INVESTIGATING NONNATIVE CONTACTS IN PROTEIN FOLDING

A Thesis

Presented to

The Graduate Faculty of The University of Akron

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

Chong Chen

May, 2009
ABSTRACT

Proteins are organic compounds, consisting of amino acids (residues) bound by peptide bonds into polypeptide chains, and can fold into their unique functional structures (native state) without any chaperones. There is more than one simple model of proteins, among which $G\delta$ model, HP model and “perturbed homopolymer model” are widely used. Amino acids in these models are considered as beads lying on the sites of a lattice. By extracting information from the system with different kind of computer algorithms, one hopes to predict the folding process.

How and why do protein fold in this particular way is still not clear although it has been extensively investigated for more than half a century. Experimental techniques (NMR, X-ray crystallography, etc) as well as computer aided theoretical works (energy landscape, homology, etc) have been used to understand how a protein starts from its amino acids sequence into a functional structure.

Nonnative contacts are defined as contacts (nearest-neighbor) not found in the native state, and can be used to study the process of folding. According to the recent literature, nonnative contacts affect protein folding rates, and make contributions to the unfolding pathways. Folding intermediates are also found rich in nonnative contacts.
We have investigated nonnative contacts for proteins in a modified HP model by generating all possible conformations (enumeration) which enable us to carry out exact calculation for the nonnative contact density $n_c(e)$ as a function of the energy $e$ as well as thermodynamically average nonnative contacts $\overline{n}_c(T)$ as a function of temperature of $T$. We have investigated different sequences to understand the sequence dependence of $n_c(e)$ and $\overline{n}_c(T)$. These results provide us with new understandings of the role for nonnative contacts play in the protein folding process, as are listed below:

1) The density $n_c(e)$ is always monotonically increasing in the standard model. In weakly interacting and strongly interacting model, the density $n_c(e)$ is spread out, but generally increasing.

2) The $\overline{n}_c(T)$ for all models is generally monotonically increasing. A few violations can be seen.

3) The protein property is sequence dependent.
### TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
</tbody>
</table>

### CHAPTER

I. INTRODUCTION

1.1 Protein structures and functions

1.1.1 Amino Acids

1.1.2 Primary structure or sequence

1.1.3 Secondary structure

1.1.3.1 The α helix

1.1.3.2 The β sheet

1.1.4 Tertiary structure

1.1.4.1 Disulfide bridges

1.1.4.2 The hydrophobic effect

1.1.4.3 Charge-charge interactions

1.1.4.4 Hydrogen bonding

1.1.4.5 Van der Waals interactions

1.1.5 Quaternary structure

1.2 The protein folding problem
1.2.1 Anfinsen’s experiment ........................................ 12
1.2.2 Levinthal’s paradox ........................................... 14
1.2.3 Energy landscape ............................................. 14
1.3 Protein structure determination and theoretical models ........ 15
   1.3.1 Methods used for determining protein structure .......... 15
   1.3.2 Theoretical Models ........................................ 16
       1.3.2.1 The $\bar{G_o}$ model ................................ 17
       1.3.2.2 The Gaussian model ................................. 18
       1.3.2.3 The HP model ....................................... 18
II. EXACT STATISTICAL MECHANICAL INVESTIGATION OF
     A FINITE PROTEIN ........................................... 20
   2.1 Folded and unfolded protein .................................. 20
   2.2 The protein model used for our research ................. 20
       2.2.1 Protein as a small system ............................. 21
       2.2.2 Structures involved in our research ............... 23
       2.2.3 Anchored Protein ...................................... 24
       2.2.4 Interaction energies ................................... 26
           2.2.4.1 The standard model ............................. 28
           2.2.4.2 Weakly perturbed model (B) .................... 28
           2.2.4.3 Strongly perturbed model (C) ................. 29
   2.3 Small system thermodynamics ................................ 30
       2.3.1 Microcanonical entropy ............................... 30
2.3.2 Canonical partition function ........................................... 31

2.4 The definition of distance between conformations .................... 32

III. CURRENT WORK: NONNATIVE CONTACTS IN PROTEIN FOLDING .... 34

3.1 Folded and unfolded state .................................................. 34

3.1.1 The unfolded (denatured) state ....................................... 34

3.1.2 The folded (native) state ............................................... 34

3.1.3 The definition of nonnative contacts ................................ 35

3.2 The role played by nonnative contacts ................................ 36

3.2.1 Nonnative contacts affect protein folding rates .................... 36

3.2.2 Nonnative contacts make great contributions to the unfolding pathways ................................................................. 37

3.2.3 Folding intermediates are rich in nonnative contacts ............ 37

3.3 Statistical Mechanical properties of Nonnative contacts ............. 37

3.3.1 Nonnative contacts density with respect of energy for a standard sequence ................................................................. 38

3.3.2 Nonnative contact density as a function of temperature ...... 43

3.3.3 Sequence dependence ..................................................... 47

3.4 Summary ................................................................................. 54

3.5 Future Work ............................................................................ 55

BIBLIOGRAPHY .............................................................................. 56
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Possible Models and their Parameters</td>
<td>28</td>
</tr>
</tbody>
</table>

viii
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Dehydration synthesis (condensation) reaction forming an amide...</td>
<td>4</td>
</tr>
<tr>
<td>1.2 Three- and single-letter codes for amino acid sequence of a primary structure</td>
<td>5</td>
</tr>
<tr>
<td>1.3 A regular $\alpha$ helix structure</td>
<td>6</td>
</tr>
<tr>
<td>1.4 A regular $\beta$ sheet structure</td>
<td>8</td>
</tr>
<tr>
<td>1.5 A schematic diagram of Anfinsen's experiment</td>
<td>13</td>
</tr>
<tr>
<td>1.6 The protein folding process</td>
<td>13</td>
</tr>
<tr>
<td>1.7 A sketch of an energy landscape</td>
<td>15</td>
</tr>
<tr>
<td>2.1 A compact protein with water surrounded on a square lattice.</td>
<td>22</td>
</tr>
<tr>
<td>2.2 A 2-d model of a finite protein on a square lattice</td>
<td>23</td>
</tr>
<tr>
<td>2.3 Two kinds of extreme conformations in our model</td>
<td>24</td>
</tr>
<tr>
<td>2.4 Log-log plot of the number of conformation with respect to number of residues</td>
<td>25</td>
</tr>
<tr>
<td>2.5 A hairpin turn</td>
<td>27</td>
</tr>
<tr>
<td>2.6 The behavior of $s(e)$ as a function of the protein size M</td>
<td>31</td>
</tr>
<tr>
<td>3.1 A schematic plot for native and nonnative state for a protein with 36 residues...</td>
<td>35</td>
</tr>
<tr>
<td>3.2 $n_\epsilon(e)$ for a 20 residue protein in the standard model</td>
<td>39</td>
</tr>
<tr>
<td>3.3 $n_\epsilon(e)$ for a 20 residue protein in the weakly interacting model</td>
<td>40</td>
</tr>
</tbody>
</table>
3.4 $n_c(e)$ for a 11 residue protein in the strongly interacting model ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ......
CHAPTER I
INTRODUCTION

