Effects of Confinement and Macromolecular Crowding on
Protein Stability and Protein Folding Dynamics

A Thesis
Submitted to the Faculty
of
Drexel University
by
Guanghui Ping
in partial fulfillment of the
requirements for the degree
of
Doctor of Philosophy
June 2005
Acknowledgements

The completion of this thesis could not be achieved without the help of my advisor, Prof. Jian-Min Yuan, who not only dedicates himself as an excellent mentor helping me find fresh idea, correct the wordings of my papers, but also always provides as much help as he can to tackle difficulties that my family encountered. It is the former that guides me to the success of my Ph. D. study, and it the latter who encourages me to never give up.

I would like to thank Dr. Michel Vallieres for his zealous readiness to help me in real technical questions. Especially, he is the one who triggered my great interest in C language and parallel computation. Dr. Yen Wei and Dr. Guoliang Yang are excellent experimentalists who provided many academic supports to my research. Specifically I will thank these two professors for their wonderful work with which I can compare my theoretical research. It is a real pleasure to work with them. I have been working as a teaching assistant for long time with Dr. Frank Ferrone, from whom I learnt many teaching techniques, particularly in physics for life sciences. I appreciate the valuable teaching experiences gained through working with him that I can benefit from for my future career.

Finally I will thank my wife, Dong Han, who never stops encouraging me in my study and makes my life much easier.
Table of Contents

LIST OF FIGURES ........................................................................................................... v
ABSTRACT ...................................................................................................................... xi

1: Introduction ............................................................................................................... 1
  1.1 Background ........................................................................................................... 1
  1.2 Thesis construct ................................................................................................. 2

2: Review of the protein folding problem ................................................................. 5
  2.1 Protein structure and protein function ............................................................... 5
  2.2 Forces that drive protein folding ....................................................................... 7
  2.3 Diseases related to protein misfolding .............................................................. 10
  2.4 Protein folding in silico .................................................................................... 11

3: Effects of confinement on proteins ...................................................................... 14
  3.1 Background ........................................................................................................ 14
  3.2 2-Dimensional HP lattice model of protein ...................................................... 16
    3.2.1 Move set ...................................................................................................... 17
    3.2.2 The Metropolis algorithm ......................................................................... 18
  3.3 Simulation procedures ...................................................................................... 19
  3.4 The histogram method ..................................................................................... 22
  3.5 Results ............................................................................................................... 25
    3.5.1 Equilibrium properties ............................................................................. 25
    3.5.2 Mean first-passage times .......................................................................... 36
4: Effects of macromolecular crowding on protein

4.1 Brief review of research on macromolecular crowding
4.2 Brute force 2-Dimensional HP lattice simulation
  4.2.1 Simplified construct of the protein/crowder system
  4.2.2 The multihistogram method
  4.2.3 The replica exchange method
  4.2.4 Simulation procedure
  4.2.5 Results

5: Crowding effects studied using polymer physics

5.1 Review of the scaled particle theory
5.2 The concepts of the excluded volume and the depletion force
5.3 Crowding effects based on a simple protein model
5.4 Evaluation of excluded volume using a 2-D lattice model
5.5 Evaluation of excluded volume using a 3-D off-lattice model

6: Conclusions

Bibliography
Vita
List of figures

2.1: Amino acid structure and a chain of amino acids. Each monomer (or residue) represents an amino acid group. A water molecule is released during the condensation process in polymerization............................................ 6

2.2: Depiction of the motifs of secondary structure: α-helix and β-strand .............. 7

3.1: Comparison between the pseudo 3-D structure of protein 1BIJ (left) and 2-D HP lattice model (right). The structure of 1BIJ is plotted using software Rasmol. To the right is the native conformation of a 2-D model protein with sequence HHHHPHHPHHHHPHH. The simplification is maximal while the 2-D HP lattice model remains nontrivial ............................................. 17

3.2: Move set adopted. From the left to right are the end monomer swing, the three-bead flip and the crankshaft turnover. The gray circles represent the possible new monomer positions after the corresponding move and arrows represent the move directions ........................................................................... 18

3.3: Diagram to show the hindrance of box walls to the trial Monte Carlo move of protein. Monomers of protein cannot penetrate the hard wall. For example, the position of the end monomer (shown in gray) resulted from a trial end-flip is prohibited .................................................................................. 20

3.4: A typical normalized histogram p(E) of protein energy obtained from simulation run at $T = 400$ K. Sum of the 10 p(E) values is 1 ......................... 24

3.5: Protein energy diagram..................................................................................... 26

3.6: Average energy $\langle E \rangle$ as a function of temperature and box size. The abscissa is the absolute temperature in Kelvin and the ordinate is energy in $\varepsilon$. $\langle E \rangle$ increase with temperature but decrease with box size. From the top to bottom, the box sizes are: $20 \times 20$, $16 \times 16$, $14 \times 14$, $10 \times 10$, $8 \times 8$, and $7 \times 7$.................................................................................................................... 27

3.7: Average radius of gyration of protein confined in different boxes plotted as a function of temperature. As in the average energy plot, from top to bottom the box sizes are: $20 \times 20$, $16 \times 16$, $14 \times 14$, $10 \times 10$, $8 \times 8$, and $7 \times 7$, respectively. The dashed line is for $14 \times 14$, which is a little higher than that in $16 \times 16$, we believe this is due to the statistical error .......................... 28
3.8: S-E phase diagram obtained from a typical Monte Carlo simulation run at temperature $T = 400$ K and box size $20 \times 20$. Relative weight of points in the diagram due to the temperature and box size is revealed by the gray level. .................................29

3.9: Heat capacity as a function of temperature for the model protein in boxes with different sizes. From the left to right, the box sizes are: $20 \times 20$, $16 \times 16$, $14 \times 14$, $10 \times 10$, $8 \times 8$, and $7 \times 7$. .............................................30

3.10: Melting temperature $T$ plotted as a function of inverse box size $1/a$. From left to right the points represent cases of box $20 \times 20$, $18 \times 18$, $16 \times 16$, $14 \times 14$, $12 \times 12$, $10 \times 10$, $8 \times 8$ and $7 \times 7$, respectively. Vertical axis $T$ is in Kelvin.................................................................31

3.11: Diagram of a silica matrix pore within which an enzyme molecule is immobilized. $D$ is the cavity size and $d$ is the pore size, controllable by the amount of template molecules at the beginning of sol-gel process .............33

3.12: Percentage remaining activity of the free and immobilized acid phosphatase (ACP) samples assayed at pH 6.2 and $[pNPP] = 5.0 \text{mmol/L}$ after heating at $50^\circ\text{C}$ for 1 hour. The vertical axis is the percentage of the remaining activity compared to that of the solution sample at room temperature. The numbers, 1 through 8, label different samples used for the assay: sample 1 has zero wt% of glucose content in the original sample, sample 2 has 10 wt% of glucose content, 3, 4, 5, 6, 7, correspond respectively, to 20, 30, 36, 42, and 50 wt% of glucose content, and 8 denotes protein in solution .............................................34

3.13: Thermal stability of free and immobilized horseradish peroxidase. Samples were treated at $65^\circ\text{C}$ for 15 minutes per each cycle for 5 consecutive cycles. Activity of free and immobilized HRP at room temperature is 100%. Note that group 1 is for FH0 (<1.7 nm), group 2 is for FH42 (3.2 nm) and group 3 is for free solution .................................35

3.14: A rugged funnel-like energy surface of protein. Protein folding maps to a diffusion along the energy surface towards the lowest point, the native conformation. One route of protein diffusion towards the native state is shown ..........................................................37

3.15: Mean first-passage time for folding directly plotted as a function of temperature for the model protein in boxes with different sizes. The curve with solid dots is for $10 \times 10$, that with solid squares for $14 \times 14$, solid diamonds for $16 \times 16$, solid triangles for $20 \times 20$, and the stars for solution protein. Solid lines are drawn to connect data points. The temperature is in unit of kK, kilo Kelvin and the time is in arbitrary unit ......................................39
3.16: The natural logarithm of mean first-passage time for folding versus the inverse temperature. T is in kilo Kelvin. V-shaped trend is seen .................. 40

3.17: The fitted asymptotic lines of natural logarithm of mean first-passage time for folding. Note that the reversing phenomenon is clearly seen. Cross stands for $10 \times 10$ box case, solid triangle for $14 \times 14$, open diamond for $16 \times 16$, solid circle for $20 \times 20$, and open square for free protein. T is in kK... 41

3.18: Natural logarithm of mean first-passage time for unfolding vs. inverse temperature, fitted with linear lines. Open circles are for box $10 \times 10$, squares for $16 \times 16$, and triangles for free protein. T is in kK........................... 43

4.1: Simplification from crowded environment in cells to a simulation system. Within the right box, the black solid squares correspond to a collection of crowder agents and the 2-D HP lattice structure corresponds to the protein.... 47

4.2: A typical Brownian motion of a particle starting from point (0, 0)................. 48

4.3: Diagram to show the replica exchange method. Each protein conformation represents a simulation copy maintained at a certain fixed temperature. Two copies (may not necessarily be adjacent ones) will be swapped according to a well defined transition probability to satisfy the detailed balance. As an example a pair of nearby copies (linked by dashed arrows) is shown to be swapped.......................................................... 53

4.4: The series of temperatures at which the system replicas run. In principle any temperature profile can be taken, here we adopt the exponentially distributed temperature profile. 20 replicas are shown with the minimum temperature being 190K and the maximum temperature being 1500K............ 55

4.5: Heat capacity $C_v$ versus temperature T. The $C_v$ is in arbitrary unit and the temperature is in kilo-Kelvin. Counting from left to right except the leftmost curve, the crowder number is 100, 200, 300, 500, and 700, and the leftmost is for protein in dilute solution. Here the crowder size is $4 \times 4$..... 58

4.6: Averaged end-to-end distance S of a protein plotted as a function of the size of crowders a at $T = 800$ K, when crowders number is $N = 500$.............. 59

4.7: Mean first-passage time for folding at $T = 500$ K increases with the number of crowders. The crowder size is set to be $4 \times 4$................................. 60

4.8: The mean first-passage time for unfolding as a function of inverse temperature and crowder number. The curves are: solid line for $N = 1$, dot line for $N = 400$, dashed line for $N = 600$, long dashed line for $N = 800$, and dot-dashed line for $N = 1000$. Temperature T is in kK......................... 61
4.9: Equilibrium constant as a function of the number of crowders. K is defined to be a ratio of two mean first-passage times. \( T = 500 \, \text{K} \) ........................................ 64

5.1: A hard spherical particle is to be inserted into the solution composed of several types of hard spherical particles, each of which is maintained at its own concentration ........................................................................................................ 68

5.2: Pressure of a fluid of hard molecules plotted against the volume fraction. The straight line is the van’t Hoff’s law, the top curve represents the deviation of hard-sphere pressure from van’t Hoff’s law and the middle curve is that of the hard square pressure. The dashed curve comes from Carnahan and Starling, which is very close to the result of the scaled particle theory ........................................................................................................... 69

5.3: Depletion zone and excluded volume. Horizontal dot-dashed line represents the center to center distance between two immersed molecules and vertical line in the middle is the projection of the cross sectional area \( S \) ........................................ 71

5.4: The protein is modeled as a solid sphere that can deform into an ellipsoid with the same volume. The regions enclosed in the dashed lines are the excluded volume. The more the protein deviates from spherical shape the more the true surface that encloses the excluded volume deviates from the dashed ellipsoidal surface. Obvious difference between the ellipsoidal surface and true surface is depicted in the bottom shape................................. 74

5.5: Normalized extra free energy and extra force plotted as functions of \( b/R_0 \) ratio at fixed radius of crowder. Thicker line represents the normalized extra free energy ........................................................................................................ 76

5.6: Depletion force (in pN) as a function of protein long axial length and dextran concentration. From top to bottom, the curves correspond to 300g/l, 250g/l, 200g/l, 150g/l, 100g/l and 50g/l ................................................................. 78

5.7: Peak extra force (in pN) as a function of dextran concentration ............... 78

5.8: Mechanical force to unfold ubiquitin molecule using AFM plotted as a function of dextran concentration. Horizontal is dextran concentration in unit of g/l, and vertical represents the pulling force measure in pN. In each group four increasing pulling speeds (nm/s) are presented: 50, 250, 1000, and 5000 ......................................................................................................................... 79

5.9: An example of a short 2-D lattice polymer and its envelope layer of square macromolecules of size \( 3 \times 3 \). The region enclosed within the complete piece-wise line is the exclude volume. Note that a free region is also counted ......................................................................................................................... 81
5.10: Excluded volume $V_{ex}$ (in lattice unit square) increases with the end-to-end distance $S$ of the protein polymer (100 monomers long) in a nonlinear way. From the bottom up the square crowder sizes are respectively, $1 \times 1$, $3 \times 3$, $5 \times 5$, $7 \times 7$, $9 \times 9$, and $11 \times 11$. Data dispersion due to insufficient sampling can be seen near the long end-to-end distance side.......................... 82

5.11: The quadratic coefficient of excluded volume as a function of protein end-to-end distance plotted versus crowder size $R_c$. The trend becomes stronger with the increase of polymer size. $k$ here is the coefficient of the 2nd order term in fitted quadratic form......................................................... 83

5.12: Depletion force along the lattice polymer end-to-end distance. Panel a is for square crowder of size $1 \times 1$, panel b for $5 \times 5$, panel c for $9 \times 9$, and panel d for $13 \times 13$. In each panel four curves at increasing crowders volume fraction, 10%, 20%, 30%, and 40% are shown from top down. The vertical axis stands for the normalized depletion force $F/ k_B T$ in unit of reverse lattice spacing, and the horizontal is end-to-end distance. If lattice spacing is $l = 3.4$ nm, an interval of 0.1 along force axis represents 1.217 pN at $T = 300$ K ................................................................................................ 85

5.13: Depletion potential energy along the lattice polymer end-to-end distance. Panel (a) is for square crowder of size $1 \times 1$, panel (b) for $5 \times 5$, panel (c) for $9 \times 9$, and panel (d) for $13 \times 13$. In each panel four curves at increasing crowder volume fraction, 10%, 20%, 30%, and 40% are shown from bottom up. The vertical axis stands for the depletion energy in $k_B T$, and the horizontal axis stands for polymer end-to-end distance ......................................................... 86

5.14: Calculation of 2-D depletion area and 3-D depletion zone. Dashed circles represent the depletion area or depletion zone around an individual monomer. To the left an irregular polygon is produced and will be triangulated. The dot-pattern rectangle to the right is the smallest rectangle parallelepiped that accommodates the 3-D depletion zone. Random points are drawn from it for the second layer Monte Carlo approximation ............... 88

5.15: Principle working mechanism of the pivot algorithm. The monomer labeled by an arrow is the pivot about which the shorter arm of the polymer can rotate an angle triplet $(\alpha, \beta, \gamma)$ randomly selected, if sterically permitted. The curved arrow shows the rotation direction.............................. 89

5.16: Natural logarithm of excluded volume $\ln(V_{ex})$ as a function of end-to-end distance $S$ of the 3-D off-lattice 76mer. $S$ is in unit of 0.1 nm. The minimum possible end-to-end distance is a bond length $b = 0.34$ nm and the maximum is 25.5 nm. Increasing from bottom up, the ten curves are calculated in simulation for crowders of size 0.5 nm, 1.0 nm, 1.5 nm, 2.0 nm, 2.5 nm, 3.0 nm, 3.5 nm, 4.0 nm, 4.5 nm, and 5.0 nm ......................... 91
5.17: The logarithm excluded volume $V_{ex, relative}$ (relative to volume of monomer volume) increases with crowder size. Number of monomers is $N = 76$ .......... 92

5.18: Depletion force-extension profile. Shown here are for crowders with sizes 1 nm (thicker line) and 3 nm (thinner line) respectively. The force is in unit of pN and the extension (end-to-end distance of polymer) is in 0.1 nm. The temperature is $T = 300$ K and the volume fraction is 40% ......................... 94

5.19: Bar diagram of maximum depletion force $F_d$ plotted as a function of both the crowder volume fraction and crowder size. Each number along the abscissa is the crowder size in unit of nm and each size includes four increasing crowder volume fraction, 10%, 20%, 30%, and 40%, respectively. Ten increasing sizes are shown. The vertical label represents the force in pN. The temperature for this plot is set to be $T = 300$ K ......................... 96

5.20: Maximum depletion force versus the volume fraction of crowder with radius 2 nm as an example to show the increasing trend of force ......................... 97

5.21: Bar diagram of maximum depletion force plotted as function of both number density of crowders and crowder size. Each number along the abscissa is the crowder size in unit of nm. Each bar group includes four increasing number density of crowders $7.639 \times 10^{23} \text{ m}^{-3}$, $1.146 \times 10^{24} \text{ m}^{-3}$, $1.528 \times 10^{24} \text{ m}^{-3}$, and $1.910 \times 10^{24} \text{ m}^{-3}$ respectively. Ten increasing sizes are shown. The vertical label represents the force in pN. The temperature for this plot is set to be $T = 300$ K ................................................. 97

5.22: Depletion free energy (in unit of $k_B T$) plotted against end-to-end distance of polymer at fixed volume fraction of crowders, $\varphi = 28\%$. From top down the crowder size goes from 0.5 nm to 5.0 nm uniformly. This trend is reversed compared to that of the excluded volume, seen in figure 5.16.......... 99

5.23: Cartoon of free energy surface of two-state protein folding................................. 101
Abstract
Effects of Confinement and Macromolecular Crowding on Protein Stability and Folding Dynamics
Guanghui Ping
Jian-Min Yuan, Ph.D.

