anti-parallel β -sheet (unpublished works). In this case, the energy term becomes:

$$E = -g_1K_1 - g_2K_2 - g_3K_3$$

$$K_3 = \text{Electrostatic energy} = k_e \frac{q_1 \cdot q_2}{r}$$

Equation 2.7

g_1 , g_2 and g_3 are the weightings of each term. A higher weighting of g_2 tends to fold into α -helices; a higher weighting of g_3 tends to fold β -sheets.

Though CSAW model is yet to be a fully developed model to study the folding of native protein structure from its amino acids sequence, the idea of CSAW provides an insight on the folding mechanism of protein secondary structure from a different angle. W. Sun improved the model by including all side chain atoms to the backbone. He has successfully folded protein 3AIT⁽⁹⁰⁾ to a small RMSD value⁽⁸⁷⁾.

2.2 The Polypeptide Cranks Model

As mentioned in preceding chapter, the protein chain consists of a sequence of standard amino acids chosen from a pool of 20. Each of these amino acids centers at a carbon atom called C_α and differ from one another only in the side chains bonded to C_α . Atoms in a protein chain are bonded together by covalent bonds. These bonds possess high vibration frequencies which could be analyzed via Nuclear Magnetic Resonance (NMR)⁽³⁷⁾ or Infrared Spectroscopy (IR) techniques⁽³³⁾. These techniques are useful to extract information on the structure and dynamics of proteins^{(91) (92) (93) (94) (95) (96) (97) (98) (99)}. For example, the identification of amide bands arising from stretching of C=O covalent bonds, can be used to determine the type of secondary structure (Table 2.1).

Table 2.1: Assignment of amide I band positions to secondary structure based on experimental data⁽¹⁰⁰⁾.

Secondary Structure	Average amide I band position in $^1\text{H}_2\text{O}/\text{cm}^{-1}$	Average amide I band position in $^2\text{H}_2\text{O}/\text{cm}^{-1}$
α -helix	1654	1652
β -sheet	1633	1630
β -sheet	1684	1679
Turns	1672	1671
Disordered	1654	1645

Figure 2.1 shows the standard geometry of a polypeptide chain that is used in our computation. The main constituents in our chain model are the backbone atoms, namely carbon (C_α , C), nitrogen (N), hydrogen (H) and oxygen (O). Due to the strong bonding nature of covalent bonds, the molecular structure of protein chain is rather rigid. There is negligible change in the dimension of bond length and bond angle. The structural change of a protein chain is mainly due to the change in dihedral

angles. Considering that all bond lengths and bond angles between molecules in a protein to be constant, except the dihedral angles (ϕ , ψ), we redefine the basic unit of a protein chain as crank, instead of looking at the basic unit of protein as residue. When the amino acids are joined into a chain, they become interlocked "residues". From a dynamical point of view, the independent units of the chain are "cranks" made up of coplanar chemical bonds, which connect one C_α to the next, as shown in Figure 2.2. Looking at Figure 2.1, from the perspective of residue unit, the first residue on the left involving position vector R_0 is not a complete residue, same happens for the last residue on the right involving position vector $R_{\alpha 3}$; from the perspective of crank unit, it is a complete crank unit. Thus we will ignore the incomplete portion of the residue in our computation.

Figure 2.1: A chain of polypeptide cranks. Two cranks are interconnected by the C_α .

Figure 2.2: Upper panel shows the crank that connects the center of one residue to the next. The vectors a , b , c , d , and e represent covalent bonds, which all lie in the same plane. (Side chains have been omitted for clarity.) Lower panel shows how the cranks are bonded to form the backbone of the protein. The vector position of C_α is denoted by $R_{C\alpha}$. S denotes the side chain molecule being modeled as a hard sphere. The bond length of the hard-sphere side chain can be obtained from ref. (101). The angle between cranks is fixed; the relative orientation of successive cranks is specified by two dihedral angles ϕ and ψ . The conformation of the backbone chain is completely specified by a set of dihedral angles. Data for bond lengths and angles are given in Table 2.2.⁽¹⁰²⁾

Table 2.2: Bond lengths and bond angles. Data are obtained from Protein Data Bank website⁽¹⁰²⁾.

	Bond length (Å)		Bond angle (deg)
$C_\alpha-C$	1.525	$\angle C_\alpha CN$	116.2
$C-N$	1.329	$\angle CNC_\alpha$	121.7
$N-C_\alpha$	1.458	$\angle NC_\alpha C$	109.5
$C=O$	1.231	$\angle C_\alpha CO$	120.8
$N-H$	1.000	$\angle C_\alpha NH$	114.0

Shown in Figure 2.2, the $C_\alpha-N-C-C_\alpha$ vectors plane is non-rotatable. The structure of our polypeptide chain is governed by the ϕ (rotation about $C-N$ bond axis) and ψ (rotation about $N-C_\alpha$ bond axis) dihedral angles as labeled in Figure 2.1 and Figure 2.2. We label the atoms by R_{Ci} , R_{Ci} , R_{Ni} , R_{Oi} and R_{Hi} (to represent the position vector of C_α , C , N , O and H atom of the i th crank). The coordinates of all the backbone atoms could be located by referring to vector $\vec{a}$, $\vec{b}$, $\vec{c}$, $\vec{d}$ and $\vec{e}$, which represent the bond vectors of $C_\alpha-C$, $C-N$, $N-C_\alpha$, $C=O$ and $N-H$ respectively with reference to an origin labeled as R_0 . Hence, two successive C_α atoms is defined by a vector of $\vec{\rho} = \vec{a} + \vec{b} + \vec{c}$. We take into consideration only the vector position of C_α as a function of ϕ and ψ dihedral angles in our computation. The position vector for each of the C_α atoms, labeled R_1 , R_2 , R_3 and so on is defined with reference to R_0 . The position vector of the origin R_0 could be any arbitrary point. Figure 2.1 shows the indexing of vectors in our protein model.

The backbone of protein chain is thus a sequence of cranks. The angle between two adjacent cranks is fixed at the tetrahedral angle $\cos^{-1}(1/3) \approx 109.5^\circ$ ⁽¹⁰²⁾. Thus the orientation of one crank with respect to its predecessor is specified by a set of dihedral angles $\{\phi, \psi\}$, as illustrated in Figure 2.2.

We write a standard procedure to define each of the crank vectors. We first initiate five bond vectors that define the first crank ($i=1$, on the leftmost end of Figure 2.1, a.k.a. standard crank), $\vec{a}_1$, $\vec{b}_1$, $\vec{c}_1$, $\vec{d}_1$ and $\vec{e}_1$. Hence,

$$\begin{aligned}
 \vec{R}_{\alpha 1} &= \vec{R}_0 + \vec{a}_1 + \vec{b}_1 + \vec{c}_1 \\
 \vec{R}_{C1} &= \vec{R}_0 + \vec{a}_1 \\
 \vec{R}_{N1} &= \vec{R}_0 + \vec{a}_1 + \vec{b}_1 \\
 \vec{R}_{O1} &= \vec{R}_0 + \vec{a}_1 + \vec{d}_1 \\
 \vec{R}_{H1} &= \vec{R}_0 + \vec{a}_1 + \vec{b}_1 + \vec{e}_1
 \end{aligned}$$

Equation 2.8

For the subsequent cranks ($i \geq 2$), each of the successive crank vectors as well as the position vector of each atom are labeled with reference to its preceding crank, i.e.

$$\begin{aligned}
\vec{a}_i &= \text{Rot}(\vec{c}_{i-1}, \phi_{i-1}) \cdot \text{Rot}(\vec{b}_{i-1} \times \vec{a}_{i-1}, 61.5^\circ) \cdot \text{Rot}(\vec{a}_{i-1}, \psi_{i-1} - \pi) \cdot \vec{a}_{i-1} \\
\vec{b}_i &= \text{Rot}(\vec{c}_{i-1}, \phi_{i-1}) \cdot \text{Rot}(\vec{b}_{i-1} \times \vec{a}_{i-1}, 61.5^\circ) \cdot \text{Rot}(\vec{a}_{i-1}, \psi_{i-1} - \pi) \cdot \vec{b}_{i-1} \\
\vec{c}_i &= \text{Rot}(\vec{c}_{i-1}, \phi_{i-1}) \cdot \text{Rot}(\vec{b}_{i-1} \times \vec{a}_{i-1}, 61.5^\circ) \cdot \text{Rot}(\vec{a}_{i-1}, \psi_{i-1} - \pi) \cdot \vec{c}_{i-1} \\
\vec{d}_i &= \text{Rot}(\vec{c}_{i-1}, \phi_{i-1}) \cdot \text{Rot}(\vec{b}_{i-1} \times \vec{a}_{i-1}, 61.5^\circ) \cdot \text{Rot}(\vec{a}_{i-1}, \psi_{i-1} - \pi) \cdot \vec{d}_{i-1} \\
\vec{e}_i &= \text{Rot}(\vec{c}_{i-1}, \phi_{i-1}) \cdot \text{Rot}(\vec{b}_{i-1} \times \vec{a}_{i-1}, 61.5^\circ) \cdot \text{Rot}(\vec{a}_{i-1}, \psi_{i-1} - \pi) \cdot \vec{e}_{i-1}
\end{aligned}$$

Equation 2.9

And

$$\begin{aligned}
\vec{R}_{\alpha i} &= \vec{R}_{\alpha(i-1)} + \vec{a}_i + \vec{b}_i + \vec{c}_i = R_0 + \sum_j^i (\vec{a}_j + \vec{b}_j + \vec{c}_j) \\
\vec{R}_{Ci} &= \vec{R}_{\alpha(i-1)} + \vec{a}_i = R_0 + \sum_j^{i-1} (\vec{a}_j + \vec{b}_j + \vec{c}_j) + \vec{a}_i \\
\vec{R}_{Ni} &= \vec{R}_{\alpha(i-1)} + \vec{a}_i + \vec{b}_i = R_0 + \sum_j^{i-1} (\vec{a}_j + \vec{b}_j + \vec{c}_j) + \vec{a}_i + \vec{b}_i \\
\vec{R}_{Oi} &= \vec{R}_{\alpha(i-1)} + \vec{a}_i + \vec{d}_i = R_0 + \sum_j^{i-1} (\vec{a}_j + \vec{b}_j + \vec{c}_j) + \vec{a}_i + \vec{d}_i \\
\vec{R}_{Hi} &= \vec{R}_{\alpha(i-1)} + \vec{a}_i + \vec{b}_i + \vec{e}_i = R_0 + \sum_j^{i-1} (\vec{a}_j + \vec{b}_j + \vec{c}_j) + \vec{a}_i + \vec{b}_i + \vec{e}_i
\end{aligned}$$

Equation 2.10

where $\text{Rot}(\vec{v}, \theta)$ is a rotation matrix about the axis of vector $\vec{v}$ for angle θ (see Appendix A).

From the equations shown, the $\vec{a}$, $\vec{b}$, $\vec{c}$, $\vec{d}$ and $\vec{e}$ crank vectors of i th crank are function of (ϕ_i, ψ_i) with reference to the crank vectors of the previous $(i - 1)$ th crank. Let k_i be $\{\vec{a}_i, \vec{b}_i, \vec{c}_i, \vec{d}_i, \vec{e}_i\}$, we can define

$$k_i = F(\phi_{i-1}, \psi_{i-1})k_{i-1}$$

Equation 2.11

where $F(\phi_{i-1}, \psi_{i-1})$ is a combined function (Equation 2.8, Equation 2.9 and Equation 2.10) that acts on vectors k_{i-1} to define vectors k_i . Referring to Equation 2.10, in a nutshell, each of the position vectors of C_ω , C , N , O and H is a function of all pairs of ϕ and ψ dihedral angles. Using the above mathematical formulation, we could model the structure of any desired polypeptide, such as the α -helix chain and β -sheet strand by inputting the corresponding ϕ and ψ dihedral angles.

2.3 The Theory of Our Polypeptide Mathematical Model

Our research is a continuation from CSAW research. Instead of working towards folding an amino acid sequence to its native structure, we study the statistical properties of protein secondary structure phase transition. The common feature of secondary structure is the presence of hydrogen bonds. Though the formation of different secondary structures could be initiated by different factors, hydrogen bonds play an important role in maintaining the stability of a secondary structure. Different types of secondary structures possess different hydrogen bonding network, which leads to different stability for different secondary structures. Thus, in specific, we focus our research on the statistical properties of hydrogen bonding in protein secondary structures transition.

Statistical physics involves computation of partition functions. In protein science, computing the partition functions of a full-atom protein chain is an extremely tedious task due to its massive number of possible conformations. In other words, the greater the number of cranks, the larger the partition functions. Such large scale computation cannot be performed by hand, and could only be solved using computer. Nevertheless, it is still constrained by limited computation power available in terms of processing speed and memory to compute the partition functions of a full-atom protein chain folding to all regions of permissible dihedral angles. To overcome these constraints, we reduce the complexity of our model as described in the followings.

In real facts, different side chain molecules change the statistical properties of a protein chain. However, addition of complex side chains would increase the number of independent variables in the model, which would further complicates the theory. As a starter, we first look at a simple model by studying hard sphere side chain though smaller side chain such as alanine can be easily represented by full atom model.

Furthermore, the folding of a protein largely depends on the amino acid sequence. Each amino acid sequence carries a unique characteristic, where the combination of these unique characteristics determines the native structure of a protein. The variation of these amino acid characteristics add up the complication of protein folding. Hence, in order to reduce the complication involved, we formulate our mathematical model based on homogenous polypeptide. To be specific, we have selected the protein polyalanine, due to the smaller size of its side chain. Omitting various independent variables would highlight the role of hydrogen bonding in the formation of protein secondary structure and its transition.

In a generic protein folding model, an amino acid residue doesn't fold to all dihedral angles region from -180° to 180° . Such constraint occurs due to steric effects between residues. The Ramachandran⁽⁸⁹⁾ plot displays the distribution of backbone dihedral angles ψ against ϕ of amino acid residues in protein structure (Figure 2.3). Through this plot, we could identify the allowed ϕ and ψ conformational regions that residues are allowed to fold into. Ramachandran plot also provides the analysis of different secondary structure regions, which helps us to identify the dihedral angles region that fold to a α -helix or β -sheet, and other types of secondary structures.

(a)⁽¹⁰³⁾

(b)⁽¹⁰⁴⁾

Figure 2.3: (a) Ramachandran plot for the general case. (b) Ramachandran plot for Glycine. (c) Ramachandran plot for Proline.

In our research, we consider only the sets of dihedral angles that matter the most. Those are the dihedral angles that fold into prominent secondary structures. The most abundant secondary structures of protein are the α -helix and β -sheet. Other secondary structures include 3_{10} -helix and π -helix. β -sheet itself could be sub-divided into parallel and anti-parallel sheet. β -turn is also an important secondary structure as it joins two β -strands in opposite direction together in most cases, which form an anti-parallel β -sheet. The random coil is not really a secondary structure, but it is a class of conformation with absence of regular secondary structure. Figure 2.4 shows different types of secondary structures.

(a)

(b)

(c)

Figure 2.4: (a) α -helix (b) Anti-parallel β -sheet (c) 3_{10} -helix. Figures are produced using Visual Molecular Dynamics (VMD)
(107)

We have selected only the two most abundant secondary structures that occur in protein as our subjects, i.e. α -helix, β -sheet, β -turn and random coil. In protein folding, as shown by the Ramachandran plot, certain dihedral angles lead to the formation of certain secondary structures. So, in our case, we compute only the partition functions covering these regions of dihedral angles. This would narrow down the scale of computation (still a large one though), but still capturing the important features. The mathematical derivation of such assumption will be further detailed in next chapter.

To further simplify our analysis, we ignore the solvent effects such as water molecules and assume the polypeptide model to be in gas phase. In our current work, we consider the covalent bonds to be

rigid and will not consider the high vibration frequencies such as those amide I frequencies due to the stretching of $C=O$ bonds. Since our main focus is on hydrogen bonding network, we will assume hydrogen bonding as the only non-covalent interactions with stretching and bending potentials which contribute to the slow vibration modes of secondary structures.

2.4 Scopes of Work

Based on the assumptions set herein, we first derive the normal modes of a simple polyalanine, folded in both α -helix, β -sheet and random coil. Using dihedral angles (ϕ , ψ) as the only degrees of freedom, we derive the Hamiltonian of a protein chain, consisting of kinetic energy, vibrational potential of hydrogen bond, and chemical potential of hydrogen bond. We then perform normal modes analysis (NMA) to compute the associated slow normal modes of the secondary structure as a whole. We find that all normal frequencies are positive definite, and this shows that the α -helix and β -sheet are mechanically stable under hydrogen bonding alone. This leads us to expect that, in the unfolded protein chain, which is subject to random bombardments from water molecules, these secondary structures should have transient existence. Our computed results agree with those obtained from classical NMA.

With verification that the physical assumptions that we stated above are sufficient to compute the normal modes of a protein secondary structure, we proceed to derive the partition function of protein, which is then used to analyze the transition of secondary structures as a function of temperature. We study the transition temperature from two approaches. Both approaches are derived based on the same physical assumptions stated above, but with different methodologies.

In our first approach, we calculate the partition functions of a pure α -helix and a pure β -sheet, which are treated as collections of harmonic oscillators with given frequency distributions. For comparison, we also calculate the partition function of a random coil. We compare the free energies of these conformations as functions of temperature. By identifying the one with least free energy as the equilibrium conformation, we find the critical temperatures at which phase transitions occur. With choices restricted to pure α -helix, pure β -sheet, or random coil, we find that the α -helix has the lowest free energy at low temperatures. From our plotting of free energy against temperature, we are able to identify the transition temperature from α -helix to β -sheet, and from β -sheet to random coil.

Our first approach includes α -helix, β -sheet and random coil, and the transition temperature was represented by Helmholtz Free Energy to temperature plotting. The partition function calculated for each secondary structure is independent from each other. In the context of partition function, this approach is not able to calculate the grand partition function of the whole ensemble of protein molecules. Out of the many different possible conformation that a protein could fold into (even using just five sets of dihedral angles per crank), the three structures (pure α -helix, β -sheet and random coil) covered in the first approach are just a few terms in the total summation of the grand partition function.

Hence, in our second approach, with different number of hydrogen bonds formed in different conformations, we calculate grand partition function of the protein chain by summing up partition functions of all possible conformations in an ensemble of protein molecules. Using the grand partition function, we compute the ensemble average properties of a protein chain and plot the temperature of structural transition from the ensemble average. In this thesis, we only study the ensemble average number of hydrogen bonds formation and heat capacity.

In both models, we have obtained interesting and intriguing findings. We make a comparison of our findings with other literature works, and conclude that our results are in good range of agreement.

Chapter 3: Hamiltonian Formalism and Normal Mode Analysis

3.1 Lagrangian and Hamiltonian Formulation of Polypeptide Secondary Structures

Consider a crank consisting of C_α, C, N, O, H atoms and S (hard-sphere side chain) molecules. We may define $\vec{R}_i$ as $\{\vec{R}_{\alpha i}, \vec{R}_{Ci}, \vec{R}_{Ni}, \vec{R}_{Oi}, \vec{R}_{Hi}\}$. As described in the previous section, for a polypeptide with n number of cranks, the position vector of each atom is a function of all pairs of dihedral angles, i.e.

$$\vec{R}_i = F\{(\phi_1, \psi_1), (\phi_2, \psi_2), (\phi_3, \psi_3), (\phi_4, \psi_4), \dots, (\phi_{n-1}, \psi_{n-1}), (\phi_n, \psi_n)\}$$

Equation 3.1

Ignoring the hard sphere side chain, for a chain of n cranks, this gives a total of $6n$ atoms. We re-label each atom by an index j . Let

$$R_j = \text{position vector of atom } j,$$

$$j = 1, \dots, 6n.$$

Equation 3.2

Let the set of dihedral angles be $\{\phi_i, \psi_i\}$ ($i = 1, \dots, n-1$) which makes up to a total of $2(n-1)$ degrees of freedom. Assuming the cranks to be perfectly rigid, it places constraints on the vector positions. The constraints are solved by using the dihedral angles as generalized coordinates, which we denote by the notation

$$q_{ik} \quad (i = 1, \dots, n-1; k = 1, 2).$$

Equation 3.3

For example,

$$q_{11} = \phi_1, q_{12} = \psi_1,$$

$$q_{21} = \phi_2, q_{22} = \psi_2, \text{ etc.}$$

Equation 3.4

Thus, we have

$$\{(\phi_1, \psi_1), (\phi_2, \psi_2), (\phi_3, \psi_3), \dots, (\phi_n, \psi_n)\} = \{q_{11}, q_{12}, q_{21}, q_{22}, q_{31}, q_{32}, \dots, q_{n1}, q_{n2}\}$$

Equation 3.5

We also use the notation q_α , where $\alpha = \{i, k\}$. We are to regard R_j as functions of $\{q_\alpha\}$.