From the point of view of polymer physics, proteins are special kinds of macromolecules that consist of mostly linear arrangement of 20 kinds of amino acids, called residues, connected by covalent bonds. The chain order formed by linked amino acids is usually called sequence or primary structure. Structural information embedded in the amino acid sequence of proteins leads to its complicated patterns of relatively fixed three-dimensional conformations which have characteristic and interesting functions. Different conformations of a certain protein share the same covalent structure, but are differentiated solely by bond rotations. It is known that a naturally occurring functional protein, such as enzymes, always has the conformation of the lowest free energy (native state). Therefore, a denatured protein which, by definition, is away from its native conformation can spontaneously fold back to its functional state. Questions about how and why proteins can adopt their specific native conformation automatically give rise to the so called protein folding problem.

A complete understanding of protein folding process is still unknown to us, though decades ago, biologists started to realize it was a spontaneous reaction. Beginning
as a biophysical-chemistry problem, the study of protein folding problem is now related
with varied fields such as physics, cell biology medicine and biotechnology \(^1\).

Over the last decade, it has been realized that protein folding plays a vital role
in significance cellular processes in areas such as medicine, diseases, human health and
life. All these areas had been studied previously in isolation so that basic principles are
now emerging together, experimentally and conceptually. Research such as the analysis
of protein folding related diseases or abnormalities has added additional twists in the
study of protein folding, and resulted in the discovery of folding catalysts that assists
protein folding in vivo \(^2\). Increasing interest in protein folding is stimulated by the
development of powerful new techniques for their analysis both in vitro and in vivo.
Sophisticated cellular machinery of molecular chaperones which are enzymes found to
help results in the significant transformation in this field; thus our knowledge of the
physical principles behind protein folding and its stability is improving.

New therapies that affect cellular regulation of protein folding can be designed
by knowing how proteins fold. The ultimate aim of solving this problem would be the
production of the protein pharmaceuticals so that scientists can design proteins with
desired functions.

1.1 Protein structures and functions

Observed from different length scales, a protein is found to have different
structures, each having a unique position occupied in its structural hierarchy. Not all
amino acid sequences will be found in naturally occurring functional proteins, which have subtle departures from random chains.

For a typical chemical polymer, such as polyethylene, the large extended molecule forms a matrix. However, protein as a special polymer has the ability of self-assembling and folding into its functional structure which is relative compact and self-contained. Proteins found in nature have specific three dimensional structures formed from the sequence of the polypeptide chains. The sequence is also known as the primary structure. Such folding behavior that a particular linear arrangement of amino acids can automatically achieve its unique three-dimensional protein structure is one of the key issues in the protein folding problem [3, 4, 5].

One of the major concerns of biochemical research is the relationship between information imbedded in a sequence and the folding behavior of proteins. However, we still do not have enough knowledge of mesoscopic principles needed to predict the 3D structure of proteins connecting the microscopic world of atoms or the amino acids scale with the macroscopic world of proteins.

Spanning from a molecule scale to the long chain scale and requiring knowledge not only from biology but also from physical sciences, the research of protein folding continues to be an intriguing area. What is more, computational and statistical methods are becoming more important to sift through the mass of accumulated data of protein structures.
1.1.1 Amino Acids

The building blocks for protein are twenty different amino acids which can be treated as dipolar ions in solutions. Both the amino and carboxyl groups in protein are charged which affects the solubility in water, the ability to act as electrolytes and its crystallinity. The amino group NH$_2$- is also called as N group for short; likewise, the carboxyl group -COOH is called as C group. Condensation reactions which eliminate water during the process are involved in the process of peptide bonds forming. Two amino acids can be covalently linked by a peptide bond to form a dipeptide along with water. An average sized polypeptide chain which may contain 1000 peptide bonds can be obtained as more amino acids link together by the repetitive formation of peptide bonds in this process$^5$.

![Figure 1.1 Dehydration synthesis (condensation) reaction forming an amide.](image)

However it is not enough to elucidate beautiful three-dimensional structures of proteins solely by the two dimensional polypeptide chain represented by only amino acids linked together which lacks important three-dimensional structural information. Despite this shortcoming, the simplicity of the two dimensional model allows us to learn a lot about real proteins. The ultimate characteristic functional properties of proteins are not yet achieved by the formation of regular secondary structure. It emerges only when it forms into complicated patterns during protein folding.
1.1.2 Primary structure or sequence

The primary structure is defined as the linear order of amino acid residues along the polypeptide chain which is formed by covalent linkage of individual amino acids via peptide bonds. Thus, the primary structure of a protein is merely the amino acid sequence from the N to C terminals of the protein. We can use a three letter or one letter abbreviations to describe an amino acid, so for the primary sequence we simply translate the three or single letter codes from left to right, from amino to carboxyl terminals. Thus in the sequence below two alternative representations of the same part of the polypeptide chain are given, starting with glutamate at residue 1, serine at position 2 and extending to threonine as the 10th residue:[2].

\[ NH_3 – Glu – Ser – Ser – Ala – Lys – Ala – Val – Tyr – Tyr – Tyr \ldots \]

Figure 1.2 Three- and single-letter codes for amino acid sequence of a primary structure.

Every protein consists of a unique sequence of residues, from which and all other higher levels of hierarchy (secondary, tertiary and quaternary structures) rely on this primary structure. However, proteins with similar functions are found related to one another more in primary structures rather than other structures.

1.1.3 Secondary structure

The complicated secondary structures are represented by three basic units of secondary structure; the \( \alpha \) helix, the \( \beta \) strand and turns. All other structures represent variations on one of these basic themes are in association with the primary structures.
During the transition from randomly coiled amino acid chain to certain regular repeating secondary structure, various interesting chemical and physical properties of polypeptides will appear. As for a dipeptide, hundreds possible mutations can be obtained from 20 amino acid residues. It is considerable for an average-sized polypeptide of ~100 residues, for which the number of potential sequences will be astronomical. And this is one of the reasons why in early days, many fundamental studies on secondary structure were done on homopolymers of amino acids in which either all the residues were identical or just some repeating units such as Ala-Gly dipeptides are repeated. The homopolymer nature of those proteins makes them somewhat regular and ordered. Consequently, they were helpful in early days.[2]