In cells, proteins perform specific tasks in crowded and confined environments; these environments influence their stability and folding dynamics. To investigate these effects of confinement and macromolecular crowding on proteins based on statistical mechanical methods, we have carried out numerical simulations using minimalist models (2-Dimensional HP lattice, Monte Carlo, Brownian dynamics) and 3-Dimensional off-lattice polymer models. For the confinement effects, our results, based on heat capacity calculations, show that the folding temperature increases when box size decreases, indicating that the protein is stabilized. These results are consistent with the experimental observations obtained in Dr. Wei’s group. We have also investigated the effects of confinement on the kinetics of refolding and unfolding as a function of temperature and box size. The unfolding time increases as box size shrinks, however, the folding time behaves in a more complicated way.

To investigate the effects of macromolecular crowding, Brownian motions of crowders are included inside a virtual box with periodic boundary conditions. Besides temperature, the concentration of crowders and crowder size are also varied in our simulations. Simulated results indicate that folding temperature increases with the crowder concentration; the protein is thus stabilized by the presence of crowders. However, this increase is not as large as that observed in the case of confinement. Our dynamic studies show that both folding and unfolding times increase with the
concentration of crowders in such a way that the equilibrium is shifted towards the folded state. Furthermore, our simulations show that the activation energy of unfolding remains approximately constant as the number of crowders increases.

Based on the concept of depletion force we have also calculated the enhancement of the mechanical force required to unfold ubiquitin molecules in a solution of dextran, the crowding agent. We have employed a 3-D polymer model using the pivot algorithm to calculate the depletion zone and have applied the scaled particle theory for the osmotic pressure of the crowders. Our results are in reasonable agreement with recent measurements carried out in Dr. Yang’s laboratory.
Chapter 1: Introduction

1.1 Background

Modern biological sciences do not follow traditional lines any more. In a broader sense, to understand biological phenomena better, scientists find it indispensable to integrate efforts from traditional biology, chemistry, physics, mathematics, engineering, and computer science to widen our perspectives at many different levels. Rich phenomena to understand can range from nanosecond (protein folding) to millions of years (evolution) in time, and from nanometer (protein structure) to the whole globe (ecology) in space.

Proteins refer to a large family of bio-molecules that play key roles in life, from producing genetic materials to carrying out vital metabolic and enzymatic functions in cells. Widespread experimental and theoretical studies on protein structure-function relationships have revealed that a nascent amino acid chain needs to fold into specific 3-dimensional structure in order to function properly. The proteins primary structure, namely, the amino acid sequence, uniquely determines the protein 3-D structure. Studies on the effects of various environmental factors on protein behaviors have been widely conducted. Physical and chemical factors, e.g., the pressure, the pH value, and the temperature, will drastically affect the protein properties. Recently, there is a growing interest in the study of proteins confined in limited space and/or in crowded environments, because of their fundamental biological and practical importance. Proteins confined in small spaces are not merely an interesting academic problem, but shed lights on some properties of new drug delivery systems. In the real
cells proteins exist in crowded environments and this crowdedness affects their behaviors, such as the folding rates and the thermal stability, due to the inherent complexity of the interactions (even we assume only hard-sphere interactions) between the protein and the surrounding macromolecules. Despite a long history of research, the study of macromolecular crowding effects remains difficult and needs more advanced theories and experimental techniques.

1.2 Thesis construct

The research in this thesis draws in large parts from long-term simulations, our published papers [1, 2], and extensive literature search. Six chapters of the thesis are organized as follow:

Following this introduction chapter, we review, in chapter 2, a wide spectrum of research on the protein folding problem. Being one of the central problems in biological sciences, many interesting topics are involved, including the topics that form the core of this thesis, namely, protein folding in confinement and in crowded environments. Lack of simulation studies on these topics will be addressed.

Chapter 3 focuses on the confinement effects on protein folding and protein stability. We address the biological and practical importance of this problem and briefly review the methodologies, including model building and simulation techniques. We have performed Monte Carlo simulations of a 2-D HP lattice model of protein (a 16mer) confined in square boxes with varying sizes. The average energy and size of the model protein was collected and the heat capacity information was computed based on the histogram method. We have demonstrated that the peak
temperature of heat capacity, defined as melting temperature of protein, is raised and the thermal stability of the protein is enhanced when the protein is confined in smaller boxes. Through dynamic studies we monitored the first-passage times for folding and unfolding processes with the help of parallel computational techniques. Our results reveal that protein folds faster in small confinement at high temperature and slower at low temperature, compared to the bulk protein. At any temperature the unfolding rate is always retarded, due to the blocking function of the hard wall against the protein conformation changes within the box. The conclusion of this confinement study is that protein is thermally stabilized.

In chapter 4 we discuss our studies on protein properties in a crowding environment. By employing the same HP model with minor modifications, we investigate protein properties in a solution of 2-D square crowders. As the foregoing, we have computed heat capacity and folding/unfolding times. Serious trapping problems need multihistogram and replica exchange methods. The results reveal the existence of the stabilizing effects of dense crowding environments on protein native conformation, but effects are less pronounced as the counterpart in the confinement case. According to simulated data, the folding and unfolding times are both retarded, but they are retarded in different rates, so that the equilibrium constant, defined as the ratio of the two mean first-passage times, is shifted in favor of the native conformation. From this perspective, protein is more stabilized at native conformation than extended structures in crowded solutions.

In chapter 5 we carry out homopolymer simulations rather than that on HP protein. Assuming only hard-sphere interactions, all effects must come from the excluded
volume due to the presence of crowders. We therefore focus on the excluded volume calculations for both a 2-D lattice homopolymer and a 3-D off-lattice homopolymer using lattice-counting and double Monte Carlo simulation techniques, respectively. We show that a depletion force acting on the polymers exists due to the presence of macromolecular crowding in solution. The calculated depletion force and depletion potential energy indicate that polymer tends to recoil into a compact structures in crowded solution, and this depletion force should correspond to the extra crowder-induced force required to unfold a single protein molecule observed in an AFM protein-pulling experiment carried out in Dr. Yang’s laboratory.

In chapter 6 we give a summary of the simulational and theoretical results on the two main topics, and point out several future research topics related to current ones.
Chapter 2: Review of the protein folding problem

2.1 Protein structure and protein function

All living organisms, from viruses to bacteria and to human beings, produce proteins that play crucial and diverse biological roles in maintaining their life characteristics. For examples, viruses, the simplest form of life on earth, have a coat of proteins that enclose their genetic molecules inside; being another example, hemoglobin is a type of protein in red cells that carries and delivers oxygen molecules from lungs to tissues in blood circulation. It was measured that various protein molecules make up more than one half of the dry weight of human cell [3]. The central dogma of biology [3] asserts that DNA, the fundamental genetic material, can replicate itself, and the segments of it, the genes, are transcribed into messenger RNA molecules. mRNA’s are translated to synthesize a string of amino acids, the building blocks of more than $10^5$ different proteins, the terminal products in this process.

An amino acid is composed of an amino group $\text{NH}_2$, which is electrochemically basic, an acidic carboxyl group COOH, an alpha-carbon atom linking the two groups, a hydrogen atom H, and a side chain R connected to the alpha-carbon that carries the individual characteristic of each amino acid. The general structure of the amino acid is depicted schematically in figure 2.1. It is the side group R that determines the individual amino acid. In nature there are (at least) 20 different amino acids in total, and this forms the reason why a large number of protein sequences and structures exist, considering the combinatorics. Based on their structures and chemical
properties, the 20 amino acids can be categorized into two groups: those that have stronger interactions with water molecules, called polar or hydrophilic, and those do not, called nonpolar or hydrophobic. A string of amino acids, namely, a polypeptide chain, is formed via a condensation process in which a water molecule is removed, when a new monomer is added into the chain.

![Amino acid structure and a chain of amino acids](image)

Figure 2.1: Amino acid structure and a chain of amino acids. Each monomer (or residue) represents an amino acid group. A water molecule is released during the condensation process in polymerization.

Protein structures are formed in a hierarchical manner. The specific sequence of amino acids of a protein is called its primary structure. To describe amino acids linked by peptide bonds, degrees of freedom such as bond length $b$ between two residues, bond angle $\theta$, and dihedral angle pair $(\phi, \psi)$ are used. The large number of conformations of a protein is largely due to the flexibility of the $(\phi, \psi)$ angle pair. Extensive experimental and theoretical research has been done in revealing the structures of proteins. It turns out that, from the arrangement of local amino acids, a set of common secondary structures appears. Generally speaking, these are the $\alpha$-helix and $\beta$-strand motifs. Amazingly the $\alpha$-helix secondary structures were firstly
proposed by Pauling et al. in their theoretical studies of the stabilizing hydrogen-bonding energy [4-10], several years earlier than the X-ray crystallography verification of such structures done by Perutz and Kendrew [11]. The tertiary structure of proteins is the unique 3-dimensional structure the protein has to correctly acquire in order to function properly. In figure 2.2 an example of secondary structures of protein, α-helix and β-strand, are shown.

Figure 2.2: Depiction of the motifs of secondary structure: α-helix and β-strand.

2.2 Forces that drive protein folding

It is now widely believed that the information of the unique 3-dimensional structure a nascent chain of amino acids must fold into is stored in the primary structure of the protein. In early 1960’s and 1970’s Christian Anfinsen [12-14] proposed and verified that “The native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment” [15]. However, how the specific sequence of protein determines the unique 3-D functional structure and governs the folding process still remains a problem to be solved. To maintain life
characteristics, nascent chain of amino acids must fold into a correct 3-D structure quickly and be transported to designated locations in a cell.

In 1968, Levinthal raised a paradox in studying the pathways of protein folding [16]. If a sequence of amino acids is approximated as a freely jointed chain, the minimum size of conformation space will be in the order of $2^N$ where $N$ is the number of amino acids and the base 2 is assumed to be degree of freedom of each residue. This is an astronomically huge number. At first glance, to visit all points in this large conformation space before folding into its unique native structure, a protein with 75 amino acids will need approximately at least one year to find the lowest free energy state (native conformation), if it is assumed that any conformation needs only $10^{15}$ s to transform into any other conformation. This conclusion contradicts the fact that proteins can fold into their native structures reliably and quickly within much shorter time that is in order of milliseconds, or at most minutes. In Levinthal’s reasoning, the conformations of protein are mutually independent and the searching of native conformation can be mapped to a simple random walk in a wide phase space with an extremely small probability to hit one single specified point, native conformation of protein.

Ramachandran, et al. [17, 18] systematically computed the dihedral angle pair $(\psi, \phi)$ for small polypeptides through finding the energetically favorable conformations. The contour diagram that locates the dihedral angle pairs of stable conformations is referred as the Ramachandran plot. It demonstrates that dipeptides only occupy limited and distinct regions on the $(\psi, \phi)$ plane. Rose [19] et al. reexamined the Ramachandran plot by performing statistical analysis on the exhaustive conformation
sampling of short and simple polypeptides and verified that protein can not be
considered a chain of mutually independent amino acids due to steric hindrance
between protein backbone monomers. The conformation space of a protein is thus
much smaller than that Levinthal proposed. It was also proved that a protein does not
have to sample most of the conformation states during a folding process. In other
words, protein folding is actually a cooperative process [20, 21]. During folding, the
preceding conformation of a protein can logically direct and improve the successive
folding process, creating narrower conformation space to help the protein fold
rapidly.

A decade ago, Wolynes and Onuchic [22, 23] developed a funnel energy
landscape theory in explaining protein folding. Also, Chan and Dill [24] proposed the
energy landscape to be funnel-like with small energy frustrations and with small non-
native local minima. The higher energy part of the funnel is wide and stands for a
large conformation space in which the protein takes extended (unfolded) structures.
The funnel becomes thinner when the free energy is lowered. In protein folding, the
free energy funnel directs the protein towards a well-defined global minimum point in
a non-smooth way due to the energy frustration along the funnel surface. Starting
from any open conformation, an amino acid chain will first collapse into a globular
structure, largely because of the existence of hydrophobic force between water and
hydrophobic amino acids. Hydrophobic monomers of proteins tend to be buried
inside and form a core and polar monomers tend to stay on the protein surface,
forming a shield that partly separates hydrophobic monomers from water. The
globular structure is then geared to form the necessary hydrogen bonds between
residues. During folding, the enthalpy gain competes with the loss of entropy [25] and as a result, there will be a folded potential well, unfolded potential wells, and transition state barriers that separate them.

### 2.3 Diseases related to protein misfolding

Some human and animal diseases are related to improper protein folding. These result from the fact that some responsible proteins, particularly the neuronal membrane proteins, do not fold into their correct 3-dimensional structures, a phenomenon called protein misfolding. Often misfolded proteins can lead to aggregation of proteins and produce insoluble amorphous or fibrillar structures, called amyloid fibrils. These fibrils can deposit and form spongy-like structures in a variety of tissues, particularly, in the brain. Alzheimer’s disease is such an example. Another type of protein misfolding disease, prion disease [26], which is usually termed spongiform encephalopathy, too, is induced by so called prion protein that infects responsible proteins and drives them to misfold. The most famous example is the Mad Cow disease. The common feature of these diseases is protein misfolding and aggregation, either triggered by prion protein, or simply by aging. As a result, these diseases damage the cognitive and motive ability of animals or humans, usually leading to death. It was pointed out that the fibrillar aggregates are a generic form of polypeptide structure under some specific circumstances [27]. A great deal of work has been done in revealing the misfolding and aggregation mechanisms and in ways to prevent the transition process from normal protein to misfolded ones. There are still many open questions that are vital in the understanding of the diseases and in the
treatment of such diseases. Protein folding, misfolding, and aggregation are important research topics with a great practical significance. Our studies of the effects of confinement and crowding on protein stability and dynamics are stimulated by this perspective.

2.4 Protein folding in silico

Besides experimental and theoretical studies, an ensemble of computer simulation approaches plays increasingly important roles in protein folding research. Various computational techniques have been implemented depending on the characteristics or aspects of the problem under study.

The earliest and simplified simulation approaches dated back to more than 40 years ago are based on Ising models [28]. Among More recent simplified models are lattice models, which can be categorized into two basic tracks. The first was invented by Go and coworkers [29] who aimed at extracting fundamental physics that determines the protein folding process from their lattice models. Many authors contributed along this track. The energy funnel idea was proposed by Wolynes, Onuchic, and Thirumalai [23], based on analyzing results of lattice models. Chan, Dill, and coworkers [24] focused on hydrophobic and hydrophilic interactions on lattice models and concluded that the driving force is mainly due to hydrophobicity of the polypeptide chain. Another track of lattice simulations uses potential energy derived from existing protein structures obtained from Protein Data Bank (PDB), in a statistical way. Authors along this track include Skolnick and coworkers [30], Miyazawa and Jernigan [31].
The usage of simple lattice model stems from its simplicity and the effective conformation sampling, so that ensemble averages of many protein properties can be calculated. Under certain circumstances, even all exhaustive enumeration of the protein conformations could be obtained which makes possible exact evaluation of the partition function, the most important function in calculating protein properties. On the other hand, the lattice model is a coarse-grained model that lacks detailed description of parameters in protein folding. A natural extension of lattice models is an off-lattice model, in which protein residues or atoms of residues are not confined on lattice sites, but can move continuously with the steric hindrance between residues considered. The bead model is such an off-lattice example. In this model, the amino acid residues are represented by beads of distinct types, and some physical energy profile is assigned between residues. In contrast to a lattice model which needs to adopt a specific move set in sampling the protein conformation space, off-lattice models require numerical integration, within the time domain, to develop dynamic trajectories in the phase space of a protein molecule.