The velocity of each atom is given by

$$\begin{aligned}\dot{R}_j &= \frac{dR_j}{dt} \\ &= \frac{\partial R_j}{\partial q_{11}} \cdot \frac{\partial q_{11}}{\partial t} + \frac{\partial R_j}{\partial q_{12}} \cdot \frac{\partial q_{12}}{\partial t} + \frac{\partial R_j}{\partial q_{21}} \cdot \frac{\partial q_{21}}{\partial t} + \frac{\partial R_j}{\partial q_{22}} \cdot \frac{\partial q_{22}}{\partial t} + \dots \\ &= \sum_{i=1}^{n-1} \sum_{k=1}^2 \frac{\partial R_j}{\partial q_{ik}} \dot{q}_{ik} = \sum_{\alpha} \frac{\partial R_j}{\partial q_{\alpha}} \dot{q}_{\alpha}\end{aligned}$$

Equation 3.6

The total kinetic energy is

$$\begin{aligned}K(q, \dot{q}) &= \frac{1}{2} \sum_{i=1}^n (m_C \dot{R}_{Ci}^2 + m_C \dot{R}_{Ci}^2 + m_N \dot{R}_{Ni}^2 + m_O \dot{R}_{Oi}^2 + m_H \dot{R}_{Hi}^2) \\ K(q, \dot{q}) &= \frac{1}{2} \sum_{\alpha, \beta} \dot{q}_{\alpha} \left(\sum_{j=1}^{6n} m_j \frac{\partial R_j}{\partial q_{\alpha}} \cdot \frac{\partial R_j}{\partial q_{\beta}} \right) \dot{q}_{\beta} \\ &= \frac{1}{2} \dot{q}^T M \dot{q}\end{aligned}$$

Equation 3.7

where we term M as the “mass matrix”, i.e.

$$M_{\alpha\beta} = \sum_{j=1}^{6n} m_j \frac{\partial R_j}{\partial q_{\alpha}} \cdot \frac{\partial R_j}{\partial q_{\beta}}$$

Equation 3.8

This is a symmetric matrix, with $M^T = M$. m_C , m_N , m_O and m_H are the atomic mass of C, N, O and H atoms respectively, as given in Table 3.1. Since we do not consider solvent effects, we consider the polypeptide chain model in a gas phase and do not include the mass of the solvent molecules. The mass of the side chain molecules can be added into the equation easily depending on the type of side chain involved. The size of the side chain and its respective mass would affect the numerical end results of the mass matrix computation.

Table 3.1: Standard atomic weight and atomic mass of polypeptide backbone atoms. ⁽¹⁰⁸⁾

Name	Symbol	Standard Atomic Weight	Atomic Mass (kg)
Carbon	C	12.01	1.9943×10^{-26}
Nitrogen	N	14.01	2.3264×10^{-26}
Oxygen	O	16.00	2.6569×10^{-26}
Hydrogen	H	1.008	1.6738×10^{-27}

The Lagrangian of the backbone chain is given by

$$L(q, \dot{q}) = K(q, \dot{q}) - U(q) = \frac{1}{2} \dot{q}^T M \dot{q} - U(q)$$

Equation 3.9

Hence, the canonical momentum is

$$p = \frac{\partial L}{\partial \dot{q}} = M \dot{q}$$

Equation 3.10

and the generalized force is

$$\frac{\partial L}{\partial q} = \frac{1}{2} \dot{q}^T \frac{\partial M}{\partial q} \dot{q} - \frac{\partial U}{\partial q}$$

Equation 3.11

Substituting Equation 3.10 and Equation 3.11 into the Lagrange equation of motion,

$$\begin{aligned} \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}} \right) &= \frac{\partial L}{\partial q} \\ \frac{d}{dt} (M \dot{q}) &= \frac{1}{2} \dot{q}^T \frac{\partial M}{\partial q} \dot{q} - \frac{\partial U}{\partial q} \\ \dot{M} \dot{q} + M \ddot{q} &= \frac{1}{2} \dot{q}^T \frac{\partial M}{\partial q} \dot{q} - \frac{\partial U}{\partial q} \\ \dot{q}^T \frac{\partial M}{\partial q} \dot{q} + M \ddot{q} &= \frac{1}{2} \dot{q}^T \frac{\partial M}{\partial q} \dot{q} - \frac{\partial U}{\partial q} \end{aligned}$$

Equation 3.12

This leads to

$$M \ddot{q} = -\frac{1}{2} \dot{q}^T \frac{\partial M}{\partial q} \dot{q} - \frac{\partial U}{\partial q}$$

Equation 3.13

On the other hand, we could obtain the same equation of motion via Hamiltonian formalism approach. The Hamiltonian is

$$H(p, q) = K(p, q) + U(q) = \frac{1}{2} \dot{q}^T M \dot{q} + U(q)$$

Equation 3.14

From Equation 3.10 we have $\dot{q} = M^{-1}p$ and $\dot{q}^T = p^T (M^{-1})^T$. Therefore,

$$H(p, q) = \frac{1}{2} p^T M^{-1} p + U(q)$$

Equation 3.15

The canonical equations of motion,

$$\dot{p} = -\frac{\partial H}{\partial q}, \dot{q} = \frac{\partial H}{\partial p}$$

Equation 3.16

takes the forms

$$\begin{aligned} \dot{p} &= -\frac{1}{2} p^T \frac{\partial M^{-1}}{\partial q} p - \frac{\partial U}{\partial q} \\ \dot{q} &= M^{-1} p \end{aligned}$$

Equation 3.17

Hence, from Equation 3.10, we get

$$\dot{p} = M\ddot{q} + \dot{M}\dot{q}$$

Equation 3.18

Substituting Equation 3.17 into Equation 3.18,

$$-\frac{1}{2} p^T \frac{\partial M^{-1}}{\partial q} p - \frac{\partial U}{\partial q} = M\ddot{q} + \dot{M}\dot{q}$$

Equation 3.19

Since

$$\begin{aligned} \frac{\partial M M^{-1}}{\partial q} &= 0 \\ \Rightarrow M \frac{\partial M^{-1}}{\partial q} &= -\frac{\partial M}{\partial q} M^{-1} \end{aligned}$$

Equation 3.20

Equation 3.19 becomes

$$\begin{aligned} -\frac{1}{2} \dot{q}^T M^T \frac{\partial M^{-1}}{\partial q} M \dot{q} - \frac{\partial U}{\partial q} &= M\ddot{q} + \dot{M}\dot{q} \\ \Rightarrow \frac{1}{2} \dot{q}^T \frac{\partial M}{\partial q} M^{-1} M \dot{q} - \frac{\partial U}{\partial q} &= M\ddot{q} + \dot{q}^T \frac{\partial M}{\partial q} \dot{q} \\ \Rightarrow M\ddot{q} &= -\frac{1}{2} \dot{q}^T \frac{\partial M}{\partial q} \dot{q} - \frac{\partial U}{\partial q} \end{aligned}$$

Equation 3.21

using $p^T = \dot{q}^T M^T$, $p = M\dot{q}$ and $M^T = M$, that M is a symmetric matrix. These are, of course, the same as the Lagrangian equation of motion as given by Equation 3.13.

3.2 Effective Potential of the Polypeptide Conformation

The partition function of the system is, up to a constant scale factor, given by

$$Z = \iint e^{-\beta[K(p,q)+U(q)]} d\vec{p} d\vec{q}$$

Equation 3.22

where $\beta = (k_B T)^{-1}$ is the inverse temperature. The p integration is Gaussian and can be immediately carried out, and the result generally depends on q (see Appendix B),

$$\begin{aligned} Z &= \int \left[\int e^{-\beta K(p,q)} d\vec{p} \right] e^{-\beta U(q)} d\vec{q} \\ &= \int \left[\int e^{-\beta(1/2)p^T M^{-1} p} dp^{2(n-1)} \right] e^{-\beta U(q)} d\vec{q} \\ &= \left(\frac{2\pi}{\beta} \right)^{n-1} \int \sqrt{\det M} e^{-\beta U(q)} d\vec{q} \end{aligned}$$

Equation 3.23

This gives rise to an effective potential $V_{\text{eff}}(q)$, which is defined through the relation

$$e^{-\beta V_{\text{eff}}(q)} \equiv \left(\frac{\beta}{2\pi} \right)^{n-1} \int e^{-\beta K(p,q)} dp^{2(n-1)}$$

Equation 3.24

Thus,

$$\begin{aligned} Z &= \int \rho_{\text{con}}(q) d\vec{q} \\ \rho_{\text{con}}(q) &\equiv \left(\frac{2\pi}{\beta} \right)^{n-1} e^{-\beta(U+V_{\text{eff}})} \end{aligned}$$

Equation 3.25

where ρ_{con} is the configurational probability density. That is, $\rho_{\text{con}} d\vec{q}$ is the relative probability of finding the system in $d\vec{q}$, regardless of the momentum $\vec{p}$. If the kinetic energy is independent of q , the effective potential is a constant.

In a canonical ensemble, the relative probability of finding the state in the element $d\vec{p}d\vec{q}$ in phase space is given by $d\vec{p}d\vec{q}(e^{-\beta H})$. If we are only interested in the probability of finding the state in $d\vec{q}$, we integrate the above over $\vec{p}$ and obtain

$$\begin{aligned} d\vec{q} \int d\vec{p} \exp(-\beta H) &= d\vec{q} \exp(-\beta U) \int d\vec{p} \exp(-\beta K) \\ &= d\vec{q} \left(\frac{2\pi}{\beta}\right)^{n-1} \exp[-\beta(U + V_{eff})] \end{aligned}$$

Equation 3.26

This is the probability used, for example, in the Metropolis Monte Carlo algorithm in the CSAW model.

From Equation 3.23 and Equation 3.24, we have

$$\beta V_{eff}(q) = -\frac{1}{2} \ln(\det M) = -\frac{1}{2} \text{Tr}(\ln M)$$

Equation 3.27

Since $V_{eff}(q)$ depends on all the dihedral angles, it is a function of the chain conformation. Our calculations show that it is sensibly constant for almost all different conformations. The percentage change of the effective potential for a three-crank model is less than 0.2%. hence, we shall assume that it is a constant in our subsequent derivation. Representative results are shown in Fig. 3.1. (See Appendix C)

Figure 3.1: Effective potential of a three-crank chain at different $\{\phi, \psi\}$ angles between the second and third cranks, in a three cranks polypeptide system. The percentage change of the effective potential is less than 0.2%.

3.3 Potential Energy of Hydrogen Bonding

In our model, the main stabilizing agents for protein secondary structures such as α -helix and β -sheet are hydrogen bonds, which exist between $N-H$ and $C=O$ groups from different residues⁽²⁶⁾⁽¹⁰⁹⁾. We assume that a hydrogen bond is formed when the distance between the H and O atoms is 2.0 ± 1.0 Å, and the bond angle between $N-H$ and $C=O$ is $180^\circ \pm 45^\circ$ ⁽²⁶⁾. For different conformation of the same length of polypeptide chain, different number of hydrogen bonds are formed.

The α -helix, also known as the 4_{13} helix, is the most abundant secondary structure due to its tight conformation⁽¹⁰⁹⁾. In this configuration, a hydrogen bond connects the $C=O$ group of the i th crank to the $N-H$ group of the $(i+3)$ th crank.

The β -sheet is a two-dimensional mat made up of backbone strands stitched together by hydrogen bonds⁽¹⁰⁹⁾. The participating strands may be parallel or anti-parallel. In this research, only the latter case will be studied.

We wish to study the normal modes of small vibrations about an equilibrium configuration. The potential energy U is assumed to be minimum and taken to be zero, at this configuration. The equilibrium is assumed to be maintained by hydrogen bonds. Deviations from equilibrium arise from the stretching and bending of these bonds.

Let b_i be the bond vector of the i th hydrogen bond, i.e. the vector between O and the bonded H , in the equilibrium position. Let b'_i be the same vector when the configuration is displaced from equilibrium. The displacement vector is given by (see Figure 3.2)

$$u_i = b'_i - b_i$$

Equation 3.28

Figure 3.2: Definition of deviated hydrogen bond from equilibrium position, i.e. displacement vector u_i .

For small displacements, we take the potential energy to be

$$U = \frac{1}{2} \kappa_1 \sum_i (|\hat{b}_i \cdot u_i|)^2 + \frac{1}{2} \kappa_2 \sum_i (|\hat{b}_i \times u_i|)^2$$

Equation 3.29

where $\hat{b}_i = \frac{b_i}{|b_i|}$, and κ_1 and κ_2 are the force constants associated with the stretching and bending of hydrogen bonds, respectively⁽¹¹⁰⁾,

$$\kappa_1 = 13 \text{ N/m}$$

$$\kappa_2 = 3 \text{ N/m}$$

Equation 3.30

Let the generalized coordinates be denoted by

$$q = q_0 + \lambda$$

Equation 3.31

where q_0 corresponds to equilibrium and λ represents a small deviation. We can write

$$\begin{aligned}
u_i &= \left[(b'_i)_0 + \sum_{\alpha} \left(\frac{\partial b'_i}{\partial q_{\alpha}} \right)_0 \lambda_{\alpha} + \mathcal{O}(\lambda^2) \right] - b_i \\
u_i &= b_i + \sum_{\alpha} \left(\frac{\partial b'_i}{\partial q_{\alpha}} \right)_0 \lambda_{\alpha} + \mathcal{O}(\lambda^2) - b_i \\
u_i &\approx \sum_{\alpha} \left(\frac{\partial b'_i}{\partial q_{\alpha}} \right)_0 \lambda_{\alpha}
\end{aligned}$$

Equation 3.32

where the subscript 0 indicates evaluation at equilibrium. Therefore, substituting Equation 3.32 into Equation 3.29, the potential energy term becomes

$$\begin{aligned}
U &= \frac{1}{2} \kappa_1 \sum_i \left\{ \left| \hat{b}_i \cdot \sum_{\alpha} \left(\frac{\partial b'_i}{\partial q_{\alpha}} \right)_0 \lambda_{\alpha} \right| \cdot \left| \hat{b}_i \cdot \sum_{\beta} \left(\frac{\partial b'_i}{\partial q_{\beta}} \right)_0 \lambda_{\beta} \right| \right\} \\
&\quad + \frac{1}{2} \kappa_2 \sum_i \left\{ \left| \hat{b}_i \times \sum_{\alpha} \left(\frac{\partial b'_i}{\partial q_{\alpha}} \right)_0 \lambda_{\alpha} \right| \cdot \left| \hat{b}_i \times \sum_{\beta} \left(\frac{\partial b'_i}{\partial q_{\beta}} \right)_0 \lambda_{\beta} \right| \right\} \\
&= \frac{1}{2} \kappa_1 \sum_{\alpha, \beta} \lambda_{\alpha} \sum_i \left\{ \left| \hat{b}_i \cdot \sum_{\alpha} \left(\frac{\partial b'_i}{\partial q_{\alpha}} \right)_0 \right| \cdot \left| \hat{b}_i \cdot \sum_{\beta} \left(\frac{\partial b'_i}{\partial q_{\beta}} \right)_0 \right| \right\} \lambda_{\beta} \\
&\quad + \frac{1}{2} \kappa_2 \sum_{\alpha, \beta} \lambda_{\alpha} \sum_i \left\{ \left| \hat{b}_i \times \sum_{\alpha} \left(\frac{\partial b'_i}{\partial q_{\alpha}} \right)_0 \right| \cdot \left| \hat{b}_i \times \sum_{\beta} \left(\frac{\partial b'_i}{\partial q_{\beta}} \right)_0 \right| \right\} \lambda_{\beta}
\end{aligned}$$

Equation 3.33

This leads to the quadratic form

$$\begin{aligned}
U &= \frac{1}{2} \lambda^T (\kappa_1 D + \kappa_2 C) \lambda \\
D_{\alpha\beta} &= \sum_i \left| \hat{b}_i \cdot \frac{\partial b'_i}{\partial q_{\alpha}} \right|_0 \cdot \left| \hat{b}_i \cdot \frac{\partial b'_i}{\partial q_{\beta}} \right|_0 \\
C_{\alpha\beta} &= \sum_i \left| \hat{b}_i \times \frac{\partial b'_i}{\partial q_{\alpha}} \right|_0 \cdot \left| \hat{b}_i \times \frac{\partial b'_i}{\partial q_{\beta}} \right|_0
\end{aligned}$$

Equation 3.34

3.4 Normal Modes

For small oscillations about equilibrium, the linear equation of motion is

$$M\ddot{\lambda} = -\frac{\partial U}{\partial q}$$

Equation 3.35

From Equation 3.34 we have

$$\frac{\partial U}{\partial q} = (\kappa_1 D + \kappa_2 C)\lambda$$

Equation 3.36

Thus,

$$M\ddot{\lambda} + (\kappa_1 D + \kappa_2 C)\lambda = 0$$

Equation 3.37

The solution to the second order differential Equation 3.37 is $\lambda = Ae^{-i\omega t}$, and its second time derivative is $\ddot{\lambda} = -\omega^2 Ae^{-i\omega t} = -\omega^2 \lambda$, where ω is the angular frequency in *rad/s*. Substituting λ and $\ddot{\lambda}$ into Equation 3.37, the following equation is obtained:

$$-\omega^2 M\lambda + (\kappa_1 D + \kappa_2 C)\lambda = 0$$

Equation 3.38

The normal frequencies ω and normal modes λ are eigenvalues and eigenvectors of the equation

$$M^{-1}(\kappa_1 D + \kappa_2 C)\lambda = \omega^2 \lambda$$

Equation 3.39

Note that our model's validity is the subject to the following conditions:

- (1) We treat small oscillation about a presumed equilibrium configuration q_0 . Whether q_0 indeed corresponds to equilibrium can be verified through the requirement that all normal frequencies are nonzero and positive.
- (2) We ignore electrostatic and other interactions. Our results can serve as a test on whether the structures investigated can maintain equilibrium purely through hydrogen bonding. Inclusion of other interactions will introduce corrections.
- (3) Actual α and β structures are embedded inside a protein molecule in solution and are subject to other forces not considered here, particularly those arising from Brownian motion in the

solution, the hydrophobic effect, and interaction with other atoms in the protein. These forces will give rise to corrections and may even destroy the stability of the structure.

In view of the limitations of the model, we only examine normal modes in a frequency range corresponding to wave numbers 10^{-1} – 10^3 cm^{-1} . This is because, in a real protein, the very low-frequency end will be dominated by binding effects to the rest of the protein, while the very high frequency region will be dominated by bond oscillations⁽³⁸⁾.

In a nutshell, we can view the vibration potential of hydrogen bonds in protein secondary structure equivalent to a spring system. Such representation will be discussed with more details in this thesis.

3.5 The α -Helix

In this study, we have selected polyalanine as our subject for a generic α -helix. Since the CH_3 side chain group of alanine residue is not a big molecule, it is easily expressed in terms of a fully atomic model instead of using the implicit hard-sphere model. Formally, it is straightforward to generalize this all-atom side chain model.

The dihedral ϕ , ψ of polyalanine⁽¹¹¹⁾ are given by

$$\{\phi, \psi\} = \{-57.4^\circ, -47.5^\circ\}$$

The equivalent spring system is illustrated in Figure 3.3. In this example, there are seven cranks, but only four hydrogen bonds. In general, for n cranks, the number of hydrogen bonds is $n-3$. The number of degrees of freedom from stretching and bending of the hydrogen bonds is thus $2(n-3)$. The total number of degrees of freedom of the system, however, is $2(n-1)$. Thus, we expect to have four zero modes, apart from rigid translations and rotations. Furthermore, from Figure 3.3, we see that the springs are isolated from each other which we know we would have zero modes which do not correspond to true normal modes. These will not be included in our results.

Figure 3.3: The mechanical system corresponding to small oscillations of the α -helix (solid lines). Springs are hydrogen bonds (dashed lines).

Figure 3.4 shows the distributions of normal modes as a function of wave number, for different crank numbers n . All calculated frequencies are positive. The distributions exhibit four peaks associated with various types of deformation, which can be ascertained by examining the corresponding eigenvectors. The results are listed in Table 3.2.

Figure 3.4: Normal frequency distributions for the α -helix, with different numbers of cranks n . Types of distortion corresponding to the peaks are listed in Table 3.2

Table 3.2: Normal modes of α -helix. The matching modes were an estimate observed from a self-written simulation program.

Frequency (cm^{-1})	Mode
0-10	Twisting
30-40	Stretching and bending
90-100	Bending
120-130	Bending

The significance of these four dominant frequencies is more pronounced at a higher number of cranks. We normalize the plot by rescaling the number of modes by the total number of modes for each length. The results are shown in Figure 3.5. In the figure, we notice that all curves have similar pattern, in which the peak points occur within the same range of wavenumber.

Figure 3.5: Normalized number of modes of generic α -helix at different lengths from 10-crank to 100-crank size.

3.6 The β -Sheet

We model a generic anti-parallel β -sheet⁽⁹³⁾ by setting the dihedral angles in each strand to

$$\{\phi, \psi\} = \{-139^\circ, 135^\circ\}$$

The connectivity of hydrogen bonds for the anti-parallel β -sheet is shown in Figure 3.6. An extra crank is included to join two adjacent strands in a turn. The dihedral angles to form the turn are:

$$\{\phi_1, \psi_1\} = \{315^\circ, 110^\circ\}$$

$$\{\phi_2, \psi_2\} = \{105^\circ, 330^\circ\}$$

Figure 3.6: Schematic diagram of anti-parallel β -sheet, illustrating the connectivity of hydrogen bonds.

Compared to the α -helix, the β -sheet of same length has fewer hydrogen bonds formed within the structure. Thus, we expect that in our model there will be more zero modes compared to the α -helix; but we ignore them for reasons stated previously. Otherwise, all calculated frequencies are positive.

Normal frequencies are computed for varying numbers of strands, and cranks per strand. We display representative distributions in Figure 3.7 for β -sheet with 5 cranks per strand. We see that the frequencies are concentrating on the infrared frequency region. This is consistent with calculations on real protein with β -sheet structure^{(39) (92) (97) (112)}. In general, the peak positions of the distributions depend only on the number of cranks per strand and are independent of the number of strands. The peaks tend to widen with increasing crank number. We put in the wavenumber with peak number of modes into a self-written simulation program and did not observe any visible mode of motion, unlike α -helix where we observed some significant mode of motion (See Table 3.2). We carry out a normalization by scaling the number of modes by the total number of modes for each β -sheet sample. The results are shown in Figure 3.8.

Figure 3.7: Normal frequency distributions of β -sheets, for the same number of cranks per strand, but different number of strands.

Figure 3.8: Normalized number of modes of β -sheet with different number of strands.

3.7 Comparison with Other Research Work

Our calculation of normal modes for both α -helix and β -sheet is in agreement with results calculated by ben-Avraham⁽³⁸⁾. ben-Avraham found that the density of states of globular proteins follows a characteristic universal curve (c.f. Fig. 1 of reference (38)). The number of modes for these two secondary structures converges to their respective characteristic curves that are independent of the number of cranks for the α -helix and the number of strands for the β -sheet. In addition, our results show that the main contribution to the normal modes of these secondary structures comes from physical effects due to the hydrogen bonds.

In order to further validate our results, we obtain the normal-mode distribution of myoglobin (1MBD)⁽¹¹³⁾ (see Figure 3.9), which is made up of eight α -helices. We found that our result given by Figure 3.9 is in good agreement with those obtained by Krimm and Reisdorf, Jr. (c.f. Fig. 7(b) of reference (111)), who derived their normal-mode distribution from a different approach. In their calculation using empirical force field, the only degree of freedoms are ϕ and ψ dihedral angles, while maintaining other covalent bonds at a much higher force field than hydrogen bonds.

Figure 3.9: Normal-frequency distribution for myoglobin, consisting of eight α -helices.

3.8 α - β Phase Transition

The transition between α -helix and β -sheet is an important subject in view of the existence of proteins with ambivalent structures⁽¹¹⁴⁾. From our results, we now proceed to making the following calculation. Based on Equation 3.15 and Equation 3.34, we can write the partition function as follows:

$$\begin{aligned}
 Z_i &= \int \exp\{-\beta[K_i(p) + U_i(q) + U_i^e]\} d\vec{p} d\vec{q} \\
 &= \int \exp\left\{-\beta\left[\frac{1}{2}p^T M_i^{-1}p + \frac{1}{2}q^T(\kappa_1 D_i + \kappa_2 C_i)q + U_i^e\right]\right\} d\vec{p} d\vec{q}
 \end{aligned}$$

Equation 3.40

with the index i being α , β , or c depending on whether the configuration is a α -helix, a β -sheet, or a random coil, respectively. Note that U_i^e is the total chemical potential of the hydrogen bonds formed within a particular polypeptide configuration. U_i^e can also be viewed as the energy minima of the potential well and, since it is independent of the dihedral angles under small displacement evaluated at equilibrium state, it does not affect the earlier normal mode analysis.

After evaluating the integral in Equation 3.40, the partition function takes the following form:

$$\begin{aligned}
 Z_i &= \frac{(2\pi k_B T)^{n-1}}{\sqrt{\det M_i^{-1}}} \frac{(2\pi k_B T)^{N_i}}{\sqrt{\det_p(\kappa_1 D_i + \kappa_2 C_i)}} (2\pi)^{2(n-1)-2N_i} e^{-\beta U_i^e} \\
 &= \frac{(2\pi k_B T)^{n-1}}{(\bar{\lambda}_i^M)^{2(n-1)}} \frac{(2\pi k_B T)^{N_i}}{(\bar{\lambda}_i^K)^{2N_i}} (2\pi)^{2(n-1)-2N_i} e^{-\beta U_i^e}
 \end{aligned}$$

Equation 3.41

where

$$\begin{aligned} (\bar{\lambda}_i^M)^{2(n-1)} &= \left(\omega_i^{M1} \omega_i^{M2} \omega_i^{M3} \dots \omega_i^{M2(n-1)} \right)^{1/2} \\ (\bar{\lambda}_i^K)^{2N_i} &= \left(\omega_i^{K1} \omega_i^{K2} \omega_i^{K3} \dots \omega_i^{K2N_i} \right)^{1/2} \end{aligned}$$

Equation 3.42

Note that ω_i^{Mj} and ω_i^{Kj} are the j th eigenvalues of the matrices M_i^{-1} and $(\kappa_1 D_i + \kappa_2 C_i)$, respectively. In Equation 3.41, n is the number of cranks, while N_i is the number of hydrogen bonds in the configuration. Since N_i is less than n , zero eigenvalues are to be expected in the potential matrix $(\kappa_1 D_i + \kappa_2 C_i)$. Hence, we define a "pseudo" determinant term " $\det_p()$ " which takes in only the non-zero eigenvalues.

Next, we calculate the Helmholtz free energy of a particular protein configuration from Equation 3.41,

$$F_i = -k_B T \ln Z_i = U_i^e - k_B T S_i$$

Equation 3.43

where

$$S_i = (n-1) \ln(2\pi k_B T) - 2(n-1) \ln \bar{\lambda}_i^K + N_i \ln(2\pi k_B T) - 2N_i \ln \bar{\lambda}_i^K + N_i^Z \ln(2\pi)$$

Equation 3.44

We observe that the zero modes now play a significant role in the determination of the free energy of the protein configuration. In other words, the zero modes are directly related to the entropy of the protein configuration with $N_i^Z = 2(n-1) - 2N_i$ zero modes giving rise to N_i^Z multiples of 2π entropy. Furthermore, the third and fourth terms on the right-hand side of Equation 3.44 give the competition between the entropic effects of the environment and the internal stretching and bending energy of the hydrogen bonds, respectively. These results imply that the smaller the number of hydrogen bonds in the protein configuration, the higher the entropy. In this particular comparison, we model a random coil as flat chain where no hydrogen bond is formed (i.e. $N_i = 0$), thus it has the highest entropy.

Figure 3.10 shows the plot of Helmholtz free energy of a 15-crank polyalanine in the α -helix, in the two-strand β -hairpin, and the random coil configuration as a function of temperature, based on

Equation 3.44. The α -helix has been chosen as a reference state in the figure. Hence the normalized plot is presented in the following way:

$$F_i = (U_i^e - U_\alpha^e) - k_B T(S_i - S_\alpha)$$

Equation 3.45

Note that the values of the quantities used in the plot are given in Table 3.3, with $U_i^e = -N_i \epsilon_{hb}$, where $\epsilon_{hb} = 5 \text{ kcal/mol}$ is the potential energy of a hydrogen bond⁽¹¹⁵⁾ and N_i is the number of hydrogen bond formed in the particular conformation; the values of $\bar{\lambda}_i^M$ and $\bar{\lambda}_i^K$ can be calculated from Equation 3.42. The plot shows that the random coil has the steepest curve, followed by the β -hairpin and then the α -helix. This is to be expected since without the constraint of hydrogen bonds, the random coil has the highest entropy. On the other hand, the structure of the α -helix is stabilized by 12 hydrogen bonds, six more than the β -hairpin. Hence, it has the least configurational freedom and the flattest curve of the three.

Figure 3.10: Normalized Free-energy curves (using α -helix as a reference state) of α -helix, β -sheet, and random coil, based on Equation 3.43, Equation 3.44 and Equation 3.45, with the variables substituted by data given in Table 3.3. The intersection of the curves indicates phase transition. The transition temperatures are found to be $T_{\alpha\beta} = 475\text{K}$, $T_{\alpha c} = 535\text{K}$ and $T_{\beta c} = 600\text{K}$.

Table 3.3: Quantities employed in the calculation of the free energy versus temperature curve for the α -helix, β -sheet, and random coil configurations.

i	$U_i^e \times 10^{-19}(\text{J})$	N_i	$\bar{\lambda}_i^M \times 10^{22}(\text{kg}^{-1/2}\text{m}^{-1})$	$\bar{\lambda}_i^K \times 10^{-10}(\text{kg}^{1/2}\text{ms}^{-1})$
α -helix	-4.17	12	1.165	5.878
β -hairpin	-2.09	6	1.190	6.960
Random coil	0.00	0	1.714	Not Applicable

The critical temperature at which protein phase transition occurs can be determined through the free-energy curves in Figure 3.10. For example, by examining the intersection between the free-energy curves of the α -helix and the random coil, the critical temperature of 535 K is observed for the α -coil transition. Similarly, the critical temperatures for the β -coil and α - β transitions are found

to occur at 600 K and 475 K, respectively. As a protein adopts its stable configuration by minimizing its free energy, we expect the 15-residue polyaniline to form an α -helix structure at low temperature. As temperature increases, entropy begins to gain importance against the dominance of the internal energy of the protein at low temperature. At the α - β transition temperature, the more entropic β -hairpin emerges as the more stable structure with a lower free energy. A further increase in temperature will eventually lead to the critical state of β -coil transition, upon which the β -hairpin denatures into a random coil.

Interestingly, our results correspond to those obtained by Ding *et al.* ⁽⁷⁵⁾ through molecular dynamics simulations. However, we notice that they have overestimated their degrees of freedom in their analytical estimates. Since the degree of freedom in their model is also based on the ϕ and ψ dihedral angles, the total degree of freedom is $2N$ for a pair of dihedral angles per residue, with N being the number of residues. Thus, a more reasonable entropy calculation should be $S_i = A \ln \langle rms d_r \rangle_i + S_0$ with $A=2N$ instead of $A=3N$. Note that $\langle rms d_r \rangle_i$ is the expected root-mean-square deviation for configuration i from a chosen reference structure. Remarkably, this correction leads to results which corroborate with ours in Figure 3.10 (see Figure 3.11), with α - β , α -coil, and β -coil transition temperatures of 427 K, 504 K, and 616 K, respectively, by employing the data given in ref. (75). Further support of our approach is given by computational results from Yasar and Demir for a hydrophobic homogeneous polypeptide chain with a helix-coil transition temperature of 550 K ⁽⁷⁷⁾. Our result is also in agreement with computational work performed by Lee *et al.* ⁽¹¹⁶⁾, whose analysis has found a helix-coil transition temperature of about 475 K for a 15-residue polyaniline (refer to Fig. 2(a) of ref (116)).

Figure 3.11: Modification of the free-energy curves of Ding *et al.* ⁽⁷⁵⁾ based on $S_i = 2N \ln \langle rms d_r \rangle_i + S_0$. The transition temperatures are found to be $T_{\alpha\beta} = 427$ K, $T_{\alpha c} = 504$ K, and $T_{\beta c} = 616$ K.

By assuming a simple hydrogen bond potential, we have considered the polyaniline to be in the gas phase in this assumption. However, if we were to include the effects of solvent, we would expect a lowering of the Helmholtz free energies as the polyaniline folds toward a new optimal conformation⁽⁷⁷⁾. Indeed, if the solvent is water, the new conformation is non- α -helical since polyaniline is a purely hydrophobic polypeptide.

Finally, it should be noted that the three curves shown in Figure 3.10 are actually independent of each other. Such plots give only a rough estimate of phase transition such that α -helix appears to have the lowest free energy at low temperature, thus a more stable structure, followed by β -sheet at mid range temperature and coil at extreme high temperature. If we were to include the free energy curves of other conformations besides the three shown into the figure, we would have multiple lines and intersection points. It is not apparent how each of these intersection points can constitute a phase transition since all these curves are calculated independently and are not mathematically related to each other. Hence, in the next chapter, we will present a theory which relates all the independent conformations into a single plot using one single equation.

Chapter 4: Phase Transition in α - β -Coil Protein Secondary Structure

4.1 Canonical Formalism of Protein Chain

In previous chapter, we have defined the Hamiltonian as:

$$H(p, q) = K(p, q) + U(q) + U^e$$

Equation 4.1

where $K(p, q)$ is the kinetic energy of the molecules, $U(q)$ is the effective potential of the hydrogen bonds due to vibrations, and U^e is the chemical potential of the hydrogen bonds in the protein chain.

The formation of a hydrogen bond carries a "chemical" potential. Furthermore, the stability of each secondary structure is determined by its configuration of hydrogen bonds. Assuming that the chemical potential of hydrogen bond associated with the α -helix and β -sheet have different statistical weights, we may write the of chemical potential as follow:

$$U^e = NU_h^e + N_\alpha U_\alpha^e + N_\beta U_\beta^e$$

Equation 4.2

where N gives the total number of hydrogen bonds, while N_α and N_β are the number of hydrogen bonds associated with the α -helix and β -sheet respectively. Note that U_h^e is the chemical potential of one hydrogen bond. U_α^e and U_β^e denote the additional statistical weights of the chemical potential of the hydrogen bond associated with the α -helix and β -sheet structures respectively. The introduction of U_α^e and U_β^e allows us to identify the number of hydrogen bonds that are associated to α -helix and β -sheet respectively. Without the different weightings, the hydrogen bonds may appear all the same for either structure. Since we only consider hydrogen bonding in our computation, there could be other factors (e.g. Van der Waals, electrostatic) which determine the most probable conformation that a polypeptide should fold into, thus the different weightings could in a way account for these effects implicitly. It should be noted that the formation of hydrogen bond can be associated to other type of secondary structure which is not discussed in the current context. Such type of hydrogen bond would then possess a chemical potential of U_h^e without additional weighting added to it. However, we would expect a different finding when we include the association of hydrogen bond to other type of secondary structure, besides α -helix and β -sheet.

In our coarse-grained model, we shall restrict the dihedral angles to five distinct sets of angles. The selection is made according to the Ramachandran plot, which indicates regions in terms of dihedral angles where stable secondary structures can form, namely α -helix, β -sheet, β -turn⁽⁸⁹⁾. Though random coil is not considered as a definite secondary structure, it is a class of conformation where secondary structure is absent. Hence the selected states or angles are: $\{-57.4^\circ, -47.5^\circ\}$, $\{-139^\circ, 135^\circ\}$, $\{315^\circ, 110^\circ\}$, $\{105^\circ, 330^\circ\}$, and $\{180^\circ, 180^\circ\}$. A combination of these angles leads to the formation of α -helix (Figure 2.4a), β -sheet and β -turn (Figure 2.4b), or any combination of these secondary structures. A schematic on the configuration of hydrogen bonds that give rise to α -helix and β -sheet is shown in Figure 3.3 and Figure 3.6 respectively. The random coil may also form hydrogen bond in its structure which may or may not be associated to α -helix and β -sheet.

4.2 Grand Partition Function of Protein Chain

If a protein has n cranks, the restriction of the sets of dihedral angles for each crank to five distinct states implies that the chain can form into a total of 5^n possible structures. Let each structure be i , then the partition function of the i th structures is given as follow:

$$Z_i = \int \exp\{-\beta \cdot [K_i(p) + U_i(q) + U_i^e]\} d\vec{p} d\vec{q}$$

Equation 4.3

where $\beta = 1/k_B T$, with k_B the Boltzmann constant and T the temperature. Note that (from now on) the subscript i means that the value of the quantity it attaches to is to be determined via the set of n dihedral angles or states which corresponds to the i th structure. In Equation 4.3, K_i is written to depend only on p because the effective potential is found to be essentially independent of the chain conformations. Since $\dot{q} = M^{-1}p$, Equation 3.7 can also take the following form:

$$K_i(p) = \frac{1}{2} p^T M_i^{-1} p$$

Equation 4.4

with the mass matrix M_i independent of q .

Substituting Equation 3.34, Equation 4.2 and Equation 4.4 into Equation 4.3, we obtain

$$Z_i = \int \exp \left\{ -\beta \cdot \left[\frac{1}{2} p^T M_i^{-1} p + \frac{1}{2} q^T (\kappa_1 D_i + \kappa_2 C_i) q + (N_i U_h^e + N_{ai} U_a^e + N_{bi} U_b^e) \right] \right\} d\vec{p} d\vec{q}$$

Equation 4.5

After evaluating the integral in Equation 4.5, we arrive at the following expression for the partition function:

$$\begin{aligned} Z_i &= \frac{(2\pi k_B T)^{n-1}}{\sqrt{\det M_i^{-1}}} \frac{(2\pi k_B T)^{N_i}}{\sqrt{\det_p(\kappa_1 D_i + \kappa_2 C_i)}} (2\pi)^{2(n-1)-2N_i} e^{-\beta U_i^e} \\ &= \frac{(2\pi k_B T)^{n-1}}{(\bar{\lambda}_i^M)^{2(n-1)}} \frac{(2\pi k_B T)^{N_i}}{(\bar{\lambda}_i^K)^{2N_i}} (2\pi)^{2(n-1)-2N_i} \times e^{-\frac{1}{k_B T}(N_i U_h^e + N_{ai} U_a^e + N_{bi} U_b^e)} \end{aligned}$$

Equation 4.6

By summing over the partition function Z_i of each structure i , we form the grand partition function of a protein chain as follow:

$$\begin{aligned} \mathbb{Z} &= \sum_{i=1}^{5^n} Z_i \\ &= \sum_{i=1}^{5^n} \left\{ \frac{(2\pi k_B T)^{n-1}}{(\bar{\lambda}_i^M)^{2(n-1)}} \frac{(2\pi k_B T)^{N_i}}{(\bar{\lambda}_i^K)^{2N_i}} (2\pi)^{2(n-1)-2N_i} \times e^{-\frac{1}{k_B T}(N_i U_h^e + N_{ai} U_a^e + N_{bi} U_b^e)} \right\} \end{aligned}$$

Equation 4.7

In view of the fact that the term $(\bar{\lambda}_i^M)^{2(n-1)}$ is large during computation, each Z_i is very small. It turns out that the value of the grand partition function is approximately zero even after performing the 5^n summation. In order to avoid numerical inaccuracies, we have rescaled Equation 4.7 by $\bar{\lambda}_M^{2(n-1)}$, with $\bar{\lambda}_M = \sum_i \bar{\lambda}_i^M / 2^n$, as follow:

$$\mathbb{Z} = \frac{(2\pi k_B T)^{n-1}}{\bar{\lambda}_M^{2(n-1)}} \sum_{i=1}^{5^n} \left\{ \frac{\bar{\lambda}_M^{2(n-1)}}{(\bar{\lambda}_i^M)^{2(n-1)}} \frac{(2\pi k_B T)^{N_i}}{(\bar{\lambda}_i^K)^{2N_i}} (2\pi)^{2(n-1)-2N_i} \times e^{-\frac{1}{k_B T}(N_i U_h^e + N_{ai} U_a^e + N_{bi} U_b^e)} \right\}$$

Equation 4.8

Note that the rescaling has no consequence on our investigation of the statistical properties of the phase transition in the next section since the term $\bar{\lambda}_M^{2(n-1)}$ will be cancelled out during the ensemble averaging operation.

4.3 The Statistical Physics of Protein Phase Transition

We shall employ the grand partition function to calculate the ensemble average of different statistical properties of a protein chain. The first property that we investigate is the average number of hydrogen bonds in the protein as a function of temperature. This average number $\langle N \rangle$ is obtained from the partition function in the following manner:

$$\langle N \rangle = \frac{\sum_{i=1}^{5^n} (N_i Z_i)}{\sum_{i=1}^{5^n} Z_i}$$

Equation 4.9

Note that Z_i is given by Equation 4.6. In the case of a small protein, $\langle N \rangle$ gives a good estimate on the type of its secondary structure.

4.4 α -helix to Coil Phase Transition

With the development of the preceding theories, we now proceed to analyze a trivial two-phase transition test case. Consider the dihedral angles of a crank unit that can adopt only either the $\{-57.4^\circ, -47.5^\circ\}$ or $\{180^\circ, 180^\circ\}$ state. The total number of possible conformations of a n crank polypeptide would reduce to just 2^n . All formulations remain the same as described in the preceding sections. Equation 4.2, Equation 4.8 and Equation 4.9 can thus be re-written as:

$$U^e = N U_h^e + N_a U_a^e$$

Equation 4.10

$$\mathbb{Z} = \frac{(2\pi k_B T)^{n-1}}{\bar{\lambda}_M^{2(n-1)}} \sum_{i=1}^{2^n} \left\{ \frac{\bar{\lambda}_M^{2(n-1)}}{(\bar{\lambda}_i^M)^{2(n-1)}} \frac{(2\pi k_B T)^{N_i}}{(\bar{\lambda}_i^K)^{2N_i}} (2\pi)^{2(n-1)-2N_i} \times e^{-\frac{1}{k_B T}(N_i U_h^e + N_{ai} U_a^e)} \right\}$$

Equation 4.11

$$\langle N \rangle = \frac{\sum_{i=1}^{2^n} (N_i Z_i)}{\sum_{i=1}^{2^n} Z_i}$$

Equation 4.12

with $N_b=0$. Since a crank could only form two different states, we assume that all hydrogen bonds formed are associated to α -helix, i.e. $U_a^e = U_h^e$.

Figure 4.1 shows the phase transition of a seven-crank two-state polyaniline, looking at the number of hydrogen bonds against temperature, i.e. $\langle N \rangle$ which is the ensemble average number of hydrogen bonds formed. In this figure, we observe a phase transition from α -helix to coil as temperature rises. The critical temperature for such transition is seen to be around 773 K.

Figure 4.1: Helix-coil phase transition of a seven-crank two-state polyaniline. We plot a graph of ensemble average number of hydrogen bonds formed against temperature change. At low temperature, the ensemble forms four hydrogen bonds. As it is a short chain with only two allowable state per crank, formation of four hydrogen bonds can be associated to formation of a full α -helix. As temperature rises, hydrogen bonds destabilize and the helix is fully denatured into coil at extreme temperature.

4.5 α -Helix to β -Sheet to Coil Phase Transition

Next, we plot the α - β -coil phase transition for a five-state polypeptide. In order to proceed, we require the relative weight of U_a^e and U_b^e . Let us assume that $U_a^e = U_h^e$ and $U_b^e = 2U_h^e$, which is reasonable according to reference (6) and (117). Based on this assumption, we have computed the number of hydrogen bonds against temperature for a seven crank polyaniline, which is plotted in Figure 4.2. From Figure 4.2, we observe that the number of hydrogen bonds decreases as temperature increases. We found that $\langle N \rangle$ in Figure 4.2 is the sum of the hydrogen bonds of the α -helical and β -sheet structures. At low temperature, we observe four hydrogen bonds and the protein is assumed to be an α -helix. When the temperature is increased to a critical value, a transition occurs with the protein containing a mixture of α -helical and β -sheet structures, before turning into a full β -sheet. As this happen, the number of hydrogen bond reduces. A further increase in temperature eventually leads to the next critical value at which the β -sheet denatures into a random coil.

Figure 4.2: The average number of hydrogen bonds in a seven crank polyaniline versus temperature obtained from an ensemble of 5^7 configurations. The three lines are: total number of hydrogen bonds (solid line), number of hydrogen bonds associated with α -helix (dashed line), and number of hydrogen bonds associated with β -sheet (dotted line).

4.6 Computation of the Change of Heat Capacity in Polypeptide

During phase transition, a sharp peak in the heat capacity is to be expected against a change in temperature⁽⁷³⁾. This results from a sudden change in the internal energy of the system⁽¹⁰⁹⁾. To verify this, we proceed to compute the ensemble average of our second property of interest: the heat capacity of the protein polyaniline as a function of temperature and then observe the possible presence of sharp peaks.

The definition of the heat capacity is given as follow:

$$C(T) = k_B T \frac{\partial^2 T \ln \mathbb{Z}}{\partial T^2}$$

Equation 4.13

By using Equation 4.8, the heat capacity of our model can be expressed in the following manner:

$$\begin{aligned}
C(T) &= k_B T \left[\frac{2}{\mathbb{Z}} \frac{\partial \mathbb{Z}}{\partial T} - \frac{T}{\mathbb{Z}^2} \left(\frac{\partial \mathbb{Z}}{\partial T} \right)^2 + \frac{T}{\mathbb{Z}} + \frac{T}{\mathbb{Z}} \frac{\partial^2 \mathbb{Z}}{\partial T^2} \right] \\
&= k_B T \left\{ \frac{2}{(2\pi k_B T)^{n-1} \mathbb{A}} [(n-1)(2\pi k_B T)^{n-1} (2k_B) \mathbb{A} + (2\pi k_B T)^{n-1} \mathbb{B}] \right. \\
&\quad - \frac{T}{(2\pi k_B T)^{2(n-1)} \mathbb{A}^2} [(n-1)(2\pi k_B T)^{n-1} (2k_B) \mathbb{A} + (2\pi k_B T)^{n-1} \mathbb{B}]^2 \\
&\quad + \frac{T}{(2\pi k_B T)^{n-1} \mathbb{A}} [(n-1)(n-2)(2\pi k_B T)^{n-3} (2\pi k_B)^2 \mathbb{A} \\
&\quad \left. + 2(n-1)(2\pi k_B T)^{n-2} (2\pi k_B) \mathbb{B} + (2\pi k_B T)^{n-1} \mathbb{C}] \right\}
\end{aligned}$$

Equation 4.14

where

$$\begin{aligned}
\mathbb{A} &= \sum_{i=1}^{5^n} \left\{ \frac{\bar{\lambda}_M^{2(n-1)}}{(\bar{\lambda}_i^M)^{2(n-1)}} \frac{(2\pi k_B T)^{N_i}}{(\bar{\lambda}_i^K)^{2N_i}} (2\pi)^{2(n-1)-2N_i} e^{-\frac{U_i^e}{k_B T}} \right\} \\
\mathbb{B} &= \sum_{i=1}^{5^n} \left\{ \frac{\bar{\lambda}_M^{2(n-1)}}{(\bar{\lambda}_i^M)^{2(n-1)}} \frac{(2\pi)^{2(n-1)-2N_i}}{(\bar{\lambda}_i^K)^{2N_i}} \left[N_i (2\pi k_B T)^{N_i-1} (2\pi k_B) e^{-\frac{U_i^e}{k_B T}} + (2\pi k_B T)^{N_i} e^{-\frac{U_i^e}{k_B T}} \left(-\frac{U_i^e}{k_B T^2} \right) \right] \right\} \\
\mathbb{C} &= \sum_{i=1}^{5^n} \left\{ \frac{\bar{\lambda}_M^{2(n-1)}}{(\bar{\lambda}_i^M)^{2(n-1)}} \frac{(2\pi)^{2(n-1)-2N_i}}{(\bar{\lambda}_i^K)^{2N_i}} e^{-\frac{U_i^e}{k_B T}} \left[N_i (N_i - 1) (2\pi k_B T)^{N_i-2} (2\pi k_B)^2 \right. \right. \\
&\quad \left. \left. + 2N_i (2\pi k_B T)^{N_i-1} (2\pi k_B) \left(-\frac{U_i^e}{k_B T^2} \right) + (2\pi k_B T)^{N_i} \left(\left[-\frac{U_i^e}{k_B T^2} \right]^2 + \frac{2U_i^e}{k_B T^3} \right) \right] \right\}
\end{aligned}$$

Equation 4.15

for a five-state polypeptide.

The problem can be scaled down to 2^n conformations for a two-state polypeptide. Following the above derivations, we plot a graph of heat capacity against temperature in Figure 4.3 for a seven-crank two-state polyalanine.

Figure 4.3: Helix-coil phase transition for a two-state seven-crank polypeptide, observed in a heat capacity against temperature plot. The peak at $T_{\alpha\text{-coil}} = 773$ K corresponds to the critical temperature as observed in Figure 4.1.

Similarly, we plot heat capacity against temperature for a five-state seven-crank polyaniline in Fig. 4.4, where we observe the occurrence of two peaks which confirm the presence of two phase transitions. The first peak represents the transition from α -helix to β -sheet at a temperature of around 276K. The second peak arises from the transition from β -sheet to random coil at a temperature of around 932K.

Figure 4.4: The heat capacity of a seven-crank polyaniline versus temperature obtained from an ensemble of 5^7 configurations. The first phase transition occurs at a temperature around 276K with the α -helix transforming into a β -sheet. The second phase transition occurs at around 932K with the β -sheet turning into a random coil.

4.7 Analysis of Results and Comparison with Literature

In summary, our results as displayed in Figure 4.3 and Figure 4.4 illustrate that the physics of protein phase transitions is captured by the grand partition function $\mathbb{Z}$. The results in these figures imply two important facts:

- (1) The network of hydrogen bonds stabilizes the protein secondary structure.
- (2) Hydrogen bonds loses its hold on the structure when the temperature reaches a critical value, and as this happens new secondary structure appears.

With $\mathbb{Z}$, we can explore into various aspects of the protein dynamics through its statistical physics, and this makes our theory more comprehensive than classical two-state models^{(70) (71)}. Our theory differs from others that also take a partition function approach, such as theories due to Zimm-Bragg, Lifson-Roig, and Lei *et. al.*. In these theories, the residue of a protein is assigned a state from a discrete set, and each state is given a statistical weight. In our model, we assign five states to each crank unit, but in the form of five distinct sets of dihedral angles. Furthermore, instead of assigning different statistical weight to each state, we assign the statistical weight based on the types of the hydrogen bond, i.e., depending on whether the hydrogen bond is associated with the α -helix or β -sheet. Our approach has the advantage of being less restrictive because a residue can adopt more than just the pairs of dihedral angles that form helix, sheet or coil. However, it requires greater computational power due to the need to sum through a larger number of configurations during the calculation of the grand partition function of the protein chain. This limitation has essentially restricted our study to at most seven cranks with a maximum of five distinct states for each crank.

Comparing against our earlier results in Chapter 3 where the α - β and the β -coil transition temperatures are 475 K and 600 K respectively, we observe that the α - β transition temperature has lowered to 276 K while the β -sheet transition temperature has increased to 932 K. This results from our assumption in this thesis that the statistical weighting of the hydrogen bonds in the β -sheet is greater than that in the α -helix. With the hydrogen bonds being relatively more stable in the β -sheet than in the α -helix, we anticipate the β -sheet to form at lower temperature than before. Furthermore, the increased stability of the hydrogen bonds in the β -sheet implies that a higher temperature is needed to break all its hydrogen bonds and converts it into a random coil. In fact, the α - β transition temperature of 276 K that we obtained is rather close to the 315 K predicted by Ding *et. al.*⁽⁷⁵⁾ through discrete molecular dynamics simulation, although β -coil transition temperature of

932 K is vastly different from the 327 K obtained by them. Figure 4.2 also indicates that the α - β transition temperature occurs at about 276 K when the number of hydrogen bonds from the α -helix drop to zero. A comparison of this temperature against the predictions made by the models of Yasar & Demir⁽⁷⁷⁾ and Lee *et. al.*⁽¹¹⁶⁾ have further validated our approach to be a reliable concept.

4.8 Revisiting Phase Transition with Respect to Helmholtz Free Energy

We have plotted in Figure 4.5 the Helmholtz free energy of the seven crank polyalanine in the α -helix, the β -sheet and the random coil configuration as a function of temperature (see also Figure 3.10). In the calculation of the Helmholtz free energy (which is based on the approach in Chapter 3), we have considered the different statistical weights for the hydrogen bonds in the α -helix and in the β -sheet structure as introduced in this section. From Figure 4.5, we observe an α - β transition temperature of about 400K and a β -coil transition temperature of 1,900K. This is to be expected based on the argument above with the β -sheet being relatively more stable than the α -helix. The results here are however different from that obtained in Figure 4.2 and Figure 4.4 because the free energy approach does not take into account the mechanism of the phase transition, unlike the approach of the grand partition function.

Figure 4.5: Free energy curves of α -helix, β -sheet, and a random coil for a seven-crank polyalanine. Note that the intersection of the curves indicate phase transition.

The transition temperatures are $T_{\alpha\beta} = 400$ K, $T_{\alpha\text{-coil}} = 1,100$ K and $T_{\beta\text{-coil}} = 1,900$ K.

4.9 Further Analysis and Discussion on Protein Phase Transition

By assigning five states to each crank such that each of the 5^n configurations is associated to the partition function Z , as given by Equation 4.3, the grand partition function $\mathbb{Z}$ describes a free energy landscape consisting of each of these 5^n configurations as local minimum. In other words, we have assumed that the 5^n structures are relatively more stable compared to their neighboring structures in the free energy space in view of our approach in evaluating the Gaussian integral given by Equation 4.5 based on the stationary phase approximation. While we expect our assumption to remain true as n increases under the above scenario (which is validated for a two-state polypeptide below), the situation would be different if we were to increase the number of states per crank or to consider sequence dependent protein structure, since a configuration under these circumstances may not constitute a local minimum. Nonetheless, the inclusion of additional states or information per crank, may lead to new stable structural phases that emerge out of the extra statistical information hidden in the grand partition function.

Let us now employ the above argument to explain the results obtained in Figure 4.5, where we have come to understand that a single state of either α -helix, β -sheet or random coil has been allocated to each crank. This implies the presence of a single minimum for the free energy landscape, and the curves in Figure 4.5 plot the Helmholtz free energy of the minimum for the corresponding conformation. In order to determine the stable configuration at a particular temperature, the three separate minimum free energies are then compared at the given temperature with the configuration having the lowest free energy being chosen as the stable conformation. This approach differs from that of the grand partition function where the statistical information of all the possible configurations is contained within the function itself. The presence of more information in the grand partition function approach leads to a more comprehensive treatment of phase transition. Nevertheless, both approaches reach the same conclusion of α -helix being formed at low temperature, followed by β -sheet being the stable structure as temperature increases, and the polypeptide becomes denatured at high temperature.

At current stage, our ability to develop further analysis on our theory using a longer polypeptide with more than five states per crank is severely limited by the available computing resources. As a result, in order to investigate the impact of a larger number of cranks, we have to reduce the number of states per crank. In this respect, we have considered a two-state polypeptide, i.e. a crank with dihedral angles $\{-57.4^\circ, -47.5^\circ\}$ and $\{180^\circ, 180^\circ\}$. We observe that the α -helix to random coil

phase transition occurs at the same critical temperature for different n . This confirms our expectation that the phase transition in a polypeptide is independent of its size. We observe that the energies of the local minimum of the stable phases remain significant as n increases. Figure 4.6 illustrates the α -helix transition of the two-state polypeptide when the length ranges from $n = 7$ to $n = 15$. We notice that phase transition occurs within a narrower window of temperature change as n becomes larger. Figure 4.7 shows a better comparison of the curves after normalization.

Figure 4.6: The ensemble average number of hydrogen bonds versus temperature of a two-state polypeptide at different lengths.

Figure 4.7: Normalized ensemble average number of hydrogen bonds versus temperature of a two-state polypeptide at different lengths. The normalization is performed by rescaling each curve by the maximum number of hydrogen bonds formed in that particular length of polypeptide chain.

We can continue to investigate the role of the total number of distinct states on protein phase transition by considering the case with three and four states per crank. By adding the dihedral angle $\{315^\circ, 110^\circ\}$ to the set $\{-57.4^\circ, -47.5^\circ\}$ and $\{180^\circ, 180^\circ\}$ to form a three-state seven-crank polypeptide, we notice the occurrence of an α -helix to random coil phase transition as the two-state polypeptide that we have discussed above, albeit with a reduced critical temperature of about 687K (see Figure 4.8), comparing to Figure 4.1 and Figure 4.3.

Figure 4.8: Plot of average number of hydrogen bonds and its corresponding heat capacity change as a function of temperature, for a three-state polypeptide comprising dihedral angle sets of $\{315^\circ, 110^\circ\}$, $\{-57.4^\circ, -47.5^\circ\}$ and $\{180^\circ, 180^\circ\}$.

However, if we were to use the set $\{-139^\circ, 135^\circ\}$, $\{315^\circ, 110^\circ\}$ and $\{105^\circ, 330^\circ\}$ instead, we observe two transitions (see Figure 4.9). The first transition at 282 K is not a secondary structure phase transition since it involves the denaturing of a single hydrogen bond not associated with any secondary structure. On the other hand, the second transition at 1100 K is a β -coil phase transition.

Figure 4.9: Plot of average number of hydrogen bonds and its corresponding heat capacity change as a function of temperature, for a three-state polypeptide comprising dihedral angle sets of $\{-139^\circ, 135^\circ\}$, $\{315^\circ, 110^\circ\}$ and $\{105^\circ, 330^\circ\}$. We observe the presence of an additional hydrogen bond at lower temperature, which is associated to neither of the α -helical nor β -sheet structure.

Similar transition scenarios occurs when we add the dihedral angle $\{180^\circ, 180^\circ\}$ to the previous set of angles to form a four-state polypeptide. We again detect a phase transition from the β -sheet to the random coil, although it now arises at a lower critical temperature of 1002 K (see Figure 4.10). The first peak of phase transition occurs at 265 K.

Figure 4.10: Plot of average number of hydrogen bonds and its corresponding heat capacity change as a function of temperature, of a four-state polypeptide comprising dihedral angle sets of $\{315^\circ, 110^\circ\}$, $\{-139^\circ, 135^\circ\}$, $\{105^\circ, 330^\circ\}$ and $\{180^\circ, 180^\circ\}$. The plot is similar to Figure 4.9.

If we were to use the set of angles $\{-57.4^\circ, -47.5^\circ\}$, $\{-139^\circ, 135^\circ\}$, $\{315^\circ, 110^\circ\}$ and $\{105^\circ, 330^\circ\}$ instead to form a four-state polypeptide, we observe both the α - β and β -coil phase transition as our original five-state polypeptide, with the transition temperatures now happen at 283 K and 990 K respectively (see Figure 4.11).

Figure 4.11: Plot of average number of hydrogen bonds and its corresponding heat capacity change as a function of temperature, for a four-state polypeptide comprising dihedral angle sets of $\{-57.4^\circ, -47.5^\circ\}$, $\{-139^\circ, 135^\circ\}$, $\{315^\circ, 110^\circ\}$ and $\{105^\circ, 330^\circ\}$. While the plot is analogous to Figure 4.9 and Figure 4.10, depicting a β -coil phase transition, the above figure illustrates the occurrence of the α - β -coil phase transition similar to Figure 4.2.

While we have let $U_b^e = \gamma U_a^e$ with $\gamma = 2.0$ in the evaluation of our results in Figure 4.2 and Figure 4.4, it is interesting to explore the effect of varying the relative statistical weight between U_b^e and U_a^e as represented by γ , on the structural phase transition (see Figure 4.12 to Figure 4.20). When $\gamma = 1.0$ and 1.2, we observe a transition from α -helix to random coil as temperature increases. A stable phase of β -sheet starts to appear only at $\gamma = 1.4$, and becomes apparent after $\gamma = 1.6$. More specifically, we observe two phase transitions from α -helix to β -sheet, and then from β -sheet to random coil, at $\gamma = 1.6, 1.8, 2.0$ and 2.2. Subsequently, at $\gamma = 2.4$, we observe the unexpected presence of three stable phases: a phase with four hydrogen bonds (α -helix); a phase with three hydrogen bonds (β -sheet and an additional uncategorized hydrogen bond); a phase with two hydrogen bonds (β -sheet); and a phase with no hydrogen bond (coil). In fact, the new phase of three hydrogen bonds is similar to the one shown in Figure 4.9 and Figure 4.10, and it begins to emerge at $\gamma = 2.2$ according to our heat capacity calculation. When γ is further increased to 2.6, the phase of α -helix disappears, and phase transition now occurs between the three phases of the mixture of α -helix and β -sheet, and the random coil. These outcome can be understood to result from the progressive enhancement in stability of the β -sheet structure as the weighting of hydrogen bond

energy associated to the β -sheet becomes larger than that associated to the α -helix. We also plot the results of heat capacity change as a function of temperature together with the corresponding figures.

Figure 4.12: The ensemble average number of hydrogen bonds and heat capacity change as a function of temperature in a seven-crank polypeptide obtained from an ensemble of 5^7 conformations with $U_a^e = U_b^e = U_h^e$.

Figure 4.13: The ensemble average number of hydrogen bonds and heat capacity change as a function of temperature in a seven-crank polypeptide obtained from an ensemble of 5^7 conformations with $U_a^e = U_h^e$, $U_b^e = 1.2U_a^e$.

Figure 4.14: The ensemble average number of hydrogen bonds and heat capacity change as a function of temperature in a seven-crank polypeptide obtained from an ensemble of 5^7 conformations with $U_a^e = U_h^e$, $U_b^e = 1.4U_a^e$.

Figure 4.15: The ensemble average number of hydrogen bonds and heat capacity change as a function of temperature in a seven-crank polypeptide obtained from an ensemble of 5^7 conformations with $U_a^e = U_h^e$, $U_b^e = 1.6U_a^e$.

Figure 4.16: The ensemble average number of hydrogen bonds and heat capacity change as a function of temperature in a seven-crank polypeptide obtained from an ensemble of 5^7 conformations with $U_a^e = U_h^e$, $U_b^e = 1.8U_a^e$.

Figure 4.17: The ensemble average number of hydrogen bonds and heat capacity change as a function of temperature in a seven-crank polypeptide obtained from an ensemble of 5^7 conformations with $U_a^e = U_h^e$, $U_b^e = 2U_a^e$.

Figure 4.18: The ensemble average number of hydrogen bonds and heat capacity change as a function of temperature in a seven-crank polypeptide obtained from an ensemble of 5^7 conformations with $U_a^e = U_h^e$, $U_b^e = 2.2U_a^e$.

Figure 4.19: The ensemble average number of hydrogen bonds and heat capacity change as a function of temperature in a seven-crank polypeptide obtained from an ensemble of 5^7 conformations with $U_a^e = U_h^e$, $U_b^e = 2.4U_a^e$.

Figure 4.20: The ensemble average number of hydrogen bonds and heat capacity change as a function of temperature in a seven-crank polypeptide obtained from an ensemble of 5^7 conformations with $U_a^e = U_h^e$, $U_b^e = 2.6U_a^e$.

Chapter 5: Conclusion

We began our research by investigating the well established protein folding problem. After improving Prof. Kerson Huang's CSAW^{(86) (85)} model which defines the basic unit of polypeptide as crank, we have successfully folded both the α -helix and β -sheet using the modified model. While the construction of the CSAW model is simple, it has provided us with a theoretical framework to produce the key research results presented in this thesis.

In this thesis, we have first formulated the mechanics of protein chains in terms of the Hamiltonian formalism. We have applied this formalism to determine the normal modes of the α -helix and the β -sheet, and the phase transition of these secondary structures. By modeling the protein as a sequence of cranks and making small deviations from the equilibrium potential of the hydrogen bonds, we have obtained normal mode distributions that correspond to those computed based on classical NMA⁽⁴⁰⁾. This has served to validate our formalism. Unlike prior models that employed a complete range of force fields, results from our simplified model reveal that hydrogen bonding contributes to the slow modes observed in the α -helix and β -sheets.

Following the success of our preceding findings, we have pursued the subject of protein phase transition from a different perspective by formulating the partition function and the Helmholtz free energy based on the equations derived from our canonical formalism and normal mode analysis. Remarkably, our analytical results on the free energy for a 15-crank polypeptide have reaffirmed previous numerical prediction⁽⁷⁵⁾ of an α - β followed by a β -coil transition, as temperature rises. While we had performed our analysis in the gas phase, which leads to a high critical temperature for the α - β transition, we expect such a transition to occur under normal physiological conditions if there are present enzymatic influences in a hydrated environment, such as the seeding of PrP^{Sc} (scarpie PrP) converts PrP^C into amyloid fibrils of recPrP (recombinant PrP)⁽⁸¹⁾. Our crude model could add on to the existing studies of structural transition in protein secondary structures^{(70) (76) (75) (77) (116)}.

Based on the same Hamiltonian formalism, we have derived the grand partition function of the protein chain. In our model, we have assumed that the dominant non-covalent potential is due to the hydrogen bond interaction. In addition, we have introduced different statistical weights for the

strength of the hydrogen bonds within the α -helix and the β -sheet structure. By considering five distinct states for the dihedral angles, we have computed the average number of hydrogen bonds and the heat capacity of the protein polypeptide using the grand partition function at different temperatures. We observe a sudden change in the average number of hydrogen bonds at two critical temperatures, signifying an α - β and a β -coil phase transition. Our results based on the heat capacity confirm the occurrence of phase transition. This finding indicates that the transition in protein secondary structure is an abrupt change, which implies difficulties in detecting any precursor of misfolding events.

Based on just a five-state polypeptide model which gives us the significant result of α - β -coil phase transition, we are confident that if we were to increase the number of states for the dihedral angles, we would be able to perform a more complex analysis to extract additional features out of the polypeptide model. However, our works have been limited by the computing resources available. The computation of partition functions is a time consuming and memory intensive process. All our computations are carried out using MATLAB computational software. The benchmark of our computation is shown in Figure 5.1. At the current stage, we are only able to analyze a polypeptide up to a length of seven cranks, and five states (five dihedral angle sets) per crank. A longer length of polypeptide or having more states would demand more computing power. The required computing power is a function of the total number of possible conformations that can be folded out of a polypeptide. Figure 5.2 and Figure 5.3 show the exponential increase of total number of conformations as either length or number of states increases.

Figure 5.1: Time consumed per computation versus number of CPU cores, for the computation of grand partition function in a seven-crank polypeptide model.

Figure 5.2: Total number of conformations for a seven crank polypeptide versus number of dihedral angle sets per crank.

Figure 5.3: Total number of conformations for a five-state polypeptide at different polypeptide lengths.

All the numerical results presented in this report, such as the critical temperature of protein secondary structure phase transition are only theoretically true, and are not practically reliable. Our model is still far from complete and would not be able to predict the exact transition temperature. Nevertheless, given a polypeptide chain, we are certain that α -helix is the most stable conformation at very low temperature, followed by β -sheet as temperature rises, and denatures into coil at extreme high temperature, though there could be other more stable structures at a particular temperature which were not discussed in the current context due to the limited range of dihedral angles that each crank can adopt. Due to our simplifying assumption, we have shown the role of hydrogen bonding network in protein phase transition, however it is still far from complete in the

consideration of all the factors that contribute to protein phase transition which lead to protein misfolding. We perceive that by improving the polypeptide model through the incorporation of full-atom side chains and other non-covalent interactions such as the electrostatic force and hydrophobic potential, the revised model will give a closer correspondence to the real protein. However, we expect the modelling of the additional features to be tedious and the resulting partition function may become analytically unsolvable while numerical computation would require the use of powerful supercomputer.

Although we have derived a mathematical model for polypeptide which can be solved analytically, the grand partition function is in the form of a sum of discrete terms (Equation 4.8). The completeness of the grand partition function can only be improved by adding more terms into the summation equation, i.e. by increasing the number of states per crank. Again, such calculations will be limited by the computing resources available. One way to resolve this is to formulate a continuous form of grand partition function which can eliminate the need to increase the number of terms for the partition function, and may eventually resolve the issue on the demand of excessive computing power.

Nonetheless, despite restriction by computing resources, our results presented in this thesis have defined a new physical potential for hydrogen bonding (U^e term) and have proven the importance of hydrogen bonding network in protein secondary structure phase transition. Instead of looking from a conventional chemical or biological point of view, we have presented a different assumption on hydrogen bond formation in protein. We simply use a criteria based assumption in forming hydrogen bond (distance between the H and O atoms is $2.0 \pm 1.0 \text{ \AA}$, and the bond angle between $N-H$ and $C=O$ is $180^\circ \pm 45^\circ$) which has proven to be pertinent in predicting the phase transition in protein secondary structure. The different weightings of hydrogen bonds associated with different secondary structure play a major role in determining the most probable phase at a particular temperature. The different weightings can be interpreted as an implicit potential such as the electrostatic interaction between $N-H$ and $C=O$ bonds. Our concept of defining different weightings for different hydrogen bonds is in fact in agreement with results of computer simulation from reference (117). Such implementation can be applied in other science research which involves hydrogen bonding. Our findings also show that hydrogen bonding is a major contributor to protein phase transition.

Finally, we would like to highlight that the physical model derived in this report is not a trivial solution to the polypeptide problems. The idea of Gaussian integral in radian space is the key to solving the grand partition function of our polypeptide model in an analytical way. We believe the analytical formula based on Gaussian integral in radian space has a larger implication and can be applied to other scientific/engineering problems.

In a nutshell, we hope that our research has contributed to a major finding in the study of protein science. We are confident that the ideas and results presented here can be further explored despite certain limitations which can be overcome by an improvement in the technology of computation. As suggested earlier, we hope that one day we can attain a unified theoretical viewpoint and solution to the protein folding problem.

Bibliography

1. **Creighton, Thomas E.** *Protein Folding*. Oxford : W. H. Freeman, 1992. ISBN 0-7167-7027-X.
2. *The Protein Folding Problem*. **Dill, Ken A., et al.** 2008, Annual Review of Biophysics, Vol. 37, pp. 289-316.
3. **YassineMrabet.** Structure générale d'un acide aminé. *Wikipedia*. [Online] Wikimedia Foundation Inc., August 12, 2007. [Cited: April 27, 2012.] <http://en.wikipedia.org/wiki/File:AminoAcidball.svg>.
4. Peptide bond formation. *Wikipedia*. [Online] Wikimedia Foundation Inc., August 12, 2007. [Cited: April 2012, 27.] <http://en.wikipedia.org/wiki/File:Peptidformationball.svg>.
5. **LadyofHats.** Main protein structures levels. *Wikipedia*. [Online] October 7, 2008. [Cited: May 7, 2012.] http://en.wikipedia.org/wiki/File:Main_protein_structure_levels_en.svg.
6. *Effective Potentials for Folding Proteins*. **Chen, Nan-Yow, Su, Zheng-Yao and Mou, Chung-Yu.** Physical Review Letters : s.n., 2006, Vol. 96, p. 078103.
7. *Molecular dynamics of native protein. I. Computer Simulation of trajectories*. **Levitt, M.** 1983, Journal of Molecular Biology, Vol. 168, p. 595.
8. *Dynamics of folded proteins*. **McCammon, J. A., Gelin, B. R. and Karplus, M.** 1977, Nature, Vol. 267, p. 585.
9. **Thijssen, J. M.** *Computational Physics*. Cambridge : Cambridge University Press, 1999.
10. *Studies in Molecular Dynamics. I. General Method*. **Alder, B. J. and Wainwright, T. E.** 2, 1959, Journal of Chemical Physics, Vol. 31, p. 459.
11. *Introduction to Monte Carlo Methods*. **Attig, Norbert, et al.** Julich : s.n., 2004, Computational Soft Matter: From Synthetic Polymers to Proteins, Lecture Notes, Vols. NIC Series, Vol. 23, p. 29.
12. *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules*. **Cornell, W. D., et al.** 1995, Journal of the American Chemical Society, Vol. 117, p. 5179.
13. *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations*. **Brooks, B. R., et al.** 2, 1983, Journal of Computational Chemistry, Vol. 4, p. 187.
14. *CHARMM: The Energy Function and Its Parameterization with an Overview of the Program*. **MacKerell, A. D. Jr., et al.** The Encyclopedia of Computational Chemistry, Vol. 1, p. 271.
15. *CHARMM: The biomolecular simulation program*. **Brooks, B. R., et al.** 10, 2009, Journal of Computational Chemistry, Vol. 30, p. 1545.

16. *Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94.* **Halgren, Thomas A.** 5-6, 1996, Journal of Computational Chemistry, Vol. 17, p. 490.
17. *Scalable molecular dynamics with NAMD.* **Phillips, James C., et al.** 2005, Journal of Computational Chemistry, Vol. 26, p. 1781.
18. *GROMACS: fast, flexible, and free.* **Van Der Spoel, D., et al.** 16, 2005, Vol. 26, p. 1701.
19. *Replica-exchange molecular dynamics method for protein folding.* **Sugita, Yuji and Okamoto, Yuko.** 1999, Chemical Physics Letters, Vol. 314, p. 141.
20. Table of standard amino acid abbreviations and properties. *Wikipedia.* [Online] Wikimedia Foundation, Inc., April 24, 2012. [Cited: April 27, 2012.] http://en.wikipedia.org/wiki/Amino_acids#Table_of_standard_amino_acid_abbreviations_and_properties.
21. **Hausman, Robert E. and Cooper, Geoffrey M.** *The Cell: a molecular approach.* Washington, D.C. : ASM Press, 2004. ISBN 0-87893-214-3.
22. *A simple method for displaying the hydropathic character of a protein.* **Kyte, J. and Doolittle, RF.** 1, 1982, Journal of Molecular Biology, Vol. 157, p. 105.
23. *Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution.* **Radzicka, Anna and Wolfenden, Richard.** 5, 1988, Vol. 27, p. 1664.
24. *Experimentally determined hydrophobicity scale for proteins at membrane interfaces.* **Wimley, William C. and White, Stephen H.** 1996, Nature Structural Biology, Vol. 3, p. 842.
25. *Recognition of transmembrane helices by the endoplasmic reticulum translocon.* **Hessa, Tara, et al.** 2005, Vol. 433, p. 377.
26. **Karshikoff, Andrey.** *Non-Covalent Interactions in Proteins.* s.l. : World Scientific Publishing Company, 2006.
27. *Energy Parameters in Polypeptides. 10. Improved Geometrical Parameters and Nonbonded Interactions for Use in the ECEPP/3 Algorithm, with Application to Proline-Containing Peptides.* **Nemethy, George, et al.** 1992, Journal of Physical Chemistry, Vol. 96, p. 6472.
28. *Intermolecular Potentials from Crystal Data. III. Determination of Empirical Potentials and Application to the Packing Configurations and Lattice Energies in Crystals of Hydrocarbons, Carboxylic Acids, Amines, and Amides.* **Momany, F. A., et al.** 16, 1974, Journal of Physical Chemistry, Vol. 78, p. 1595.
29. *Intermolecular Potentials from Crystal Data. 6. Determination of Empirical Potentials for O-H...O=C Hydrogen Bonds from Packing Configurations.* **Sippl, Manfred J., Nemethy, George and Scheraga, Harold A.** 1984, Journal of Physical Chemistry, Vol. 88, p. 6231.
30. *Infrared spectra and resonance interaction of amide-I vibration of the antiparallel-chain pleated sheet.* **Chirgadze, Yu. N. and Nevskaya, N. A.** 4, 1976, Vol. 15, p. 607.

31. *Infrared spectra and resonance interaction of amide-I vibration of the parallel-chain pleated sheet.* **Chirgadze, Yu. N. and Nevskaya, N. A.** 4, 1976, Vol. 15, p. 627.
32. *Infrared spectra and resonance interactions of amide-I and II vibrations of alpha-helix.* **Nevskaya, N. A. and Chirgadze, Yu. N.** 4, 1976, Vol. 15, p. 637.
33. **Jabs, Andreas.** Determination of Secondary Structure in Proteins by Fourier Transform Infrared Spectroscopy (FTIR). *Jena Library of Biological Macromolecules*. [Online] Leibniz Institute of Age Research - Fritz Lipmann Institute (FLI). [Cited: May 8, 2012.] http://www.imb-jena.de/ImgLibDoc/ftir/IMAGE_FTIR.html.
34. *Examination of the secondary structure of proteins by deconvolved FTIR spectra.* **Byler, D. Michael and Susi, Heino.** 3, 1986, Biopolymers, Vol. 25, p. 469.
35. *Vibrational Spectroscopy and Conformation of Peptides, Polypeptides, and Proteins.* **Krimm, Samuel and Bandekar, Jagdeesh.** 1986, Advances in Protein Chemistry, Vol. 38, p. 181.
36. *Far-Infrared Spectra of Polyalanines with Alpha-Helical and Beta-Form Structures.* **Itoh, K., Nakahara, T. and Shimanouchi, T.** 1968, Biopolymers, Vol. 6, p. 1759.
37. **Wuthrich, Kurt.** *NMR of Proteins and Nucleic Acids.* New York : Wiley-Interscience, 1986.
38. *Vibrational normal-mode spectrum of globular proteins.* **ben-Avraham, Daniel.** 1993, Physical Review B, Vol. 47, p. 14559.
39. *Normal vibrations of crystalline polyglycine I.* **Abe, Yasuaki and Krimm, S.** 9, 1972, Biopolymers, Vol. 11.
40. *Normal modes analyses of macromolecules.* **ben-Avraham, D. and Tirion, M. M.** 1998, Physica A, Vol. 249, p. 415.
41. *Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor.* **Brooks, B. and Karplus, M.** 21, 1983, Proceedings of the National Academy of Sciences of the United States of America, Vol. 80, p. 6571.
42. *Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme.* **Levitt, Michael, Sander, Christian and Stern, Peter S.** 3, 1985, Journal of Molecular Biology, Vol. 181, p. 423.
43. *Normal mode analysis of protein dynamics.* **Case, David A.** 2, 1994, Current Opinion in Structural Biology, Vol. 4, p. 285.
44. *Investigating protein dynamics in collective coordinate space.* **Kitao, Akio and Go, Nobuhiro.** 2, 1999, Current Opinion in Structural Biology, Vol. 9, p. 164.
45. *Building-block approach for determining low-frequency normal modes of macromolecules.* **Tama, Florence, et al.** 1, 2000, Proteins, Vol. 41, p. 1.
46. *Theoretical Studies of Protein Folding.* **Go, N.** 1983, Annual Review of Biophysics and Bioengineering, Vol. 12, p. 183.

47. *Dynamics of a small globular protein in terms of low-frequency vibrational modes.* **Go, N., Noguti, T. and Nishikawa, T.** 12, 1983, Proceedings of the National Academy of Sciences of the United States of America, Vol. 80, p. 3696.
48. *Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis.* **Tirion, Monique M.** 9, 1996, Physical Review Letters, Vol. 77, p. 1905.
49. *Normal-Modes-Based Prediction of Protein Conformational Changes Guided by Distance Constraints.* **Zheng, W. J. and Brooks, B. R.** 5, 2005, Biophysical Journal, Vol. 88, p. 3109.
50. *Membrane interactions control residue fluctuations of outer membrane porins.* **Besya, A. B., Mobasheri, H. and Ejtehadi, M. R.** 2010, Physical Review E, Vol. 81, p. 051911.
51. *Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential.* **Bahar, Ivet, Atilgan, Ali Rana and Erman, Burak.** 3, 1997, Folding and Design, Vol. 2, p. 173.
52. *Gaussian Dynamics of Folded Proteins.* **Haliloglu, Turkan, Bahar, Ivet and Erman, Burak.** 16, 1997, Physical Review Letters, Vol. 79, p. 3090.
53. *Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model.* **Atilgan, A. R., et al.** 1, 2001, Biophysical Journal, Vol. 80, p. 505.
54. *Go-ing for the prediction of protein folding mechanisms.* **Takada, Shoji.** 21, 1999, Proceedings of the National Academy of Sciences of the United States of America, Vol. 96, p. 11698.
55. *Normal Mode Analysis with Simplified Models to Investigate the Global Dynamics of Biological Systems.* **Tama, F.** 2, 2003, Protein and Peptide Letters, Vol. 10, p. 119.
56. *Refinement of protein dynamic structure: normal mode refinement.* **Kidera, A. and Go, N.** 10, 1990, Proceedings of the National Academy of Sciences of the United States of America, Vol. 87, p. 3718.
57. *Normal modes as refinement parameters for the F-actin model.* **Tirion, M. M., et al.** 1, 1995, Biophysical Journal, Vol. 68, p. 5.
58. *Collective Motions in HIV-1 Reverse Transcriptase: Examination of Flexibility and Enzyme Function.* **Bahar, Ivet, et al.** 3, 1999, Journal of Molecular Biology, Vol. 285, p. 1023.
59. *Cooperative Fluctuations and Subunit Communication in Tryptophan Synthase.* **Bahar, I. and Jernigan, R. L.** 12, 1999, Biochemistry, Vol. 38, p. 3478.
60. *Collective NMR relaxation model applied to protein dynamics.* **Bruschweiler, Rafael and Case, David A.** 6, 1994, Physical Review Letters, Vol. 72, p. 940.
61. *Calculation of nuclear magnetic resonance order parameters in proteins by normal mode analysis.* **Sunada, Shinji, Go, Nobuhiro and Koehl, Patrice.** 1996, The Journal of Chemical Physics, Vol. 104, p. 4768.

62. *Dynamics of large proteins through hierarchical levels of coarse-grained structures.* **Doruker, Pemra, Jernigan, Robert L. and Bahar, Ivet.** 1, 2002, Journal of Computational Chemistry, Vol. 23, p. 119.
63. *Twenty-four near-instabilities of Caspar-Klug viruses.* **Englert, Francois, Peeters, Kasper and Taormina, Anne.** 2008, Physical Review E, Vol. 78, p. 031908.
64. *Prions.* **Prusiner, Stanley B.** 23, 1998, Proceedings of the National Academy of Sciences of the United States of America, Vol. 95, p. 13363.
65. *Prion Diseases and the BSE Crisis.* **Prusiner, Stanley B.** 1997, Science, Vol. 278, p. 245.
66. *Prion Protein Biology.* **Prusiner, Stanley B., et al.** 1998, Cell, Vol. 93, p. 337.
67. **Centers for Disease Control and Prevention.** CJD (Creutzfeldt-Jakob Disease, Classic). *Centers for Disease Control and Prevention.* [Online] Centers for Disease Control and Prevention, December 13, 2010. [Cited: May 17, 2012.] <http://www.cdc.gov/ncidod/dvrd/cjd/index.htm>.
68. *Bovine spongiform encephalopathy and variant Creutzfeldt-Jakob disease.* **Brown, Paul.** 2001, British Medical Journal, Vol. 322, p. 841.
69. *Evolution in the conceptualization of dementia and Alzheimer's disease: Greco-Roman period to the 1960s.* **Berchtold, N. C. and Cotman, C. W.** 3, 1998, Neurobiology of Aging, Vol. 19, p. 173.
70. *Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains.* **Zimm, B. H. and Bragg, J. K.** 1959, Journal of Chemical Physics, Vol. 31, p. 526.
71. *On the Theory of Helix-Coil Transition in Polypeptides.* **Lifson, Shneior and Roig, A.** 6, 1961, Journal of Chemical Physics, Vol. 34, p. 1963.
72. *Phase transition in polypeptides: a step towards the understanding of protein folding.* **Yakubovich, A. V., et al.** 3, 2006, The European Physical Journal D, Vol. 40, p. 363.
73. *Ab initio theory of helix-coil phase transition.* **Yakubovich, A. V., et al.** 2, 2008, The European Physical Journal D, Vol. 46, p. 215.
74. *Alpha-helix-random coil phase transition: analysis of ab initio theory predictions.* **Solov'yov, I. A., et al.** 2008, European Physics Journal D, Vol. 46, p. 227.
75. *Mechanism for the Alpha-Helix to Beta-Hairpin Transition.* **Ding, Feng, et al.** 2003, Proteins, Vol. 53, p. 220.
76. *Statistical mechanical model for helix-sheet-coil transitions in homopolypeptides.* **Hong, Liu and Lei, Jinzhi.** 5, 2008, Physical Review E, Vol. 78, p. 051904.
77. *The study of helix-coil transition of polyalanine with single histogram method.* **Yasar, F. and Demir, K.** 2006, Computer Physics Communications, Vol. 175, p. 604.
78. *Theory of Helix-Coil Transitions in Polypeptides.* **Gibbs, J. H. and DiMarzio, E. A.** 1958, Journal of Chemical Physics, Vol. 28, p. 1247.

79. *Solvation Model Dependency of Helix-Coil.* **Peng, Yong and Hansmann, Ulrich H. E.** 2002, Biophysical Journal, Vol. 82, p. 3269.
80. *On the Mechanism of Alpha-Helix to Beta-Sheet Transition in the Recombinant Prion Protein.* **Morillas, Manuel, Vanik, David L. and Surewicz, Witold K.** 2001, Biochemistry, Vol. 40, p. 40.
81. *Mechanisms of prion protein assembly into amyloid.* **Stohr, Jan, et al.** 7, 2008, Proceedings of National Acedemy of Sciences, Vol. 105, p. 2409.
82. *Structural diversity of sequentially identical subsequences of proteins: Identical octapeptides can have different conformations.* **Sudarsanam, Sucha.** 1998, Proteins, Vol. 30, p. 228.
83. *Helix versus sheet formation in a small peptide.* **Peng, Yong and Hansmann, Ulrich H. E.** 2003, Vol. 68, p. 041911.
84. *Beta-Hairpin conformation of fibrillogenic peptides: Structure and Alpha-Beta transition mechanism revealed by molecular dynamics simulations.* **Daidone, Isabella, et al.** 1, 2004, Proteins, Vol. 57, p. 198.
85. *Conditioned Self-Avoiding Walk (CSAW): Stochastic Approach to Protein Folding.* **Huang, Kerson.** 2, 2007, Biophysical Reviews and Letters (BRL), Vol. 2, p. 139.
86. *Protein Folding As a Physical Stochastic Process.* **Huang, Kerson.** 1-2, 2008, Biophysical Reviews and Letters (BRL), Vol. 3, p. 1.
87. *Protein folding simulation by all-atom CSAW method.* **Sun, Weitao.** 2007. IEEE International Conference on Bioinformatics and Biomedicine Workshops, 2007. p. 45.
88. *Vibrational Dynamics of Folded Proteins: Significance of Slow and Fast Motions in Relation to Function and Stability.* **Bahar, Ivet, et al.** 12, 1998, Physical Review Letters, Vol. 80, p. 2733.
89. *Stereochemistry of polypeptide chain configurations.* **Ramachandran, G. N., Ramakrishnan, C. and Sasisekharan, V.** 1963, Journal of Molecular Biology, Vol. 7, p. 95.
90. *Restrained Energy Refinement with Two Different ALgorithms and Force Fields of the Structure of the Alpha-Amylase Inhibitor Tendamistat Determined by NMR in Solution.* **Billeter, M., et al.** 1990, Biopolymers, Vol. 29, p. 695.
91. *Vibrational Analysis of Conformation in Peptides, Polypeptides, and Proteins.* **Krimm, S.** 1983, Biopolymers, Vol. 22, p. 217.
92. *Vibrational Analysis of Peptides, Polypeptides, and Proteins. I. Polyglycine I.* **Moore, W. H. and Krimm, S.** 1976, Biopolymers, Vol. 15, p. 2439.
93. *Vibrational Analysis of Peptides, Polypeptides, and Proteins. II. Beta-Poly (L-alanine) and Beta-Poly (L-alanlglycine).* **Moore, W. H. and Krimm, S.** 1976, Biopolymers, Vol. 15, p. 2465.
94. *Vibrational Analysis of Peptides, Polypeptides, and Proteins. 3. Alpha-Poly (L-alanine).* **Rabolt, J. F., Moore, W. H. and Krimm, S.** 5, Macromolecules, Vol. 10, p. 1065.

95. *Vibrational Analysis of Peptides, Polypeptides, and Proteins. V. Normal Vibrations of Beta-Turns.* **Krimm, S. and Bandekar, Jagdeesh.** 1980, *Biopolymers*, Vol. 19, p. 1.
96. *Vibrational Analysis of Peptides, Polypeptides, and Proteins. X. Poly(glycine I) and Its Isotopic Derivatives.* **Dwivedi, Anil M. and Krimm, S.** 1982, *Macromolecules*, Vol. 15, p. 177.
97. *Vibrational Analysis of Peptides, Polypeptides, and Proteins. XI. Beta-Poly(L-alanine) and Its N-Deuterated Derivative.* **Dwivedi, Anil M. and Krimm, S.** 1982, *Macromolecules*, Vol. 15, p. 186.
98. *Vibrational Analysis of Peptides, Polypeptides, and Proteins. XV. Crystalline Polyglycine II.* **Dwivedi, A. M. and Krimm, S.** 1982, *Biopolymers*, Vol. 21, p. 2377.
99. *Vibrational Analysis of Peptides, Polypeptides, and Proteins. XVIII. Conformational Sensitivity of the Alpha-Helix Spectrum: Alpha and Alpha-Poly(L-alanine).* **Dwivedi, Anil M. and Krimm, S.** 1984, *Biopolymers*, Vol. 23, p. 923.
100. *What vibrations tell us about proteins.* **Barth, Andreas and Zscherp, Christian.** 4, 2002, *Quarterly Reviews of Biophysics*, Vol. 35, p. 369.
101. *Effect of short- and long-range interactions on protein folding.* **Anderson, J. S. and Scheraga, H. A.** 4, 1982, *Journal of Protein Chemistry*, Vol. 1, p. 281.
102. *The Protein Data Bank.* **Berman, Helen M., et al.** 1, 2000, *Nucleic Acids Research*, Vol. 28, p. 235.
103. **Dcrjsr.** Ramachandran plot general. *Wikipedia*. [Online] Wikimedia Foundation Inc., January 29, 2003. [Cited: June 4, 2012.] http://en.wikipedia.org/wiki/File:Ramachandran_plot_general_100K.jpg.
104. Ramachandran plot Gly. *Wikipedia*. [Online] Wikimedia Foundation Inc., May 15, 2011. [Cited: June 4, 2012.] http://en.wikipedia.org/wiki/File:Ramachandran_plot_Gly.jpg.
105. Ramachandran plot Pro. *Wikipedia*. [Online] Wikimedia Foundation Inc., May 15, 2011. [Cited: June 4, 2012.] http://en.wikipedia.org/wiki/File:Ramachandran_plot_Pro.jpg.
106. *Structure validation by C-Alpha geometry: Phi, Psi and C-Beta deviation.* **Lovell, Simon C., et al.** 3, 2003, *Proteins: Structure, Function, and Bioinformatics*, Vol. 50, p. 437.
107. *VMD - Visual Molecular Dynamics.* **Humphrey, W., Dalke, A. and Schulten, K.** 1996, *Journal of Molecular Graphics*, Vol. 14, p. 33.
108. *Atomic weights of the elements 2007 (IUPAC Technical Report).* **Wieser, Michael E. and Berglund, Michael.** 11, 2009, Vol. 81, p. 2131.
109. **Finkelstein, A. V. and Ptitsyn, O. B.** *Protein Physics: A Course of Lectures.* London : Academic, 2002.
110. *Vibrational frequencies and modes of Alpha-Helix.* **Itoh, Koichi and Shimanouchi, Takehiko.** 4, 1970, *Biopolymers*, Vol. 9, p. 383.

111. *Understanding normal modes of proteins.* **Krimm, Samuel and Reisdorf, William C.** 1994, Faraday Discussions, Vol. 99, p. 181.
112. *Normal mode spectrum of the parallel-chain Beta-Sheet.* **Bandekar, Jagdeesh and Krimm, S.** 6, 1988, Biopolymers, Vol. 27, p. 909.
113. *Neutron diffraction reveals oxygen-histidine hydrogen bond in oxymyoglobin.* **Phillips, Simon E. V. and Schoenborn, Benno P.** 1981, Nature, Vol. 292, p. 81.
114. *The sequence TGAAKAVALVL from glyceraldehyde-3-phosphate dehydrogenase displays structural ambivalence and interconverts between Alpha-Helix and Beta-hairpin conformations mediated by collapsed conformational states.* **Patel, Sunita, Balaji, Petety V. and Sasidhar, Yellamraju U.** 5, 2007, Journal of Peptide Science, Vol. 13, p. 314.
115. **Schulz, George E.** *Principles of Protein Structure.* New York : Springer-Verlag, 1979.
116. *Investigations on the alpha-helix to coil transition in HP heterogeneous polypeptides using network rigidity.* **Lee, M. S., Wood, G. G. and Jacobs, D. J.** 44, 2004, Journal of Physics: Condensed Matter, Vol. 16, p. S5035.
117. *The mechanism of antiparallel Beta-Sheet formation based on conditioned self-avoiding walk.* **Goh, Boon Chong, et al.** 2012, European Physical Journal E, Vol. 35, p. 27.

Appendix A: Rotation Matrix

Given a vector $\vec{v} = v_x \hat{i} + v_y \hat{j} + v_z \hat{k}$,

Rotation of vector $\vec{v}$ about X-axis is given as:

$$\vec{v}'_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{bmatrix}$$

Rotation of vector $\vec{v}$ about Y-axis is given as:

$$\vec{v}'_y = \begin{bmatrix} \cos \phi & 0 & \sin \phi \\ 0 & 1 & 0 \\ -\sin \phi & 0 & \cos \phi \end{bmatrix}$$

Rotation of vector $\vec{v}$ about Z-axis is given as:

$$\vec{v}'_z = \begin{bmatrix} \cos \alpha & -\sin \alpha & 0 \\ \sin \alpha & \cos \alpha & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Given another vector $\vec{w} = w_x \hat{i} + w_y \hat{j} + w_z \hat{k}$, the rotation of vector $\vec{w}$ about vector $\vec{v}$ for angle given by θ can be written as:

$$\alpha = \tan^{-1} \frac{v_y}{v_x}, \quad \phi = \tan^{-1} \frac{v_z}{\sqrt{v_x^2 + v_y^2}}$$

$$\vec{w}' = \begin{bmatrix} \cos \alpha & \sin \alpha & 0 \\ -\sin \alpha & \cos \alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \cos \phi & 0 & -\sin \phi \\ 0 & 1 & 0 \\ \sin \phi & 0 & \cos \phi \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{bmatrix} \cdot \begin{bmatrix} \cos \phi & 0 & \sin \phi \\ 0 & 1 & 0 \\ -\sin \phi & 0 & \cos \phi \end{bmatrix}$$

$$\cdot \begin{bmatrix} \cos \alpha & -\sin \alpha & 0 \\ \sin \alpha & \cos \alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} w_i \\ w_j \\ w_k \end{bmatrix}$$

We define the above equation in short form as:

$$\vec{w}' = Rot(\vec{v}, \theta) \cdot \vec{w}$$

Thus, $Rot(\vec{v}, \theta)$ is a rotation matrix about the axis of vector $\vec{v}$ for angle θ .

Appendix B: Analytical Solution of Partition Function for polypeptide cranks

Detail workings to integrate Equation 3.22 over p :

$$\begin{aligned}\int e^{-\beta \frac{1}{2} p^T M^{-1} p} dp^{2(n-1)} &= \int e^{-\frac{1}{2} p^T M_{\beta} p} dp^{2(n-1)} \\ &= \sqrt{\frac{(2\pi)^{2(n-1)}}{\det(M_{\beta})}} \\ &= \sqrt{\frac{(2\pi)^{2(n-1)}}{\beta^{2(n-1)} \det(M^{-1})}} \\ &= \sqrt{\frac{(2\pi)^{2(n-1)}}{\beta^{2(n-1)}}} \sqrt{\det M} \\ &= \left(\frac{2\pi}{\beta}\right)^{n-1} \sqrt{\det M}\end{aligned}$$

Appendix C: Effective Potential of Protein Chain Configuration

OVERVIEW

For a given n number residues of protein chain,

Figure 1: Protein Structure

The kinetic partition function of a given protein chain configuration is

$$Z = \int dp_1 \cdots dp_{2(n-1)} \exp\left(-\frac{\beta}{2} p^T M^{-1} p\right) = \left(\frac{4\pi}{\beta}\right)^{n-1} \sqrt{\det M} \quad (1)$$

This gives rise to an effective potential V_{eff} in the configurational partition function:

$$\beta V_{eff} = \frac{1}{2} \ln(\det M) = \frac{1}{2} \text{Tr}(\ln M) \quad (2)$$

β is equals to $1/k_B T$ where k_B is Boltzmann constant ($k=1.38 \times 10^{-23} \text{JK}^{-1} \text{mol}^{-1}$) and T is temperature in Kelvin. M is the “mass matrix” of the protein chain defined as

$$M_{\alpha\beta} = \sum_{i=1}^{n-1} \frac{\partial R_i}{\partial q_\alpha} \frac{\partial R_i}{\partial q_\beta} \quad (3)$$

Where R_i is the position vector of i th C_α with respect to the origin $(0, 0, 0)$ in a protein chain, and q_α & q_β (or q_{ik}) are representing $\{\phi_i, \psi_i\}$, the i th pair of torsional angles in a protein chain. q_{ik} is a vector that could be represented as

$$q_{ik} = \left\{ \begin{array}{c} \phi_1 \\ \psi_1 \\ \phi_2 \\ \psi_2 \\ \phi_3 \\ \psi_3 \\ \vdots \\ \phi_{2(n-1)} \\ \psi_{2(n-1)} \end{array} \right\} \quad (4)$$

Mass matrix $M_{\alpha\beta}$ is represented as:

[illegible]

A program has been written using MATLAB (originally developed from C) to construct the mass matrix and calculate its determinant. The flow chart of the program algorithm is as following:

MASS MATRIX DETERMINANT

2 residues: variation of mass matrix determination with respect to variation in ϕ and ψ angles

A simple test has been conducted using only 2 residues protein chain to observe the variation of mass determination with respect to variation in ϕ and ψ angles. As there are only two residues, therefore the variation only depends on one ϕ and one ψ angles.

(0,0,0)

R_1

Figure 3: 2 Residues Protein

The result is as shown below:

Figure 3: Variation of Mass Matrix Determinant for 2 Residues Protein with Respect to variation of ϕ and ψ angles

It is observed that any change in ϕ angle doesn't affect the magnitude of mass matrix determinant for a fixed value of ψ angle. The variation of ψ angle from 0° to 360° gives a periodic curve with determinant equals zero at $\psi = 0^\circ, 180^\circ$ and 360° .

3 residues: variation of mass matrix determination with respect to variation in ϕ and ψ angles

A protein containing 3 residues is constructed (Figure 4).

Figure 4: Protein with only 3 residues

The variation of mass matrix determinant is observed with respect to the variation of ϕ_1, ψ_1, ϕ_2 and ψ_2 angles. Mass matrix determinant is plotted against torsional angles from 0° to 360° (with 10° interval).

Figure 5(a) shows the variation of mass matrix determinant with respect to variation in ψ_1, ϕ_2 and ψ_2 angles. Different layers of surface plots are due to a variation of ψ_1 angle. For simplification, the variation of mass matrix magnitude with respect to ψ_1 angle is further elaborated in figure 5(b).

Figure 5(a): Variation of Mass Matrix Determinant with respect to variation of ϕ_2 , ψ_2 and ψ_1 angles at a fixed value of ϕ_1 angle

Figure 5(b): Variation of Mass Matrix Determinant with respect to ψ_2 angle at various ψ_1 angle (ϕ_1 and ϕ_2 angles are held fix).

Figure 6(a) shows the variation of mass matrix determinant with respect to ψ_1 , ϕ_2 and ψ_2 angles. The different layers of surface plots are due to the variation in ψ_2 angle. The variation of mass matrix determinant with respect to ψ_2 angle is further elaborated in figure 6(b) with ψ_1 and ϕ_2 angles being held constant.

Figure 6(a): Variation of Mass Matrix Determinant with respect to variation of ϕ_2 , ψ_1 and ψ_2 angles at a fixed value of ϕ_1 angle

Figure 6(b): Variation of Mass Matrix Determinant with respect to ψ_1 angle at various ψ_2 angle (ϕ_1 and ϕ_2 angles are held constant).

Figure 7(a) shows the variation of mass matrix determinant with respect to ϕ_1 , ϕ_2 and ψ_2 angles at a fixed value of ψ_1 angles. The different layers are built up due to a variation in ψ_2 angle. Similar to previous plots of mass matrix determinant vs. ψ angle, the magnitude of mass matrix determinant increases with an increment of ψ_2 angle from 0° to 90° , followed by a decrease with angle increment from 90° to 180° . The angle increment from 180° to 360° performs a period change in mass matrix determinant as in from 0° to 180° .

Figure 7(a): Variation of Mass Matrix Determinant with respect to variation in ϕ_1 , ϕ_2 and ψ_2 angles at a fixed value of ψ_1 angle

Figure 7(b) shows the variation of mass matrix determinant with respect to a variation in all torsional angles. Layers of mass matrix determinant are built up due to variation in ψ_1 and ψ_2 angles.

Figure 7(b): Variation of Mass Matrix Determinant with respect to variation in of ϕ_1 , ψ_1 , ϕ_2 and ψ_2 angles

Figure 8 shows the variation of mass matrix determination with respect to variation in ψ_1 and ψ_2 angles.

Figure 8: Variation of Mass Matrix Determination with respect to variation in ψ_1 and ψ_2 angles (ϕ_1 and ϕ_2 being held constant)

Figure 9(a) shows the variation of mass matrix determination with respect to ϕ_1 and ψ_2 angles at various ϕ_2 angles with the ψ_1 being held constant. The magnitude of the mass matrix determinant differs slightly at different ϕ_2 angles. Figure 9(b) shows the variation of Mass Matrix Determinant with respect to ϕ_1 and ψ_2 angles at various ψ_2 angles with the ϕ_2 angle being held constant. The magnitude of mass matrix determinant increases with an increment of ψ_2 angle from 0° to 90° , followed by a decrease with angle increment from 90° to 180° . The angle increment from 180° to 360° performs a period change in mass matrix determinant as in from 0° to 180° . Figure 9(c) is the variation of mass matrix determinant with respect to different torsional angles configuration, i.e. combination of figure 9(a) and 9(b).

Figure 9(a): Variation of Mass Matrix Determinant with respect to ϕ_1 and ψ_2 angles at various ϕ_2 angles with the ψ_1 angle being held constant

Figure 9(b): Variation of Mass Matrix Determinant with respect to ϕ_1 and ψ_2 angles at various ψ_2 angles with the ϕ_2 angle being held constant

Figure 9(c): Variation of Mass Matrix Determinant with respect to ϕ_1 and ψ_2 angles at various ψ_2 and ϕ_2 angles

Figure 10(a) shows the variation of mass matrix determinant with respect to ϕ_1 and ψ_1 angles at various ϕ_2 angles with the ψ_2 being held constant. Figure 10(b) shows the variation of mass matrix determinant with respect to ϕ_1 and ψ_1 angles at various ψ_2 angles with the ϕ_2 being held constant. Figure 10(c) is a variation of Mass Matrix Determinant with respect to different torsional angles configuration.

Figure 10(a): Variation of mass matrix determinant with respect to ϕ_1 and ψ_1 angles at various ϕ_2 angles with the ψ_2 being held constant

Figure 10(b): variation of mass matrix determinant with respect to ϕ_1 and ψ_1 angles at various ψ_2 angles with the ϕ_2 being held constant

Figure 10(c): variation of mass matrix determinant with respect to ϕ_1 and ψ_1 angles at various ψ_2 and ϕ_2 angles configuration

ANALYSIS OF EFFECTIVE POTENTIAL

From equation (2), effective potential of a protein chain configuration is derived to be related to:

$$\beta V_{eff} = \frac{1}{2} \ln(\det M) \quad (6)$$

The mass matrix determinant of a protein chain configuration has been presented in the previous section. The βV_{eff} of a protein chain configuration could be determined by taking half the natural logarithm of mass matrix determination.

Effective Potential of 2 Residues Protein

Figure 15 shows the βV_{eff} variation of 2 residues protein. The curtain shape seen at ψ equals 0° , 180° and 360° is infinity value due to a natural logarithm of zero value mass matrix determinant.

Figure 15: Variation of effective potential of 2 residues protein with respect to ϕ and ψ angles

Effective Potential of 3 Residues Protein

The variation of βV_{eff} of 3 residues protein is plotted against its torsional angles from 0° to 360° (at 10° interval).

Figure 16 shows the variation of βV_{eff} with respect to variation in ψ_1 , ϕ_2 and ψ_2 angles. Different ψ_1 angle gives rise to different layers of surface mesh. Infinity values occur at ψ_2 equals 0° , 180° and 360° . The bottom layer with large fluctuation is infinity due to natural logarithm of zero value mass matrix determinants at ψ_1 equals 0° , 180° and 360° .

Figure 16: Variation of Effective Potential with respect to ψ_1 , ϕ_2 and ψ_2 angles (ϕ_1 is held constant)

Figure 17 shows the variation in βV_{eff} with respect to variation in ψ_1 , ϕ_2 and ψ_2 angles. Different ψ_2 angle value gives rise to layers of surface mesh with different magnitude. The effective potential is infinity at ψ_1 equals 0° , 180° and 360° . The most bottom layer is infinity due to natural logarithm of zero value mass matrix determinants at ψ_2 equals 0° , 180° and 360°

Figure 17: Variation of Effective Potential with respect to ψ_1 , ϕ_2 and ψ_2 angles (ϕ_1 is held constant)

Figure 18 shows the variation of the βV_{eff} with respect to variation in ψ_1 and ψ_2 angles. Certain plots are due to infinity value at ψ_1 and ψ_2 equals 0° , 180° and 360° .

Figure 18: Variation of Effective Potential with respect to ψ_1 and ψ_2 angles at constant ϕ_1 and ϕ_2 angles

Figure 19 shows the variation of βV_{eff} with respect to ϕ_1 and ϕ_2 angles at different ψ_1 and ψ_2 angles configuration. Figure 19(a) is plotted at a constant ψ_1 angle value. The blue fluctuating layer is infinity value due to natural logarithm of zero value mass matrix determinant at ψ_1 angle equal 0° , 180° and 360° . The yellow layers in figure 19(b) are infinity values due to same reason.

Figure 19(a): Variation of effective potential with respect to ϕ_1 and ϕ_2 angles at different ψ_2 angle (with a constant ψ_1 angles)

Figure 19(b): Variation of effective potential with respect to ϕ_1 and ϕ_2 angles at different ψ_1 and ψ_2 angles

Figure 20 shows the variation of βV_{eff} with respect to ϕ_1 and ψ_2 angles at different ϕ_2 and ψ_1 angles. The effective potential is infinity at ψ_1 or ψ_2 equals 0° , 180° and 360° . The yellow layers seen in figure 20(c) are infinity value (when ψ_1 equals 0° , 180° and 360°).

Figure 20(a): Variation of effective potential with respect to ϕ_1 and ψ_2 (constant ψ_1 value)

Figure 20(b): Variation of effective potential with respect to ϕ_1 and ψ_2 (constant ϕ_2 value)

Figure 20(c): Variation of effective potential with respect to ϕ_1 and ψ_2 (at different ϕ_2 and ψ_1 value)

Figure 21 shows the variation of βV_{eff} with respect to ϕ_1 and ψ_1 angles. The effective potential is infinity at ψ_1 or ψ_1 equals 0° , 180° and 360° . The yellow layers in figure 21(c) are infinity values.

Figure 21(a): Variation of effective potential with respect to ϕ_1 and ψ_1 angles (constant ψ_2 value)

Figure 21(b): Variation of effective potential with respect to ϕ_1 and ψ_1 angles (constant ϕ_2 value)

Figure 21(c): Variation of effective potential with respect to ϕ_1 and ψ_1 angles at different ψ_2 and ϕ_2 values

GENERAL OBSERVATION

Generally, for a given n number residues of protein, a change in ϕ_1 angle doesn't affect the magnitude of its mass matrix determinant or the effective potential. The mass matrix determinant of a protein chain configuration is not constant for any given torsional angles. However, for 2 residues protein and 3 residues as observed above, a periodic mass matrix determinant variation is seen. The variation in ψ angle is believed to be responsible for this periodic determinant (from 0° to 360°). In the given 2 residues and 3 residues cases, mass matrix determinant always equal zero for any ψ equals 0° , 180° or 360° .

Although there is a large difference in mass matrix determinant for any protein chain of different torsional angles, the difference in their effective potentials is not large (except for those values in infinity due to natural logarithm of zero value mass matrix determinants). This is observed in the case for 2 residues and 3 residues protein.

Appendix D: MATLAB Programming Codes

The computation carried out in this thesis is done by MATLAB software. All equations presented in this thesis were coded into MATLAB. All codes were self written without using any MATLAB toolbox.

The following is the programming codes used to calculate the ensemble average number of hydrogen bonds and heat capacity change as a function of temperature (results presented in Chapter 4):

```
tic

%% clear memory and clear screen

clear;
clc;

%% Initialization

n=7;

Total=5^n-1;

%% Constants

kB=1.3806503E-23;
h=6.62606876E-34;
Avo=6.02214199e23;

multiple=5.0;

eneH=-20929/Avo;
eneHA=-20929/Avo;
eneHB=-2*20929*multiple/Avo;

a0=1.525; %bond length Ca-C
b0=1.329; %bond length C-N
c0=1.458; %bond length N-Ca
o0=1.231; %bond length C-O
h0=1.00; %bond length N-H

Atet=109.5*pi/180; %angle between C-Ca and Ca-N
Aab=116.2*pi/180; %angle between Ca-C and C-N
Abc=121.7*pi/180; %angle between Ca-N and C-N
Aoa=120.8*pi/180; %angle between Ca-C and C-O
Ahc=114*pi/180; %angle between Ca-N and N-H
del=Abc-Aab;
dell=Ahc+del;

%% Standard

%standard crank

aa=[a0,0,0];
bb=[b0*cos(pi-Aab),b0*sin(pi-Aab),0];
cc=[c0*cos(del),c0*sin(del),0];
oo=[-o0*cos(Aoa),-o0*sin(Aoa),0];
hh=[h0*cos(dell),h0*sin(dell),0];

%crank vector of alanine
AvCaCb=[0.584454158937995      -0.678156805418597      -1.26701447638317];
AvCaHa=[0.415906136537049      -0.482586311680055      0.901626051847364];
```

```

AvCbHb1=[0.109699818889522      -1.0832785525391      0.164412673916544];
AvCbHb2=[0.989877991623565      0.421226892558287      -0.232970956741365];
AvCbHb3=[-0.663967445979111      0.158928330989027      -0.863858215385498];

%% Main Program

Tmax=2000;

Zi=zeros(Total+1,Tmax+1);
dZdT=zeros(Total+1,Tmax+1);
d2ZdT2=zeros(Total+1,Tmax+1);
hbond_count_i=zeros(Total+1,Tmax+1);
hbond_alpha_count_i=zeros(Total+1,Tmax+1);
hbond_beta_count_i=zeros(Total+1,Tmax+1);
detM=zeros(1,Ttotal+1);
meanMi=zeros(1,Ttotal+1);
meanKi=zeros(1,Ttotal+1);
helix_count_i=zeros(1,Ttotal+1);
hbond_i=zeros(1,Ttotal+1);
hbond_alpha_i=zeros(1,Ttotal+1);
hbond_beta_i=zeros(1,Ttotal+1);

parfor run=0:Total
    run
    phi=zeros(1,n);
    psi=zeros(1,n);

    temp=dec2base(run,5);

    sizeoftemp=size(temp);
    state=zeros(1,n);
    for i=1:sizeoftemp(2)
        state(n-sizeoftemp(2)+i)=base2dec(temp(i),5);
        %state(n-sizeoftemp(2)+i)=base2dec(temp(i),2);
    end

    for i=1:n

        if state(i)==1
            phi(i)=-139;
            psi(i)=135;
        end
        if state(i)==2
            phi(i)=315;
            psi(i)=110;
        end
        if state(i)==3
            phi(i)=105;
            psi(i)=330;
        end
        if state(i)==4
            phi(i)=-57.4;
            psi(i)=-47.5;
        end
        if state(i)==0
            phi(i)=180;
            psi(i)=180;
        end
    end

    phi=phi*pi/180;
    psi=psi*pi/180;

    k0=standardcrank(aa,bb,cc,oo,hh,AvCaCb,AvCaHa,AvCbHb1,AvCbHb2,AvCbHb3);
    chain=flatchain(k0,n,Atet,del);

    iphi=zeros(1,n);
    ipsi=zeros(1,n);

    for i=1:n
        iphi(i)=pi;
        ipsi(i)=pi;
    end

    origin=[0 0 0];
    [R,Rc,Rn,Ro,Rh,CB,HA,HB1,HB2,HB3]=position(chain,n,origin);

```

```

dphi=zeros(1,n);
dpsi=zeros(1,n);

for i=1:n
    dphi(i)=phi(i)-iphi(i);
    dpsi(i)=psi(i)-ipsi(i);
end

[R,Rc,Rn,Ro,Rh,CB,HA,HB1,HB2,HB3,chain]=define(R,Rc,Rn,Ro,Rh,CB,HA,HB1,HB2,HB3,chain,d
phi,dpsi,n);

hbond=0;
hbond_alpha=0;
hbond_beta=0;

for i=1:n
    for j=1:n
        if ((norm(Ro(i,:)-Rh(j,:))-2.0)<=1.0 && (norm(Ro(i,:)-Rh(j,:))-2.0)>=-1.0)
&& (pi-acos(dot((Rh(j,:)-Ro(i,:)),(Rh(j,:)-Rn(j,:)))/(norm(Rh(j,:)-
Ro(i,:))*norm(Rh(j,:)-Rn(j,:))))<=45*pi/180)
            hbond=hbond+1;
            if j==i+3
                hbond_alpha=hbond_alpha+1;
            end
            if (i~=1 && j~=n) && (((norm(Ro(j+1,:)-Rh(i-1,:))-2.0)<=1.0 &&
(norm(Ro(j+1,:)-Rh(i-1,:))-2.0)>=-1.0) && (pi-acos(dot((Rh(i-1,:)-Ro(j+1,:)),(Rh(i-
1,:)-Rn(i-1,:)))/(norm(Rh(i-1,:)-Ro(j+1,:))*norm(Rh(i-1,:)-Rn(i-1,:))))<=45*pi/180))
                hbond_beta=hbond_beta+1;
            end
        end
    end
end

hbond_i(run+1)=hbond;
hbond_alpha_i(run+1)=hbond_alpha;
hbond_beta_i(run+1)=hbond_beta;

delta=0.001*pi/180;

dRdphi=zeros(n,3);
dRdpsi=zeros(n,3);
dodphi=zeros(n,3);
dodpsi=zeros(n,3);
dcdphi=zeros(n,3);
dcdpsi=zeros(n,3);
dndphi=zeros(n,3);
dndpsi=zeros(n,3);
dhdpfi=zeros(n,3);
dhdpfi=zeros(n,3);
dCBdphi=zeros(n,3);
dCBdpsi=zeros(n,3);
dHAdphi=zeros(n,3);
dHAdpsi=zeros(n,3);
dHB1dphi=zeros(n,3);
dHB1dpsi=zeros(n,3);
dHB2dphi=zeros(n,3);
dHB2dpsi=zeros(n,3);
dHB3dphi=zeros(n,3);
dHB3dpsi=zeros(n,3);

dRP=zeros(n,3);
dRN=zeros(n,3);

Q=zeros(2*(n-1),3,n);
Qc=zeros(2*(n-1),3,n);
Qo=zeros(2*(n-1),3,n);
Qn=zeros(2*(n-1),3,n);
Qh=zeros(2*(n-1),3,n);
QCB=zeros(2*(n-1),3,n);
QHA=zeros(2*(n-1),3,n);
QHB1=zeros(2*(n-1),3,n);
QHB2=zeros(2*(n-1),3,n);
QHB3=zeros(2*(n-1),3,n);

```

```

QD=zeros(hbond,2*(n-1));
QC=zeros(hbond,2*(n-1));

for i=2:n

    k0=standardcrank(aa,bb,cc,oo,hh,AvCaCb,AvCaHa,AvCbHb1,AvCbHb2,AvCbHb3);
    chain=flatchain(k0,n,Atet,del);
    [RP,RcP,RnP,RoP,RhP,CBP,HAP,HB1P,HB2P,HB3P]=position(chain,n,origin);

[RP,RcP,RnP,RoP,RhP,CBP,HAP,HB1P,HB2P,HB3P,chain]=define(RP,RcP,RnP,RoP,RhP,CBP,HAP,HB
1P,HB2P,HB3P,chain,dphi,dpsi,n);

    for k=6:10
        chain(k,:,i-1)=rot(chain(3,:,i-1),delta)*chain(k,:,i-1)';
    end
    CBP(i-1,:)=RP(i-1,:)+chain(6,:,i-1);
    HAP(i-1,:)=RP(i-1,:)+chain(7,:,i-1);
    HB1P(i-1,:)=CBP(i-1,:)+chain(8,:,i-1);
    HB2P(i-1,:)=CBP(i-1,:)+chain(9,:,i-1);
    HB3P(i-1,:)=CBP(i-1,:)+chain(10,:,i-1);
    for j=i:n-1
        for k=1:10
            chain(k,:,j)=rot(chain(3,:,i-1),delta)*chain(k,:,j)';
        end
        RcP(j,:)=RP(j-1,:)+chain(1,:,j);
        RnP(j,:)=RP(j-1,:)+chain(1,:,j)+chain(2,:,j);
        RP(j,:)=RP(j-1,:)+chain(1,:,j)+chain(2,:,j)+chain(3,:,j);
        RoP(j,:)=RP(j-1,:)+chain(1,:,j)+chain(4,:,j);
        RhP(j,:)=RP(j-1,:)+chain(1,:,j)+chain(2,:,j)+chain(5,:,j);
        CBP(j,:)=RP(j,:)+chain(6,:,j);
        HAP(j,:)=RP(j,:)+chain(7,:,j);
        HB1P(j,:)=CBP(j,:)+chain(8,:,j);
        HB2P(j,:)=CBP(j,:)+chain(9,:,j);
        HB3P(j,:)=CBP(j,:)+chain(10,:,j);
    end

    for k=1:5
        chain(k,:,n)=rot(chain(3,:,i-1),delta)*chain(k,:,n)';
    end
    RcP(n,:)=RP(n-1,:)+chain(1,:,n);
    RnP(n,:)=RP(n-1,:)+chain(1,:,n)+chain(2,:,n);
    RP(n,:)=RP(n-1,:)+chain(1,:,n)+chain(2,:,n)+chain(3,:,n);
    RoP(n,:)=RP(n-1,:)+chain(1,:,n)+chain(4,:,n);
    RhP(n,:)=RP(n-1,:)+chain(1,:,n)+chain(2,:,n)+chain(5,:,n);

    k0=standardcrank(aa,bb,cc,oo,hh,AvCaCb,AvCaHa,AvCbHb1,AvCbHb2,AvCbHb3);
    chain=flatchain(k0,n,Atet,del);
    [RN,RcN,RnN,RoN,RhN,CBN,HAN,HB1N,HB2N,HB3N]=position(chain,n,origin);

[RN,RcN,RnN,RoN,RhN,CBN,HAN,HB1N,HB2N,HB3N,chain]=define(RN,RcN,RnN,RoN,RhN,CBN,HAN,HB
1N,HB2N,HB3N,chain,dphi,dpsi,n);

    for k=6:10
        chain(k,:,i-1)=rot(chain(3,:,i-1),-delta)*chain(k,:,i-1)';
    end
    CBN(i-1,:)=RN(i-1,:)+chain(6,:,i-1);
    HAN(i-1,:)=RN(i-1,:)+chain(7,:,i-1);
    HB1N(i-1,:)=CBN(i-1,:)+chain(8,:,i-1);
    HB2N(i-1,:)=CBN(i-1,:)+chain(9,:,i-1);
    HB3N(i-1,:)=CBN(i-1,:)+chain(10,:,i-1);
    for j=i:n-1
        for k=1:10
            chain(k,:,j)=rot(chain(3,:,i-1),-delta)*chain(k,:,j)';
        end
        RcN(j,:)=RN(j-1,:)+chain(1,:,j);
        RnN(j,:)=RN(j-1,:)+chain(1,:,j)+chain(2,:,j);
        RN(j,:)=RN(j-1,:)+chain(1,:,j)+chain(2,:,j)+chain(3,:,j);
        RoN(j,:)=RN(j-1,:)+chain(1,:,j)+chain(4,:,j);
        RhN(j,:)=RN(j-1,:)+chain(1,:,j)+chain(2,:,j)+chain(5,:,j);
        CBN(j,:)=RN(j,:)+chain(6,:,j);
        HAN(j,:)=RN(j,:)+chain(7,:,j);
        HB1N(j,:)=CBN(j,:)+chain(8,:,j);
        HB2N(j,:)=CBN(j,:)+chain(9,:,j);
        HB3N(j,:)=CBN(j,:)+chain(10,:,j);
    end
end

```



```

for k=1:5
    chain(k,:,n)=rot(chain(3,:,i-1),-delta)*chain(k,:,n)';
end
RcN(n,:)=RN(n-1,:)+chain(1,:,n);
RnN(n,:)=RN(n-1,:)+chain(1,:,n)+chain(2,:,n);
RN(n,:)=RN(n-1,:)+chain(1,:,n)+chain(2,:,n)+chain(3,:,n);
RoN(n,:)=RN(n-1,:)+chain(1,:,n)+chain(4,:,n);
RhN(n,:)=RN(n-1,:)+chain(1,:,n)+chain(2,:,n)+chain(5,:,n);

for k=1:n
    dRdphi(k,:)=(Rp(k,:)-Rn(k,:))/(2*delta);
    dcdphi(k,:)=(RcP(k,:)-RcN(k,:))/(2*delta);
    dodphi(k,:)=(RoP(k,:)-RoN(k,:))/(2*delta);
    dndphi(k,:)=(RnP(k,:)-RnN(k,:))/(2*delta);
    dhdp(k,:)=(RhP(k,:)-RhN(k,:))/(2*delta);
    dCBdp(k,:)=(CBP(k,:)-CBN(k,:))/(2*delta);
    dHADp(k,:)=(HAP(k,:)-HAN(k,:))/(2*delta);
    dHB1dp(k,:)=(HB1P(k,:)-HB1N(k,:))/(2*delta);
    dHB2dp(k,:)=(HB2P(k,:)-HB2N(k,:))/(2*delta);
    dHB3dp(k,:)=(HB3P(k,:)-HB3N(k,:))/(2*delta);
end

for k=1:n
    Q(2*(i-1)-1,:,k)=dRdphi(k,:);
    Qc(2*(i-1)-1,:,k)=dcdphi(k,:);
    Qo(2*(i-1)-1,:,k)=dodphi(k,:);
    Qn(2*(i-1)-1,:,k)=dndphi(k,:);
    Qh(2*(i-1)-1,:,k)=dhdp(k,:);
    QCB(2*(i-1)-1,:,k)=dCBdp(k,:);
    QHA(2*(i-1)-1,:,k)=dHADp(k,:);
    QHB1(2*(i-1)-1,:,k)=dHB1dp(k,:);
    QHB2(2*(i-1)-1,:,k)=dHB2dp(k,:);
    QHB3(2*(i-1)-1,:,k)=dHB3dp(k,:);
end

k=1;
for ii=1:n
    for jj=1:n
        if ((norm(Ro(ii,:)-Rh(jj,:))-2.0)<=1.0 && (norm(Ro(ii,:)-Rh(jj,:))-
2.0)>=-1.0) && (pi-acos(dot((Rh(jj,:)-Ro(ii,:)),(Rh(jj,:)-Rn(jj,:)))/(norm(Rh(jj,:)-
Ro(ii,:))*norm(Rh(jj,:)-Rn(jj,:))))<=45*pi/180)
            b=Rh(jj,:)-Ro(ii,:);
            unitb=b/norm(b);
            dRP=RhP(jj,:)-RoP(ii,:);
            dRN=RhN(jj,:)-RoN(ii,:);
            QD(k,2*(i-1)-1)=norm(dot(unitb,(dRP-dRN)/(2*delta)));
            QC(k,2*(i-1)-1)=norm(cross(unitb,(dRP-dRN)/(2*delta)));
            k=k+1;
        end
    end
end

end

for i=2:n

    k0=standardcrank(aa,bb,cc,oo,hh,AvCaCb,AvCaHa,AvCbHb1,AvCbHb2,AvCbHb3);
    chain=flatchain(k0,n,Atet,delta);
    [RP,RcP,RnP,RoP,RhP,CBP,HAP,HB1P,HB2P,HB3P]=position(chain,n,origin);

[RP,RcP,RnP,RoP,RhP,CBP,HAP,HB1P,HB2P,HB3P,chain]=define(RP,RcP,RnP,RoP,RhP,CBP,HAP,HB
1P,HB2P,HB3P,chain,dphi,dpsi,n);

    for j=i:n-1
        for k=1:10
            chain(k,:,j)=rot(chain(1,:,i),delta)*chain(k,:,j)';
        end
        RcP(j,:)=RP(j-1,:)+chain(1,:,j);
        RnP(j,:)=RP(j-1,:)+chain(1,:,j)+chain(2,:,j);
        RP(j,:)=RP(j-1,:)+chain(1,:,j)+chain(2,:,j)+chain(3,:,j);
        RoP(j,:)=RP(j-1,:)+chain(1,:,j)+chain(4,:,j);
        RhP(j,:)=RP(j-1,:)+chain(1,:,j)+chain(2,:,j)+chain(5,:,j);
        CBP(j,:)=RP(j,:)+chain(6,:,j);
        HAP(j,:)=RP(j,:)+chain(7,:,j);
        HB1P(j,:)=CBP(j,:)+chain(8,:,j);

```

```

        HB2P(j,:)=CBP(j,:)+chain(9,:,j);
        HB3P(j,:)=CBP(j,:)+chain(10,:,j);
    end

    for k=1:5
        chain(k,:,n)=rot(chain(1,:,i),delta)*chain(k,:,n)';
    end
    RcP(n,:)=RP(n-1,:)+chain(1,:,n);
    RnP(n,:)=RP(n-1,:)+chain(1,:,n)+chain(2,:,n);
    RP(n,:)=RP(n-1,:)+chain(1,:,n)+chain(2,:,n)+chain(3,:,n);
    RoP(n,:)=RP(n-1,:)+chain(1,:,n)+chain(4,:,n);
    RhP(n,:)=RP(n-1,:)+chain(1,:,n)+chain(2,:,n)+chain(5,:,n);

    k0=standardcrank(aa,bb,cc,oo,hh,AvCaCb,AvCaHa,AvCbHb1,AvCbHb2,AvCbHb3);
    chain=flatchain(k0,n,Atet,del);
    [RN,RcN,RnN,RoN,RhN,CBN,HAN,HB1N,HB2N,HB3N]=position(chain,n,origin);

[RN,RcN,RnN,RoN,RhN,CBN,HAN,HB1N,HB2N,HB3N,chain]=define(RN,RcN,RnN,RoN,RhN,CBN,HAN,HB
1N,HB2N,HB3N,chain,dphi,dpsi,n);

    for j=i:n-1
        for k=1:10
            chain(k,:,j)=rot(chain(1,:,i),-delta)*chain(k,:,j)';
        end
        RcN(j,:)=RN(j-1,:)+chain(1,:,j);
        RnN(j,:)=RN(j-1,:)+chain(1,:,j)+chain(2,:,j);
        RN(j,:)=RN(j-1,:)+chain(1,:,j)+chain(2,:,j)+chain(3,:,j);
        RoN(j,:)=RN(j-1,:)+chain(1,:,j)+chain(4,:,j);
        RhN(j,:)=RN(j-1,:)+chain(1,:,j)+chain(2,:,j)+chain(5,:,j);
        CBN(j,:)=RN(j,:)+chain(6,:,j);
        HAN(j,:)=RN(j,:)+chain(7,:,j);
        HB1N(j,:)=CBN(j,:)+chain(8,:,j);
        HB2N(j,:)=CBN(j,:)+chain(9,:,j);
        HB3N(j,:)=CBN(j,:)+chain(10,:,j);
    end

    for k=1:5
        chain(k,:,n)=rot(chain(1,:,i),-delta)*chain(k,:,n)';
    end
    RcN(n,:)=RN(n-1,:)+chain(1,:,n);
    RnN(n,:)=RN(n-1,:)+chain(1,:,n)+chain(2,:,n);
    RN(n,:)=RN(n-1,:)+chain(1,:,n)+chain(2,:,n)+chain(3,:,n);
    RoN(n,:)=RN(n-1,:)+chain(1,:,n)+chain(4,:,n);
    RhN(n,:)=RN(n-1,:)+chain(1,:,n)+chain(2,:,n)+chain(5,:,n);

    for k=1:n
        dRdpsi(k,:)=(RP(k,:)-RN(k,:))/(2*delta);
        dcdpsi(k,:)=(RcP(k,:)-RcN(k,:))/(2*delta);
        dodpsi(k,:)=(RoP(k,:)-RoN(k,:))/(2*delta);
        dndpsi(k,:)=(RnP(k,:)-RnN(k,:))/(2*delta);
        dhdpai(k,:)=(RhP(k,:)-RhN(k,:))/(2*delta);
        dCBdpsi(k,:)=(CBP(k,:)-CBN(k,:))/(2*delta);
        dHADpsi(k,:)=(HAP(k,:)-HAN(k,:))/(2*delta);
        dHB1dpsi(k,:)=(HB1P(k,:)-HB1N(k,:))/(2*delta);
        dHB2dpsi(k,:)=(HB2P(k,:)-HB2N(k,:))/(2*delta);
        dHB3dpsi(k,:)=(HB3P(k,:)-HB3N(k,:))/(2*delta);
    end

    for k=1:n
        Q(2*(i-1),:,k)=dRdpsi(k,:);
        Qc(2*(i-1),:,k)=dcdpsi(k,:);
        Qo(2*(i-1),:,k)=dodpsi(k,:);
        Qn(2*(i-1),:,k)=dndpsi(k,:);
        Qh(2*(i-1),:,k)=dhdpai(k,:);
        QCB(2*(i-1),:,k)=dCBdpsi(k,:);
        QHA(2*(i-1),:,k)=dHADpsi(k,:);
        QHB1(2*(i-1),:,k)=dHB1dpsi(k,:);
        QHB2(2*(i-1),:,k)=dHB2dpsi(k,:);
        QHB3(2*(i-1),:,k)=dHB3dpsi(k,:);
    end

    k=1;
    for ii=1:n
        for jj=1:n

```

```

        if ((norm(Ro(ii,:)-Rh(jj,:))-2.0)<=1.0 && (norm(Ro(ii,:)-Rh(jj,:))-
2.0)>=-1.0) && (pi-acos(dot((Rh(jj,:)-Ro(ii,:)),(Rh(jj,:)-Rn(jj,:)))/(norm(Rh(jj,:)-
Ro(ii,:))*norm(Rh(jj,:)-Rn(jj,:))))<=45*pi/180)
            b=Rh(jj,:)-Ro(ii,:);
            unitb=b/norm(b);
            dRP=RhP(jj,:)-RoP(ii,:);
            dRN=RhN(jj,:)-RoN(ii,:);
            QD(k,2*(i-1))=norm(dot(unitb,(dRP-dRN)/(2*delta)));
            QC(k,2*(i-1))=norm(cross(unitb,(dRP-dRN)/(2*delta)));
            k=k+1;
        end
    end
end

end

RealMass=zeros(2*(n-1),2*(n-1));
mass=zeros(2*(n-1),2*(n-1));
DR=zeros(2*(n-1),2*(n-1));
CR=zeros(2*(n-1),2*(n-1));
RR=zeros(2*(n-1),2*(n-1));

for i=1:2*(n-1)
    for j=1:2*(n-1)
        for k=1:n-1
            RealMass(i,j)=1.9934e-26*dot(Q(i,:),k),Q(j,:),k))+1.9934e-
26*dot(Qc(i,:),k),Qc(j,:),k))+2.6578e-26*dot(Qo(i,:),k),Qo(j,:),k))+2.3256e-
27*dot(Qn(i,:),k),Qn(j,:),k))+1.6611e-27*dot(Qh(i,:),k),Qh(j,:),k))+1.9934e-
26*dot(QCB(i,:),k),QCB(j,:),k))+1.6611e-27*dot(QHA(i,:),k),QHA(j,:),k))+1.6611e-
27*(dot(QHB1(i,:),k),QHB1(j,:),k))+dot(QHB2(i,:),k),QHB2(j,:),k))+dot(QHB3(i,:),k),QHB3(j,:
,k)))+RealMass(i,j);
            mass(i,j)=dot(Q(i,:),k),Q(j,:),k))+mass(i,j);
        end
        RealMass(i,j)=1.9934e-26*dot(Q(i,:),n),Q(j,:),n))+1.9934e-
26*dot(Qc(i,:),n),Qc(j,:),n))+2.6578e-26*dot(Qo(i,:),n),Qo(j,:),n))+2.3256e-
27*dot(Qn(i,:),n),Qn(j,:),n))+1.6611e-27*dot(Qh(i,:),n),Qh(j,:),n))+RealMass(i,j);
    end
end

for i=1:2*(n-1)
    for j=1:2*(n-1)
        for k=1:hbond
            DR(i,j)=QD(k,i)*QD(k,j)+DR(i,j);
            CR(i,j)=QC(k,i)*QC(k,j)+CR(i,j);
        end
    end
end

MassDet=det(mass);
RealMassDet=det(RealMass);

k1=13;      %force constant 1

k2=3.0;     %force constant 2

MK=RealMass\((k1*DR+k2*CR); %Real mass
M=RealMass;
K=k1*DR+k2*CR;

detM(run+1)=det(M);

[EVEC EVAL]=eig(MK);

omega=sqrt(eig(MK));

frequency=omega/(2*pi);

wavenumber=1/3e8*(frequency);

eigM=eig(inv(M));
omegaM=sqrt(eigM)/3e8;
meanM=prod(omegaM)^(1/(2*(n-1)))*3e8;

eigK=eig(K);
prodK=1;
omegaK=sqrt(eigK);

```

```

for i=2*(n-1)-2*hbond+1:2*(n-1)
    prodK=prodK*omegaK(i);
end

if hbond~=0
    meanK=prodK^(1/(2*hbond));
    meanK=meanK*(1e-10);
    meanKi(run+1)=meanK;
end

T=zeros(1,Tmax+1);
for i=1:Tmax+1
    T(i)=i-1;
end

meanM=meanM/(1e-10);
meanMi(run+1)=meanM;
end
%%

mM=mean(meanMi(:));

parfor run=0:Total
    Z=zeros(1,Tmax+1);
    dZdT=zeros(1,Tmax+1);
    d2ZdT2=zeros(1,Tmax+1);
    hbond_count=zeros(1,Tmax+1);
    hbond_alpha_count=zeros(1,Tmax+1);
    hbond_beta_count=zeros(1,Tmax+1);

    T=zeros(1,Tmax+1);
    for i=1:Tmax+1
        T(i)=i-1;
    end

    if hbond_i(run+1)~=0
        for i=2:Tmax+1
            Z(i)=1/(meanMi(run+1)/mM)^(2*(n-1))*
            (2*pi*kB*T(i))^hbond_i(run+1)/(meanKi(run+1))^(2*hbond_i(run+1))*(2*pi)^(2*(n-1)-2*hbond_i(run+1))*exp(-
            (hbond_alpha_i(run+1)*eneHA+hbond_beta_i(run+1)*eneHB+hbond_i(run+1)*eneH)/(kB*T(i)));
            dZdT(i)=(mM/meanMi(run+1))^(2*(n-1))*(2*pi)^(2*(n-1)-2*hbond_i(run+1))/(meanKi(run+1))^(2*hbond_i(run+1))*(hbond_i(run+1)*(2*pi*kB*T(i))^(hbond_i(run+1)-1)*(2*pi*kB)*exp(-
            (hbond_alpha_i(run+1)*eneHA+hbond_beta_i(run+1)*eneHB+hbond_i(run+1)*eneH)/(kB*T(i)))+(2*pi*kB*T(i))^hbond_i(run+1)*exp(-
            (hbond_alpha_i(run+1)*eneHA+hbond_beta_i(run+1)*eneHB+hbond_i(run+1)*eneH)/(kB*T(i)))*
            (-
            (hbond_alpha_i(run+1)*eneHA+hbond_beta_i(run+1)*eneHB+hbond_i(run+1)*eneH)/(kB*T(i)^2)
            ));
            d2ZdT2(i)=(mM/meanMi(run+1))^(2*(n-1))*(2*pi)^(2*(n-1)-2*hbond_i(run+1))/(meanKi(run+1))^(2*hbond_i(run+1))*(hbond_i(run+1)*(hbond_i(run+1)-1)*(2*pi*kB*T(i))^(hbond_i(run+1)-2)*(2*pi*kB)^2*exp(-
            (hbond_alpha_i(run+1)*eneHA+hbond_beta_i(run+1)*eneHB+hbond_i(run+1)*eneH)/(kB*T(i)))+(hbond_i(run+1)*(2*pi*kB*T(i))^(hbond_i(run+1)-1)*(2*pi*kB)*exp(-
            (hbond_alpha_i(run+1)*eneHA+hbond_beta_i(run+1)*eneHB+hbond_i(run+1)*eneH)/(kB*T(i)))*
            (-
            (hbond_alpha_i(run+1)*eneHA+hbond_beta_i(run+1)*eneHB+hbond_i(run+1)*eneH)/(kB*T(i)^2)
            )+hbond_i(run+1)*(2*pi*kB*T(i))^(hbond_i(run+1)-1)*(2*pi*kB)*exp(-
            (hbond_alpha_i(run+1)*eneHA+hbond_beta_i(run+1)*eneHB+hbond_i(run+1)*eneH)/(kB*T(i)))*
            (-
            (hbond_alpha_i(run+1)*eneHA+hbond_beta_i(run+1)*eneHB+hbond_i(run+1)*eneH)/(kB*T(i)^2)
            )+(2*pi*kB*T(i))^hbond_i(run+1)*exp(-
            (hbond_alpha_i(run+1)*eneHA+hbond_beta_i(run+1)*eneHB+hbond_i(run+1)*eneH)/(kB*T(i)))*
            (-
            (hbond_alpha_i(run+1)*eneHA+hbond_beta_i(run+1)*eneHB+hbond_i(run+1)*eneH)/(kB*T(i)^2)
            )^2+(2*pi*kB*T(i))^hbond_i(run+1)*exp(-
            (hbond_alpha_i(run+1)*eneHA+hbond_beta_i(run+1)*eneHB+hbond_i(run+1)*eneH)/(kB*T(i)))*
            ((2*(hbond_alpha_i(run+1)*eneHA+hbond_beta_i(run+1)*eneHB+hbond_i(run+1)*eneH)/(kB*T(i)^3))));

            hbond_count(i)=hbond_i(run+1)*Z(i);
            if hbond_alpha_i(run+1)~=0
                hbond_alpha_count(i)=hbond_alpha_i(run+1)*Z(i);
            end
        end
    end
end

```

```

        if hbond_beta_i(run+1)~=0
            hbond_beta_count(i)=hbond_beta_i(run+1)*Z(i);
        end
    end
else
    for i=2:Tmax+1
        Z(i)=1/(meanMi(run+1)/mM)^(2*(n-1))*(2*pi)^(2*(n-1));
        hbond_count(i)=hbond_i(run+1)*Z(i);
        hbond_alpha_count(i)=hbond_alpha_i(run+1)*Z(i);
        hbond_beta_count(i)=hbond_beta_i(run+1)*Z(i);
    end
end

Zi(run+1,:)=Z(:);
dZdT(run+1,:)=dZdT(:);
d2ZdT2(run+1,:)=d2ZdT2(:);
hbond_count_i(run+1,:)=hbond_count(:);
hbond_alpha_count_i(run+1,:)=hbond_alpha_count(:);
hbond_beta_count_i(run+1,:)=hbond_beta_count(:);
end

%%

T=zeros(1,Tmax+1);
parfor i=1:Tmax+1
    T(i)=i-1;
end

HeatCapacity=zeros(1,Tmax+1);
ZTotal=zeros(1,Tmax+1);
dZdTTotal=zeros(1,Tmax+1);
d2ZdT2Total=zeros(1,Tmax+1);

ZTotal(:)=sum(Zi(:,:));
dZdTTotal(:)=sum(dZdT(:,:));
d2ZdT2Total(:)=sum(d2ZdT2i(:,:));

HTotal=zeros(1,Tmax+1);
HATotal=zeros(1,Tmax+1);
HBTTotal=zeros(1,Tmax+1);
HTotal(:)=sum(hbond_count_i(:,:));
HATotal(:)=sum(hbond_alpha_count_i(:,:));
HBTTotal(:)=sum(hbond_beta_count_i(:,:));

hbond_count=zeros(1,Tmax+1);
hbond_alpha_count=zeros(1,Tmax+1);
hbond_beta_count=zeros(1,Tmax+1);
parfor i=1:Tmax+1
    a=2*((n-1)*(2*pi*kB*T(i))^(n-2)*(2*pi*kB)*ZTotal(i)+(2*pi*kB*T(i))^(n-1)*dZdTTotal(i));
    b=T(i)/((2*pi*kB*T(i))^(n-1)*ZTotal(i))*((n-1)*(2*pi*kB*T(i))^(n-2)*(2*pi*kB)*ZTotal(i)+(2*pi*kB*T(i))^(n-1)*dZdTTotal(i))^2;
    c=T(i)*((n-1)*(n-2)*(2*pi*kB*T(i))^(n-3)*(2*pi*kB)^2*ZTotal(i)+2*(n-1)*(2*pi*kB*T(i))^(n-2)*(2*pi*kB)*dZdTTotal(i)+(2*pi*kB*T(i))^(n-1)*d2ZdT2Total(i));
    HeatCapacity(i)=kB*T(i)/((2*pi*kB*T(i))^(n-1)*ZTotal(i))*(a-b+c);
    hbond_count(i)=HTotal(i)/ZTotal(i);
    hbond_alpha_count(i)=HATotal(i)/ZTotal(i);
    hbond_beta_count(i)=HBTTotal(i)/ZTotal(i);
end

ElapsedTime=toc;

```