1.1.3.1 The $\alpha$ helix

Figure 1.3 A regular $\alpha$ helix structure. It has 3.6 residues per turn with distance between residues 0.15nm. The distance per turn is defined the translation per residue distance with the value of 0.54 nm. (Picture from [http://courses.cm.utexas.edu/jrobertus/ch339k/overheads-1/ch6_A-hlx.jpg](http://courses.cm.utexas.edu/jrobertus/ch339k/overheads-1/ch6_A-hlx.jpg)).
Largely predicted from theoretical studies, $\alpha$ helix which occurs in proteins is a right-handed coiled conformation is probably the best known and most common motif in the secondary structure of proteins. In globular proteins, over 30 percent of all residues are found in helices.\(^3\) Resembling like a spring, hydrogen bonds are formed between backbone N-H groups and backbone C=O groups of amino acid that are four residues apart. Helices observed in proteins can range from four to over forty residues long, but a typical helix contains about ten amino acids. The translation distance between any two corresponding atoms on the helix is called a pitch. In real proteins, hydrogen bonds in the helices have different lengths and angles with respect to helix axes. Hydrogen bonds are polarized as it involves electronegative atoms such as fluorine, oxygen, or nitrogen. Together with the peptide bonds which also have polarity, $\alpha$ helices have dipole moments. For this reason, distorted conformations are always found in short $\alpha$ helix.

1.1.3.2 The $\beta$ sheet

The $\beta$ sheet is a generally twisted, pleated sheet of beta strands connected laterally by three or more hydrogen bonds. It is so called because it was the second unit of secondary structure predicted by Linus Pauling and Robert Corey in 1951. A $\beta$ strand is fully extended stretch of peptide backbone typically 5–10 amino acids long when compared with the $\alpha$ helix. The $\beta$ strand is helically arranged adjacent to other strands although in extremely elongated form with two residues per turn and a translation distance of 0.34 nm between similar atoms in neighbor residues. An extensive hydrogen bond network is formed by N-H groups in the backbone of one strand connecting with
the C=O groups in the backbone of the adjacent strands as hydrogen bonds. The β sheets are large structures stabilized by hydrogen bonds between strands, while a single β strand can not be large because of the limited number of local stabilizing interactions.

Figure 1.4 A regular β sheet structure. (Pictures from http://courses.cm.utexas.edu/jrobertus/ch339k/overheads-1/Figure4-7.jpg)

1.1.4 Tertiary structure

The tertiary structure is the spatial arrangement of amino acid sequence as the overall topology formed by the polypeptide. As of June 24 2008, the database such as Protein Data Bank has been deposited with 51,491 released atomic coordinate entries (or "structures"), 47,526 of that are proteins, the rest being nucleic acids, nucleic acid-protein complexes, and a few other molecules. Each year, about 5,000 new structures from
increasingly diverse organisms are released. It is estimated that the size of the PDB archive will triple to 150,000 structures by the year 2014. Most of those data are obtained by X-ray crystallography or NMR spectroscopy and submitted by biologists and biochemists from around the world, and are released into the public domain, and can be accessed for free. Rutgers University currently maintains this data base with several mirrors found at other sites across the world.

PDB is a repository for 3-D structural data of proteins and nucleic acids, containing information of the exact location in space of the atoms making up a protein, by which the whole molecule can be defined. The PDB files also contain header information about the method used to get the structure, the primary sequence, the elements of secondary structure as well as the authors [6].

The most typical conformation of a protein in cellular environment is generally referred as the native state or native conformation of the folded polypeptide chain, which is also the most thermodynamically stable conformation attainable for a given primary sequence. The tertiary aspect of the compact molecule involves secondary structural motifs with little irregular structure which is normally confined to the N and C terminals. Various interactions not only hydrogen bonds, as in β sheets, but also disulfide bridges, electrostatic interactions, Van der Waals interactions, hydrophobic contacts and hydrogen bonds between non-backbone groups play a vital role. All these interactions between secondary structure themes make the protein to fold into a compact globular molecule and form its tertiary structures [7].
1.1.4.1 Disulfide bridges

Cysteines which are normally separated by at least five other residues in primary sequence contain sulfur atoms. A strong covalent bond as a result of the oxidation of the sulfhydryl (SH) groups of two cysteine molecules is known as a disulfide bridge. It is important in linking polypeptide chains in proteins. It cannot be formed between consecutive Cysteine residues. Due to high bond strength, disulfide bonds are only broken at high temperatures, acidic pH or in the presence of reducing agent.

1.1.4.2 The hydrophobic effect

Intramolecular force such as charged interactions and those forming hydrogen bonds are not strong in water, because the polar water molecules compete significantly with them. The importance of the hydrophobic effect is that in water it becomes one of the major interactions leading the proteins to fold. Water is a poor solvent for non-polar molecules which cannot form hydrogen bonds with water molecule. Hydrophobic molecules tend to decrease entropy for the water. To counter this, the water molecules force hydrophobic molecules together to decrease the surface area. This would increase the entropy. For example, dissolving cyclohexane in water, preventing cyclohexane molecules dissolving extensively in aqueous solutions, interactions between non-polar molecules and water are weak or not exist. As a consequence, interactions between non-polar molecules are enhanced and hydrophobic clusters within water are formed, which are the basis for the hydrophobic effect. Since most of the side chains of amino acids are
hydrophobic, the hydrophobic effect which can also be treated as the preference of non-polar atoms for non-aqueous environments as can not be ignored.

1.1.4.3 Charge-charge interactions

Between oppositely charged residues in side chains as well as between the \( NH_3^+ \) and \( COO^- \) groups at the ends of polypeptide chains, there are charge-charge interactions.

1.1.4.4 Hydrogen bonding

A hydrogen bond which is a kind of dipole-dipole force that exists between an electronegative atom (Nitrogen, Oxygen and Fluorine) and a hydrogen atom, contributes greatly to the stability of \( \alpha \) helices and the interaction of \( \beta \) strands to form parallel or antiparallel \( \beta \) sheets. These hydrogen bonds are between main chain N-H and C=O groups. With its bond strength close to 155kJ/mol, it is about 1/20 of a typical covalent bond; while unlike the covalent bond, it is kind of an intermolecular bond.

1.1.4.5 Van der Waals interactions

Van der Waals forces are believed to be caused by the induction of dipoles due to fluctuating charge densities within atoms (a consequence of quantum dynamics), and surfaces. Other than covalent and ionic bonding, attractions between atoms and molecules are also very important in protein folding and must be included. These interactions occur between uncharged and non-bonded atoms. Since atoms are
continually oscillating the induction of dipoles is a constant phenomenon. They are also intermolecular interactions.

1.1.5 Quaternary structure

Proteins could contain more than one polypeptide chain. Quaternary structure results from the interactions between those polypeptide chains. All those interactions namely disulfide bonds, hydrophobic interactions, charge-pair interactions and hydrogen bond can be found in tertiary structure. The difference is that tertiary structure is referred as 3D structure for one polypeptide chain.

1.2 The protein folding problem

For briefness, protein folding problem centers on why and how a protein can find its unique functional structure in a short time and how we can utilize those special properties to benefit human lives.