At the top level of the in silico simulations of protein folding are the all-atom models using molecular dynamics (MD). There exist several versions of all-atom MD models. In the first, the earliest ones, all atoms in each protein monomer are represented and inter-atomic interaction potentials are included, but no solvent effect is enclosed. In the second version, the protein is represented using an all-atom approach, and in order to include the effects of solvent (water), a continuum model of solvent is added in. The third version uses all-atom method to represent both protein molecule and the solvent molecules, usually leading to a large number of atoms, for
example an order of magnitude $10^5$, in systems under study. These all-atom models need accurate descriptions of interaction potentials between different types of atoms, which can only be obtained through experiments and high-accuracy quantum mechanical computations. The large number of atoms in the system and the short time step ($10^{-15}$ s) required in the numerical integration make the tracking of a complete protein folding process, typically in a time scale of at least $10^{-6}$ s, difficult due to the limited power of current computer systems. Currently all-atom models are often used to study the folding of short polypeptides or proteins for shorter durations. Another direction of all-atom MD models is to track the unfolding process and establish the free-energy landscape [32-34]. The advantage of all-atom models is their accurate description of interaction potentials for the study of protein folding and protein unfolding processes. To answer the question ‘How does the amino acid sequence determine the protein 3-D structure?’ in detail, all-atom protein models with explicit solvent are the generally accepted tools. Positive results have been obtained for short polypeptides [35, 36]. With increasing computer power, it will become possible to track the complete protein folding process. Recently, Pande et al. at Stanford University developed a distributed system termed Folding@home [37] that can integrate contributions of computers worldwide to study the protein folding process. Recently, Blue Gene supercomputers at IBM have been used to produce interesting results of protein folding using ensembles of all-atom MD trajectories.
Chapter 3: Effects of confinement on proteins

3.1 Background

Inside a cell a protein binds and interacts with other molecules, modifies its conformations via folding/unfolding to function well in a crowded and/or confined environment. The free volume available to the functional protein is limited either by the dense surrounding biomolecules, or by small confinements. As a result the reduced free volume availability will affect the protein stability and folding dynamics, in a nontrivial way. Using circular dichroism in the UV region, Eggers and Valentine [38] showed that the thermal stability of an enzyme, α-lactalbumin, when encapsulated in a silica matrix, is enhanced. Chaperones (Chaperonin folding machines), a class of proteins found in all organisms, can help nascent proteins fold into their correct native conformations [39]. Two mechanisms were proposed in understanding chaperonin action: the Anfinsen cage model and iterative annealing model. In distinguishing these two working mechanisms, Brinker [40] found that the folding of denatured protein in narrow space is accelerated compared to that in free solution, supporting the Anfinsen cage model. Encapsulated in small space, the thermal stability, chemical reactivity and folding dynamics of proteins are expected to be affected.

Confinement or entrapment of proteins is also important for practical applications which drive people to understand the folding and stability in confinement better. Using entrapped proteins as biocatalysts or biosensors [41] is such an example. Individual biomolecules can be encapsulated, preventing possible self-aggregation
and protecting the biomolecules from microbial degradation attacks from bacteria. Smaller substrates that can react with the immobilized biomolecules can move freely in and out of the confinement, rendering possible to create sensitive biosensors. A good example is to monitor the blood glucose using enzyme glucose oxidase (GOx) immobilized. Other advantages of using encapsulated enzymes reside in the continuous usage and easy separation and recovery of these heterogeneous biocatalysts from the reaction mixture. Much work has been done in this direction [42]. For best result in biosensing or biocatalysis the size of pores in which the biomolecules are immobilized should be controllable and narrowly distributed within a range so that substrate molecules can easily move into the pores and the leaking out of pores is minimized. Technically this remains a problem [41]. Recently, Dr. Wei’s group at Drexel University has successfully obtained mesoporous silica matrices with controllable and narrowly distributed pore sizes, using sugars or other non-surfactant compounds as template in the sol-gel process [43-47]. The particular advantage of their synthesis is that technically it can be achieved under mild, biocompatible conditions (e.g., pH 7 and room temperature), making it practically easy to encapsulate various biomolecules into the silica matrices.

Along the theoretical track of studies, Minton [48, 49] and Zhou [50] have applied statistical thermodynamics and diffusional polymer models, respectively, to investigate the effects of confinement on protein properties. In their work, much attention is concentrated on the protein equilibrium properties, such as change of free energy and chemical potential and protein activity.
In this chapter, we aim to provide a study of kinetics of protein entrapped in confinement in addition to the protein stability using a minimalist protein model and present qualitative comparison between the theoretical and experimental results.

### 3.2 2-Dimensional HP lattice model of protein

Lattice models are extensively used throughout polymer and protein sciences. Flory [51] successfully applied square (or cubic) lattice models in studying the statistical properties of polymers. Each solvent molecule and residue of the polymer is modeled as a lattice site, and the bond between two linked monomers is modeled as the center to center distance between two adjacent lattice sites. Advanced mathematics and statistics on lattice models in polymer physics have been developed [52]. A variety of lattice models have been devised and employed to represent the proteins, too. Dill, Karplus, Shaknovich, and Skolnick are pioneers in applying the lattice models in protein folding. In our study on confinement effects we found that a 2-D HP lattice model is a good candidate due to the limits of the computer resources. This HP model was firstly introduced by Chan and Dill [53] and it has been successfully and fruitfully applied and extended to a great number of simulation studies [54-56].

Given a 2-D square lattice, a protein is represented as a series of connected lattice sites with the number of sites being the number of amino acids of the protein. The bond length between any two consecutive lattice sites is fixed. In HP model the H-bead represents a hydrophobic monomer which tends to be buried in the core of protein, and the P-bead represents a polar monomer that tends to stay on the protein surface, interacting with water. Only non-bonded nearest neighboring H-H contact is assigned a negative energy unit in order for the protein to fold and the hydrophobic
residues to be buried inside, where the idea that the main driving force for protein folding is the hydrophobic interaction is assumed. We assign this energy to be $\epsilon = 1.38944 \times 10^{20}$ J. In so doing we have already included the effects of solvent molecules implicitly in the simulation setup. Figure 3.1 depicts the structural difference between a real protein with a pseudo 3-D structure obtained using molecular graphics software Rasmol and a 2-D square HP lattice entity, which is the minimalist model adopted in our simulation studies.

![Figure 3.1: Comparison between the pseudo 3-D structure of protein 1BIJ (left) and 2-D HP lattice model (right). The structure of 1BIJ is plotted using software Rasmol. To the right is the native conformation of a 2-D model protein with sequence HHHHHPPHPPHHHHHPHH. The simplification is maximal while the 2-D HP lattice model remains nontrivial.](image)

### 3.2.1 Move set

To simulate how a lattice polymer or lattice protein may change its conformations, we need to define a move set, i.e., a set of movement on a lattice that may allow the lattice polymer or protein to modify its conformations by steps. To study protein folding, either its equilibrium or dynamic properties, it is important that all conformations of the polymer can be reached by executing the move set. This is called the ergodic condition. Socci and Onuchic showed that for a short HP lattice
chain three types of moves are enough to satisfy the ergodic condition provided that
the number of simulation steps is long enough [57]. They are, respectively, the end
swing in which the end monomer can rotate 90 degrees clockwise or
counterclockwise, the three-bead flip in which the middle bead in three linking
monomers that forms a triangle can flip in the diagonal direction, and the crankshaft
turnover in which a crankshaft-like structure can turn over. These moves are shown in
figure 3.2. There exists a fourth move which involves the rigid rotation about a
randomly selected monomer. We did not include it in our simulations for no specific
reason but simplicity.

![Figure 3.2: Move set adopted. From the left to right are the end monomer swing, the
three-bead flip and the crankshaft turnover. The gray circles represent the possible
new monomer positions after the corresponding move and arrows represent the move
directions.](image)

3.2.2 The Metropolis algorithm

In a broad sense any simulation involving the application of random numbers can be
called a Monte Carlo simulation. An algorithm for Monte Carlo simulations in the
statistical mechanics was originally designed by Metropolis [58] in 1953. From the
perspective of statistical physics, properties of a system in equilibrium can be
calculated as the weighted averages of the system in a canonical ensemble and the results correspond to the measured quantities in experiments. The distribution of the ensemble of microstates is determined by the well-known Boltzmann’s factor $e^{-E_i/k_BT}$, where $E_i$ is the system’s energy of any microstate $i$, $k_B$ is Boltzmann’s constant and $T$ is the absolute temperature. The Metropolis algorithm determines the transition probability between any two microstates. In the context of our 2-D HP lattice model, for any two consecutive conformations of the protein, $\mu$ and $\nu$, we label their energies by $E_{\mu}$ and $E_{\nu}$. The Metropolis algorithm is described by the following criteria:

$$
\begin{cases}
  e^{-(E_\nu-E_\mu)/k_BT} > \chi & \text{state } \nu \text{ is accepted,} \\
  \text{otherwise} & \text{remain at state } \mu
\end{cases}
$$

where $\chi$ is a random number drawn uniformly from 0 to 1. It is proved that this criterion guarantees the detailed balance condition and Boltzmann’s distribution. Metropolis algorithm is the building block of many fancy Monte Carlo algorithms.

### 3.3 Simulation procedures

The protein confined in a 2-D box can move freely if no collision with the walls occurs. A Brownian-like motion is assumed to mimic the translational behavior of protein in confined space. Each step of the Brownian-like motion is arbitrarily taken to be one lattice site translation of the protein’s center of mass. Besides the Metropolis algorithm in free volume, we treat the walls of confinement box as hard edges so that any trial change of conformation of the protein that penetrates into the wall will be rejected.
Since we are basically interested in folding and unfolding processes, there will be two simulation procedures. For protein folding, one is interested in the folding dynamics and the folding time [59], which is defined as the Monte Carlo steps in a lattice model, starting from a set of random unfolded conformations. Even for a short HP sequence the number of nonnative conformations is very large and the protein folding time starting from any fixed conformation fluctuates drastically, a reason why lattice models work. To study the behavior of protein folding as a function of temperatures, we discretized the temperature domain into a series of values with equal intervals. The following procedures are implemented.

1. At high enough temperature the native state unfolds for a certain number of Monte Carlo steps in the absence of a box; save 599 nonnative structures with the collecting intervals large enough to minimize the correlation between conformations.
2. At a selected box size and temperature, randomly choose a structure from
the conformation pool of 599 as an initial conformation. Propagate the
trajectory until the protein reaches the native state. Each move is followed
by a Brownian-like translational move. When a trial residue move collides
with the walls of the box, it is simply rejected.
3. Record the number of MC steps as the folding time and repeat Step 2,
until exhausting all 599 conformations; perform statistics calculation.
4. Move to higher temperature and repeat Step 2 and 3.
5. Choose another box size and repeat Step 2, 3, and 4.

For an unfolding process, all simulations initiate from the native conformation. We
are particularly interested in the equilibrium properties of protein, such as the average
energy and the heat capacity at various temperatures. For this purpose, we should cut
off a transient period before sampling the conformations. It turns out the length of the
transient period increases as temperature decreases. For simplicity we take the longest
transient period of all trial simulations. Another problem is the difficulty in defining
the unfolded states. In principle we can argue that any structure that is different from
the native structure is unfolded. It is thus not strictly defined when a protein is
considered unfolded. This situation leaves us with the possibility to subdivide the
ensemble into sub-ensembles and there is freedom in choosing a sub-ensemble of
structures as the unfolded ensemble. After examining the fluctuations of protein
radius of gyration $R_g$, the end-to-end distance $L$, and the energy $E$ in a long single
run, we choose the unfolding truncating criteria to be: $R_g$ larger than 1.9 lattice
spacing and $E$ less than or equal to one energy unit, $\epsilon$. Here $L = |r_i - r_N|$ is defined to
be the distance between the first and the last residue of the protein, and
\[ R_g = \sqrt{\frac{\sum_{i=1}^{N} (r_i - r_c)^2}{N}}, \]
where \( r_i \) is the location of \( i^{th} \) residue on the lattice, \( r_c \) is the location of center of mass, and \( N \) is the number of residues of protein. For the unfolding simulation we follow the procedure below:

1. Fix a box size, start with the native conformation at a selected temperature and propagate the trajectory according to the Metropolis algorithm; whenever a move hits the hard walls of box, reject it but count as one MC step.
2. Stop the run when the protein satisfies the unfolding conditions; record the MC steps as the unfolding time.
3. Repeat Step 1 and 2 10,000 times for the statistical calculation of unfolding time.
4. After the 10,000\(^{th}\) successful reach of the unfolding conditions, continue running the program for many MC steps until the common transient period is passed and equilibrium is assumed; record the protein statistical properties such as energy, radius of gyration, and end-to-end distance.
5. Vary the temperature and box size, repeat Step 1 through 4.

### 3.4 The histogram method

In calculating the heat capacity of protein, \[ C_V = \frac{\partial E}{\partial T}, \]
we need the energy \( E \) as a continuous function of temperature \( T \). The traditional way to approximate the first derivative of \( E \) with respect to \( T \) is to simulate the system in the discretized
temperature domain, and the heat capacity is calculated as \( C_v = \frac{\Delta E}{\Delta T} \). To overcome this shortcoming of discretization, in their 1988 paper [60], Ferrenberg and Swendsen designed a histogram method which can extract much more information from a single Monte Carlo simulation, particularly useful for phase transition problems. We briefly review their algorithm below.

The histogram, the probability for a canonical system to be in states with fixed energy \( E \) at temperature \( T \) is defined by:

\[
p_\beta(E) = \frac{N(E)e^{-\beta E}}{Z(\beta)} \tag{2.1}
\]

where \( N(E) \) is the density of state at energy \( E \), \( k_B \) is Boltzmann’s constant, \( \beta \) is \( 1/k_B T \), and \( Z(\beta) \) is the partition function, given by:

\[
Z(\beta) = \sum_E N(E)e^{-\beta E} \tag{2.2}
\]

The idea of the histogram method is to evaluate the probability for the system to be in states with energy \( E \) at an arbitrary temperature:

\[
p_\beta'(E) = \frac{N(E)e^{-\beta'E}}{Z(\beta')}
= p_\beta(E)e^{\beta E} \frac{Z(\beta)}{Z(\beta')}
\tag{2.3}
\]

but:

\[
\sum_E p_\beta(E)e^{\beta E}e^{-\beta'E} = \sum_E N(E)e^{-\beta E} / Z(\beta)
= Z(\beta') / Z(\beta) \tag{2.4}
\]

Substituting equation 2.4 into equation 2.3, we get the following identity:
\[ p_\beta(E) = \frac{p_\beta(E) e^{\beta E} e^{-\beta E}}{\sum_E p_\beta(E) e^{\beta E} e^{-\beta E}} \]  

(2.5)

Therefore in principle for a fixed energy \( E \) the probability for the system to be at any temperature \( T' \) can be expressed in terms of the probability for it to be at the simulated temperature \( T \). Any energy-dependent observable \( O \) will then be calculated as the weighted sum given by:

\[ \langle O(\beta') \rangle = \sum_E p_\beta(E) O(E, \beta') \]

Using the histogram method, the average energy at any temperature can be extracted from a single simulation at a trial temperature, for example, the peak of heat capacity, which is our main interest. It turns out that the simulated temperature can not be

![Figure 3.4: A typical normalized histogram \( p(E) \) of protein energy obtained from simulation run at \( T = 400 \text{ K} \). Sum of the 10 \( p(E) \) values is 1.](image)
too far from the temperature of interest, so a trial- and-error approach is adopted to find the peak location. A typical normalized histogram of the protein energy is shown in figure 3.4.

3.5 Results

3.5.1 Equilibrium properties

Under normal physical and chemical conditions protein can fold from a newly synthesized amino acid chain into native structure. Depending on the temperature the protein structure fluctuates around its native conformation. The change between the folded and unfolded states corresponds to a phase transition [61]. To describe such a phase transition the heat capacity of a protein is computed through the following equation based on energy fluctuations:

\[ C_v = \frac{\partial E}{\partial T} = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_B T^2}, \]

where \( E \) is the protein energy at any temperature, \( T \) is the absolute temperature and \( k_B \) Boltzmann’s constant.

In figure 3.5 a representative energy fluctuation diagram is shown. Each dot in the diagram is an energy evaluation averaged over 1000 MC sweeps. In a typical simulation, \( 10^7 \) MC sweeps were collected. After a transient phase when the averaged energy fluctuates in larger amplitude, it steadily approaches an equilibrated value. The line representing the average energy \( E \) that goes through the data set shows this trend.
After the transient phase in the simulation we calculate the energy according to the number of H-H contacts for each succeeding Monte Carlo step. The averaged energy at any temperature is computed according to the histogram method, when enough MC steps are collected at a trial temperature. It was found that at low temperatures the simulation time is much longer than that at high temperatures due to the retardation from the small Boltzmann’s factor. Another serious problem is that at low temperatures protein tends to be trapped in local minima for long times. This makes the simulation time very long at low temperatures. In figure 3.6 we have plotted the averaged energy as a function of temperature; there the increasing trend is clearly shown. Also noticeable is the decreasing trend of the energy with the reducing of box size. At a fixed temperature the protein tends to reduce its energy to a lower value when box size is reduced. This, on the other hand, implies that more energy is needed to unfold the protein residing in a small box. In other words, the protein is stabilized.
Figure 3.6: Average energy $<E>$ as a function of temperature and box size. The abscissa is the absolute temperature in Kelvin and the ordinate is energy in $\varepsilon$. $<E>$ increase with temperature but decrease with box size. From the top to bottom, the box sizes are: $20 \times 20$, $16 \times 16$, $14 \times 14$, $10 \times 10$, $8 \times 8$, and $7 \times 7$.

We also plotted the average protein size as a function of the size of confinement. This is seen in figure 3.7, where the average radius of gyration of protein in boxes of different size is plotted versus temperature. Similar to the energy trend, the averaged protein radius of gyration $<S>$ increases with the temperature as expected, but it shrinks as the box size shrinks. This is easy to interpret: the confinement reduces the protein conformation space and biases it towards the native conformation. Only sub-domain of the entire conformation space that is around the native state will be available for the protein to sample. In principle a protein can be confined in boxes of any shape; for example it could be restricted between two close hard slabs and extends along the slab surface, which is not close to the native state. In such a case, the protein will be stabilized among a set of unfolded states.
Figure 3.7: Average radius of gyration of protein confined in different boxes plotted as a function of temperature. As in the average energy plot, from top to bottom the box sizes are: 20×20, 16×16, 14×14, 10×10, 8×8, and 7×7, respectively. The dashed line is for 14×14, which is a little higher than that in 16×16, we believe this is due to the statistical error.

The similar trend exhibited by both average energy and average size of protein implies a correlation between them. This is shown in figure 3.8, which gives not only the correlation between energy and size, but also the relative weight of visits in the energy-size phase diagram.
Figure 3.8: $S$-$E$ phase diagram obtained from a typical Monte Carlo simulation run at temperature $T = 400$ K and box size $20 \times 20$. Relative weight of points in the diagram due to the temperature and box size is revealed by the gray level.

In practice we made use of a trial-and-error method, as mentioned in section 3.3, to obtain a reliable heat capacity curve. Interestingly the melting temperature, peak temperature on the heat capacity curve, is approached in an oscillatory way, that is, from above and below the true melting temperature. Usually we need only four or five iterations to obtain a convergent value with error being within $\pm 0.5$ K. The convergent peak temperature gives the melting temperature. The confinement size-dependent heat capacity as a function of absolute temperature is plotted in figure 3.9.
Figure 3.9: Heat capacity as a function of temperature for the model protein in boxes with different sizes. From the left to right, the box sizes are: 20×20, 16×16, 14×14, 10×10, 8×8, and 7×7.

Since a 2-D square lattice model is a great simplification, we do not expect a quantitative comparison between the experimental results and the simulated ones, and the heat capacity is thus expressed in an arbitrary unit. Despite this fact, two very interesting phenomena are clearly seen in figure 3.9. The first noticeable feature is that the peak temperature, or the melting temperature, shifts towards higher value when the confinement size reduces. The difference between the maximum melting temperature and the minimum is about 250K. Box of size 7×7 is a highly constrained case. Thus the difference between the minimum and maximum melting temperature drops to $\Delta T_m = 65$ K if the smallest box is considered an outlier. This is not far from the experimental value of 30 K obtained by Eggers and Valentine [38]. In any case,
both simulations and experiments yield a large confinement-induced stabilizing effect. Another feature is concerned with the width of the heat capacity profile. The width of the heat capacity curve becomes larger, when the box size becomes smaller, indicating the boundary between the two phases, folded and extended, becomes blurred.

Figure 3.10: Melting temperature $T$ plotted as a function of inverse box size $1/a$. From left to right the points represent cases of box $20 \times 20$, $18 \times 18$, $16 \times 16$, $14 \times 14$, $12 \times 12$, $10 \times 10$, $8 \times 8$ and $7 \times 7$, respectively. Vertical axis $T$ is in Kelvin.

We have plotted the melting temperature as a function of inverse box size in figure 3.10. Besides the trend of increasing melting temperature with decreasing of box size, we see that the melting temperature shows three different behaviors as a function of inverse box size. From the $20 \times 20$ box to the $12 \times 12$ box we see a linear increase of melting temperature, from $12 \times 12$ to $8 \times 8$ the melting temperature still increases
linearly but with a smaller slope, and from \( 8 \times 8 \) to \( 7 \times 7 \) a very sharp rise of melting temperature appears. While the difference between the first two domains may be due to simulation error, the sharp rise can only be explained by noting that the smallest possible box we have is \( 7 \times 7 \), which corresponds to an extremely limited confinement within which the protein finds that the achievable conformation space is greatly compressed towards the native state. The box size cannot be varied in a continuous way, because of the discrete nature of a lattice model, which limits the possibility to investigate the sharp rise region. But this plot already shows the strong stabilizing effect on protein folding, particularly near the smallest possible confinement size, thus we expect a rapid increase of melting temperature when, the box size approaches the size of a native protein.

Our simulated results are in consistent with the experimental results obtained in Dr. Wei’s group [43-47]. Their experiments are briefly described here. To entrap enzymes in a small confinement, a sol-gel process was used. Non-surfactant templates, such as glucose and fructose, were mixed into solution with enzymes. The gel thus created will be flushed with water to dissolve the template molecules, leaving the enzyme trapped in the silica matrices. The size of the pores of silica matrix is controllable by varying the amount of template molecules used in the sol-gel process, and can be measured by applying the standard nitrogen adsorption-desorption isotherm methods. Figure 3.11 depicts a trapped enzyme in a silica cavity.
Figure 3.11: Diagram of a silica matrix pore within which an enzyme molecule is immobilized. D is the cavity size and d is the pore size, controllable by the amount of template molecules at the beginning of sol-gel process.

Two enzymes, acid phosphatase (ACP) and horseradish peroxidase (HRP) were entrapped in silica matrices using a sol-gel process. The remaining activity of the enzyme after a heating process can be assayed either with p-nitrophenyl phosphate (pNPP, for ACP) or through a dye generating reaction using hydrogen peroxide, phenol, and 4-aminoantipyrine to produce N-antipyryl-p-benoquinoneimine (for HRP). Specifically, to measure the thermal stability of entrapped ACP or HRP, the enzyme, both in solution and in confinement samples were subjected to a cycle or cycles of heating processes between 50°C and 65°C for different duration of time. The results are summarized in figure 3.12 And 3.13.

In figure 3.12 the percentage of the remaining activity of ACP clearly exhibits a decreasing trend with the increasing pore size of the silica matrices where the enzyme molecules are localized. Enzyme in free solution thus shows minimal remaining activity. It seems that heating could denature a larger population of enzymes in free buffer but only a smaller population in smaller nanopores. Confinement shields
enzyme from heat-denaturing via blocking the swelled conformations from the hard wall. This is also seen in figure 3.13, where, five cycles of reheating was applied on HRP, a slower decline of remaining activity was observed for HRP confined in smaller pores. This again supports the hypothesis that confinement stabilizes the native proteins.

![Figure 3.12](image)

**Figure 3.12**: Percentage remaining activity of the free and immobilized acid phosphastase (ACP) samples assayed at pH 6.2 and [pNPP] = 5.0mmol/L after heating at 50°C for 1 hour. The vertical axis is the percentage of the remaining activity compared to that of the solution sample at room temperature. The numbers, 1 through 8, label different samples used for the assay: sample 1 has zero wt% of glucose content in the original sample, sample 2 has 10 wt% of glucose content, 3, 4, 5, 6, 7, correspond, respectively, to 20, 30, 36, 42, and 50 wt% of glucose content, and 8 denotes protein in solution.
Figure 3.13: Thermal stability of free and immobilized horseradish peroxidase. Samples were treated at 65°C for 15 minutes per each cycle for 5 consecutive cycles. Activity of free and immobilized HRP at room temperature is 100%. Note that group 1 is for FH0 (<1.7 nm), group 2 is for FH42 (3.2 nm) and group 3 is for free solution.

Zhou and Dill in their paper [50] presented a theoretical calculation of the stabilizing effect, in terms of ΔΔG, the change of free energy between native and unfolded states of protein, due to the confinement of various geometrical shapes. Their model is based on random-flight Gaussian chain, different from the self-avoiding lattice model. Their results did show a large decrease of free energy and a conclusion that confinement shifts the folding equilibrium toward the native state was drawn. Our simulation study complements the theoretical approaches of Zhou and Dill in the sense that their Gaussian chain polymer models work for large proteins, and our lattice simulations work for short proteins. Both make qualitative predictions about confinement-induced stability of proteins. Besides equilibrium properties, we also investigate protein folding and unfolding dynamics in our simulations. This is the
The topic of next section. The random-flight Gaussian chain models, which are extensively applied in various problems, has the drawback that a polymer chain is allowed to cross itself, that is, is not self-avoiding. As far as we know, there does not exist a proper theory for self-avoiding polymers which can be used to predict the confinement effect more accurately. Efficient simulations can be used to compensate this shortage.

3.5.2 Mean first-passage times

The mean first-passage time is the average time taken for a stochastic system to reach a set of predefined conditions, starting from a common state or set of states [62]. This quantity is an important descriptor that describes how fast an inherently stochastic system evolves into new states. Typical examples of stochastic processes include random walks and diffusion phenomena. Specifically at normal conditions, the protein moves along a free energy surface, which is globally biased towards a unique native conformation with the lowest free energy. But due to the inherent ruggedness of the free energy surface, which is generally believed to shape like a funnel as a whole, the protein folding and unfolding processes can be mapped into a particle diffusion driven by an external force field [63, 64]. The mean first-passage time idea can be readily applied to calculate the folding time, which is the time for protein to fold into native conformation from any initial conformation, fluctuating in a large scale due to the large ensemble of open conformations and the ruggedness of the free energy surface. In principle, if the energy surface is explicitly given, the first-passage probability \( F(S, t) \), which is defined to be the probability for system to reach a specified state \( S \) after time \( t \), can be calculated via generating function techniques [65,
In reality, it is not practical to calculate the entire free energy surface for real proteins. In figure 3.14 a rugged funnel-like free energy surface of protein is shown with one route towards the native conformation given as an example. Computer simulations provide us with a powerful tool to evaluate the first-passage time for folding and unfolding under the circumstances where theoretical calculations become hard.

![Rugged funnel-like energy surface of protein](image)

Figure 3.14: A rugged funnel-like energy surface of protein. Protein folding maps to a diffusion along the energy surface towards the lowest point, the native conformation. One route of protein diffusion towards the native state is shown.

We define two mean first-passage times. Starting from any non-native conformation, the protein seeks a path along the potential surface towards the folded native structure. This process is defined as a first passage process for folding. On the other hand, a process during which the protein starts from its native structure and
moves until it reaches a certain ensemble of extended conformations is called a first passage process for unfolding. Correspondingly we define the Monte Carlo sweeps in these processes as first-passage time for folding (folding time for short) and unfolding (unfolding time), respectively. The magnitudes of the first-passage times usually fluctuate within some range and are distributed according to the inherent probability characteristics of the systems under study. In our simulations we did observe large fluctuations which are comparable with the magnitude of the mean first-passage times, though we did not set out to extract the distributions of these quantities. As described before, an ensemble of 599 distinct unfolded conformations is used to initiate folding trajectories and the mean first-passage time for folding is the average over such an ensemble.

In figure 3.15 the mean first-passage time for folding is plotted. The common feature of the folding time, namely, the bell-like shape, is in agreement with previous theoretical and simulation studies [59, 67, 68]. Three noticeable features are shown. First, the fast folding processes are roughly temperature-independent and box-size-independent over a range of temperatures in the middle. Second, at low temperature the folding time increases in an exponential way with the decrease of temperature and increases with the decrease of the size of confinement. Third, at high temperature, the folding time increases exponentially with the reciprocal temperature as well and increases as box size increases. An interesting point that should be emphasized here is that although the folding time increases on both extreme domains of temperature, the confinement size dependence is reversed, namely, at lower temperature protein folds more slowly in smaller boxes, while at high temperature protein folds more slowly in
larger boxes. To see this feature more clearly the natural logarithm of folding time is computed and plotted again versus the inverse temperature in figure 3.16. As a function of inverse temperature the logarithm of folding time shows linear relation meaning that the folding time itself follows a power law.

Figure 3.15: Mean first-passage time for folding directly plotted as a function of temperature for the model protein in boxes with different sizes. The curve with solid dots is for 10×10, that with solid squares for 14×14, solid diamonds for 16×16, solid triangles for 20×20, and the stars for solution protein. Solid lines are drawn to connect data points. The temperature is in unit of kK, kilo Kelvin and the time is in arbitrary unit.
Figure 3.16: The natural logarithm of mean first-passage time for folding versus the inverse temperature. T is in kilo Kelvin. V-shaped trend is seen.

The further the temperature reaches the low or high domains, the more the logarithm of folding time approaches linear relation, which is fitted to straight lines. The fitted lines, given in figure 3.17, show how the box-size-dependent reversing phenomenon becomes clearer. Not only is there an increasing trend of folding time with box size at high temperature and decreasing trend with box size at low temperature are shown, but also the box-size-dependence becomes stronger when the temperature is pushed towards both extremes.

When box size shrinks at low temperature, the free space within the box is reduced and protein will easily collide with the walls of box which hinders protein’s trial Monte Carlo moves and retards the folding process. Among all possible folding
pathways some are faster than others. The protein, starting from a relatively compact conformation, may be required to extend somehow in order to find the fast folding paths, which are possibly blocked by a smaller confinement. Thus a larger portion of longer folding paths are kept compared to fast folding ones, leading to a longer folding time. At high temperature we believe the box hindrance still certainly exists but this effect is overcome by the fact that a high temperature forces the protein to sample and resides at much larger conformation space before folding into native structure, and it is the latter effect that dominates the first.

![Figure 3.17: The fitted asymptotic lines of natural logarithm of mean first-passage time for folding. Note that the reversing phenomenon is clearly seen. Cross stands for 10×10 box case, solid triangle for 14×14, open diamond for 16×16, solid circle for 20×20, and open square for free protein. T is in kK.](image)

The mean first-passage time for unfolding of the confined protein is also recorded as a function of temperatures. Understandably, the unfolding time becomes shorter
when the temperature is higher, as seen in figure 3.18. The wall encloses a limited free space for protein to move, often blocks the trial move of protein and quenches the probability for protein to open up. As anticipated, the unfolding time increases when the free space shrinks. At the same temperature, a free protein takes on the average the shortest time to unfold into extended conformations compared to the confined protein. We implemented three cases, namely the free protein, protein confined in a $16 \times 16$ box, and protein confined in a $20 \times 20$ box for comparison. As the confinement becomes small and comparable to the size of the confined protein, we expect that the unfolding time will increase drastically. Too small a confinement will drive the probability for protein to unfold towards zero because the protein is forced to stay at native conformation. A noticeable feature of the lines shown in figure 3.18 is that they are roughly parallel to each other. We will discuss this feature in section 4.2.5 where a similar property for unfolding time of a protein in a solution of macromolecular crowders was also obtained.

In figure 3.12 and figure 3.13 we have shown that the remaining activities of immobilized acid phosphatase and horseradish peroxidase are higher and the decrease of the remaining activity in the heating cycle is slower in smaller silica matrix pores. This can be understood with the help of folding/unfolding times. Smaller confinement induces a longer unfolding time, or the time to denature, but a shorter refolding time within a proper temperature range. Statistically, as a result, the proportion of proteins that stay at the native state will be larger, leading to higher remaining activity and/or a slower drop of the same quantity.
Figure 3.18: Natural logarithm of mean first-passage time for unfolding vs. inverse temperature, fitted with linear lines. Open circles are for box 10×10, squares for 16×16, and triangles for free protein. T is in kK.
Chapter 4: Effects of macromolecular crowding on protein

4.1 Brief review of research on macromolecular crowding

In real living cells proteins do not exist in a very dilute solution but in a very crowded cytoplasm environment with many types of other biomacromolecules, which shall be called crowding agents [69-76]. Jointly these macromolecules occupy a large percentage of cell volume, ranging typically from 5% to 40%. Biochemical reactions and biophysical processes in the cells, for instances protein folding, protein-protein interaction, and transportation of biomacromolecules in cell matrix, are expected to differ from their counterparts in dilute solutions in rather complex ways, constituting the subject of many recent theoretical and experimental studies [69-76]. Though minor discrepancies exist among different groups, the consensus belief in this field converges to conclusions that crowding agents can speed up chemical reactions limited by specific macromolecular interactions such as protein folding and protein aggregation, and that crowding can slow down the chemical reactions limited by diffusion at high crowder concentration, as well,

In the folding process, segments of protein interact with themselves and with other molecules such as water. Some interactions are long-ranged (electrostatic) and some are short-ranged (Van de Walls). These forces can be expressed explicitly using formulae with adjustable parameters, making the evaluation of such interactions straightforward. Another type of interaction is not due to the real existence of physical fields but comes from principle of maximum entropy. Taking ideal gas as an example, we see that although no interaction is assumed between any gas molecules,
to confine all molecules in limited region we have to apply pressure, an entropic force; otherwise gas molecules will fly off in all directions to maximize the entropy. Another good example is perhaps rubber. Stretching a rubber will reduce the conformation space of polymers of rubber and the rubber tends to contract. This retractive force is also entropic in nature. In highly crowded environment the macromolecular crowding agents have a tendency to occupy more free space in order to maximize the system’s entropy and help drive the system into a stabilized state. We then believe that the entropic force plays very important roles in protein folding and protein stability. Several research groups have tried to understand this effect from different perspectives. Because a complete self consistent theory starting from first principles seems to be very hard to obtain, simple theories [71-77] are applied trying to obtain at least qualitative results that conform to real experiments. Another track of research is to use computer simulations [1, 2, 76] to extract information on the subtle effects.