1.2.1 Anfinsen’s experiment

It can be dated back to the 1960s since Chris Anfinsen started the modern work on the mechanism of protein folding problem. He proposed his "Thermodynamic Hypothesis", which states that an unfolded (denatured) polypeptide chain could spontaneously refold into its functional structure. It strikingly shows that sufficient information has already embedded in the protein sequence to guarantee correct folding from any of a large number of unfolded states \[8\].
In the first process, denatured ribonuclease is obtained by breaking four disulfide bonds in ribonuclease to thioles by mercaptoethanol in the presence of urea. While ribonuclease is then oxidized and the urea first moved by dialysis, the activity of this enzyme can be recovered as a result of protein folding. However, if the denatured ribonuclease is oxidized first, only less than 1 percent of the activity can be recovered.

Figure 1.5 A schematic diagram of Anfinsen's experiment. (http://dwb.unl.edu/Teacher/NSF/C10/C10Links/www.bio.cmu.edu/Courses/03231/LecF99/Lec08/lec08.html)

Figure 1.6 The protein folding process. (From http://commons.wikimedia.org/wiki/Image:Protein_folding.png)

The problem can be restated by asking how the amino acid sequence determines a protein’s three-dimensional structure. To be more specific, our job is to be clear of (1) The mechanism of the folding process. (2) The way to predict the three-
dimensional structure of a protein from its amino acid sequence. and (3) The relation between the folding process in vivo and in vitro? \[1\]

1.2.2 Levinthal’s paradox

In 1968, Levinthal’s paradox known as a bombshell to the field was released by Cyrus Levinthal. In order to predict of the 3D structures of proteins from its sequences, the time needed to fold by searching all possible conformations randomly can be longer than the life of universe. For instance, let us assume a short protein (100 amino acids) with 3 conformations per peptide bonds. The overall conformation number is \(3^{100}\). Assuming the time required for exploration for each conformation is \(10^{-6}\) seconds, it will take \(5 \times 10^{41}\) seconds = \(1.6 \times 10^{34}\) years to finish. In order to fit into the time for protein folding in realistic, Levinthal’s conclusion was that there must be folding intermediates and pathways.\[9\] The paradox is never true since a protein does not explore all those conformations. With an addition free energy bias as the driving force for folding, the time needed to search all conformations by a random search process was reduced to a few seconds \[10\].

1.2.3 Energy landscape

To avoid the excesses of Levinthal’s paradox, there are themes to restrict the conformational spaces of the protein. One way to restrict the conformational numbers is represented by the contoured energy landscape that directs the unfolded high free energy state to the native state which is believed to have the minimum free energy. The funnel
shaped three dimensional plots suggest several possible pathways of folding by avoiding sampling all the possible conformations.

Figure 1.7 A sketch of an energy landscape. The energy is on the vertical axis with other axes representing conformation factors. (From http://www.lsbu.ac.uk/water/protein2.html)

1.3 Protein structure determination and theoretical models

Progresses in both experiments and computer speed help the determination of protein structures. As a consequence, information about proteins can be gotten from Protein Data Bank easily these days. Special softwares, such as Visual Molecular Dynamics (VMD), can also provide 3D pictures of protein structures.

1.3.1 Methods used for determining protein structure

The development in both instrumentation as well as basic biochemistry theories result in cutting-edged experimental methods which provides atomic level resolution of protein structure determination. Sophisticated pictures of proteins are established by knowing the exact position of most atoms in the protein such as carbon, nitrogen and oxygen. Among those techniques, nuclear magnetic resonance (NMR), cryoelectron microscopy, X-ray crystallography, neutron scattering and optical spectroscopic
techniques are widely used nowadays. All these experimental techniques that provide information on different regions of the electromagnetic spectrum, help understand the structure as well as dynamics of proteins. X-ray crystallography and multi-dimensional NMR yield details of protein structure at atomic level, so the structural differences between native state and denatured of a certain protein can be provided. Time-resolved measurements allow measurements of protein mobility.

Computational methods can also help in the structure prediction. Homology methods adopt the idea that similar protein structures from primary to superior will result in similar functions and similar functions might be due to similar proteins structures. \textit{Ab initio} methods provide protein structures from first principles such as energy, geometry and kinematics. Different models are used in the calculation.

1.3.2 Theoretical Models

So far there are only simplified descriptions of proteins, for we still lack soluble realistic models. Treating a protein as a homopolymer or copolymer and assigning different interactions between species and local conformations, different models presented to elucidate the folding phenomenon can explain certain aspects of this problem.
1.3.2.1 The $G\bar{o}$ model\textsuperscript{[11]}

It treats a protein as a self-avoiding homopolymer chain with fixed native conformation. This model is being used increasingly to describe the folding reaction, with only native contact interactions in the energy function.

It makes use of a potential function based on the knowledge of the native structure of a protein, which is obtained from experiment. The model has the virtue of making the native state to be the global energy minimum of the system aside from making computationally feasible the description of the folding process. The potential function is the sum of two-body terms contributing with -1 if a native contact is formed and zero otherwise. The Hamiltonian of certain conformation can be written as

$$H = \sum_{i<j} \alpha_{ij} \Delta_{ij}$$  \hspace{1cm} (1.1)

$$\Delta_{ij} = 1 \quad \text{if the monomer } i \text{ and } j \text{ are in contact}$$

$$\Delta_{ij} = 0 \quad \text{otherwise}$$

$$\alpha_{ij} = -1 \quad \text{if the monomer } i \text{ and } j \text{ are in contact in the native conformation}$$

$$\alpha_{ij} = 0 \quad \text{otherwise}$$

This model treats all amino acids on equal footing, neglecting the chemical properties of the different types of amino acids, and also the N and C terminals of a protein.

Some modified $G\bar{o}$ type models have been developed. Adding repulsion to the Hamiltonian based on nonnative contacts, the new ones can avoid the shortcoming of the standard $G\bar{o}$ model, noticeably that the properties of a protein are solely determined by its geometry.
1.3.2.2 The Gaussian model

Gaussian model, \cite{12}, which is an off-lattice model, also treats proteins as homopolymers. A protein molecule is modeled as a linear chain of \( n \) beads. The interactions are governed by quadratic potentials, which are either attractive or repulsive. All interactions are treated as spring forces between monomer beads. That is why the model is also called the bead-spring model.

A protein molecule is modeled as a \( n \)-bead linear chain. There are covalent potentials between neighboring beads. There are additional nonbonded potentials between pairs of non-neighboring beads.