Our research on the macromolecular crowding is developed along two directions. The first is based on computer simulations similar to the study of confinement effects; and the second direction is a theoretical approach in which we apply excluded volume concept and scaled particle theory to study the crowding effects.
4.2 Brute force 2-Dimensional HP lattice simulation

Besides experimental and theoretical studies on macromolecular crowding effects, there exists a whole range of computer simulation approaches [78]. The finest are all-atom molecular dynamics simulations, either with explicit water molecules or without. In each run the motion of each atom is integrated out according to specifically defined force field and a trajectory of the system in phase space is given. This method does give dynamics and time-averaged properties, but may not provide rich information of ensemble-averaged properties. Most importantly, it demands very large amount of CPU time because a large number of energy terms is monitored. Different levels of coarse grained approaches have been established to reduce the computation time and gain more information, while still keep the most significant factors. Yuan and Li [79] have applied a bead model to simulate protein under external pulling force. In such a model each amino acid group in the chain is represented by a bead that can interact with other beads based via the designed force field. An even simpler coarse grain model is a lattice model where the continuity of the system under study is eliminated and replaced by discretized lattice. The amino acid groups are represented by lattice sites and the motion of protein is confined on the lattice. The lattice model, though simplified, is never trivial when the number of lattice sites that represent the system under study is large. The lattice model requires much less computation time compared to MD and bead model, yet it gives more ergodic sampling in the system phase space. We turned to implementing HP lattice simulation which is considerably less expensive but can provide us with rich valuable information specifically needed in crowding effects study.
4.2.1 Simplified construct of the protein/crowder system

The complexity of the cell environment where proteins interact with other types of macromolecules is dramatically simplified to a collection of 2-D square crowders and a single 2-D HP lattice protein. This simplification neglects most protein-crowder interaction and replace it by hard square interaction similar to ideal gas model. In figure 4.1 we have shown the comparison between the complex environment in a real cell and the dramatically simplified 2-D counterpart. The macromolecular crowders are hard square molecules of certain size (under control) and the protein is the 16mer we adopted in the confinement effect study. Both types of molecules are confined in a virtual box of size 201 by 201, across whose walls the periodic condition is imposed.

![Figure 4.1: Simplification from crowded environment in cells to a simulation system. Within the right box, the black solid squares correspond to a collection of crowder agents and the 2-D HP lattice structure corresponds to the protein.](image)

The protein and crowder agents can move freely in the virtual box according to Brownian motion [80] described by equation 4.1, where the position of the molecule a time interval later is the current position plus a random displacement observing
Gaussian distribution. \( \xi_i = 6 \pi a_i \eta \) is the frictional coefficient for the \( i^{th} \) molecule of radius \( a_i \) where \( \eta \) is viscosity, and \( z_i \) is the random unit vector.

\[
 r_i(t + \Delta t) = r_i(t) + \sqrt{2k_B T \Delta t / \xi_i} z_i 
\] (4.1)

Mathematically the Brownian motion is continuous. For molecules limited to a lattice we truncate the random displacement around its nearest integer to get a pseudo-Brownian motion which is confined in two directions, horizontal and vertical. A typical example of Brownian motion is described in figure 4.2.

![Brownian motion](image)

Figure 4.2: A typical Brownian motion of a particle starting from point (0, 0).

We assume there is no specific interaction between crowder and crowder, and between crowder and protein. The interaction is expressed in the following profile:

\[
 E_{i,j} = \begin{cases} 
 0, & \text{no overlapping between molecule } i \text{ and } j \\
 \infty, & \text{otherwise}
 \end{cases}
\]
The control parameters in this picture are temperature, crowder size and crowder concentration. They necessitate the application of a series of efficient algorithms.

4.2.2 The multihistogram method

We have briefly reviewed the idea of histogram method, using which much more information can be extracted in single Monte Carlo simulation. This is particularly beneficial when critical point region is the study objective. Beyond this region we may be interested in a wide range of parameters domains to obtain global understanding. Thanks to Ferrenberg and Swendsen again, a new optimizing multihistogram method was devised to reach this goal [81].

The most important essence of Metropolis Monte Carlo simulations is to approximate the energy structure of the system under study as well as possible. If the energy structure of the system is discrete, people need the exact density of states in the partition function calculation. While it may be possible to enumerate all lattice structures of a small system, in most cases the system is too large to allow the complete enumeration. The number of distinct configurations of our crowded model system, which is composed of a 16mer protein and hundreds of crowders, will be astronomical.

Suppose a Monte Carlo simulation is run at a fixed temperature $T_n$ for a length of $n_n$ sweeps that are mutually independent, a single histogram that encloses the simulated energy distribution will be collected, with the energy level $E$ being visited $N_n(E)$ times. $N_n(E)$ is the single histogram at temperature $T_n$. The true probability $p_n(E)$ for the system maintained at temperature $T_n$ to be at energy $E$ is:
\[ p_n(E) = \frac{W(E) e^{-E/k_B T_n}}{\sum_E W(E) e^{-E/k_B T_n}} = \frac{W(E) e^{-E/k_B T_n}}{Z(T_n)} \] (4.2)

where \( W(E) \) is the true density of states, \( Z(T_n) = \sum_E W(E) e^{-E/k_B T_n} \) is the partition function at temperature \( T_n \). The probability \( p_n(E) \) can also be approximated after the simulation by: \( p_n(E) = N_n(E)/n_n \). It is easy to get the density of states approximately as follow:

\[ W(E) = N_n(E)/n_n e^{E/k_B T_n} Z(T_n) \] (4.3)

Define a free energy profile obtained from this simulation as:

\[ f_n(T) = -k_B T \ln Z(T_n) \] (4.4)

then the density of states can be further expressed as:

\[ W(E) = N_n(E)/n_n e^{E/k_B T_n - f_n/k_B T} \] (4.5)

The basic idea of multihistogram is to combine multiple single histogram information and to get a good approximation of the true density of states after a proper optimization procedure. Geometrically speaking it is to combine multiple histogram curves into a smooth single histogram that mimics the true energy distribution. Let the weighting factor from the \( n^{th} \) simulated histogram among total of \( R \) simulations be \( q_n(E) \), a linear combination of all histograms is then constructed as:

\[ W(E) = \sum_{n=1}^{R} q_n(E) N_n(E)/n_n e^{E/k_B T_n - f_n/k_B T} \] (4.6)

with the normalization condition \( \sum_{n=1}^{R} q_n(E) = 1 \). Minimizing the following defined error quantity [82]:

\[ \delta = \sum_{m=1}^{R} \left( \sum_{n=1}^{R} q_n(E) N_n(E)/n_n e^{E/k_B T_n - f_n/k_B T} - N_m(E)/n_m e^{E/k_B T_m - f_n/k_B T} \right)^2 \]
the weighting factors can be calculated as follow:

\[
q_n(E) = n_n e^{f_n/k_BT - E/k_BT_n} \sum_{m=1}^{R} n_m e^{f_m/k_BT - E/k_BT_m} \quad (4.7)
\]

It is convenient to define a new quantity named multihistogram

\[
P(E, T) = W(E) e^{-E/k_BT} \quad (4.8)
\]

with the normalized condition being:

\[
e^{-f_n/k_BT_n} = \sum_E P(E, T_n) \quad (4.9)
\]

\(P(E, T)\) will be found in a self-consistent way by iterating equation (4.8) and (4.9), provided with the \(f_n\) initialization by a proper estimate. The average of any energy-related quantity is evaluated as a function of \(T\) using

\[
\langle O(E) \rangle(T) = \sum_E O(E) P(E, T) / Z(T) \quad (4.10)
\]

where the optimized partition function at any temperature is \(Z(T) = \sum_E P(E, T)\).

### 4.2.3 The replica exchange method

Due to the ruggedness of protein free energy landscape at low temperature, the protein tends to be trapped in one of many local minimum potential wells for quite a long time with a tiny probability to jump out, negatively affecting the efficiency in sampling the entire conformational space, if one simply applies straightforwardly the Metropolis algorithm. To avoid this low-temperature trap problem many algorithms have been devised and applied with demonstrated power in some aspects and some
drawbacks in other aspects. To name a few, multicanonical ensemble algorithm, simulated tempering [83, 84], smart walking [85], and more. The replica exchange method (REM) [86-89] surpasses other algorithms in efficient conformation space sampling under different temperatures.

The fundamental idea of replica exchange method is to maintain $M$ copies of the same system under study at different temperatures. A copy at one temperature, trapped in a local minimum or not, can be swapped with another copy with a prior defined probability that leads the respective copies to relax into their individual Boltzmann’s distribution of energy at the corresponding temperatures. This method leaves the users the freedom of choosing the temperature profile under which replicas are running in order to achieve the optimization of the problem at hand.

![Figure 4.3: Diagram to show the replica exchange method. Each protein conformation represents a simulation copy maintained at a certain fixed temperature. Two copies (may not necessarily be adjacent ones) will be swapped according to a well defined transition probability to satisfy the detailed balance. As an example a pair of nearby copies (linked by dashed arrows) is shown to be swapped.](image)

Other than swapping, the $M$ replicas are mutually independent, which are considered to be part of a generalized ensemble. The ensemble state of these $M$
replicas is then uniquely determined by all individual canonical states. The
distribution weight of the ensemble state obeys the product of the Boltzmann factor of
all individual state as:

\[ W(X) = e^{\sum_{m} \beta_{m} E(q')} \]  

(4.11)

where the subscript \( m \) stands for the temperature \( \beta_{m} \) which is fixed, and \( q' \) refers to the
\( i^{th} \) replica which is a permutation function of \( m \). We are interested in the transition
probabilities between two generalized ensemble states. The detailed balance condition
requires that:

\[ W(X)w(X \rightarrow X') = W(X')w(X' \rightarrow X) \]  

(4.12)

where \( w \)'s are transition probabilities between state \( X \) and \( X' \). Let’s suppose the two
replicas \([i]\) and \([j]\) originally simulated at temperature \( \beta_{a} \) and \( \beta_{b} \) will be swapped, the
relative transition probability in real simulation is then given by:

\[
w(X \rightarrow X') \cdot w(X' \rightarrow X) = e^{-\beta_{a}[E(q'^{i}) - E(q^{i})] - \beta_{b}[E(q'^{j}) - E(q^{j})]} \\
= e^{-(\beta_{b} - \beta_{a})(E(q'^{j}) - E(q^{j}))} \\
= e^{-\Delta}
\]

(4.13)

where \( \Delta \equiv (\beta_{b} - \beta_{a})(E(q'^{j}) - E(q^{j})) \).

The routine Metropolis-like criteria can then be applied:

\[
\begin{cases} 
  e^{-\Delta} > \chi & \text{state } X' \text{ is accepted,} \\
  \text{otherwise} & \text{remain at state } X
\end{cases}
\]

where \( \chi \) is random number uniformly drawn between 0 and 1.

The most important advantage of REM is that when the system is trapped in a local
minimum, it has good chance to jump out of the potential well without spending too
much time sampling only the local minimum conformation space. The replica
exchange method can be easily modified to be applicable in the molecular dynamics [86-89] or Brownian dynamics which record information of both conformation and momentum, if the momentum is geared by a factor proportional to the square root of temperature, a thermal condition should be satisfied, namely, the following condition should be observed when swap replicas:

\[
\begin{align*}
    p^{ij} &= \sqrt{T_b/T_a} p^{[i]} \\
    p^{[ij]} &= \sqrt{T_a/T_b} p^{[j]}
\end{align*}
\]

### 4.2.4 Simulation procedure

The replica exchange method is suitable only for the study of the system’s properties at equilibrium, from which the statistical physical quantities, such as energy and heat capacity, can be obtained efficiently. The kinetics information of the system cannot be obtained using this method. For the calculation of the folding and unfolding rates, we use the standard Monte Carlo methods.

We maintain \( m = 20 \) replicas of our protein-crowders system at 20 distinct temperatures distributed exponentially with respect to \( T \). No particular reason is associated with such a temperature profile. A linear dependence of temperature on \( m \) could also function well.
Figure 4.4: The series of temperatures at which the system replicas run. In principle any temperature profile can be taken, here we adopt the exponentially distributed temperature profile. 20 replicas are shown with the minimum temperature being 190K and the maximum temperature being 1500K.

The adjustable parameters in our system are crowder size $a$ and crowder number $N$ within the simulation virtual box of size $l \times l = 201 \times 201$. The volume fraction (in fact area fraction) of the crowders is simply: $\varphi = V_c / V = Na^2 / l^2$. We change the crowder size from $1 \times 1$ to $5 \times 5$. Fixing the volume fraction $\varphi$, which can be up to 40%, one may still change the crowder number and the crowder size, which may affect the protein properties in a systematic way. To investigate fully the contributions from varying all parameters ($\varphi$ is dependent) will be almost impossible currently due to the limited computer facility. We therefore study only several cases of different crowder sizes and crowder numbers. Below we give the main procedure to perform the simulation at given size and number of crowders.
1. Initialize the 20 replicas with randomly selected protein conformations and assign an exponentially distributed temperature to each of them.

2. Perform the simulations in parallel for certain number of MC steps in order for all replicas to equilibrate and then collect histogram and size information for long MC sweeps.

3. Randomly select two replicas nearby to each other in temperature, check the transition probability according to the Metropolis criteria; if it is satisfied, the swapping takes place. Repeat Step 2 and 3 until enough information is collected.

4.2.5 Results

As we have discussed, the confinement has a stabilizing effect on protein. Therefore with the expectation to obtained similar stabilizing trend for the crowding study, we again calculate the heat capacity of the protein in thermal equilibrium and the result is shown in figure 4.5. Since we are more interested in the melting temperature shift, the heat capacity is plotted in an arbitrary unit versus the absolute temperature in kilo Kelvin (kK). Generally, when the number of crowders (with fixed size) or equivalently the volume fraction increases, the melting temperature shifts toward the higher end. This trend is similar to what we obtained in our confinement study, meaning the protein tends to assume more compact structures and is stabilized against the open conformations. The trend, on the other hand, seems weaker than that found in confinement, though we have assigned an energy unit \( \varepsilon = 1.40 \times 10^{-20} \) J, only with tiny difference from that has been assigned to the confined protein model, to the topological H-H contact. In rigid confinement the protein is forced to react within
small region, while in solution of macromolecular crowders the protein finds more free space when the crowder concentration is dilute. Even when the crowder volume fraction is moderately high, the protein can still find free space to fold or unfold due to the finite probability that a larger free space around the protein can be created temporarily. We predict that only when the crowder concentration is very high that insertion of more crowders with similar sizes becomes very difficult, can the protein be much more stabilized, as seen in the confinement. Currently this result is from a 2-D lattice simulation for a short HP sequence, therefore one should bear in mind that only qualitative, but not quantitative results can be expected. Another difference between the heat capacity of protein in confinement and that in crowded solution is that the broadening effect does not show up. The conformation space is dramatically reduced within a small box but is less squeezed in the crowder situation. If we would like to map the two effects onto one another, a fundamental issue that should be considered is that the confinement is hard in the former, but is kind of “soft” in the latter, because the random moving of crowders always leaves the protein nonzero chance to extend itself into the interstitial regions among crowders. In a sense the conformation space a protein can sample in crowders is still larger than the space in confinement. We believe this is part of reasons behind the nonexistence of broadening of the heat capacity curve in the crowding studies that blurs the folded and unfolded boundary.
Figure 4.5: Heat capacity $C_V$ versus temperature $T$. The $C_V$ is in arbitrary unit and the temperature is in kilo-Kelvin. Counting from left to right except the leftmost curve, the crowder number is 100, 200, 300, 500, and 700, and the leftmost is for protein in dilute solution. Here the crowder size is $4 \times 4$.

The average size of protein is also altered with the concentration of crowders. In simulations we record the end-to-end distance, radius of gyration and the largest distance between any two monomers. To see how the crowder can limit the expansion of protein, we plotted the largest distance as a function of crowder size. Not surprisingly this dimension does decrease when the crowder grows larger while keeping the number density a constant, this result agrees with results obtained by other groups [90], and the linear trend is shown in figure 4.6. We can see that the change is not as large as that observed in hard box, in accordance with the earlier discussion on the nonexistence of heat capacity broadening.
Figure 4.6: Averaged end-to-end distance $S$ of a protein plotted as a function of the size of crowders $a$ at $T = 800$ K, when crowders number is $N = 500$.

We have performed the calculations of the mean first-passage times for both folding and unfolding processes. A protein conformation is randomly selected from an ensemble of 1379 distinct conformations obtained from a long MC simulation run at high temperature without crowders, the protein folds into its native structure in solution of crowders at various temperature, crowder size and crowder number conditions. We can not apply the REM in simulating the mean first-passage time because the REM is designed for equilibrium calculations. Again a complete survey of all parameters, temperature, crowder size, and crowder number, makes the simulations computationally demanding, so only several selected cases were studied.

Since the random starting protein conformation in a folding process is drawn from an ensemble that covers a wide conformation space, and since the effect from crowders on different conformation is not uniform, it is not surprising to observe
large fluctuations (90% of averages value) in folding times. Future work will involve subdividing the ensemble into sub-ensembles and collecting the folding times distribution. In figure 4.7 the folding time at $T = 500$ K is shown to be increasing with the number of crowders (or volume fraction, virtual box size is $300 \times 300$). The irregularity but generally increasing trend of this plot demands the need to a more thorough simulation study.

Figure 4.7: Mean first-passage time for folding at $T = 500$K increases with the number of crowders. The crowder size is set to be $4 \times 4$.

On the other hand the unfolding time is much easier to simulate. One simply starts from a native conformation and terminate the MC moves when protein reaches the extended structures such as the radius of gyration $R_g = 1.9$ and energy $E = \varepsilon$. 10,000 repeated simulations are performed to average the unfolding time. Fixing the crowder
size to be $4 \times 4$, we have finished unfolding time calculations for various temperature and crowder size.

![Figure 4.8: The mean first-passage time for unfolding as a function of inverse temperature and crowder number. The curves are: solid line for $N = 1$, dot line for $N = 400$, dashed line for $N = 600$, long dashed line for $N = 800$, and dot-dashed line for $N = 1000$. Temperature $T$ is in kK.](image)

In figure 4.8 we observe that the logarithm of unfolding time follows Arrhenius trend and increases with the inverse temperature, though not in a linear way as we have already seen in confinement simulations. The unfolding time increases noticeably with the number of crowders as well. Interestingly the curves seem parallel to one another, namely the percentage increase of unfolding time for smaller crowder size to larger crowder size does not obviously depends on crowder number. Recall that we have obtained a similar parallel characteristic in showing the logarithm of unfolding time (figure 3.18) for the confined protein as a function of the reciprocal
temperature. There the slope of the lines does not apparently depend on the confinement (box) size. Therefore we propose that even when the number of crowders increases (or the box size decreases), the activation energy barrier for the unfolding process seems to remain approximately constant, if not absolutely so. Coupled with the fact that the unfolding time elongates with the number of crowders, it is the probability that a random trial move of protein in simulation to be accepted that is reduced due to the existence of crowders. In the transition state theory of a two-state protein folding, the unfolding rate of protein is related to the free energy barrier of unfolding according to the following expression [91]:

\[ k_u = \nu e^{-\Delta G_u / k_B T} \]  \hspace{1cm} (4.14)

where \( \Delta G_u \) is the free energy barrier across which protein can unfold, and \( \nu \) is the pre-exponential factor describing the trial frequency to jump out of the native potential well. We can define the unfolding rate to be inversely proportional to the unfolding time. Taking the logarithm of the above equation we have:

\[ \ln k_u(N,T) = \ln \nu(N,T) - \Delta G_u(N) / k_B T \]  \hspace{1cm} (4.15)

where we have explicitly written the rate constant, pre-exponential factor functions of both crowder number and temperature. The reason we expressed the pre-exponential factor \( \nu \) as a function depending on temperature is that the curve for any fixed number of crowders shows deviation from linearity, which is a characteristic of constant \( \nu \). Take the partial derivative with relative to \( \beta = 1/(k_B T) \) we got:
\[
\frac{\partial \ln(k_u)}{\partial \beta} = \frac{\partial \ln(\nu)}{\partial \beta} - k_b \beta \Delta G_u
\]  

(4.16)

the parallelism of the curves mean that the above partial derivative does not depend on number of crowders N, though it does depend on temperature as seen from the curvature of the curves. At a fixed temperature this derivative gives a constant value through N. Therefore in general, both the pre-exponential factor and the unfolding energy barrier change with the number of crowders in a cooperative manner. But, if the trend of the logarithm of unfolding time as a function of reverse temperature is pretty linear, in other words, if all slopes take the same value at any temperature, the temperature-dependent pre-exponential factor simply vanishes, and it is the activation energy for unfolding that change with the concentration of crowders. Note that the change of unfolding energy barrier does not necessarily mean the level of folded potential well is changed. It means the separation between the transition state and the folded state is changed. It is possible to extract the dependence of unfolding energy barrier on crowder concentration only if we can obtain explicit information of \( \nu \) as a function of temperature and crowder number. This is left for future 3-D lattice study employing more physical folding/unfolding time.
We define the equilibrium constant $K$ between folded and unfolded states of protein as followed:

$$K = \frac{[F]}{[U]} = \frac{k_f}{k_u} = \frac{\tau_u}{\tau_f}$$  \hspace{1cm} (4.17)

where the $k_f$ and $k_u$ are folding rate and unfolding rate, and $\tau_f, \tau_u$ are folding time and unfolding time, respectively. With the increase of crowder number, both folding and unfolding times increase, providing a hint that the equilibrium constant may vary. The equilibrium constant at $T = 500$ K is plotted as a function of number of crowders $N$ is given in figure 4.9. It shows that the equilibrium constant is shifted down as the volume fraction increase. Relatively speaking, the folding time increases slower than the unfolding time does, though both times become longer. In this sense we conclude that as the number of crowders increases, the folded state of protein is preferred over the unfolded one, and the protein is stabilized.
Chapter 5: Crowding effects studied using polymer physics

5.1 Review of the scaled particle theory

To the zeroth order approximation the protein (particularly the globular protein) and its surrounding macromolecular crowders can be both modeled as hard spherical molecules. A theory, dated back to late 1950’s, called the scaled particle theory [92-96] is particularly successful in calculating the equations of state for a solution composed of mixtures of hard spherical molecules of different sizes. The interaction between molecules is assumed to be that between the hard spheres given as

\[ V_{ij}(r) = \begin{cases} \infty, & r < R_{ij} = (R_i + R_j), \\ 0, & r > R_{ij} \end{cases} \]

where \( r \) is the center to center separation between two hard spherical molecules of size \( R_i \) and \( R_j \), respectively.

Given a solution of \( n \) types of hard spherical particles of different sizes maintained at different concentrations, the interesting question to ask is the required mechanical work \( w \), the equivalent free energy change of the system, \( \Delta G \), to insert one spherical particle of a finite size into the solution (see figure 5.1). The fundamental idea of the scaled particle theory is to generalize the form of mechanical work calculated for much smaller hard spheres, which is rather straightforward to compute. Imagine a hard sphere of infinitesimal radius \( r \rightarrow 0 \) is to be inserted, the probability that it collides with any solution molecules is simply the ratio between the volume of all molecules in the solution and the total solution volume. Based on this fact, the work, or the free energy change of the system can be easily calculated by considering the
difference between the two partition functions of the system, one before insertion, \( Q_1 \), and the other after insertion, \( Q_2 \), through the following equation

\[
\beta W = \beta \Delta G = -\ln(Q_2 / Q_1)
\]

(5.2)

where \( Q_1 = \int_{C_1} e^{-\beta H(C_1)} dC_1 \) in which \( C_1 \) is the configuration of the solution before inserting the molecule. To evaluate the second partition function we can decompose the new system’s configuration into two parts: one is \( C_1 \) plus a molecule falling in the void region and the other is \( C_1 \) plus a molecule falling into any one of the solvent particles, which leads to an infinite interaction potential. The contribution of the second configuration to the partition function is simply zero due to the Boltzmann’s factor. Therefore \( Q_2 \) is given by

\[
Q_2 = \int_{C_2} e^{-\beta H(C_2)} dC_2 = Q_1 p_0 = Q_1 (1 - \varphi)
\]

(5.3)

where \( p_0 = 1 - \varphi = 1 - \sum_i 4\pi / 3N_i (R_i + r)^3 / V \) is the probability that the center of inserted infinitesimal molecule doesn’t fall into the solution molecules, \( \varphi \) is the volume fraction of solvent molecules, \( R_i \) is the radius of solvent species \( i \), and \( r \) is the radius of particle to be inserted. We finally obtain

\[
\beta W = \beta \Delta G = -\ln(Q_2 / Q_1) = -\ln(1 - \varphi)
\]

(5.4)

To generalize the work inserting an infinitesimal particle to the work inserting a molecule with finite radius \( R \), Lebowitz et al [92-96] approximated the latter to be given by a cubic polynomial in \( R \) with the coefficients determined by \( W \) near \( R = 0 \), noting that when the molecules increase, they increases in radius, in surface area and in volume.
\[ \beta W(R) = \beta[W(0) + W'(0)R + \frac{1}{2}W''(0)R^2 + \frac{4\pi}{3} pR^3] \quad (5.5) \]

Suppose the fluid is composed of \( m \) species of hard spherical particles and define the following quantities:

\[ \xi_i = \frac{\pi}{6} \sum_{i=1}^{m} \rho_i (2R_i)' \]

where \( \rho_i \) is the number density of species \( i \). The expression for the work of insertion then becomes

\[ \beta W(R) = -\ln(1 - \xi_3) + \frac{6\xi_2 R}{1 - \xi_3} + \frac{12\xi_1}{1 - \xi_3} + \frac{18\xi_2^2}{(1 - \xi_3)^2} R^2 + \frac{4\pi}{3} \beta pR^3 \quad (5.6) \]

by taking derivatives of equation 4.21 at \( R = 0 \). Furthermore, if the inserted molecule has a radius equal to one of the species in the fluid, say \( i \)th species, the chemical potential of this species will be

\[ \beta \mu_i = \ln(\rho_i h^3 \beta^{3/2} / (2\lambda m_i)^{3/2}) + \beta W(R_i) \quad (5.7) \]

where \( \lambda = h\beta^{1/2} / (2\pi m)^{1/2} \) is the thermal de Broglie wavelength. The first term in the above expression comes from the fact that after insertion, the molecules can also move around thermally. Apply this expression and the approximation of the insertion work to the Gibbs-Duhem relation:

\[ \frac{\partial p}{\partial \rho_j} = \sum_{i=1}^{m} \rho_i \frac{\partial \mu_i}{\partial \rho_j} \quad (5.8) \]

Lebowitz et al were able to derive an expression of the pressure of a hard-sphere mixture by integrating the Gibbs-Duhem relation:

\[ \pi \beta p = \frac{6\xi_0}{1 - \xi_3} + \frac{18\xi_1 \xi_2}{(1 - \xi_3)^2} + \frac{18\xi_2^3}{(1 - \xi_3)^3} \quad (5.9) \]
It is this important result, or more accurately, the version modified for a square lattice model that will be applied in evaluating the crowder effects in the following sections.

It is not hard to obtain the pressure of a fluid of unique hard spherical molecules

$$\beta \rho / \rho = \frac{1 + \varphi + \varphi^2}{(1 - \varphi)^3}$$  \hspace{1cm} (5.10)

where $\varphi$ is the volume fraction of hard crowders. It is worthwhile to mention a similar result derived by Carnahan and Starling [97]

$$\beta \rho / \rho = \frac{1 + \varphi + \varphi^2 - \varphi^3}{(1 - \varphi)^3}$$  \hspace{1cm} (5.11)

If we confine the system to be 2-dimensional, the work of insertion will be truncated up to the 2nd order by noting that no 3-dimensional volume effect exists. Apply the same technique and note that for a hard square molecule with side length, $a$, the pressure of a solution composed of such molecules is
where $\varphi$ again is the volume fraction of square crowders. To see the deviation from van’t Hoff’s law of osmotic pressure $\beta p / \rho = 1$ we plotted the pressure of fluid of hard spherical molecules and 2-D square molecules, respectively, in figure 5.2. The deviations of both hard-sphere and hard-square pressure from van’t Hoff’s law are remarkable in wide range of volume fraction. Only when volume fraction is very close to 0 are all three types of pressure comparable.

Figure 5.2: Pressure of a fluid of hard molecules plotted against the volume fraction. The straight line is the van’t Hoff’s law, the top curve represents the deviation of hard-sphere pressure from van’t Hoff’s law and the middle curve is that of the hard square pressure. The dashed curve comes from Carnahan and Starling, which is very close to the result of the scaled particle theory.
5.2 The concepts of the excluded volume and the depletion force

Yang and coworkers observed, in their protein pulling experiment done in dextran solution of varying concentration using AFM, that an extra concentration-dependent force of the order of 10 pN is needed to extend an ubiquitin from the native state to an unfolded state. We believe this extra force is partially, if not largely, due to the system’s entropic effect. To verify this point, a simple protein model is designed and described below. We model the protein in native conformation as a hard sphere with a radius of gyration, \( R_b \). The protein is considered deformable in the sense that it can change the native spherical shape into an ellipsoidal shape of the same volume. The protein is immersed in a solution of dextran molecules whose radius \( R_c \) is of the same order of magnitude as the protein.

It was firstly found by Asakura and Oosawa (AO) that there exists an effective attractive force between two hard spherical molecules (radius \( R_a \)) immersed in a solution of a second type of hard spherical molecules (radius \( R_b \)) [98]. Assuming the only interaction between these two types of molecules is a hard sphere potential,

\[
E_{a,b}(r) = \begin{cases} 
0 & r > R_a + R_b, \\
\infty & \text{otherwise}
\end{cases}
\]

where \( r \) is the distance between centers of hard molecule 1 and 2. The interaction does not depend on the motions of the molecules.
The partition function (configurational partition function in particular) of the above mentioned system is easily computed assuming this simple hard sphere interaction. It is expressed as the available volume for the crowders. In the following equation, $V_{\text{ex}}$ represents the excluded volume, the region where the crowder centers cannot penetrate into, which is called the depletion zone as well.

$$Q = \int e^{-w/k_B T} dx = V - V_{\text{ex}}$$  \hspace{1cm} (5.13)

The effective attractive force between the two immersed hard particles, under the assumption that the solution is of low concentration, can be calculated according to the standard statistical mechanics formula [98, 99] as follows:

$$F = k_B T N \frac{\partial \ln(Q)}{\partial a}$$
$$= k_B T N \frac{\partial Q}{\partial a} / Q$$
$$\approx - k_B T N \frac{\partial V_{\text{ex}}}{V} \frac{1}{\partial a}$$
$$= - p_0 \frac{\partial V_{\text{ex}}}{\partial a}$$
$$= - p_0 S$$  \hspace{1cm} (5.14)
where $a$ is the center-center distance between the two molecules, $p_0$ is the osmotic pressure due to crowders, having the form of van’t Hoff’s law of osmotic pressure at lower concentration, $S$ is the cross sectional area of the depletion zone shown in figure 5.3, and $Q$ is approximated by the total volume $V$ by noting that $V >> V_{ex}$.

Equation 5.22 is derived under the assumption that the solution is dilute. When solution concentration becomes not dilute Asakura and Oosawa suggested that the pressure term, which is in van’t Hoff’s form for diluted solution, may be generalized to represent the real solution pressure. The AO model and subsequent development in depletion theories have been extensively employed in many physicochemical aspects. An example is the discussion of polymeric stabilization of colloidal dispersions [100]. On the other hand it seems that the stabilization of biopolymer by macromolecules is rarely seen.

### 5.3 Crowding effects based on a simple protein model

From a mechanical perspective we can apply the concept of the depletion force. Suppose we insert two hard particles with a same volume, one being of spherical shape and the other being of an ellipsoidal shape deformed from the first, into two identical solutions whose osmotic pressure is known as $P$. We are interested in the difference between two mechanical works, $W_1$ and $W_2$, to insert the two molecules into the system and force them to thermally equilibrate with the ambient. Each work is then nothing but the Gibbs’ free energy increase of that system, given by:
\[ \Delta \Delta G = \Delta G_2 - \Delta G_1 = W_2 - W_1 \]  

(5.15)

where \( \Delta G_1 = G_{1,f} - G_{1,i} \) and \( \Delta G_2 = G_{2,f} - G_{2,i} \) are the free energy changes of the respective system before and after the insertion of the respective molecule.