The energy of a conformation consists of two parts: the first part \( E_{ij} \) is from interactions between beads \( i \) and \( j \), and the second part \( E_i \) is due to mean-field repulsive potential on each bead.

\[
E = \sum_{ij} E_{ij} + \sum_i E_i = \frac{1}{2} \sum_{ij} a_{ij} r_{ij}^2 + \sum_i c_i R_i^2
\]  

(1.2)

The position of the \( i \)th bead is \( \mathbf{R}_i \), and, \( r_{ij} = |\mathbf{R}_i - \mathbf{R}_j| \), \( R_i = |\mathbf{R}_i| \), and \( a_{ij} \) and, \( c_i \) are coefficients.

1.3.2.3 The HP model

In this model, each amino acid is also represented as a bead, while chemical bonds are represented by bonds. The protein is put on a lattice with all other sites on the lattice occupied by water molecules.

HP model \cite{13} simplifies the driving force for folding as hydrophobic force, which minimizes the number of hydrophobic residues exposed to water. There are only two
types of monomers in the 2D self-avoiding walk chain: H for hydrophobic residues and P for hydrophilic residues. Each lattice nearest-neighbor HH contact interaction is assigned energy $\varepsilon < 0$.

The disadvantage is that the model lacks details. The resolution of protein structures and energetic are not accurately represented, and the chain length is usually short because of the computer limitation. But it does bring larger conformational changes than those from atomic level simulations.
CHAPTER II

EXACT STATISTICAL MECHANICAL INVESTIGATION OF A FINITE PROTEIN

2.1 Folded and unfolded protein

Consider cooking eggs, during which egg white, which is a common protein, coagulates. Properties of protein will undergo dramatic change upon heating or acidification, which is called the denaturation of protein. The process of renaturation, which recovers biological function, can occur occasionally as was shown in the early 1930’s. All these reactions represent expansion or contraction in the changing conformations. However, at that time, the list of protein amino acids was not yet complete and even now the intricate protein biosynthesis and the way it works are still not been totally understood.

2.2 The protein model used for our research

We used a coarse-grained HP protein which is examined on a square lattice in our calculation. Different energy penalties are assigned to different structures in the conformations generated by our computer program.
2.2.1 Protein as a small system

In our work, a general incompressible model of a finite protein of size $M$ in its environment is used. This model does not allow voids in the lattice. Each conformation of the protein on the lattice results in certain sites of the lattice being occupied by the protein. The rest of the sites will be occupied by the solvent. In this way, it ensures that each conformation of the protein is associated with only one possible distribution of the solvent molecules on the lattice. The protein is represented by a semiflexible copolymer consisting of amino acid residues classified into only two species (H and P) following Lau and Dill. Three sets of parameters for interactions presenting different energetic parts are introduced between chemically unbouned residues from the conformation and the solvent (water). All the conformations are generated on both infinite and finite lattice so that unrestricted conformations are also allowed. The enumeration allows us to investigate exactly the effects of energetic on the native state(s), the effect of small size on protein thermodynamics and the difference between conformations.

Self-assembling small proteins have been extensively investigated using lattice models by thermodynamic principles. Proteins are similar to semiflexible polymers in which semiflexibility forces an ordered (crystalline) compact structure at low temperatures, in contract with flexible polymers, which collapse to a compact disordered state.

We take an infinitely large lattice so that the boundary of the volume does not affect the behavior of the system. In all cases, we are dealing with one peptide chain in
the presence of solvent (water) on the lattice. As we only deal with a single protein, no superior structures than tertiary structures will be involved.

![Figure 2.1 A compact protein with water surrounded on a square lattice. The blue balls are hydrophilic species. Hydrophobic residues are represented as red balls. The green beads are water molecules.](image)

The behavior of a simplified protein with only one polypeptide chain which consists of $M \ll \infty$ residues focused in our work is governed by small system thermodynamics. We know that predictions of different ensembles describing a macroscopic system are the same, except at some singular points. But for a small system, different statistical ensembles differ from each other. The work on how to distinguish small system thermodynamics from a macroscopic system thermodynamics has been done already in our group’s previous work.
2.2.2 Structures involved in our research

The primary structure of the protein in our model is represented by residue sequence. All the residues are treated as hard-core beads which must lie on a lattice site and are further classified to two subsets, H (hydrophobic species) and P (hydrophilic species). Three submodels with different sets of interactions between species and for different structures are considered.\[15\]

The semiflexibility of the protein gives rise to a crystalline phase, which represents the native state of the protein at low temperatures. So we treat a protein as a semiflexible self-avoiding copolymer chain on an infinitely large lattice to study its folding behavior. We are considering an incompressible but self-avoiding chain, which means a site must and can only be occupied by either occupied by a residue or solvent. Proper model must also account for the directionality of the protein, because proteins are always biosynthesized form the N-terminus (free amino groups) to C-terminus (free carboxyl groups).

Figure 2.2 A 2-d model of a finite protein on a square lattice. The red spheres represent hydrophobic residues and the blue ones are hydrophilic residues.
We show a protein in its compact form so that all the solvent molecules (W) are expelled from inside and surround the protein. The red beads denote hydrophobic residues (H) and blue beads represent hydrophilic residues (P). The nearest-neighbor distinct pairs PP, HH, HP, PW and HW between the residues and the water are also shown, but not the contact WW. Only three out of these six contacts are independent on the lattice,[39] which we take to be HH, HW, and HP pairs.

2.2.3 Anchored Protein

In the enumeration of all possible conformation for a given sequence, we only consider a square lattice, on which the protein has no restriction on its allowed conformations, except that it is anchored at one end. The protein will be allowed to take all shapes including compact shapes and the stretched-out shapes by having it probe all allowed sites on an infinite lattice. The end of the protein is rooted and is not allowed to move. The realistic reason for doing so is that the process of folding in vivo often begins co-translationally, so that the N-terminus of the protein begins to fold while the C-terminal portion of the protein is still being synthesized by the ribosome. So in our model, the C-terminus is rooted and N terminus is free to begin folding.

Figure 2.3 Two kinds of extreme conformations in our model. The above one is the stretched-out conformation. The below one is the compact conformation.
The anchoring also provides two important benefits for our computation.

1) It reduces the number of conformations to a finite number, so the calculation is possible. For an unanchored protein that can start form anywhere of the infinite lattice, the number of conformation will also be infinite, which has no bearing on thermodynamics.

2) Anchoring allows us to uniquely define the distance between two conformations, which we will discuss later.

To save computational time, we also restrict protein conformations, so that its first bond from the root is always to be right and the first bend is always downward. In this way, any other conformation of the protein related by trivial rotation to one of the generated conformations will be neglected. In this way the total number of conformations is reduced without any loss of information, and the first generated conformation is always in stretched-out state.

The number of conformations $W$ for rooted proteins increased rapidly with the protein size, as is seen in figure.

![Figure 2.4 Log-log plot of the number of conformation ($W$) with respect to number of residues ($M$).](image)
The growth of $W$ for the rooted protein with its first bond in a specified direction on an infinite lattice can be fitted by

$$W = 0.102272 \exp(0.000933 M)$$

with $R^2 = 0.999876$

2.2.4 Interaction energies

The excluded-volume effects are accounted for by enforcing that a lattice site cannot be occupied by more than one residue or water molecule. The interaction energies are restricted between chemically unbonded residues H and P, and water molecules W that are nearest neighbors of each other. Long range interactions are neglected, but can be incorporated later if so desired. There are three species of particles (H, P, and W) in our model. We need to only consider three independent energies of interaction between three chemically unbonded pairs of species. We have decided to use the following three van der Waals energies $e_{HH}$, $e_{HW}$, and $e_{PH}$ between the three unbonded pairs HH, HW, and PH.

To account for the semiflexibility of the protein, we use a previously studied model in our group. This model contains addition energies that allow us to study crystallization and glass transition in polymers. We have extended it to include a preference for helical formation.

We consider a protein with $M$ residues in a given sequence $\chi$ of H and P associated with the residues on a square lattice. We only consider the case in which the number of H and P are equal. But we can change the number of H and P if wanted.
The number of bends $N_b$, pairs of parallel bonds $N_p$, and hairpin turns $N_{hp}$ are used to characterize the semiflexibility. A bend is where the protein deviates from its collinear path. Each hairpin turn requires two consecutive bends in the same direction (clockwise or counterclockwise); see Figure 2.5.

Two parallel bonds form a pair when they are one lattice spacing apart. We also use the number of helical turns $N_{hl}$. On a square lattice, a "helical turn" is interpreted as two consecutive hairpin turns in opposite directions as shown in Figure 2.3.

We let $\mathbf{e}$ denote the set containing all $\{e_i\}$, and $\mathbf{N}$ denote the set containing all $\{N_i\}$, where $i$ stands for $b,p, hp, hl, HH, HW, \text{ and } HP$. Thus,

$$\mathbf{e} \equiv \{e_b, e_p, e_{hp}, e_{hl}, e_{HH}, e_{HW}, e_{PH}\}$$

$$\mathbf{N} \equiv \{N_b, N_p, N_{hp}, N_{hl}, N_{HH}, N_{HW}, N_{PH}\}$$

Let $W(N)$ denote the number of protein configurations. The energy of certain configuration $\Gamma$ corresponding to the set $\mathbf{N}$ is given by

$$E(N) = \mathbf{e} \cdot \mathbf{N} = \sum_i e_i N_i$$

Different conformations for given protein may have different numbers of $b, p, hp, hl, HH, HW, \text{ and } HP$ so that the energy varies from configuration to configuration as it depends
on $N$. But it does not depend on thermodynamic state parameters such as the temperature, pressure, etc. The entropy function corresponding to configurations with a given $N$ is defined as

$$S(N) \equiv k \ln W(N)$$

(2.3)

Here, the Boltzmann constant $k$ is set to be 1. There are three sets of energies as mentioned before. Those parameters for each set are shown in Table 2.1

**Table 2.1 Possible Models and their Parameters**

<table>
<thead>
<tr>
<th></th>
<th>Standard (A)</th>
<th>Weakly (B₁)</th>
<th>Strongly (C₁)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bend</td>
<td>0</td>
<td>1/50</td>
<td>1/3</td>
</tr>
<tr>
<td>Parallel</td>
<td>0</td>
<td>−1/50</td>
<td>−1/3</td>
</tr>
<tr>
<td>Hairpin</td>
<td>0</td>
<td>−2/50</td>
<td>−1/3</td>
</tr>
<tr>
<td>Helix</td>
<td>0</td>
<td>−2/50</td>
<td>−1/3</td>
</tr>
<tr>
<td>HH</td>
<td>−1</td>
<td>−50/50</td>
<td>−3/3</td>
</tr>
<tr>
<td>HW</td>
<td>0</td>
<td>20/50</td>
<td>2/3</td>
</tr>
<tr>
<td>PH</td>
<td>0</td>
<td>5/50</td>
<td>1/3</td>
</tr>
</tbody>
</table>

2.2.4.1 The standard model

In standard model, only the interactions between hydrophobic residues are considered, which are always set to be $e_{\text{HH}} = −1$. The density $n_{\text{HH}} = N_{\text{HH}} / M$ is going to be a discrete quantity, and the energy density is presented as $e = E / M = n_{\text{HH}}$

2.2.4.2 Weakly perturbed model (B)

In this model, we allow for other energies to be non-zero, but still small compared with $e_{\text{HH}} = −1$. The model with the parameters in the above table is called B. The numerator of various energies are integers and are used to determine the energy $E$ as an integer, which makes it easy to classify energy levels in groups of a given energy.
energy is divide by the denominator at the end to ensure that $e_{HH} = -1$. In the program, we multiply those interactions with their denominators, so that all the energies are integers. After we have done the calculation, they have been divided back. Due to the strong repulsion between H and W, the only energy close to $|e_{HH}|$ is $e_{HW}$. The model will be identified as a model with weak perturbation on the standard model, for the other energies are small compared with $e_{HH} = -1$.

2.2.4.3 Strongly perturbed model (C)

In this model, we also allow for the energies to be not only non-zero, but also comparable in strength to $e_{HH} = -1$. The choice of parameters is shown in the table. Again, the numerators for various energies are integers for the same reason explained above. In the model B, most other interactions are much weaker than $|e_{HH}|$, while they are comparable to $|e_{HH}|$ in the model C. Thus the model B is closer to the model A than to the model C is. Despite this, we will see that the models B and C behave very different from A, which is because the interactions other than hydrophobic interactions have contributed to the overall Hamiltonian. It should be noted that $W$ does not depend on the model; it is its partition into $W(E)$ that depends on the model.
2.3 Small system thermodynamics

A typical thermodynamic system should be macroscopically large. However, a protein chain never consists of infinite number of amino acids. Thus we would expect the thermodynamic properties of a protein to be different from systems are macroscopic.

2.3.1 Microcanonical entropy

The dimensionless ME (Microcanonical Ensemble) entropy corresponding to configurations with a given energy $E$ is given by the Boltzmann relation. We always set the Boltzmann constant to be 1 in our work here.

In our model, a small protein is considered, which makes $E$ a discrete variable. The energy density per residue

$$ e \equiv \frac{E}{M} $$

(2.4)

is also discrete and will become continuous only when $M \rightarrow \infty$ at which point, the protein becomes infinitely large. Thus, the entropy density per residue

$$ s(e) = \frac{S(E)}{M} $$

(2.5)

is also discrete. As is shown in Figure 2.6 below.
Figure 2.6 The behavior of $s(e)$ as a function of the protein size $M$. It shows the case for the strongly perturbed model (C) of a compact protein, which is discovered by our group in previous work.

For $M = 16, 24, 32, 40,$ and $48$, we see clearly how the entropy density profile goes as $M$ increases for standard model. We also note that the discrete nature of the energy and entropy persists. There is a clear evidence of many local maxima in the entropy, each maximum surrounded by many energies of lower entropy forming an energy band. These bands are well separated by gaps in the energy, at least near the low end of the energy even for $M = 48$: It is surprising to observe the erratic form of the entropy in that the bands are highly irregular in shape, at least near the low energy end.

2.3.2 Canonical partition function

A real protein in Nature is not a closed system. Therefore, microcanonical ensemble is the most suitable to investigate. When we consider the temperature, at which
the protein interacts with its surrounding, we need to consider the canonical ensemble in which the temperature of the system and its surrounding is fixed. The description is more realistic and can be characterized by the canonical partition function given by

\[ Z(T) = \sum_{E} W(E) \exp(-\beta E) \]  

(2.6)

where \( \beta = \frac{1}{T} \)

with which a bunch of canonical average can be calculated.

The canonical probability distribution for a given conformation \( \Gamma \) is given by

\[ p(\Gamma) = e^{-\beta E(\Gamma)} / Z(T), \]  

(2.7)

so that any thermodynamically extensive quantity \( O \) has in average given by

\[ \overline{O} = \sum_{\Gamma} O(\Gamma) p(\Gamma) \]  

(2.8)

or

\[ \overline{O} = \sum_{E} O(E) p(E) W(E) \]  

(2.9)

### 2.4 The definition of distance between conformations

For monomeric systems, in which each monomer is treated as a particle, we can label all the monomers with \( \alpha (=1, 2, \ldots, M) \) so any conformation in the \( 3M \)-dimensional space \( \kappa \) can be described by the order set

\[ R = \{ r^{(1)}, r^{(2)}, r^{(3)}, \ldots, r^{(M)} \} \]

The distance between two conformations \( R \) and \( R' = \{ r'^{(\alpha)} \} \) is defined here to be the Euclidean distance

\[ d(R, R') = \sqrt{\sum_{\alpha=1}^{M} \left( r^{(\alpha)} - r'^{(\alpha)} \right)^2} \]  

(2.10)
The distance provides useful information not only about the topology of the energy landscape but may also be relevant for the dynamical description of the folding process (even though we are not presently interested in the dynamics) by introducing the concept of a neighborhood of a point in the conformation space $\kappa$: two conformations are neighbors or are “connected” in $\kappa$ if their separation is less than or equal to some chosen distance.

The distance $d(\mathbf{R}, \mathbf{R}')$ can be used as an element to define a $W \times W$ distance or neighborhood matrix $\mathbf{D}$, with diagonal elements to be 0, for the distance between two same conformations is of course zero. As $W$ is usually a large number, it is not possible to study the entire matrix. So we always consider the distance of each conformation from special conformations, namely the native state and the stretched state which is generated first. See Figure 2.3. If the native state is not unique, we pick the first generated of the native states (lowest energy state). The native state strongly depends on the energetics and is not unique as far as different energetic are concerned, while the stretched state is always unique and does not depend on the energetic. This provides us good comparisons.
CHAPTER III
CURRENT WORK: NONNATIVE CONTACTS IN PROTEIN FOLDING

3.1 Folded and unfolded state

We believe that a functional protein should be in a conformation that is relatively packed and has lowest free energy. Under certain condition, a protein can be stretched and therefore lose its functionality.

3.1.1 The unfolded (denatured) state

A protein changes from its biological active state, the word denaturation is used to describe the process. Here we will consider only conformational changes, which in principle can be treated as reversible. The term folding and unfolding have this implication.

3.1.2 The folded (native) state

The native state of a protein is a compact, lowest free energy state in which a protein is able to perform its biological functions.
3.1.3 The definition of nonnative contacts

Contacts: in our model, a coarse-grained copolymer with hydrophobic residues (H) and hydrophilic residues (P) is put in the square lattice. The term “coarse-grained” means that we treat every monomer as a bead in the chain, neglecting the structure details of the monomer. Any pair of these residues occupying nearest neighbor sites that are not connected by covalent bonds is called a contact.

Nonnative contacts: those contacts appearing in the non-native conformations that do not exist in the native conformation are called nonnative contacts.

Figure 3.1 A schematic plot for native and nonnative state for a protein with 36 residues.

In the left figure in Figure 3.1, a compact native conformation for a protein is shown: all residues are labeled in numerical order. Chemical bonds are represented by bold lines. The dash lines indicate there are contacts. For example, in the left figure, residue 4 is nearest neighbor with residue 7, 33, 5, and 3. However, the linkage between
residue 4 and residue 3, or residue 5 is a chemical bond which is not counted as contacts in our model. So pairing 4-7, and pairing 4-3 are considered as contacts.

In the right figure, a higher-energy denatured protein conformation is shown. All the residues, contacts and chemical bonds are noted in the same way as before. Nonnative contacts are shown by adding stars on the dash lines. As an illustration, residue 4 here has now contacts with residue 35 and 29, which are not found in the left figure (native conformation). So these two contacts are now nonnative contacts as we defined before. However, in the native state, residue 8 and 13 are in contact, while in the denatured state, they are still in contact. By our definition, the pairing 8-13 is not a nonnative contact.

3.2 The role played by nonnative contacts

Nonnative contact whose definition is given before are found in denatured conformations of proteins. And those nonnative contacts are believed to play a vital role in different stages of protein folding.

3.2.1 Nonnative contacts affect protein folding rates

Li and Cieplak [16] investigated nonnative contacts in an extended $G_0\bar{\alpha}$ model, in which non-native contact energies may be either attractive or repulsive. As the energy of nonnative contacts increases, the folding temperature is found to go up. Repulsive nonnative contact energies also can lead to folding at $T=0$ for certain sequences and
accelerate the folding rate by reducing the complexity of its connectivity graphs for local energy minima.

3.2.2 Nonnative contacts make great contributions to the unfolding pathways

In the work of Paci etc.\textsuperscript{[17]}, in order to investigate the validity of Go-type models, which include only native contact interactions in the energy function, they determine the role of native and nonnative interactions along folding and unfolding pathways. The results shows that nonnative contacts tend to make a significant contribution for molten globules and collapsed states along the unfolding pathways.

3.2.3 Folding intermediates are rich in nonnative contacts.

In the folding of β-lactoglobulin, the mechanism of transition from α-helix to β-sheet is discussed using the free energy landscape analysis by Chikenji and Kikuchi.\textsuperscript{[18]} As they use the HP model, the helical propensity of β-lactoglobulin is driven by conformational entropy. This suggests an “on-pathway” folding intermediate rich in nonnative structures, from which they conclude that the native structure topology plays an important role in $\alpha \rightarrow \beta$ transition.

3.3 Statistical Mechanical properties of Nonnative contacts

Our program runs in a way that given a special sequence, all possible conformations are generated in turns. The lowest energy conformation is selected as the native conformation. If more than one lowest energy conformations are found, the first
one is to be chosen. With the native state information stored, our program runs again to
collect all nonnative contacts for each conformation by comparing with the native
conformation.

3.3.1 Nonnative contacts density with respect of energy for a standard sequence

One of the properties we are interested in is the nonnative contacts density as a
function of the energy, which is defined as

\[ n_c(e) = \frac{N_c(e)}{W(e) \times M} \]  \hspace{1cm} (3.1)

\( M \) is the number of residues, \( E \) is the energy of certain conformation. Here, \( e = E / M \)
is the energy per residue, \( W(e) \) is the number of conformations that has the same \( e \)
and \( N_c(e) \) is the sum of nonnative contacts in all the conformations that have the same
\( e \). The \( n_c(e) \) describes the average number of contacts for certain energy density \( e \) per
residue.

From Figure 3.2, we find that for a standard model protein, \( n_c \) is almost linearly
increasing with \( e \).
For a weakly perturbed model protein, $n_c(e)$ distributes itself around the value of $e$ found in for the standard model, because of the slight changes in the energy. The low-energy end is erratic and discrete. See Figure 3.3.

Figure 3.2 $n_c(e)$ for a 20 residue protein in the standard model.
Figure 3.3 $n_{c}(e)$ for a 20 residue protein in the weakly interacting model.

From Figure 3.4, for a strongly perturbed protein, the energy level is fully spread out, so the line shape appears more smooth. When the number of residue $M$ increases; see Figure 3.4, 3.5 and 3.6. However, the low energy density end appears erratic and not smooth. Also, $n_{c}$ is also found to saturate when $e$ reaches its high end.
Figure 3.4 $n_c(e)$ for a 11 residue protein in the strongly interacting model.
$n_e(e)$ for a 14 residue strongly interacting model protein

Figure 3.5 $n_e(e)$ for a 14 residue protein in the strongly interacting model.
for a 20 residue strongly interacting model protein

Figure 3.6 \( n_c(e) \) for a 20 residue protein in the strongly interacting model.

3.3.2 Nonnative contact density as a function of temperature

Another interesting property we consider is the behavior of nonnative contacts with respect to temperature. It is defined as the canonical average of \( n_c(E) \)

\[
\overline{n_c}(T) = \frac{\sum_E n_c(E) W(E) \exp(-\beta E)}{\sum_E W(E) \exp(-\beta E)} = \frac{1}{M} \frac{\sum E N_c(E) e^{-\beta E}}{\sum E W(E) e^{-\beta E}} \quad \beta = \frac{1}{T}
\]  

(3.3)

For a protein that has only one lowest energy state (native state), notable in weakly perturbed model (Figure 3.7) and strongly perturbed model (Figure 3.8), \( \overline{n_c}(T) \) increases from zero as \( T \) increases. Clearly, when \( T \) is zero, the only possible state for the protein

43
is its unique compact native state, so there are surely no nonnative contacts. As the temperature goes up, $n_c(T)$ increases dramatically and finally slows down and saturates.

$$n_c(e) \text{ for a 10 residue weakly interacting model protein}$$

Figure 3.7 $n_c(e)$ for a 10 residue protein in the weakly interacting model.
Figure 3.8 $n_c(e)$ for a 10 residue protein in the strongly interacting model.
Figure 3.9 $\overline{n_c}(T)$ for a 14 residue protein in the weakly interacting model.

But for a standard model protein (Figure 3.10), the lowest energy conformation is not unique. In our research, we just pick the first lowest energy conformation as the native state for that protein. So $\overline{n_c}(T)$ does not start from zero; however, the same kind of behavior is obtained as the temperature goes up.
Figure 3.10 $\bar{n}_e(T)$ for a 14 residue protein in the standard model.

3.3.3 Sequence dependence

In order to justify that the sequence of protein (primary structure) is important in protein functionality, we tried different protein sequences.

For the sequence $\mathcal{X}_1$ (pphhpphhhhphphppphp) (Figure 3.11) and $\mathcal{X}_2$ (hphphpbphphhhphp) (Figure 3.12), which have the same number of P residues and H residues, no big difference are found in $n_e(e)$ and $\bar{n}_e(T)$ behavior. $\mathcal{X}_1$ and $\mathcal{X}_2$ are randomly generated with equal H and P type residues.
$n_c(e)$ for a 20 residue protein in weakly interacting model

Figure 3.11 $n_c(e)$ for a 20 residue protein with sequence (pphhpphhhhphppphp) in the weakly interacting model.
$n_c(e)$ for a 20 residue protein in weakly interacting model

Figure 3.12 $n_c(e)$ for a 20 residue protein with sequence (hphhphppphhhphppp) in the weakly interacting model.
Figure 3.13 $\bar{n}_e(T)$ for a 20 residue protein with sequence (pphhpphhhhphpphp) in the weakly interacting model.
Figure 3.14 $\overline{n_c}(T)$ for a 20 residue protein with sequence (hphphhppphhphppppp) in the weakly interacting model.

However for a certain sequence $\chi_s$ (hphphphppppphhphpppp), for T as high as 4, there are no other conformations but the native conformation, see Figure 3.15. Only when T goes to about 15 we start to see the change in $\overline{n_c}(T)$, see Figure 3.16. The reason is that in $n_c(e)$ chart for this protein, the energy gap between the lowest energy conformation and the second lowest conformation is huge, say about 0.07, which means that this particular protein is way more thermo-stable than other proteins, see Figure 3.17. The native conformation needs much more energy to jump to its excited state.
Figure 3.15 $\overline{n_c}(T)$ for a 20 residue protein with sequence (hhphphppphppphhhhp) in the weakly interacting model with small $T$ scale.
Figure 3.16 $\bar{n}_c(T)$ for a 20 residue protein with sequence (hhphphpppphhpphhhp) in weakly interacting model with larger $T$ scale.
Figure 3.17 $n_c(e)$ for a 20 residue protein with sequence (hhphphpphhpphhpphhhp) in the weakly interacting model.

3.4 Summary

1) The density $n_c(e)$ is always monotonically increasing in the standard model. In weakly interacting and strongly interacting model, the density $n_c(e)$ is spread out, but generally increasing.

2) The $n_c(T)$ for all models is generally monotonically increasing. A few violations can be seen.

3) The protein property is sequence dependent. As other literatures [19], [20] indicate the sequence plays more important role in the protein function.
3.5 Future Work

1) Conformations that have same amount of nonnative contacts might be very different in shape from each other. That motivates to investigate at the relationship between distance and nonnative contacts.

2) Single chain is not a thermodynamic system. In order to see the folding transition, we should look at many finite proteins. In particular, we need to study them on a recursive lattice.

3) Develop a 3D program for our enumeration of protein conformations.