We apply the scaled particle theory to predict \( W_1 \) and a simple technique to evaluate \( W_2 \). Suppose initially we also insert an identical spherical particle into the solution and somehow this spherical molecule (with initial radius \( R_0 \)) deforms into an ellipsoid with a long-axis length of \( 2b \). During this process the immersed particle will gain energy to overcome the pressure \( P \) from the surrounding solution particles. So relative to \( \Delta G_1 \), we get the extra free energy as:

\[ \Delta \Delta G = \int_{R_0}^{b} P dV_{ex}(b) \]  

(5.16)

Here we use the exclude volume \( V_{ex} \), a function of ellipsoid long axis length \( 2b \), because whenever the particle is deformed, the excluded volume changes, but not the ellipsoid volume itself (in fact it is zero based on our assumption). Take the derivative of \( \Delta \Delta G \) with relative to the long axis \( 2b \), we obtain the attractive depletion force, or equivalently the external pulling force acting at the two ends of the long axis, to maintain a deformed structure,

\[ dF = -\frac{1}{2} \frac{\partial \Delta G}{\partial b} = -\frac{1}{2} P \frac{\partial V_{ex}}{\partial b} \]  

(5.17)

We can now apply both the excluded volume idea and osmotic pressure result computed from the scaled particle theory to explain the experimental results of Yang’s laboratory. The excluded volume around a native protein is calculated straightforwardly. When protein deforms into an ellipsoid the depletion zone is in fact not just an envelope ellipsoid with larger axial lengths but a shape which is more
bulged out close to both ends of the protein. Fortunately this bulge does not differ much from the excluded volume calculated based on the ellipsoid structure, provided the protein deformation is not large, though it is obvious that this bulge becomes more important when protein becomes a very elongated structure, for which fortunately we do not care in this simple model, because the probability for a protein to be so extended is rather small.

![Figure 5.4: The protein is modeled as a solid sphere that can deform into an ellipsoid with the same volume. Regions enclosed in the dashed lines are the excluded volume. The more the protein deviates from spherical shape the more the true surface that encloses the excluded volume deviates from the dashed ellipsoidal surface. Obvious difference between the ellipsoidal surface and true surface is depicted in the bottom shape.](image)

Let the long radius of ellipsoid be $b = R_0 + \delta$, and since we have to keep the protein volume constant, the short transverse axis have radius equal to $a = R_0^3/(R_0 + \delta)^2$. The excluded volume around any such ellipsoid is approximated to be:
Given $R_c$, the radius of the crowder, the osmotic pressure $P$ is calculated to be:

$$P = \rho k_B T \frac{1 + \phi + \phi^2}{(1 - \phi)^3}$$  \hspace{1cm} (5.19)

where $\rho$ is the crowder number density and $\phi$ is the crowder volume fraction given by:

$$\phi = \frac{4}{3} \pi R_c^3 \rho$$  \hspace{1cm} (5.20)

As a result we obtain both extra free energy $\Delta \Delta G$ that stabilizes the folded protein and the extra pulling force $dF$ to maintain protein in the unfolded state, as given below:

$$\Delta \Delta G(b) = \frac{4\pi P}{3} \left[ (\sqrt{R_0^3/b + R_c})^3 (b + R_c) - (R_c + R_0)^3 \right]$$  \hspace{1cm} (5.21)

and

$$dF(b) = -PS_{\text{eff}}$$  \hspace{1cm} (5.22)

where $P$ is given by equation 4.35, and $S_{\text{eff}}$ is an effective cross sectional area given by

$$S_{\text{eff}} = \frac{2\pi}{3} R_0 \frac{R_c}{R_0} \left[ 1 - \left( \frac{R_0}{b} \right)^{\frac{2}{3}} \right] \frac{R_c}{R_0} \left[ 1 + \left( \frac{R_0}{b} \right)^{\frac{1}{3}} \right]$$  \hspace{1cm} (5.23)

Both the extra energy and the depletion force will reduce to zero when $b$ approaches to the radius of the initial sphere.

The above mentioned results are for a protein in quasi-equilibrium states. In a real experiment the pulling force is applied in order to unfold the protein in finite times. The process in principle is not quasi-static but a dynamical one. But if, by
assumption, the pulling speed is within acceptable range, namely longer than the protein relaxation time, the protein will have enough time to sample the conformation space at fixed end-to-end distance. Our simple ellipsoid model can be applied to this quasi-equilibrium system. To get an idea on the trend of the extra free energy and the force to maintain an ellipsoid protein we have plotted the normalized quantities (dividing by maximum value in normalization) as functions of the $b/R_0$ ratio in figure 5.5. We can see the extra free energy increases with the $b/R_0$ ratio nonlinearly at first while it becomes asymptotically linear later. As a result the magnitude of depletion force increases from 0 rapidly at first, reaching and maintaining a higher value slowly.

Figure 5.5: Normalized extra free energy and extra force plotted as functions of $b/R_0$ ratio at a fixed radius of crowder. Thicker line represents the normalized extra free energy.
The molecular weight of dextran used in the experiment is \( M = 40 \text{kD} \). Given the dextran concentration \( C \) in the buffer solution we can compute the number density of dextran

\[
\rho = \frac{CN_A}{M}
\]  

(5.24)

and the volume fraction

\[
\varphi = \frac{4\pi}{3} \rho R_v^3
\]  

(5.25)

We have assumed the radius of dextran molecule to be \( R_v = 3.3 \text{ nm} \) [101]. These in turn give the osmotic pressure according to equation 4.35. \( V_{ex} \) and then the partition function \( Q \) are computed as functions of the ellipsoid long axial length and dextran concentration, resulting in a concentration dependent effective attractive force \( dF \) as a function of protein length. To maintain the protein in a deformed structure, an extra force is needed to balance this effective attraction and the magnitude of this force is plotted in figure 5.6.

Assuming that in different concentrations the peak of protein pulling force occurs when the long axial length of the protein equals 6.7 nm, we have plotted the depletion force against the dextran concentration in figure 5.7. Experimentally measured result of Dr. Yang’s group is plotted in figure 5.8 where see that the concentration dependence of the force required to unfold a protein is roughly linear. Although this simple protein model yields a force curve which is more nonlinear than the observed one, the general increasing trend and order of magnitude of the force predicted by the simple model agree roughly with the observed one. This may indicate that the excluded volume effect plays an important role in macromolecular crowding experiments. The exponential increase of this force is possibly due to our assumption
that the randomly coiled (in reality) dextran molecule is assumed to be a round molecule with a hard core. The osmotic pressure profile of the hard spherical molecules is only an approximation to that of real dextran polymer solutions.

Figure 5.6: Depletion force (in pN) as a function of protein long axial length and dextran concentration. From top to bottom, the curves correspond to 300g/l, 250g/l, 200g/l, 150g/l, 100g/l and 50g/l.

Figure 5.7: Peak extra force (in pN) as a function of dextran concentration.
Figure 5.8: Mechanical force to unfold ubiquitin molecule using AFM plotted as a function of dextran concentration. Horizontal is dextran concentration in unit of g/l, and vertical represents the pulling force measure in pN. In each group four increasing pulling speeds (nm/s) are presented: 50, 250, 1000, and 5000.

5.4 Evaluation of excluded volume using a 2-D lattice model

Besides the very simple sphere-ellipsoid protein model, we have implemented the excluded volume calculation using both lattice and off-lattice unbiased athermal homopolymer models. While we believe that using lattice or off-lattice models does not affect the basic property of the model but dimensionality does, we have implemented a 2-D lattice model and a 3-D off-lattice model. As a thermodynamic entity, the protein assumes a weighted distribution of conformations according to both energy dependent Boltzmann’s factor and the conformation entropy, \( g(E)e^{BE} \). We assume the interaction between the protein and the surrounding crowders is only additive, meaning that it does not affect the interaction potentials of protein in water, for an example the interactions between protein monomers and the water molecules are not affected in presence of crowders. As a result to adopt an athermal polymer
model but not a model with complete explicit potentials makes no difference in studying the crowder effects. Comparing to the simple sphere-ellipsoid model, we can collect more complete excluded volume information over a wider range of protein size for evaluating the effects due to crowders.

Consider a square lattice homopolymer with a length of 100 lattice spacing. Here since we are only interested in the excluded volume calculation, we simply ignore any interactions between the monomers and the environment, and interactions among monomers, except for the hard square interactions. The objective is to sample uniformly enough of the conformation space of this polymer so that we can obtain an averaged excluded volume. Fundamentally there exist two types of algorithms to sample the conformation space. The first one starts from a random conformation, and the model evolves in its conformational space using some predefined move set, until the ergodic condition is believed to be satisfied. In the second method a polymer conformation is grown from a single monomer each time, and a large number of newly created conformations will be collected for statistical analysis. When the number of moves in the first algorithm and the number of created polymers in the second one both approach infinity, these two algorithms converge to an identical statistics. We have adopted the second approach for conformation space sampling. The entire algorithm is as follows and the idea of it is shown in figure 5.9:

1. Grow a self-avoiding 2-D square lattice polymer with length of 100 lattice spacing; record the size information.
2. Scan the contour of the polymer using a 2-D square macromolecule crowder; record all lattice positions around the polymer that can accommodate the crowders.

3. Connect the centers of the accommodated macromolecules to form a closed path within which the number of square lattices is counted as the 2-D excluded volume.

4. Repeat Step 1, 2, and 3 until a set of large enough number of conformations is collected and the corresponding excluded volumes are calculated; Do statistical analysis and establish the relationship between the excluded volume and polymer size.

Figure 5.9: An example of a short 2-D lattice polymer and its envelope layer of square macromolecules of size $3 \times 3$. The region enclosed within the complete piece-wise line is the exclude volume. Note that a free region is also counted.
Figure 5.10: Excluded volume $V_{ex}$ (in lattice unit square) increases with the end-to-end distance $S$ of the protein polymer (100 monomers long) in a nonlinear way. From the bottom up the square crowder sizes are respectively, 1×1, 3×3, 5×5, 7×7, 9×9, and 11×11. Data dispersion due to insufficient sampling can be seen near the long end-to-end distance side.

In figure 5.10, the excluded volume around the contour of a lattice polymer (100mer) is plotted versus the end-to-end distance of this polymer. It turns out that using the polymer-growth algorithm to get a longer polymer becomes more and more difficult. This can be seen by observing the dispersed features of the curve reaching large end-to-end distance. Therefore we truncate the end-to-end distance at around 60 lattice unit, expecting not to lose much information since usually the probability for a polymer to be extremely extended is negligible anyway. All curves show a smooth increase of excluded volume with the end-to-end distance in a nonlinear way. Detailed study indicates that these curves could be fitted well by quadratic forms with the statistical $R^2$ value all being greater than 0.95. On the other hand the quadratic
dependence on the polymer end-to-end distance becomes stronger with the increase of the crowder size, which is clearly shown in figure 5.11 where the $2^{\text{nd}}$ order coefficient of the fitted quadratic form is plotted against the crowder size. A direct result is that the linear dependence becomes less dominant for larger crowders. This could be understood by realizing that the excluded volume due to large crowder depends on the size of crowder strongly because the crowder size is comparable to the average size of the polymer.

![Graph showing quadratic coefficient vs. crowder size](image)

**Figure 5.11:** The quadratic coefficient of excluded volume as a function of protein end-to-end distance plotted versus crowder size $R_c$. The trend becomes stronger with the increase of polymer size. $k$ here is the coefficient of the $2^{\text{nd}}$ order term in fitted quadratic form.

Applying the scaled particle theory, we were able to compute the osmotic pressure of the hard-square lattice crowders in the solution where the lattice protein is in. Since the number of bombardments of crowders on the protein becomes greater because the
excluded volume becomes larger when protein is extended, this is equivalent to a kind of retractive force that compresses the protein to a more compact structure. For different crowder volume fraction, this retractive force is calculated against the polymer end-to-end distance in figure 5.13. This figure shows an increasing trend, almost linear, along the polymer extension, resembling a kind of Hooke’s law that describes a simple spring force. There the force axis is in fact \( F/k_B T \), with the unit being in 1/lattice cell. To give a rough idea on how large this force is, we state the fact that an interval of 0.1 along the force axis stands for 1.217 pN, when the temperature \( T = 300 \) K and the lattice spacing \( l = 3.4 \) nm. The depletion force increases with the volume fraction of crowders at fixed crowder size, but decreases with the crowder size at fixed volume fraction. Here in addressing the question about the effect of crowder size on protein, we have to fix either the volume fraction \( \phi \) or the crowder number density \( \rho \). One more feature in this plot that should be mentioned is that our 2-D simulation results show no peak in the depletion force. That is, the magnitude of the depletion force continues to increase even until the 2-D lattice polymer reaches its longest conformation, where we have to apply the largest external force in order to maintain a polymer in such a conformation. This may due to the inherent statistical characteristics of 2-D lattice polymer; but another possibility is that our simulation is not complete and it is possible that when the end-to-end distance exceeds 60 lattice spacing, the excluded volume levels off, approaching the largest value asymptotically. We leave this to be part for our future work.

Associated with the depletion force we can also construct a depletion potential as a function of polymer extension. This is simply the integral of the depletion force over
the extension. The result is shown in figure 5.14 where four panels similar to the force figure, 5.13, are presented. We can see that at the same volume fraction, the smaller crowder (which has higher number density) has a higher stabilization effect on the compact polymer against extended structures, though the excluded volume is smaller. This stabilizing potential energy (though calculated from homopolymer without considering specific interactions) can be added to the original free energy landscape of the protein molecule without the presence of crowders. We will come back to this point in 3-D off-lattice discussion.

Figure 5.12: Depletion force along the lattice polymer end-to-end distance. Panel a is for square crowder of size $1 \times 1$, panel b for $5 \times 5$, panel c for $9 \times 9$, and panel d for $13 \times 13$. In each panel four curves at increasing crowder volume fraction, 10%, 20%, 30%, and 40% are shown from top down. The vertical axis stands for the normalized depletion force $F/ k_B T$ in unit of reverse lattice spacing, and the horizontal is end-to-end distance. If lattice spacing is $l = 3.4$ nm, an interval of 0.1 along force axis represents 1.217 pN at $T = 300$ K.
Figure 5.13: Depletion potential energy along the lattice polymer end-to-end distance. Panel (a) is for square crowder of size $1 \times 1$, panel (b) for $5 \times 5$, panel (c) for $9 \times 9$, and panel (d) for $13 \times 13$. In each panel four curves at increasing crowder volume fraction, 10%, 20%, 30%, and 40% are shown from bottom up. The vertical axis stands for the depletion energy in $k_B T$, and the horizontal axis stands for polymer end-to-end distance.

5.5 Evaluation of excluded volume using a 3-D off-lattice model

We now move on to a simple 3-D off-lattice homopolymer model for the excluded volume calculation. In this model each monomer is represented as a hard sphere with certain radius $R_m$, and between any two successive monomers a bond with fixed length $b$ (center to center) is assigned. To sample the conformational space more efficiently we used the polymer growing algorithm as well as the trial-and-error pivot algorithms [102-109] (see page 88 for detail). We found, however, the probability to obtain a very extended polymer conformation, for example the rod-like structure, is
rather low as has been mentioned before. In fact the distribution of the end-to-end distance of an athermal polymer follows a Gaussian-like distribution positively skewed [110]. Therefore to acquire a complete estimation of the excluded volume over the entire end-to-end distance range, from one bond length, the minimum value, to $(N-1)b$, the maximum length, we have also initiated the polymer at its straight rod structure, from there on the pivot algorithm is applied, to sample the conformation space localized around the long end-to-end distance region.

Given any 3-D polymer conformation and any crowder radius, the outer surface of the depletion zone turns to be quite complex in general [111], making explicit volume calculation very difficult geometrically. Though not implemented, the 2-D off-lattice case is comparatively easier to calculate due to the relatively simple partitioning of the depletion area. The depletion area can be partitioned into a set of incomplete circles and an irregular polygon whose vertices being the subset of all the intersection points of the excluded circles that envelope any individual monomer (see figure 5.14). The excluded area can then be calculated as the sum of the area of all incomplete circles and area of the polygon, which should be triangulated beforehand. To avoid the geometric difficulty in the 3-D case we introduced another layer of Monte Carlo calculations on top of the main Monte Carlo method and the pivot algorithm that generate and modify the polymer conformations. Basically we find the smallest rectangular parallelepiped that encloses the depletion zone completely, and apply the standard Monte Carlo method to estimate the excluded volume by counting the number of random points, drawn uniformly in the rectangular parallelepiped, that fall
inside the depletion zone. This procedure, illustrated in figure 5.14, can be described as follows:

1. Given any self-avoiding off-lattice polymer conformation, find the minimum rectangular parallelepiped enveloping the depletion zone completely.

2. Draw a random point within the rectangular parallelepiped; if the distance between this point and the center of any monomer is less than the sum of monomer radius and crowder radius, it will be counted; otherwise it will not.

3. After drawing a certain large number of random points, approximate the excluded volume by the product of box volume and count ratio.
Figure 5.15: Principle working mechanism of the pivot algorithm. The monomer labeled by an arrow is the pivot about which the shorter arm of the polymer can rotate an angle triplet \((\alpha, \beta, \gamma)\) randomly selected, if sterically permitted. The curved arrow shows the rotation direction.

It was found that the rotational angles cannot be set too large or too small. The former choice will reduce the acceptance rate in a trial rotation because the shorter arm of polymer will collide with monomers along the longer arm with a high probability. The latter choice erases this problem but makes the sampling of conformation space very inefficient. After several trial and error simulations we found the optimal triplet of angles should be drawn between 0 to 0.5 degrees uniformly. Adopting such a proper angle range increases the acceptance rate of the trial rotations to about 50% while maintains the efficiency of conformation space sampling as well. Below we list the pivot algorithm as applied in our simulations:

1. Grow a 3-D off-lattice polymer with 76 monomers, or initiate it at rod-like structure

2. Randomly pick up a monomer about which the shorter arm of polymer will be rotated by an triplet of angles \((\alpha, \beta, \gamma)\) about x-, y-, and z-axis,
respectively; for each newly modified conformation, calculate approximately and record the excluded volume.

3. Repeat Step 2 for a certain number of times in order to sample the conformation space local to the created polymer.

4. Repeat Step 1, 2 and 3 for enough MC steps until good statistics is collected.

In order to test our simulation results with experimental results obtained by Dr. Yang’s group, we set the parameters as follows: the monomer radius is $R_m = 0.125$ nm, the bond length between two linked monomers is $b = 0.34$ nm, and the number of monomers is 76, which is just the number of the amino acids of ubiquitin, the protein used in Yang’s experiments. Since as the crowder size grows, the computed excluded volume becomes very large, we present the natural logarithm of the excluded volume in figure 5.15. This set of curves bear a fundamental difference from the 2-D case (figure 5.10) where the excluded area follows a quadratic form; it shows an abrupt increase around 6 or 7 nm, below and above which are two distinct domains. We recall that the depletion force depends on the changing rate of the excluded volume over the geometrical parameter of the molecule. We see, from figure 5.16, that in the small end-to-end distance region the depletion force steadily increases and in the large end-to-end distance domain, since the slope levels off towards 0, the depletion force approaches minimum value asymptotically there. We therefore expect there exist peak values in evaluating the first derivative of the excluded volume with respect to the end-end distance.
Figure 5.16: Natural logarithm of excluded volume $\ln(V_{\text{ex}})$ as a function of end-to-end distance $S$ of the 3-D off-lattice 76mer. $S$ is in unit of 0.1 nm. The minimum possible end-to-end distance is a bond length $b = 0.34$ nm and the maximum is 25.5 nm. Increasing from bottom up, the ten curves are calculated in simulation for crowders of size 0.5 nm, 1.0 nm, 1.5 nm, 2.0 nm, 2.5 nm, 3.0 nm, 3.5 nm, 4.0 nm, 4.5 nm, and 5.0 nm.

Analytically, the excluded volume of the 3-D polymer in straight rod structure can be expressed as the following mathematical form:

\[
\frac{V_{\text{ex,max}}}{\frac{4\pi}{3} R_m^3} = \left(\frac{R_c}{R_m} + 1\right)^3 + \frac{3}{4} (N - 1) \left(\frac{R_c}{R_m} + 1\right)^2 - \frac{1}{16} (N - 1) \left(\frac{b}{R_m}\right)^3
\]

which is a polynomial of factor $1+R_c/R_m$, given fixed bond length $b$ and fixed radius of monomer $R_m$. The first and third terms can be discarded when $b$ is much smaller than $R_m$ and the number of monomers $N$ is large, leading to an expression with only a quadratic form. Correspondingly we can compute the minimal excluded volume when protein is at native state, modeled as a hard sphere with radius $R_0$: 
\[
\frac{V_{\text{ex, min}}}{4\pi \frac{3}{R_m}} = \left( \frac{R_c + R_d}{R_m} \right)^3
\] (5.27)

This expression is purely cubic. From these facts we can expect that when the crowd size is much larger than the protein size, the difference of excluded volume between folded and extended states tends to be negligible. In other words the excluded volume will not change apparently with the protein size. The maximal excluded volume (relative to monomer volume) is plotted for our 76mer model in figure 5.17.

Figure 5.17: The logarithm excluded volume \( V_{\text{ex, rel}} \) relative (relative to volume of monomer volume) increases with crowder size. Number of monomers is \( N = 76 \).

Given the excluded volume at any polymer extension we can compute the depletion force according to equation 5.22. We note that in real simulations larger box enclosing the polymer (see figure 5.14) is used to evaluate the excluded volume
when the polymer extension is larger, while a fixed number of MC steps are applied to all extensions. As a direct result the fluctuation of the excluded volume at a larger polymer end-to-end distance will be larger than that around a shorter end-to-end distance. This noise-inhomogeneity makes the curve fitting complex. The way we did is to fit the excluded volume in a piecewise method. The data set is split non-uniformly into segments with different lengths. Data segment at shorter end-to-end polymer distance is generally taken shorter and data segment at longer polymer extension is longer in order to erase the Gaussian distributed noise in an equal way, tentatively. All data segments are fitted linearly. The slope of the fitted straight line is taken as the first derivative of the excluded volume with respect to the polymer extension. We have computed the depletion force for crowders of size 1 nm and 3 nm, at a crowder volume fraction 40% and room temperature $T = 300$ K. In figure 5.18 the piecewise depletion force is plotted. This force steadily increases to a peak value at around 7 nm, above which the force rapidly decreases and levels off with the polymer extension. Therefore if the depletion force is considered to be partially responsible for the extra pulling force observed using AFM to unfold a real protein (ubiquitin) in a solution of crowders, this force will be of the order of 10 pN, an extra amount added to the pulling force required for a protein in the absence of crowders.
Figure 5.18: Depletion force-extension profile. Shown here are for crowders with sizes 1 nm (thicker line) and 3nm (thinner line) respectively. The force is in unit of pN and the extension (end-to-end distance of polymer) is in 0.1nm. The temperature is $T = 300$ K and the volume fraction is 40%.

In some cases, AFM pulling being one of those, we are more interested in the maximum depletion force generated due to the excluded volume. Our figures show that when the end-to-end distance reaches around 6 or 7 nm, the derivative of excluded volume reaches its maximum value, corresponding to the maximum depletion force. We focused on a set of data with the maximum end-to-end distance less than 10 nm and the minimum larger than 5 nm for a finer curve fitting. To minimize the fitting error the data has been firstly normalized by dividing each datum by the maximum one, and then fitted by a cubic polynomial. The first order derivative is simply computed from the polynomial. For the osmotic pressure calculation the temperature is set to be at $T = 300$ K. Ten increasing crowder sizes, from 0.5 nm to 5.0 nm uniformly, are used for the pressure computation. We expect, as in the 2-D
lattice case, the maximum depletion force should increase with the crowder volume fraction. Therefore we completed the force calculation for four increasing crowder volume fractions, 10%, 20%, 30%, and 40%, respectively. The result is summarized in figure 5.19. There we can see in each bar group that corresponds to a fixed crowder size, the maximum depletion force does increase with the volume fraction of crowders in the same way as the pressure dependence on the volume fraction. An example is shown in figure 5.20 where the crowder size is chosen to be 2.0 nm. Another notable feature of figure 5.19 is the deceasing trend of the maximum depletion force with the crowder size at fixed volume fraction. This is understandable when we consider that at fixed volume fraction \( \phi \) of crowder the number density \( \rho \) in equation 4.35 is inversely proportional to the volume of a single crowder, \( V_c \), because \( \rho = N/V = \phi/V_c \). Therefore when we focus on a specified volume fraction \( \phi \), the osmotic pressure, and then the maximum depletion force will be inversely proportional to the crowder size.
Figure 5.19: Bar diagram of maximum depletion force $F_d$ plotted as a function of both the crowder volume fraction and crowder size. Each number along the abscissa is the crowder size in unit of nm and each size includes four increasing crowder volume fraction, 10%, 20%, 30%, and 40%, respectively. Ten increasing sizes are shown. The vertical label represents the force in pN. The temperature for this plot is set to be $T = 300$ K.

Similarly we are also interested in the relation between the maximum depletion force and the crowder size, when the number density of crowders in solution is fixed. We consider four increasing number density in series. They are $7.639 \times 10^{23}$ m$^{-3}$, $1.146 \times 10^{24}$ m$^{-3}$, $1.528 \times 10^{24}$ m$^{-3}$, and $1.910 \times 10^{24}$ m$^{-3}$, being uniformly separated. Figure 5.21 shows that at each crowder size, the maximum depletion force increases with the number density linearly, and globally it increases with the crowder size when the number density is a constant.
Figure 5.20: Maximum depletion force versus the volume fraction of crowder with radius 2 nm as an example to show the increasing trend of force.

Figure 5.21: Bar diagram of maximum depletion force plotted as function of both number density of crowders and crowder size. Each number along the abscissa is the crowder size in unit of nm. Each bar group includes four increasing number density of crowders $7.639 \times 10^{23}$ m$^{-3}$, $1.146 \times 10^{24}$ m$^{-3}$, $1.528 \times 10^{24}$ m$^{-3}$, and $1.910 \times 10^{24}$ m$^{-3}$ respectively. Ten increasing sizes are shown. The vertical label represents the force in pN. The temperature for this plot is set to be $T = 300$ K.
Other than the depletion force calculation the effective depletion potential is applied to investigate the stabilization of protein. When in a crowded environment the native protein has smaller probability to unfold into extended structure spontaneously than in dilute solution. There will be two main reasons for this. One is that the monomer movement of protein will be limited by the presence of crowders, even when the crowders are assumed to be at random static states without motion. This is similar to the confinement effect and it does not affect the protein energy profile along the folding-unfolding reaction coordinate. The second reason is that with the presence of depletion force, there exists a corresponding depletion potential energy. Taken the native conformation as reference zero-potential point, the unfolded conformations of protein have energy higher than zero. We are interested in the part due to the depletion force. In solutions of crowders, from folded to unfolded conformations, a protein requires extra energy to overcome the effect of depletion force. Therefore the potential energy difference between the folded and unfolded conformations will be larger than that in a solution without crowders. Let $\Delta G = G_u - G_f$, be the free energy difference between the folded and unfolded conformations in a solution without crowders, and $\Delta G_c = [G_u - G_f]_c$ the counterpart in a solution with crowders. The depletion free energy is then defined to be $\Delta \Delta G = \Delta G_c - \Delta G$ and calculated for any crowder size according to:

$$
\Delta \Delta G = -\int_{S_{\min}}^{S} F_d(S) dS \\
\approx -\sum_{i=1}^{n} F_d(S_i) \Delta S_i \\
= -P \sum_{i=1}^{n} (\frac{\partial V_{ex}}{\partial S})_i \Delta S_i \\
= -P(V_{ex,max} - V_{ex,min}) \tag{5.28}
$$
We have done the calculations for all ten crowder sizes at fixed volume fraction $\varphi = 28\%$. The magnitude of this depletion free energy is plotted in figure 5.22. As before, the depletion free energy curves for different crowder sizes show a sigmoidal shape with the turning point located at around 7 nm. Different from the bare excluded volume curves, the sequence of depletion force at fixed volume fraction is reversed, with the higher free energy (or larger effect) associated with the smaller crowder size.

Figure 5.22: Depletion free energy (in unit of $k_B T$) plotted against end-to-end distance of polymer at fixed volume fraction of crowders, $\varphi = 28\%$. From top down the crowder size goes from 0.5 nm to 5.0 nm uniformly. This trend is reversed compared to that of the excluded volume, seen in figure 5.16.

We now consider a two-state protein model. On top of the routine free energy profile along the one-dimensional folding-unfolding reaction coordinate (for example end-to-end distance) the depletion free energy maybe added, resulting in a new free energy profile. The shape of the new free energy curve depends on several factors.
The first factor is the magnitude of the depletion free energy, and the second is the relative shift between the inflection point on the sigmoidal curve of depletion free energy and the location of the transition state between the folding-unfolding wells along protein extension. If further we assume the shift is small, in other words the protein transition state location is just around the inflection point of calculated depletion potential, we can draw an artificial diagram to show the difference between the free energy landscape without crowders and the one with added depletion effect. This is presented in figure 5.23. The lower basin represents the free energy well when protein is around the native state, the higher basin represents the well when protein is unfolded, and the middle bulge represents the transition state, over which the protein can fold and unfold into the respective states. Addition of the extra depletion potential onto the original profile yields a new energy profile with the native well mostly unchanged, but the transition state and/or unfolded state are elevated. The possible elevation of the transition state may have the same amount of elevation as that of the unfolded state or a much smaller amount, depending on the distribution of depletion free energy along the protein end-end extension. Globally the free energy gap $\Delta G$ between the native and unfolded states becomes larger, destabilizing the unfolded state, or equivalently, stabilizing the native conformation by the amount of $\Delta \Delta G$. 
Figure 5.23: Cartoon of free energy surface of two-state protein folding.
Chapter 6: Conclusions

We have reviewed extensively the protein folding problem and its associated research approaches, from in vivo experiments and theoretical studies, to in silico computer simulation methods. Protein folding is an interesting problem with a long history, and many questions about it remain unanswered.

We have investigated the protein folding problem and equilibrium properties for a protein in confinement using an extended 2-D HP lattice model. The thermal stability of a protein, revealed by studying the heat capacity, is enhanced by confinement. Smaller confinement shifts the protein melting temperature to a higher value, making heat-denaturation of protein under the same physiological conditions more difficult. This finding matches the experimental results qualitatively well. In different temperature ranges the mean first-passage time of protein folding exhibits opposite trends: at lower temperature, confinement slows down the folding process possibly because that the hard wall of confinement blocks the fast folding pathways, while at higher temperature confinement speeds up the folding process. The mean first-passage time of unfolding, on the other hand, monotonically increases with a decrease of confinement size at any temperature.

Similarly, we have applied an extended HP lattice model to study how macromolecular crowding affects protein properties. To avoid persistent trappings by local minima, especially at low temperature, we have employed some advanced computational algorithms, such as multihistogram method and the replica change method in this investigation. We have found that the melting temperature is shifted,
similar to but less pronounced than in confinement, towards higher temperature as well. In this sense the protein is stabilized with respect to its native conformation. This is also shown by the calculation of the equilibrium constant, defined as the ratio of folding and unfolding times. Macromolecular crowding shifts the equilibrium towards the native state. The mean first-passage times of unfolding plotted versus the reciprocal temperature under different volume fraction of crowders seem to be parallel to one another. This fact indicates to us that the free energy surface, from the folded potential well to the transition state remains approximately the same, but the pre-exponential factor (frequency to attempt unfolding) decreases with crowding.

Introducing the depletion zone idea and the scaled particle theory, we have carried out a calculation of the excluded volume for a system of a homopolymer composed of 76 monomers. This calculation is repeated for solutions containing different number of hard-sphere crowders with variable radius, using double Monte Carlo simulations. Different from a monotonically increasing trend obtained for a 2-D lattice homopolymer model (100mer) and a simple ellipsoidal protein model, the excluded volume calculated for the 3-D off-lattice homopolymer demonstrates a phase transition along the polymer extension, hinting the existence of a peak in depletion force. This depletion force is believed to be a part of the extra pulling force needed to unfold mechanically a single protein molecule in a concentrated solution using AFM. The results are in reasonable accordance with the experiment semi-quantitatively. The depletion potential calculated from the depletion force can be in principle added to the free energy landscape of protein along the folding/unfolding reaction coordinate. Based on the specific shape of the depletion potential the resultant free energy surface
maybe modified in different ways. Activation energies for the folding and unfolding processes and even the free energy difference between the folded and unfolded states can undergo non-uniform changes, leading to stabilized folded state and affecting, in principle, the folding time as well as the unfolding time.
Bibliography


[37] [http://folding.stanford.edu/](http://folding.stanford.edu/)


Guanghui Ping was born in Baoding, southern guard city of Beijing, on July 2nd, 1972. After obtaining his high school diploma from the 17th Middle School of Baoding in 1991, he entered Hebei University to study theoretical physics and obtained his Bachelor’s degree in 1995. From 1995 to 1996, he worked as a technician for the 2nd Transformer Factory of Baoding. He then went to Shanghai Observatory, Chinese Academy of Sciences, to study the theories and the 25-m radio telescope observation techniques of astronomical maser.

In 1999, he came to the Department of Physics at Drexel University and converted to the biophysics research, tutored by Dr. Jian-Min Yuan. He obtained his Master’s degree in 2001, and completed his graduate studies in the spring term of 2005.

His publications and posters are listed below:


