MOVE-BIASED MONTE CARLO SIMULATION METHOD FOR PROTEIN NATIVE STRUCTURE PREDICTION

BY

SAMSON OLATUBOSUN AISIDA

B.Sc. (Hons.) Physics and Electronics (AAUA), M.Sc. Physics (Ibadan)

A Thesis in the Department of Physics,
Submitted to the Faculty of Science
In partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN PHYSICS

of the

UNIVERSITY OF IBADAN

February 2016

ABSTRACT

Proteins are polymers of Amino Acid (AA) which are constructed after translation of genetic code in DNA of organisms, and have functionality that depends on their Native Structure (NS). Experimental methods for protein NS determination are complicated, expensive and time-consuming. Consequently, Computational Methods (CM), including Monte Carlo (MC), aim to circumvent these challenges. However, the MC is complex and inconsistent in NS Prediction (NSP). This study was designed to develop a Move-Biased MC (MBMC) simulation algorithm that may simplify the complexity of existing MC and makes it consistent for NSP.

Protein was described as a coarse-grained structure and folding as Self-Avoiding Walks (SAW) on square lattices. Relative Probability Parameters (RPP) were introduced to determine natural probabilities of protein conformations from SAW and to simulate the desired sequence length from RPP optimal combination. Thereafter, a graphical algorithm was developed to group the SAW steps into hydrophobic and polar AA units according to the Hydrophobic-Polar (HP) model. The MBMC method was developed as a coupling of diagonal-pull Move-Biased (MB) on the lowest energy SAW conformation. The materials for testing the MBMC method included eight Benchmark Sequences (BMS) from the protein data bank such as SI-1, SI-2, SI-3, SI-4, SI-5, SI-6, SI-7, and SI-8 with sequence lengths 20, 24, 25, 36, 48, 60, 64, and 85 nm, respectively. The lowest energy (consistency in prediction of NS), computation time and algorithmic steps of the MBMC method was compared with some existing methods [such as Conventional MC (CMC), Genetic Algorithm (GA), Evolutionary MC (EMC), Ant Colony Optimization (ACO), Hybrid Elastic Net Algorithm (ENA)]. Data were analysed using inferential statistics.

The optimal combination of the RPP for the MBMC algorithm were 0.71, 0.02, 0.25 and 0.02 for up, down, left and right orientations, respectively. The energies of the NS obtained from the MBMC method were -9, -9, -8, -14, -23, -35, -42 and -52 J for the BMS, respectively. In contrast, for GA energies derived were -9, -9, -8, -12, -22, -34, -37, and no record for eighth BMS; for ACO they were -9, -9, -8, -14, -23, -34, -32, -53; for EMC they were -9, -9, -8, -14, -23, -35, -39 and -52; for ENA they were -9, -9, -8, -14, -23, -36, -39, and no record for eighth BMS; for CMC they were -9, -9, -7, -12, -20, -33, -35, and no record for the eighth BMS. Also, MBMC method consistently predicted the NS of the BMS in 8.90, 8.51, 8.37, 9.14, 9.45, 9.46, 9.52, and 12.85 seconds, respectively. In contrast the computation times for GA were only reported for the first four BMS as 5.60, 6.00, 3.66, 54.60 seconds, and no record of computational time for the CMC and EMC Benchmark sequences, respectively. Moreover, MBMC has fewer algorithmic steps and simpler simulation procedure than CMC, GA, EMC and ENA methods.

The developed Move-Biased Monte Carlo method had simpler algorithmic steps than the existing Monte Carlo methods and consistently predicted the native structure of proteins faster than existing algorithms.

Keywords: Protein primary structure, Coarse-grained model, Hydrophobic-Polar lattice model, Monte Carlo simulation.

Word count: 495

ACKNOWLEDGEMENTS

I want to express my appreciation to all members of staff in the Department of Physics, University of Ibadan for their input in my academic pursuit for the last eight years in this citadel of learning; especially, to my supervisor, Dr E. O. Oyewande, for his thoroughness, fatherly advice, encouragement and patience all through the period of this work. The Lord Almighty will bless and enlarge you in all facets of life in Jesus name (Amen).

My appreciation also goes to my wonderful parent Mr. Samuel Ajayi Aisida and Mrs. Dorcas Remilekun Aisida for their prayers, support, encouragement and love all through the period of my academic pursuit. Thanks daddy and mummy for being there for me always.

Also, my sincere appreciation goes to my siblings: Ebenezer Aisida and his family, Akinola Aisida and his family, Babatope Aisida and his family, Temidayo Aisida and his family, Olarewaju Aisida and his family, Mopelola Aisida and her family and Funmilayo Aisida and her family. Word cannot quantify your love towards me. I am indeed grateful to all of you for your support and encouragement in the course of this work. May God continually reward and bless you all in Jesus name (amen).

My heartfelt appreciation goes to my dear wife Jolaade Aisida sine qua non and my children Ayanfeoluwa and Oluwaferanmi; their encouragements and deep sacrifices in all ramifications made this work a success. You are simply the best. I love you all.

Finally, to the Almighty God the giver of life, who kept me through the period of this work. Daddy I am indeed grateful.

CERTIFICATION

I certify that this work was carried out by Mr. S.O. Aisida in the Department of Physics, University of Ibadan

……..

Supervisor
E. O. Oyewande, MInstP
B.Sc, M.Sc (Ibadan); DICTP (Trieste), Ph.D. (Gottingen)
Senior Lecturer in Theoretical Physics, Department of Physics,
University of Ibadan, Nigeria

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure **page**

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
AUG	Adenine-Uracil-Guanine
mRNA	Messenger RNA
GP	Globular Protein
FP	Fibrious Protein
MP	Membrane Protein
NMR	Nuclear Magnetic Resonance Spectroscopy
3D	Three-dimension
2D	Two-dimension
PDB	Protein Data Bank
MBMC	Move-biased Monte Carlo
H-P	Hydrophobic- Polar
tRNA	Transfer RNA
rRNA	Ribosomal RNA
PS	Primary Structure
SS	Secondary Structure
TS	Tertiary Structure
QT	Quaternary Structure
BCX	Bacillus Circulans Xylanase
ER	Endoplasmic Reticulum
AD	Alzheimer's Disease
HX	Hydrogen Exchange
F	Folded Protein
U	Unfolded Protein
N	Sequence Length
PSP	Protein Structure Prediction
CS	Computer Simulation
NOEs	Nuclear Overhauser Effects
EM	Electron Microscopy
MS	Mass Spectroscopy

FS	Fluorescence Spectroscopy
CD	Circular Dichroism
SMS	Single Molecule Experiments
AFM	Atomic Force Microscopy
SMFS	Single Molecule Force Spectroscopy
TEM	Transmission Electron Microscopy
SEM	Scanning Electron Microscopy
IRS	Infrared Spectroscopy
ESR	Electron Spin Resonance
ORD	Optical Rotatory Dispersion
REMC	Replica Exchange Monte Carlo
GA	Genetic Algorithm
GAOSS	Genetic Algorithm on Optimal Secondary Structure
TSS	Search Strategy
EMC	Evolutionary Monte Carlo
CMC	Conventional Monte Carlo
FCC	Face Centre Cubic
ACO	Ant Colony Optimization Algorithm
ENL	Hybrid Elastic Net Algorithm
CG	Coarse–Grained
MC	Monte Carlo
MD	Molecular Dynamic
H	Hydrophobic
P	Polar
BPTI	Bovine Pancreatic Trypsin Inhibitor
SQR	Square
CUB	Cubic
TN	Topological Neighboring
MCMC	Markov Chain Monte Carlo
MCMCMC	Metropolis-Coupled Markov Chain Monte Carlo
SAW	Self Avoiding Walk

PBC	Periodic Boundary Condition
E_n	New Energy
E_{ref}	Reference Energy
$E_{l.b}$	Lower Bound Energy
RI	Relative Improvements
GSC	Ground State Conformation

LIST OF SYMBOLS

Chapter 1

Symbol	Description
Å	Amstrong

Chapter 2

Symbol	Description
N	Nitrogen
H	Hydrogen
C	Carbon
O	Oxygen
α	Alpha
C_α	Carbon-alpha (central atom)
NH_2	Amine group
COOH	Carboxyl group
R	Side chain (variant)
ϕ	Conformational angle around $N-C_\alpha$ bond
ψ	Conformational angle around $C_\alpha - C'$ bond
KD	Kilo decibel
eV	Electron volt
υ	Scaling exponent
X_i	Concentration of proteins
K	Association constant
M	Number of nonpolar molecules
P	Polar molecules
nP	Number of polar molecules
E_{i-i}	Ion-ion interaction
E_{i-d}	Ion-dipole interaction

E_{d-d}	Dipole-dipole interaction
q	Charge
ε	Emissivity
ε_o	Emissivity of free space
r	Radius
K_B	Boltzmann constant
T	The absolute temperature
$\in_{hp}$	Effective hydrophobic attraction
τ	The angular dependence
ω_{ij}	The strength of attraction between amino acid i and j
η_{ij}	The geometric factor
e_{L-J}	Leonard-Jones potentials
σ	Sigma
S	Configurational entropy
R	The gas constant
F	Folded state
U	Unfolded state
C_f	The microscopic rate constants for the folding
C_u	The microscopic rate constants for the unfolding
ΔG^0	The free energy of folding
ΔS^o	The entropy
ΔH^0	The enthalpy
ΔC_p	The heat capacity difference
T_R	Reference temperature
T_f	Folding temperature
T_g	Glass temperature

Chapter 3

Symbol	Description
$\|i-j\|$	The distance along the chain
$\|r_i-r_j\|$	The spatial distances
σ_i	Amino acid sequence
N	Number of protein sequence
$\{b_i\}=\{r_{i+1}-r_i\}$	Bond vector
q_i	Amino acid position
L	Lattice
N_L	Neighborhood vector
ζ	Protein structure
ζ_i^b	Backbone monomer
ζ_i^s	Side chain monomer
$E(\sigma,\zeta)$	Energy function
E_{HH}	Energy between hydrophobic-hydrophobic
E_{HP}	Energy between hydrophobic-polar
E_{PP}	Energy between polar-polar
λ	The number of hydrophobic contacts
$\varpi_{m,n}(T)$	Temperature-dependent probability of the metropolis
ΔE	The change in the potential energy
ξ	Probability distribution
$P(n\|m)$	The transition probability density
$\Omega(n)$	State distribution
$P(\sigma)$	Probability distribution
Ω^*	The stationary distribution
Z	Canonical partition function
k	Boltzmann constant
$\zeta^{\otimes}$	Conformation with the lowest energy

Chapter 4

Symbol	Description
σ_{orig}	Original sequence
σ_{HP}	Converted sequence
$\Phi(\sigma)$	The degeneracy
α	Alpha
β	Beta
P_u	Probability of up
P_d	Probability of down
P_r	Probability of right
P_l	Probability of left

CHAPTER 1

INTRODUCTION

1.1 Biophysical backgrounds of proteins

The word "protein" stems from the Greek word "*proteios*" which means "of the first rank" which generally refers to the complete biological molecule in a stable conformation. Like other biological macromolecules such as polysaccharides, nucleic acids and lipids; proteins are essential parts of organisms and participate in virtually every process within cells. They are compact macromolecules that exist and function in aqueous environment (Michael, 2003; Zhao, 2008).

Life on earth is made of atoms and molecules, if DNA defines life; proteins are life, (Erik, 2000). Three properties guarantee life over many periods of replication: (i) The information transfer (ii) self-replication and (iii) self-repair. Without these three properties, life could not exist. Hence, molecular design of life is determined by protein with diverse functions such as enzymatic catalysis, mechanical support, immune protection, or generation and transmission of nerve impulses. It is good to know what life is all about, according to Yukawa, "life appears to work like assembling blocks" which is in agreement with what usual biological term genetics tells us i.e. genetics is one of the most essential aspects of living existence. The most essential messages from the study of genetics are twofold: the first is that life exists in two phases: information phase and an action phase. The second is that life exists as a very complex system made of parts. These two aspects are also commonly observed in complex man-made machines. These man-made machines also exist both in design books and as real machines in action. Thus, genetics provides us a view to look at life as a complex system (Broglia and Tiana, 2003; Nobuhiro, 2007).

To understand any complex system, we need to understand firstly, the component parts and their functional characteristics; and secondly, the design principles to assemble the parts into working complex systems. Relating to life on earth, parts are in the

information phase ‘genes’ and in the action phase ‘proteins’. System design is written in the way gene expression is controlled and also in protein-protein interaction. Functional characteristic of parts in action phase (i.e. Proteins) obey laws of physics, because the protein molecules are nothing but linear polymers consisting of twenty different types of amino residues (Nobuhiro, 2007).

Protein molecules exist in an interesting place between physics and biology. Being simply polymers, they are surely an object of physics that carry biological functions and work as parts (building block) assembled into complex living system (Nobuhiro, 2007).

Many proteins are enzymes that catalyse biochemical reactions and are vital to metabolism. Proteins also have structural or mechanical functions, such as actin and myosin in muscle and the proteins in the cytoskeleton, which form a system of scaffolding that maintains cell shape. Other proteins are important in cell signalling, immune responses, cell adhesion, and the cell cycle. Proteins are also necessary in animals' diets, since animals cannot synthesize all the amino acids they need and must obtain essential amino acids from food. Through the process of digestion, animals break down ingested protein into free amino acids that are then used in metabolism.

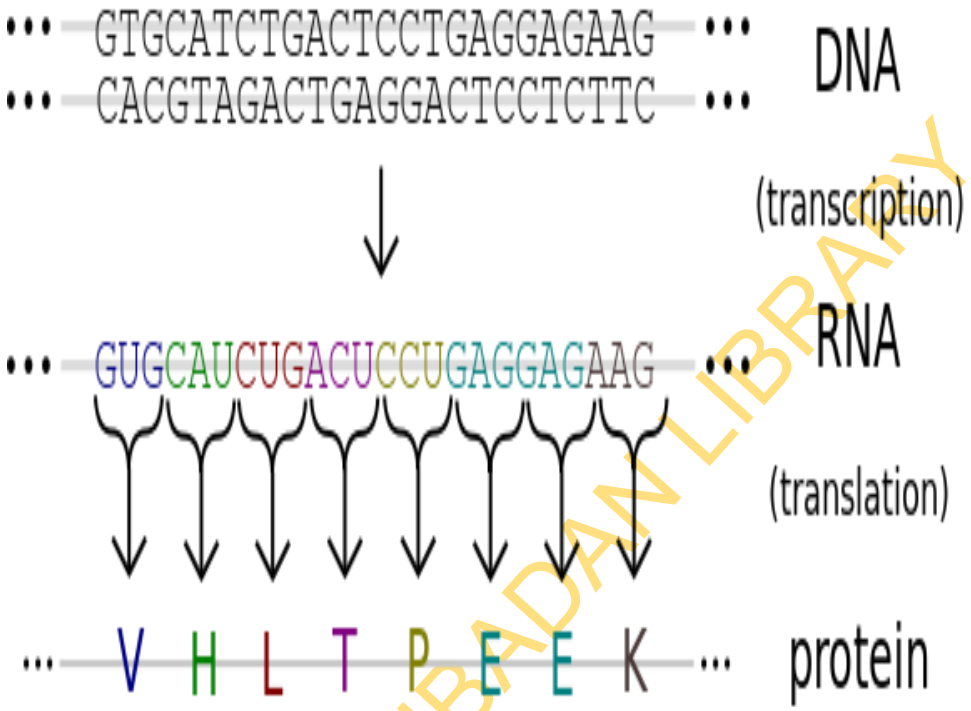


Figure1.1. The DNA sequence of a gene encodes the amino acid sequence of a protein (retrieved April 20, 2013, from http://en.wikipedia.org/wiki/Protein).

Proteins are assembled from amino acids using information encoded in genes. They are composed of basic structural units called amino acids. Each amino acid consists of an amino group, a carboxyl group, and a unique R group. All the groups are bonded to the centres $\alpha-$carbon atom. Each protein has its own unique amino acid sequence that is specified by the nucleotide sequence of the gene encoding this protein, since there exists 20 different amino acids commonly found in naturally occurring proteins, in which their diversity gives the different range of functions performed in proteins and differ only in their unique R group, which is also known as a side chain.

The genetic code is a set of three-nucleotide sets called codons and each three-nucleotide combination designates an amino acid, for example AUG (Adenine-Uracil-Guanine) is the code for methionine. Because DNA contains four nucleotides, the total number of possible codons is 64; hence, there is some redundancy in the genetic code, with some amino acids specified by more than one codon. Genes encoded in DNA are first transcribed into pre-messenger RNA (mRNA) by proteins such as RNA polymerase (as shown in figure 1.1). Most organisms then process the pre-mRNA (also known as a primary transcript) using various forms of Post-transcriptional modification to form the mature mRNA, which is then used as a template for protein synthesis by the ribosome. In prokaryotes the mRNA may either be used as soon as it is produced, or be bound by a ribosome after having moved away from the nucleotide. In contrast, eukaryotes make mRNA in the cell nucleus and then translocate it across the nuclear membrane into the cytoplasm, where protein synthesis then takes place. The rate of protein synthesis is higher in prokaryotes than eukaryotes and can reach up to 20 amino acids per second.

Proteins are complicated systems and each of them can be different from the others in size, shape and function; but all of them display a number of common features. For example, secondary motives known as β-sheets and α-helices, hydrophobic cores, etc. (Broglia and Tiana, 2003).

1.2 The physics of proteins

Life is based on biomolecules (e.g. proteins) which determine how living systems develop and what they do either as a catalyst, regulator, converter and transporter. The knowledge of the structure and function of biomolecules is essential for biology,

biochemistry, biophysics, medicine, pharmacology and as well as technological applications. In biophysics; one of the goals is to describe the physics of biological systems, to discover physical models, and possibly to find new laws that characterize biological entities. Progress in physics has often followed a path in three areas which are essential: structure, energy level and dynamics in which experimental and theoretical approaches are both needed for the progress. This sequence can be combined into two broad groups of "simple" systems such as atoms and nuclei; and "complex" systems, which is sub-divided into two; namely "passive" complex systems such as glasses and spin glasses, and "active" complex systems such as biomolecules (proteins) (Hans, 2010).

The following properties show that proteins are complex many body systems:

1. Structure:

Crystalline solids possess a periodic spatial structure, whereas glasses and protein possess a nonperiodic type. The disorder in glasses is random while that of protein selected by evolution is closer to the disorder in Beethoven's Cross Fuge. Hence, aperiodicity consequently describes the situation in a protein better than disorder.

In solids, glasses, and spin glasses, the strong forces that hold the atoms together are essentially equally strong in all three directions. In proteins, however, the bonds are strong (covalent) along the backbone, but across the side chains are weak (hydrogen bond, Van der Waals forces).

A solid is dead and an individual atom can as a rule; only vibrate around its equilibrium position. In contrast, the weak bonds in a protein can be broken by thermal fluctuations, as a result of this, protein can therefore execute large motions; it can vibrate and acts as a miniature machine (Hans, 2010).

2. Dynamics:

Solids are spatially homogeneous, apart from surface defects and effect. In glasses, inhomogeneities are random and minor. Protein in contrast, is spatially inhomogeneous; properties such as density charge and dipole moment changes from region to region within the protein.

A Solid or glass structure does not change shape on an atomic or molecular scale at a particular point. While a protein can change at any desired place at the molecular level. Through genetic engineering, the primary sequence is modified at the desired

location and this modification leads to the corresponding change in the protein. The motions in a crystal are predominantly elastic, that is atoms vibrate about their equilibrium positions, but the conformation of the crystal remains unchanged. Protein and glasses, however, show both elastic and conformational (plastic) motions, hence their conformation can change (Hans, 2010).

3. Energy landscape:

The surface energy of a crystal as a function of the configuration of its constituents (conformation) is nondegenerate and has a single minimum as shown in figure 1.2. Whereas, the energy surface of a complex system such as glass, spin glass or protein show a very large number of local minima which corresponds to many slightly different conformations that a complex system can assume. The ruggedness of the energy surface of a protein means that there are many energy barriers that have to be crossed during the folding process. This assumed complexity stands as the rationale of why it is difficult to understand how the random-coil conformation of covalently bonded residues spontaneously folds into a unique stable native conformation (Hans, 2010; Andrej et al., 1995).

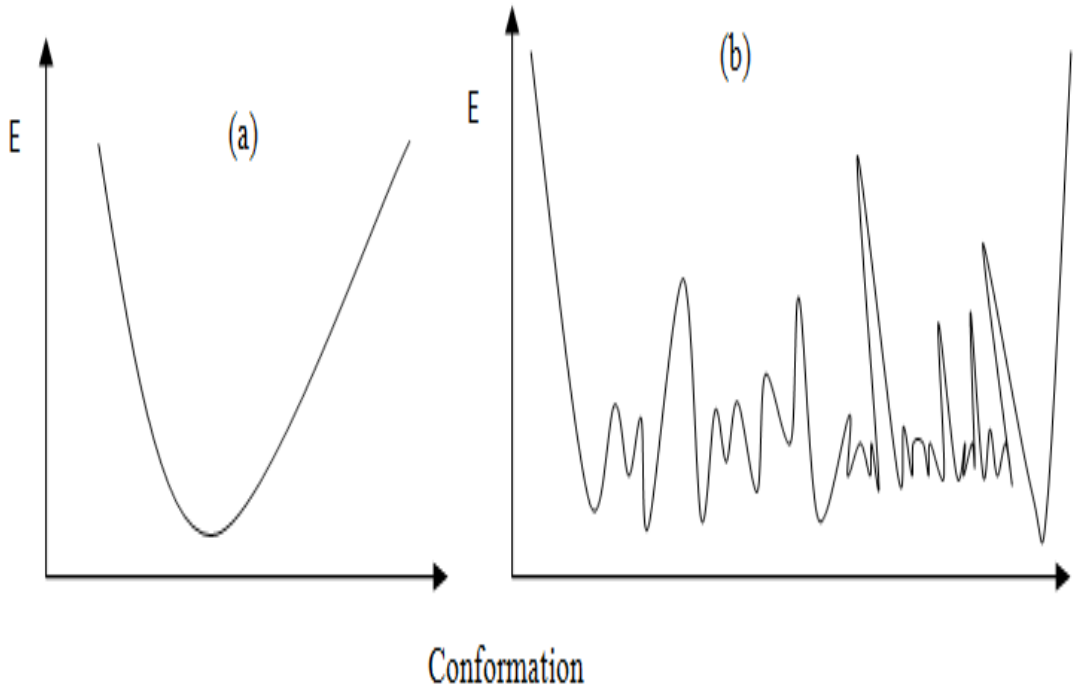


Figure1.2. Energy landscape of (a) a simple system (b) a complex system

1.3 Classes of protein

Proteins can be informally divided into three main classes based on their overall three-dimensional structure and on their functional role, which correlate with typical tertiary structures: globular, fibrous and membrane proteins.

1.3.1 Globular Proteins (GP)

Globular proteins are a diverse class of soluble proteins which have the most varied types of structures; they are soluble in aqueous solution and for this to be achieved they have hydrophobic residues on the core or interior and polar residues on the surface. Their optimal structure is maintained by interactions of the chain with itself and sometimes with various other molecules (co-factor). Many of the most heavily studied proteins belong to this class of proteins. Globular proteins often have stable structures with about 50 - 200 residues and diameter of about 25 – 40 Å; as a result of this more three-dimensional structural information is available for globular proteins than for all other classes of proteins. A good example is haemoglobin and many of them are enzymes (William and Alantha, 2001; Finkelstein and Galzitskaya, 2004) (as seen in figure 1.3).

1.3.2 Fibrous Proteins (FP)

Fibrous proteins are used to construct macroscopic structures, especially those outside of cells and they have structural roles, although some have active functions. They form a water-free aggregates. The structure is highly regular, non-compact and bounded by hydrogen bonds. Examples are: collagen, the most common animal protein and the major component of connective tissue and α - keratin, a protein component of hair, nails and skin (William and Alantha, 2001; Finkelstein and Galzitskaya, 2004).

1.3.3 Membrane Proteins (MP)

Membrane proteins comprise a unique class of proteins which often serves as receptors or provide channels for polar or charged molecules to pass through the cell membrane. They have a hydrophobic region that interacts with the nonpolar interior of membranes, highly hydrogen- bonded and regularly. In this type of protein a

significant region of the protein must be stable in a hydrophobic environment. This is typically achieved by having non-polar side-chains on specific surface regions of the protein. Because of this exposed hydrophobic surface, and because many membrane proteins are destabilized by removal from the membrane, most membrane proteins are much less well understood than those of other classes, and are difficult to work with because they are difficult to purify and study. An example is cytochrome-c-oxidase, which donates electrons to oxygen in the electron transport chain and acts as the primary oxygen-utilization enzyme in aerobic organisms. Its importance stimulated the series of studies that resulted in the solution of its three-dimensional structure (William and Alantha, 2001; Finkelstein and Galzitskaya, 2004; http://en.wikipedia.org/wiki/Protein, retrieved April 20, 2013).

Figure1.3. The molecular surface of several proteins showing their comparative sizes. From left to right are: immunoglobulin G (IgG, an antibody), haemoglobin, insulin (a hormone), adenylate kinase (an enzyme), and glutamine synthetase (an enzyme) (retrieved April 20, 2013, from http://en.wikipedia.org/wiki/Protein)

.

1.4 Statement of problem

The amino acid sequence of each protein determines how it folds into a unique three-dimensional (3D) structure which is the minimal energy state. A protein can then be unfolded or denatured by adding some denaturants like solvent, pH, temperature e.tc. This denaturants change the protein into a flexible chain that has lost its natural shape. When the denaturant is removed, the protein refolds, or renatures into its original conformation. As a result of this we infer that all the information necessary to specify the three-dimensional (native) shape of a protein is contained in its amino acid sequence. The functionality of each protein outright depends on the structure of the protein. Over the years, experimental techniques such as x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy have been veritable tools to determine the 3D structure of a protein. As potent as these techniques are they are very expensive, time-consuming, laborious and restricted to some specific conditions.

Since two decades ago, scientists have devised computational approaches which circumvent the experimental challenges to predict the 3D structure of a protein; yet the native structure of most proteins is still unknown. It has been estimated by scientists that there are at least 100,000 different proteins in the human body (Petsko, 2001). The protein data bank (PDB) is the worldwide repository where information on 3D structures is kept. Figure 1.4 shows the annual and cumulative growth in the number of structures of proteins in the PDB. As of September 2, 2014 PDB contained 103,015 structures and over 80% of them were found by X-ray Crystallography or NMR spectroscopy.

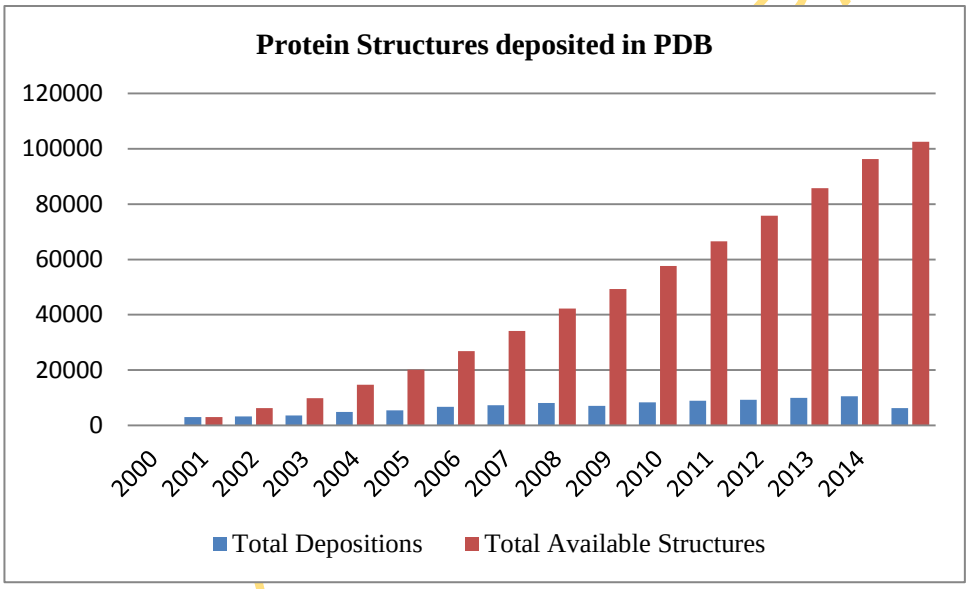


Figure1.4. Protein data bank (PDB) content growth (last updated: 13 Aug. 2014) (retrieved August 18, 2014, from http://www.rcsb.org/)

1.5 Motivation

Knowledge of 3D structure of proteins is crucial to pharmacology, medical sciences and has some technological implications as a result of these:

1. Most drugs work by attaching themselves to a protein, so that they can either stabilize the normally folded structure or disrupt the folding pathway which leads to a harmful protein. Hence, knowing the exact 3D shapes of proteins will help to design drugs.
2. The knowledge of the 3D structures of proteins helps to detect structural differences due to misfolding. Occasionally, the protein may not have the correct 3D shapes- a process known as misfolding- for reasons not yet ascertained. Accumulations of misfolded proteins are known as the cause of some neurodegenerative diseases such as Alzheimer disease, Parkinson, cystic fibrosis, mad cow disease, cancer and inherited form of emphysema.

1.6 Justification

This research work is relevant to the advancement of computational statistical mechanics approach for protein optimisation and is justified by the following reasons:

1. Protein structure prediction and design remains an unsolved problem in protein science,
2. Lattice and other simplified analytical models are the statistical mechanician's contribution to the protein folding
3. The knowledge of 3D protein structures is crucial to pharmacology and medical science, therefore a solution to these problems would have important implications for drug design.
4. The knowledge of 3D protein structures helps to detect structural differences due to misfolding, a process which causes many neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, cystic fibrosis, mad cow disease, an inherited form of emphysema, and even many cancers.

1.7 Aims

1. To develop a computational approach that can help to predict the native structure of proteins based on their amino acid sequences.

2. To use lattice protein model which abstracts from real proteins to obtain the conformation and sequence space of proteins on a level that is computationally amiable.
3. To use lattice protein model which abstract from real protein to learn about the specific physical mechanism required for the formation of unique native (3-D) conformation (structure).

1.8 Objectives

1. To develop a heuristic move-biased Monte Carlo (MBMC) algorithm based on self-avoiding walk for sequence assembly in a lattice protein model.
2. To use (MBMC) algorithm on 2D-square HP protein lattice models to predict the protein conformation.
3. To understand the general physical principles of the folding and mis-folding mechanism of any protein.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction to protein structure and folding

Proteins are heterogeneous macromolecules, consisting of hundreds of several thousands of atoms. Proteins are the workhorse of living organism, executing the genetic code inscribed in the DNA. They are the most basic biological units in a living cell. Basically, all characteristic properties of life are affected by proteins; examples are: conversion of chemical energy to mechanical energy, respiratory systems, photosynthesis, gene expression, genome replication, immune systems and senses. Proteins participate in many different ways in these processes, and the precise task they carry out vary widely: they store and transport molecules, catalyse chemical reactions, transmit information between cells, control the passage of molecules across the cell and organelle membranes, bind to specific sequences of nucleic acids in DNA molecules, and they can simply act as structural building blocks (Erik, 2000; Liu, 2009; Luca, 2005).

All proteins are constructed in the same way, despite their diverse functions. They are linear polymer formed by connecting monomers, these monomers are the 20 naturally occurring amino acids (as shown in figure 2.1). Hence, a protein is specified by its sequence of amino acids. Albeit, the chemical differences of the amino acids partly explain the functional diversity of proteins, the key to protein function is to be found in the three-dimensional structures found by the amino acid chains. How to determine the structure requires a long process, most of them are in fact obtained through X-ray crystallography, while the smaller ones via nuclear magnetic resonance (NMR). In general the latter gives structures at a lower resolution than the first one, which can provide native structures at a resolution lower than 2.0Å. Since 1958, when the first protein structure has been determined by X-ray crystallography (Kendrew et al., 1953), a number of recurrent structures and motifs have been discovered through

experimental approaches but they are more costly, laborious and time consuming. Up till September 2, 2014 more than 103,015 protein structures have been recorded in the protein data bank repository (retrieved April 20, 2013, http://www.rcsb.org/).

2.2 The Amino acids

The 20 naturally occurring amino acids are the building blocks of proteins. They constitute the essential unit of protein tertiary structure analysis. All the 20 amino acids, except proline, have the same general form (as shown in figure 2.1). The amino acids are usually divided into three different classes defined by the chemical nature of the side chain. Class one are those with strictly hydrophobic side chains: Ala (A), Val (V), Leu (L), Ile (I), Phe (F), Pro (P), and Met (M). Class two are the charged residues: Asp (D), Glu (E), Lys (K), and Arg (R). Class three are those with polar side chains: Ser (S), Thr (T), Cys (C), Asn (N), Gln (Q), His (H), Tyr (Y), and Trp (W) (Carl and John, 1999). The amino acids differ in the chemical composition of the side chain R, which vary from one in just one hydrogen atom 1 (glycine) and 18 (arginine) atoms (as shown in table 2.1). Amino acids are bound together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues to form a linear chain called polypeptide chains, the peptide bond constitutes the backbone of the structure. A peptide bond is formed by condensing the carboxyl group of the first amino acid with the amino group of the next and eliminating water in the process.

The side-chain of each amino acid is bound with the backbone C_α atom except proline that has an extra bound to the backbone N. All amino acids except glycine are chiral molecules that can exist in two different forms with different hands L-or D-form (as shown in figure 2.2). The translation machinery for protein synthesis has evolved to utilize only one of the chiral forms of amino acids, the L-form. All amino acids that occur in proteins therefore have the L-form. Although, no obvious reason why the L-form was chosen during evolution (Carl and John, 1999).

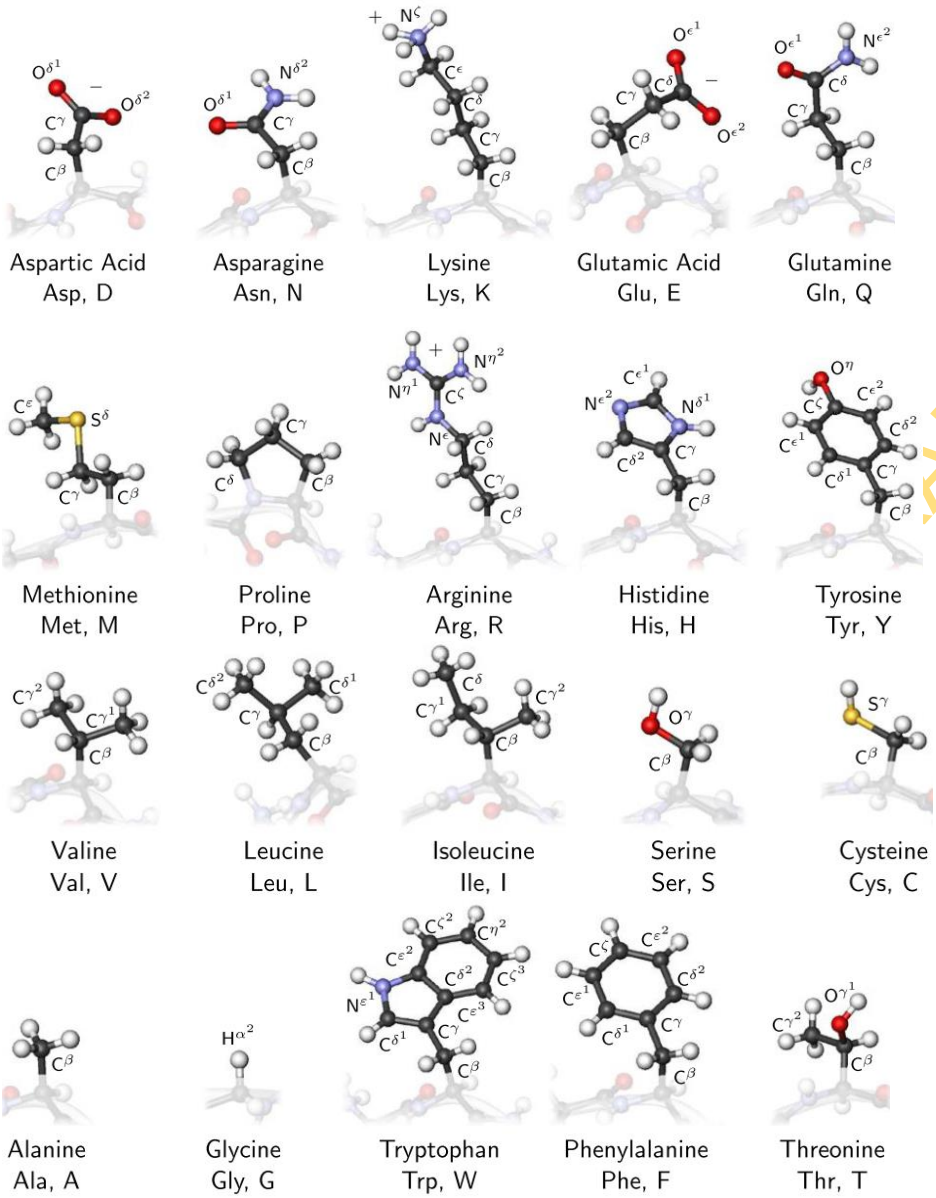


Figure 2.1. The twenty amino acids found in eukaryotes, grouped according to the properties of their side chains (Carl and John, 1999).

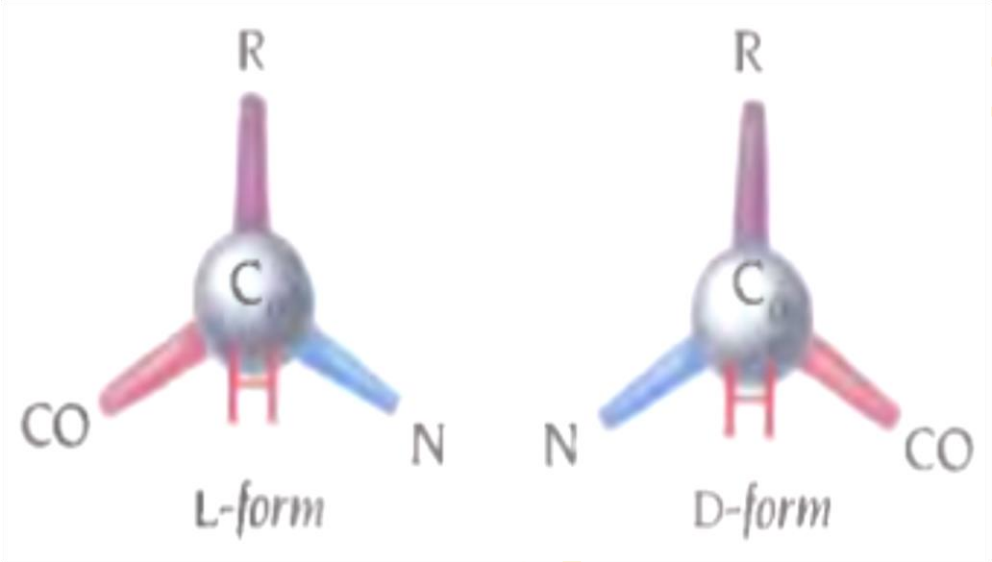


Figure 2.2. The handedness of amino acids (Carl and John, 1999).

The formation of peptide bonds generates a main chain, or backbone, consisting of the common repeating unit, $NH - C_{\alpha} H - CO$, from which the side chains extend. The order of amino acids placed along the chain is of fundamental importance, since changing it may dramatically change the interactions and destabilize the native conformation. The sequence, which is the order according to which the amino acids are placed along the protein backbone, is the first level of complexity; it can be represented by a one-dimensional string, where each letter is associated to one of the twenty types of amino acids (as shown in Table 2.1).

Table 2.1. The 20 amino acids (retrieved on October 14, 2013, from protein data bank http://www.rcsb.org).

1-letter code	3-letter code	Full name	Chemistry	Abundance(%)	H[Kcal/mol]
A	Ala	alanine		8.84	1.81
C	Cys	cysteine		1.24	1.28
D	Asp	aspartic acid		5.39	-8.72
E	Glu	glutamic aci		6.24	-6.81
F	Phe	phenylalanine		4.00	2.98
G	Gly	glycine		7.03	0.94
H	His	histidine		2.20	-4.66
I	Ile	isoleucine		5.95	4.92
K	Lys	lysine		5.2	-5.55
L	Leu	leucine		9.94	4.92
M	Met	methionine		2.37	2.35
N	Asn	asparagines		4.17	-6.64
P	Pro	proline		4.71	-
Q	Gln	glutamine		3.82	5.54
R	Arg	arginine		5.70	-14.92
S	Ser	serine		6.72	-3.40
T	Thr	threonine		5.43	-2.57
V	Val	valine		6.77	4.04
W	Trp	tryptophan		1.21	2.33
Y	Tyr	tyrosine		3.00	-0.14

The primary structure apparently does not contain much information, but one has to associate the structure of every amino acid to each letter in the sequence. By doing so a polymeric chain is obtained, which can assume in principle many different conformations, compatible with steric constraints. One needs to know amino acid sequences and how they bind together to form the peptide chain in order to understand which conformations are allowed and which are not (Erik, 2000; Liu, 2009; Luca, 2005). All amino acids have in common a central atom $\mathbf{C_\alpha}$, to which are attached a hydrogen atom H, an amino group $\mathbf{NH_2}$, a carboxyl group **COOH** and an **R** group (as seen in figure 2.3). All amino acid share the same general structure, the difference of side-chain **R** distinguishes one amino acid from another like the shape, size, polarity and their hydrophobicity which is the most important difference between them i.e, their lack of affinity for water.

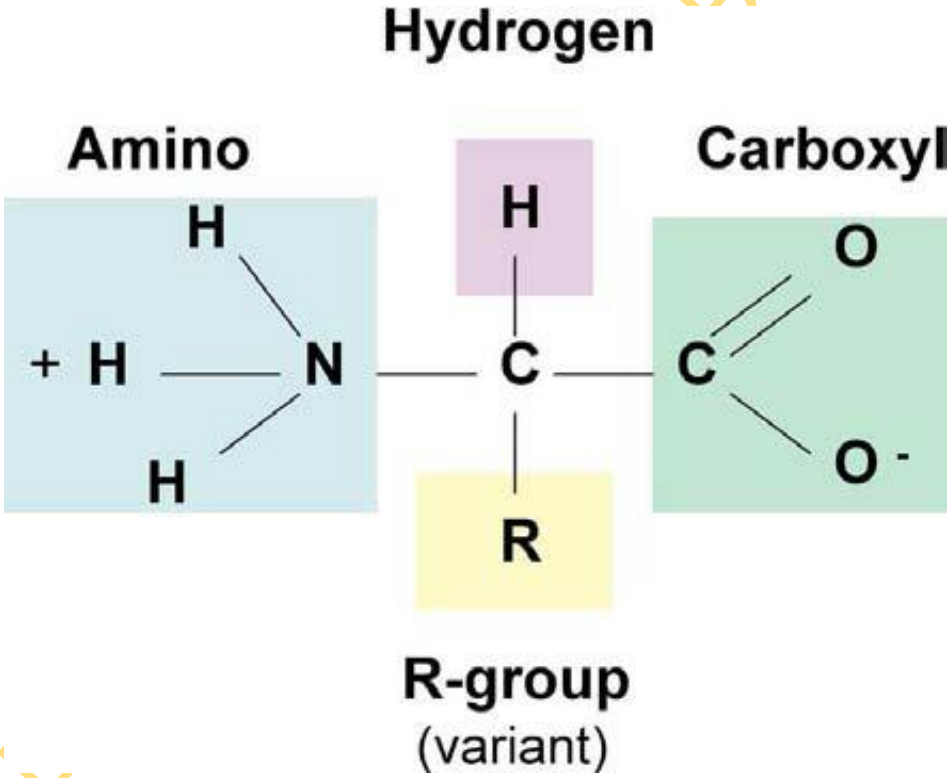


Figure 2.3. Schematic chemical structure of an amino acid (Liu, 2009)

2.3 Peptide bonds

When two amino acids are hydrolysed, the amine group and the carboxylic group of different amino acids form a covalent bond as shown in Figure 2.4. After an amino acid has lost a water molecule, it is called residue, this residue is divided into main chain atoms and side chain, the main chain part which is identical in all residues, contains a central C_α atom attached to an NH group, a COOH group, and an H atom. The side chain R, which is different for different residues, is bound to the C_α atom.

In the translation from DNA sequence into protein, amino acids are joined end-to-end by the formation of peptide bonds, which means the carboxyl group of former amino acid condenses with the amino group of the next one by eliminating a water molecule (as shown in Figure 2.4). Successive formation of peptide bonds generates the main chain, or the backbone of a protein (Liu, 2009).

Because the peptide bond between the carbonyl carbon and the nitrogen has a partial double bond character, rotation around this bond is restricted. Thus the peptide unit can be regarded as a planar rigid structure, with the bond lengths almost fixed. The only remaining freedoms are rotations around the covalent bonds $N-C_\alpha$ and $C_\alpha-C'$. The angle around $N-C_\alpha$ bond is called ϕ; while the angle around $C_\alpha-C'$ bond of the same C_α is called ψ (see figure 2.5). The angle ϕ defines the rotation of the plane containing C_i^α, C_i' and O_i (and N_{i+1}) around the N_i-C_α' bond. This is measured from the plane in which lie C_i^α, N_i and C_{i-1}'. As the peptide bond is generally planar, O_{i-1} and C_{i-1}^α are also in the plane. The angle is defined as 0 when the C_i' is in the same plane as C_i^α, N_i and C_{i-1}', and C_i' and C_{i-1}' are cis. Values of ϕ are positive when measured in a clockwise direction for rotation when viewed down the $N_i-C_i^\alpha$ bond from N to C. The angle ψ defines the rotation of the plane containing C_i', O_i and N_{i+1} around the $C_i^\alpha-C_i$ bond. This is measured from the plane in which lie C_i', C_i^α and N_i (and C_{i-1}'). The angle is defined as 0 when the N_{i+1} is in the same plane as C_i', C_i^α and N_i, and N_i and N_{i+1} are cis. Values of ψ are positive when measured in a clockwise direction for rotation when viewed down the $C_i^\alpha-C_i'$ bond from the C^α (Liu, 2009). In

this way, each amino acid is associated with two conformational angles ϕ and ψ. Since the chemical bonds are almost fixed for all peptide units, the backbone conformation of a protein will be completely determined, if ϕ and ψ angles for each amino acids are defined with high accuracy (Liu, 2009).

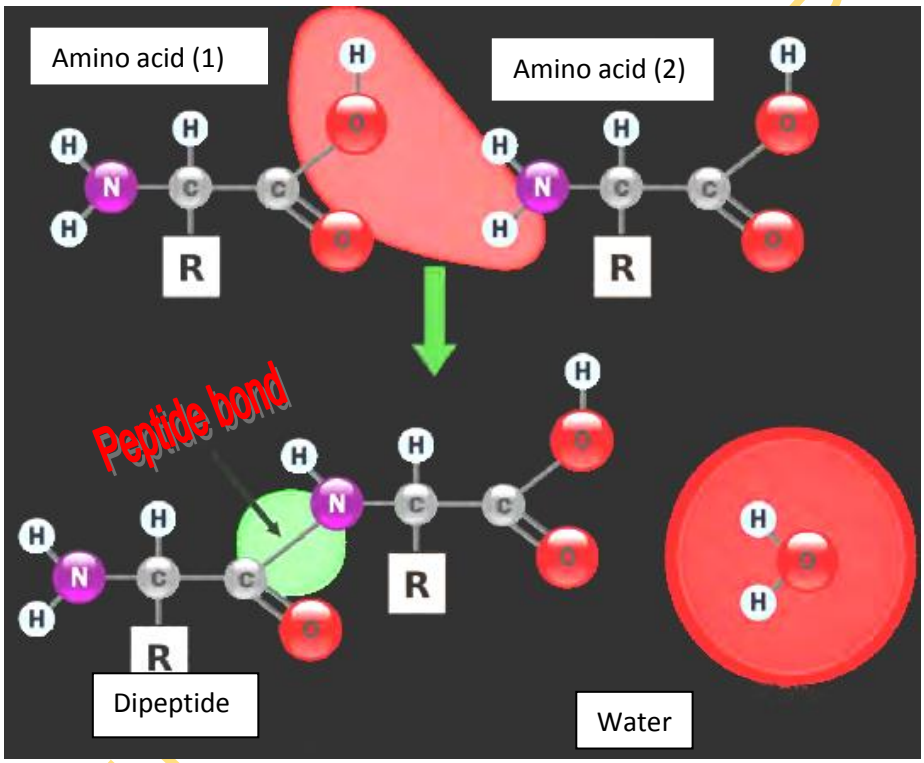


Figure 2.4. Schematic chemical structure of a peptide bond between two residues to form peptide (Martin, 2011).

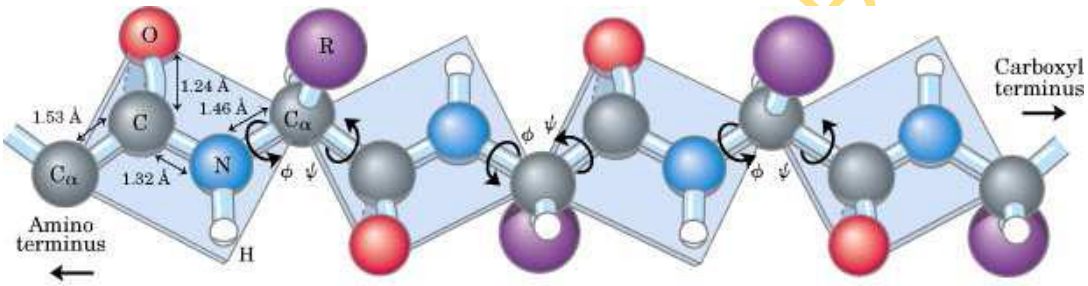


Figure 2.5. Polypeptide chain (Liu, 2009).

2.4 The Ramachandran plot

The bond between the carbon and the nitrogen is called peptide bond and, since it is a partial double-bond, rotations along this axis are forbidden (except rotation of 180^0). On the other hand, rotations, are allowed along the single bonds between C_α and N and between the two carbon atoms, as far as steric clashes do not occur: rotations along these axes are represented by two torsional angles called ϕ and ψ, respectively. Since bonds between nearest neighbouring atoms are not aligned, these rotations cause a conformational change in the polypeptide chain. However, most combinations of ϕ and ψ angles are not allowed, because of steric collisions between the side chains and main chain. G. Ramachandran first made calculations of sterically allowed regions for the angle pairs ϕ and ψ, which is later named as Ramachandran plot (as shown in figure 2.6).

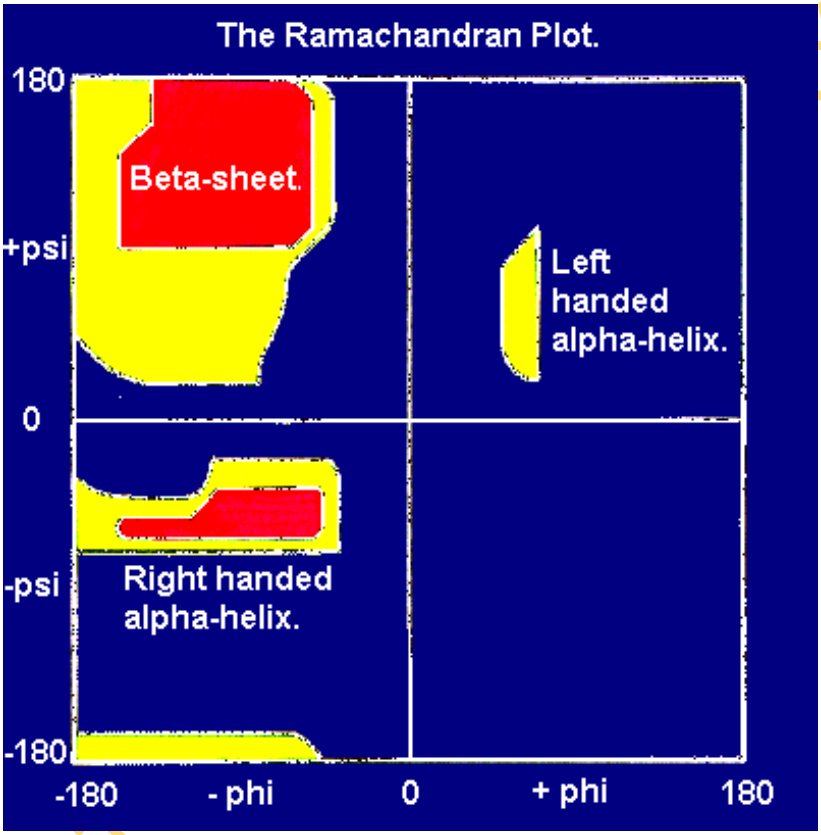


Figure 2.6. Ramachandran plot of a tripeptide, showing sterically forbidden areas for all amino acids except glycine (white) (Ramachandran and Sasisekharan, 1968).

The typical result of such calculation is seen in figure 2.6, the major allowed regions in the figure are the right-handed $\alpha-$helix cluster in the lower left quadrant; the broad region of parallel and antiparallel $\beta-$sheet in the upper left quadrant; and the small, sparsely populated left-handed $\alpha-$helical region in the upper right quadrant (Liu, 2009; Luca, 2005).

2.5 Protein synthesis

Nucleic acids are responsible for information storage and transfer consequently they direct the synthesis of proteins. The nucleic acid is basically divided into two; the DNA and RNA:

1. The DNA (deoxyribonucleic acid) which carries the genetic information.
   (a) DNA is the blueprint for protein creation.
   (b) It has a molecular weight of between 10^7 *and* $10^9 D$.
   (c) The amount of DNA in a system increases with system complexity.
   (d) DNA is a double –stranded linear helical molecule
2. The RNA (ribonucleic acid) which transfers the information and direct protein synthesis.
   (a) It comes in three forms:
      (i) tRNA transfer RNA about 25KD
      (ii) mRNA messenger RNA 100-4000KD
      (iii) rRNA ribosomal RNA 40-1600KD
   (b) tRNA and mRNA are single-stranded
   (c) rRNA exist as a nucleo-protein in the ribosome and acts as the catalytic centre for the protein synthesis
   (d) tRNA transports the correct amino acids to the growing protein chain
   (e) mRNA is involved in the protein synthesis

Nucleic acid contains four nucleotides which serve as the building blocks via DNA and RNA:

DNA: A T G C

RNA: A U G C

Where A is Adenine, T is thymine, U is Uracil, G is guamine and C is cytosine.

The four building blocks are matched in pairs, and the interaction between the partners of a pair is stronger than between two nonpartners. The matched partners are:

DNA	RNA
A=T	A=U
G≡C	G≡C

A and T (or U) are connected by two bonds while G and C are connected by three. The genetic code is a set of three nucleotide sets called codons and the total number of possible codons is 64.

The DNA is located in the nucleus of the cell while synthesis occurs in the ribosome (as shown in figure 2.7), the chromosomes in the cell nucleus contain the DNA, a single chromosome contains a single DNA molecule (Hans, 2010; Hyun-suk, 2006).

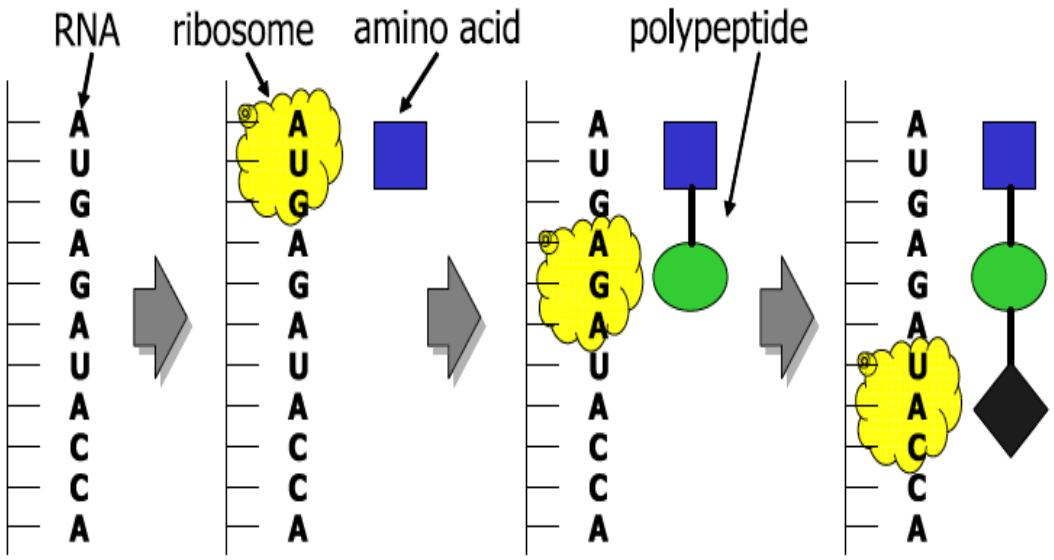


Figure 2.7. Procedure of protein synthesis (Hyun-suk, 2006).

Basically, every vital process within cells of living organisms involves proteins. The cells have sizes ranging from 10^{-7} - 10^{-4} m. In a living body the cell is immersed in an aqueous environment whose pH lies between 7.2 - 7.4. Cells that have nucleus are designated as eukaryotic which include most plants and animal cells. In this type of cell, the nucleus is bound by its own membrane wall, and contains genetic molecules collectively known as chromatin. The other types of cell which is primitive and without a nucleus is called prokaryotic, a good example is the single cell of the bacterium E. Coli, in this case, the chromatin floats free in the cytoplasm. In the human cell, among the chromatin are chromosomes, which are very long and very thin threads of diameter $4\times10^{-9}m$ and length 1.8m. Each chromosome consists of one DNA molecule, festooned at regular intervals with bead-like proteins called histones. Genes are embedded in nucleotide sequences; one gene carries the code for making one type of protein. Protein synthesis involves two stages: transcription of genetic information from DNA to messenger RNA (mRNA) by RNA polymerase and translation from mRNA to proteins. Which gives the central dogma of molecular biology, that is protein synthesis proceeds according to the sequence (as shown in figure 2.8) (Liu, 2009; Zia et al., 2011; Kerson, 2005).

Figure 2.8. Rudiment of protein synthesis. (The central dogma in molecular Biology)

The transcription occurs inside the nucleus, where a specific gene is copied to a mRNA molecule, which is transported through pores in the nuclear membrane to the cytoplasm. There, they are translated into protein molecules by a ribosome, which is made up of a special rRNA and proteins. Protein synthesis demands concerted actions by hundreds of molecules in sequential steps and typically requires a high level of regulation. Its vast demand for the energy needed to complete the reactions also establishes its crucial role in all metabolic pathways. Therefore, developing a quantitative process would be most desirable. Indeed, this task has dominated much of the recent research in molecular biology, as well as mathematics, Physics and emerging cross-disciplinary fields (Zia et al., 2011; Kerson, 2005).

2.5.1 Protein folding *In Vivo* (in the cell)

Many details of the folding process depend on the environment in which folding takes place. When polypeptides are synthesized in the cells, they fold in the cytoplasm after release from the ribosome or in other sub-cellular compartments such as endoplasmic reticula (ER) or mitochondria after they are translated through membranes (Cheolju and Myeong-Hee, 2005).

In a living cell, protein is synthesized by a ribosome that makes a protein chain from amino acids (brought by adaptor tRNAs). There are 20 main natural amino acids; positions of their residues in the protein chain are encoded by mRNA encoded by the gene. The ribosome synthesizes protein chain residue by residue from its N-to C-end and not quite uniformly: there are temporary rests of the synthesis at the "rare" codons (They correspond to tRNAs which are rare in the cell, and these Condons are rare in the cell's mRNAs, too). It is assumed that the pauses may correspond to the boundaries of structural domains that can help a quick maturation of the domain structures. The biosynthesis takes about a minutes and yielding of a "ready" folded protein lasts as long the experiment does not see any difference. Some enzymes, like prolyl-peptide or disulfide-isomerases accelerate *in vivo* folding (Finkelstein and Galzitskaya, 2004).

Within the cells, proteins in the process of synthesis encounter particular challenges imposed by the crowded macromolecules before completion of folding. As incompletely folded chains expose particular regions that are destined to be buried in

the native state, they are prone to aggregate with other molecules because they have exposed hydrophobic surfaces.

Consequently, elaborate systems have evolved to prevent proteins from being aggregated prior to folding. The first one is molecular chaperones, and the second one is ubiquitin-proteasome system, each of which is not exclusively independent of the other, and in some way they cooperate in living cells. Protein chains fold under protection of special proteins, chaperons. Chaperons are the cell's troubleshooters that fight the aggregation of nascent proteins, since in a cell, folding takes place in a highly crowded molecular environment. There is no reason to assume, though, that anything other than the amino acid sequence determines protein conformation in the cell. Chaperones are members of diverse protein families capable of binding to stabilize nonnative conformations of other proteins. They bind to the folding intermediate of polypeptides, which prevents aggregation of the intermediates and facilitates correct folding and assembly through controlled binding and release cycles. Chaperones are found in all types of cells from archaea to eukarya and various cellular compartments of the eukaryotic cell. In addition, their concentrations are increased as a response to diverse stresses such as unfolded protein response as well as the heat shock, which explains why a large number of chaperones are heat shock proteins. Overview on Hsp60s and Hsp70s, the two most studied chaperones, will provide us an insight into the mechanism of those folding assistants (Cheolju and Myeong-Hee, 2005).

In-vivo folding of large proteins usually needs the help of chaperon, which can greatly reduce the energy barrier between the misfolded state and the native state, as well as prevent the aggregation of folding intermediate (Liu, 2009).

It looks as though the protein biosynthetic machinery (ribosomes + chaperons), besides chemical synthesis of the protein chain, serves only as a kind of incubator, which does not determine the protein structure (at least if the protein is not very large) but rather provides "hothouse" conditions for its maturation, just like a usual incubator helps a nestling to develop but does not determine what will be developed, a chicken or duckling. Unfortunately, it is difficult to follow the *in vivo* folding of a nascent protein chain against the background of the huge ribosome and the other constituents of the cell (Finkelstein and Galzitskaya, 2004).

Chaperonins are a group of chaperones with a molecular weight of about 60 kDa. Members include bacterial GroEL, Hsp60 of mitochondria and chloroplasts, and the TRiC in eukaryotic cytosol. They are characterized by the barrel-shaped double-ring structure; GroEL seems to be the best studied chaperoning with regard to folding mechanism. GroEL works with GroES, a co-factor of Hsp10 family. Inside of the ring structure of GroEL, a central cavity is formed, in which an incompletely folded polypeptide is sequestered via hydrophobic interactions. The first role of GroEL is providing a protected environment to prevent the folding intermediate from sticking to one another. TRiC constitutes a different subgroup of chaperonin, because it functions independent of Hsp10 cofactor. TRiC cooperates with different upstream chaperones in the folding of distinct protein classes (Cheolju and Myeong-Hee, 2005).

2.5.2 Protein folding *In Vitro* (in the test-tube)

In about 1960, a remarkable discovery was made: it was shown that a globular protein is capable of spontaneous folding in vitro (Finkelstein and Galzitskaya, 2004). This means the following: If the protein chain has not been heavily chemically modified after the initial (*in vivo*) folding, then the protein gently (without chemical damaging) unfolded by temperature, denaturant, etc, spontaneously "renatures", that is, restores its activity and structure after solvent "normalization". It was demonstrated that the protein chain synthesized chemically, without any cell or ribosome, and placed in the proper ambient conditions, folds into a biologically active protein. The phenomenon of spontaneous folding of protein into native structures allows one to detach, at least to a first approximation, the study of protein folding physics from the study of protein biosynthesis. Protein folding *in vitro* is the most simple (and therefore, the most interesting for a physicist) case of pure self-organization: here nothing "biological" (but for the sequence!) helps protein chain to fold (Finkelstein and Galzitskaya, 2004).

2.6 The levels of protein structure

A protein is a long-chain molecule consisting of a backbone made up of amino acids connected sequentially via a peptide bond and the chain is called a polypeptide chain. To get a better insight to the three-dimensional conformation of protein, four distinct levels of protein structure are observed (as shown in figure 2.9), they are:

2.6.1 Primary Structure (PS)

The primary structure refers to the amino acid linear sequence of the polypeptide chain the number of which ranges from the order of 50 to 3000 chosen from a list of 20 naturally occurring amino acids. The structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. The two ends of the polypeptide chain are referred to as the carboxyl terminus (C-terminus) and the amino terminus (N-terminus) based on the nature of the free group on each extremity. Counting of residues always starts at the N-terminal end (NH_2-group), which is the end where the amino group is not involved in a peptide bond. The primary structure of a protein is determined by the gene corresponding to the protein (see figure 2.9) (Liu, 2009; Carl and John, 1999; Kerson, 2005).

2.6.2 Secondary Structure (SS)

Secondary structure refers to highly regular local sub-structures, which contain three main elements the alpha helix ($\alpha-$ helix), the beta strand or beta sheets ($\beta-$ sheet) and turns. These were suggested in 1951 by Linus Pauling and coworkers. These elements may be connected with each other by loops. In globular proteins, Helices are the most abundant form of secondary structure, followed by sheets and in the third place turns. These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups and provide an efficient mechanism of pairing polar groups of the polypeptide backbone by hydrogen bonds. They have a regular geometry, being constrained to specific values of the dihedral angles ψ and ϕ on the Ramachandran plot. Both the alpha helix and the beta-sheet represent a way of saturating all the hydrogen bond donors and acceptors in the peptide backbone. SS is often the first step in protein folding (see figure 2.9) (Liu, 2009; Carl and John, 1999; Kerson, 2005).

2.6.3 Tertiary Structure (TS)

Tertiary structure refers to the completely folded and compacted polypeptide chain i.e to three-dimensional structure of a single protein molecule, obtained from the secondary structures. The alpha-helices and beta-sheets are folded into a compact

globule. The folding is driven by the non-specific hydrophobic interactions (the burial of hydrophobic residues from water), but the structure is stable only when the parts of a protein domain are locked into place by specific tertiary interactions, such as salt bridges, hydrogen bonds, and the tight packing of side chains and disulfide bonds (see figure 2.9) (Liu, 2009; Carl and John, 1999; Kerson, 2005).

2.6.4 Quaternary Structure (QS)

Quaternary structure is the three-dimensional structure of the conglomeration of several protein chains into one multi-subunit of protein complexes. The Complex of several protein molecules or polypeptide chains, usually called protein subunits in this context, functions as part of the larger assembly or protein complex (Liu, 2009; Luca, 2005). In this context, the quaternary structure is stabilized by the same non-covalent interactions and disulfide bonds as the tertiary structure. Complexes of two or more polypeptides (i.e. multiple subunits) are called multimers. Specifically, it would be called a di-mer if it contains two subunits, a tri-mer if it contains three subunits and a tetra-mer if it contains four subunits (see figure 2.9) (Liu, 2009; Carl and John, 1999; Kerson, 2005).

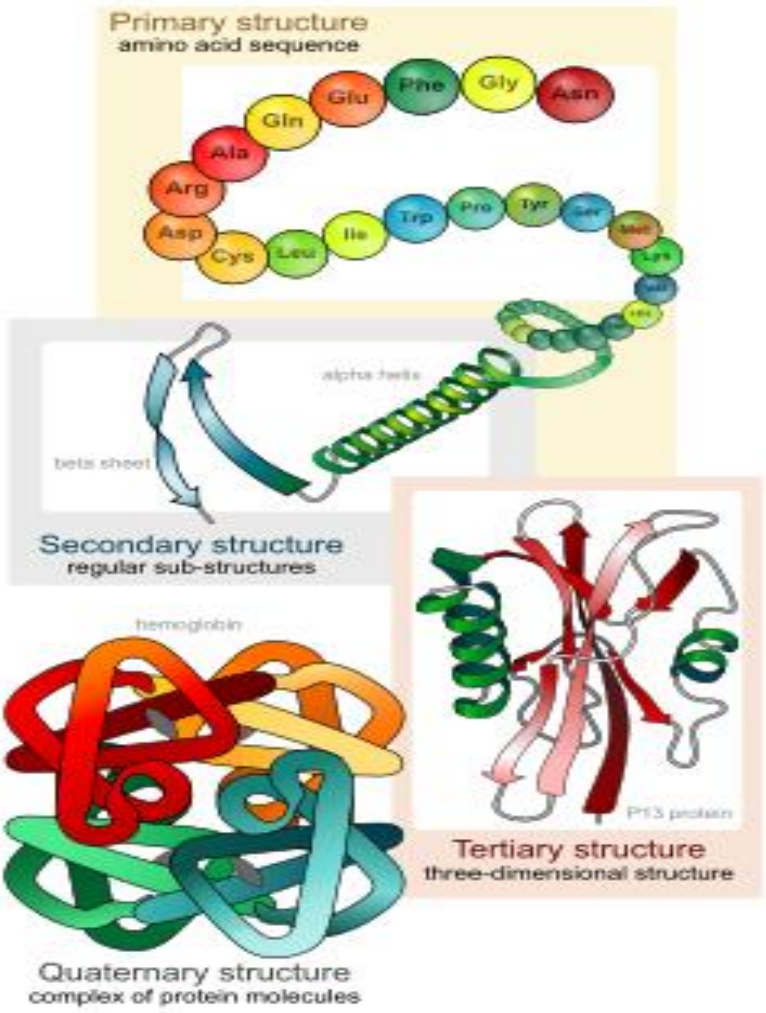


Figure 2.9. Levels in protein structure, from primary to quaternary structure (retrieved on April 20, 2013, from http://en.wikipedia.org/w/index.php?title=Protein_structure&oldid=551224293)

2.7 The native state of protein

The native state is realized when a protein molecule assumes very complex, but unique 3D structures specific to their amino acid sequence under the physiological condition. A protein molecule is a copolymer of twenty types of amino acid residues which are arranged linearly in specific sequence. When a protein molecule is placed under conditions different from the physiological one, the linear polymer may assume unfolded random structures (protein in the denatured state), or it may be folded into non-specific globular form (the molten globular state). Thus, protein molecules are known to exist typically in the three different states depending on their environmental condition. The Transition between these states is a phenomenon similar to a phase transition.

In 1961, Christian Anfinsen has shown by a conceptually simple experiment that the native state of the protein is realized as a physical equilibrium state. This simple fact has a profound biological implication and Anfinsen received a Nobel Prize (1972) for it. For this reason his finding is often referred to as Anfinsen dogma.

Protein molecules perform their biological function after they assume the specific native state (3D) structure. Since the native state structure is in the physical equilibrium state, according to Anfinsen, the state is independent of its past history. Whatever occurred to the molecule in its history does not have any influence to it in equilibrium. The state of equilibrium is completely determined by its amino acid sequence and the environmental condition. In order words the 3D structure under the physiological condition is uniquely determined by the amino acid sequence alone. Translation from amino acid sequence to 3D structure follows the law of physics completely. Therefore prediction of 3D structure from amino acid sequence should be possible in principle by using the law of Physics (Nobuhiro, 2007).

2.8 Protein Folding Problem (PFP)

One of the most important and challenging problems in Molecular Biology/Physics is to find out the native (3D) structure of a protein. This can in principle be done by experimental methods like X-ray crystallography or NMR spectroscopic analysis, but these methods are very slow and expensive. Thus, the goal is to find a computational

method for predicting the native tertiary structure of a protein, given the linear sequence of the amino acid residues. This task is called the protein folding problem. Any possible folding of a protein in the 3D space is called a conformation, and the native structure is believed to be the conformation with the lowest free energy (Hans-Joachim and Dirk, 2007). PFP is a problem of great scientific interest. How a protein spontaneously forms a well-defined biological active structure is fascinating and is still an open question in protein science.

PFP is often approached by statistical methods which, however, have their limitations. The more fundamental Physical-chemical approach is restricted by computational as well as conceptual difficulties, but important insight has been gained by studying coarse-grained (CG) models which are not meant for specific structure predictions, but rather to elucidate generic physical principles of protein folding (Erik, 2000). The problem of spontaneous folding of protein amino acid chains into compact, highly organized 3D structures continue to challenge the modern science (Ginka et al., 2011). The protein folding research has two primary objectives; the first is to be able to predict the 3D protein structure from its amino acid sequence. Protein amino acid sequences are encoded in the genes, but protein structures are key to understanding the mechanisms that control their ultimate biological functionality. Protein folding is therefore the final step in the translation of the genetic information to biological functions. As the wealth of amino acid sequence information obtained by rapid new gene sequencing methods continues to severally outpace the experimental determination of protein structures, the significance of reliable structure prediction methods is enormous.

The second main goal is to understand the mechanism of protein 3D structure formation. The protein folding mechanism may not seem as directly, biologically relevant as folding occurs spontaneously and the biological function is tied predominantly to the folded structure. However, in many cases protein folding or unfolding is an integral part of biochemical function and of other bio-cellular processes, such as translocation and degradation. Defects in folding can lead to several disorders, illness, even death (Ginka et al., 2011).

The protein folding problem can be viewed currently from three major ways:

1. The Molecular Biologist and Physical Chemists use modern experimental devices, to try to determine proteins' native structures as well as folding intermediates.
2. The Computer Scientists use programming languages with the goal of simulating the folding process efficiently.
3. Also, the Mathematician and Physicist put the folding problem in the framework of classical statistical Physics or condensed matter physics.

2.8.1 Protein folding and design

Two of the most investigated protein problems in molecular biology are protein folding and design. Both problems stem from Anfinsen's discovery (Anfinsen, 1973) that the sequence of amino acids of a naturally occurring protein uniquely specifies its thermodynamically stable native structure. The protein folding challenge consists of predicting the native state of a protein from its sequence of amino acids, while in protein design one is concerned with identifying the amino acid sequences folding into a pre-assigned native conformation. The protein design problem asks which and how many amino acid sequences fold into a given native structure. This last issue, having obvious practical and evolutionary significance, has attracted considerable attention and effort from experimentalist and theorist (Anderea et al., 2001). Folding is an essential process for proteins to acquire their biological functionality. The proper folding of proteins is critically important in cellular activities and involves substantial interesting physics and complexity.

Protein folding is the spontaneous process of assembling a poly-peptide chain into a distinct three-dimensional structure (as shown in figure 2.10). To carry out their functions, proteins must fold rapidly and reliably. They must satisfy a kinetic requirement that folding can be completed within a reasonable time and a thermodynamic requirement that the folded conformation be stable under physiological conditions (Martin, 2011). Knowledge of the details of this reaction lies at the heart of understanding some of the basic mechanisms of life, as the final conformation is normally the unique biologically active conformation (Bruno and Valerie, 2013). According to the hierarchical folding theory of Baldwin and Rose (1999), a protein folds by first forming local structural elements, namely, $\alpha-$helices and $\beta-$strands

these secondary structural elements, then interact with each other, resulting in the formation of the folded protein (Nikolas et al., 2012).

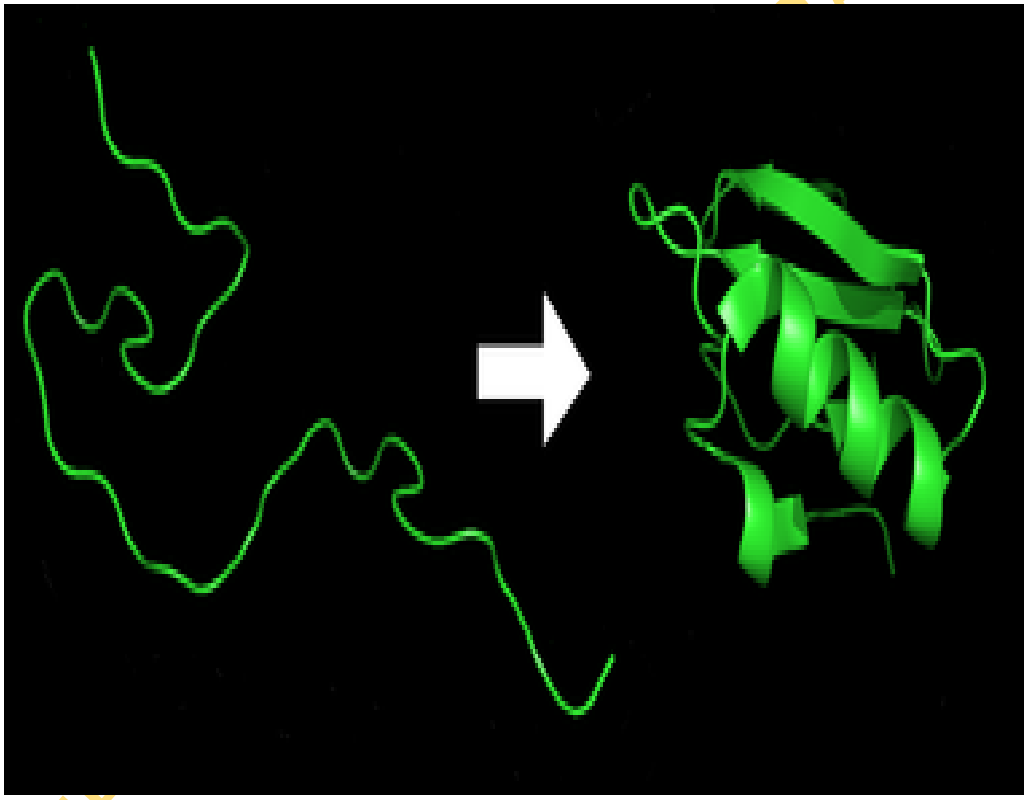


Figure 2.10. Protein before and after folding (retrieved April 20, 2013, from http://en.wikipedia.org/wiki/Protein_folding)

To understand the mechanism of protein folding, experimentalists and theoreticians have focused their efforts on analysing small, single-domain proteins, (less than ~110 amino acids) over the past decade. Experiments showed that many of these proteins folds with simple two–state kinetics in vitro. In addition, these simple proteins not only spontaneously fold to a unique structure, but can also do so amazingly quickly than that by randomly exploring all possible conformations of its unfolded state (Guo et al., 2010). According to Christodoulos Floudas**,** even a pentapeptide, a protein with just five amino acids, could fold in any of 100 billion possible structures. Amino acid sequence tells us little about what the protein does and how it does it. In order to carry out their function (e.g. as enzymes or antibodies), they must take on a particular shape, also known as"fold". Thus, proteins are truly amazing machines: before they do their work, they assemble themselves! This self-assembly is called "folding". Understanding protein folding in detail is a cornerstone for the successful design of protein inhibitors and other therapeutic proteins (Olav and Ulrich, 2008).

In the higher organism proteins are synthesized in the cytoplasm through a complex mechanism of biosynthesis. Once the sequence is synthesized the protein is not yet active. To become biologically active it has to fold into a unique 3D conformation characteristic of each protein (Luca, 2005; Sooyoung and Faming, 2011). This involves a complex molecular recognition phenomenon that depends on the cooperative action of many nonbonded interactions. As the number of possible conformations for a polypeptide chain is astronomically large, a systematic search to find the native (lowest energy) structure would require an almost infinite length of time. This is known as the "Levithal paradox" (Moret, 2011; Levinthal, 1968). Cell's life often relies on the ability of its constituent proteins to fold correctly into the three dimensional structures that are crucial for their function. In principle, the amount of folded functional protein in a cell may be dependent on several factors in addition to thermodynamic stability measure *in vitro*, for example, the rate of protein biosynthesis and degradation (Carlo et al., 2009).

Characterizing the folding and unfolding kinetics of proteins is important for understanding the energetic landscape, leading to the active native conformations of these molecules. This folding process is reversible for many proteins by changing for examples temperature or pH, the protein will unfold, but every time natural conditions

are restored it will refold into its native structure. The reversal of the folding process implies that the native structure of a protein is encoded in its amino acid sequence (Erik, 2000; Liu, 2009).

The reversible nature of the typical two-state conformational equilibrium between the folded and unfolded states of monomeric, single-domain proteins makes it possible to use classical methods, such as stopped-flow spectrofluorimetry, to characterize *in vitro* the folding kinetics of proteins by initiating folding reactions from their unfolded conformations. However, thermal or chemical-induced unfolding of many proteins, such as cystine kinase, acetylcholine Sestaerases, $\alpha-$amylases and bacillus circulans xylanase (BCX) can be irreversible under *in vitro* experimental conditions. This precludes characterization of the folding kinetics of such proteins, just as it's impossible to "un-boil" an egg. Such irreversible denaturation of proteins often result from the subsequent aggregation of the unfolded proteins in question, which typically occurs between (partially) denatured molecules in response to the exposure of their constituent hydrophobic residues to solvent (Liu, 2009; Ashlee and Hongbin, 2010). However, under some other conditions, this protein folding may also go into wrong pathways and lead to some misfolded structures. When an egg is boiled, the proteins in the white unfold and misfold into a solid mass, which will not refold or redissolve. In a similar way, irreversibly misfolded proteins, which form insoluble aggregation mass, have been found in certain tissues (Prusiner, 1991). They often are characteristic of some well known diseases (Liu, 2009; Pain, 2000).

2.9 Protein folding intermediate and aggregation

Historically, intermediates were viewed as essential stepping-stones that guide a protein through the folding process to the native state. It is a critical species in misfolding processes that lead to aggregation and diseases (Maksym and Laura, 2013). One commonly observed type of intermediate was the so-called 'molten globule', i.e. a state possessing native-like secondary structure elements, but lacking the tight packed tertiary structure of the native state. In order to populate an intermediate sufficiently to allow detailed analysis, extreme conditions such as low pH or co-solvent were used.

Computational approaches, such as lattice model, Coarse-grained models and atomistic simulation of ref. (9-11) reviewed by (Maksym and Laura, 2013) provided tremendous

insights into the physical principles underlying protein folding helps to explain why stable intermediates are not essential to the folding reaction.

Many diseases arise through loss of function caused by mutations that disrupt a binding site or active site of a protein or through elevated expression levels leading to increased activity of a protein. However, mutations can alternatively cause disease by destabilizing the native structure or stabilizing non-native conformations, resulting in the population of partly folded intermediates and leading to the side reaction of aggregation. Intermediate states tend to be aggregation prone because they expose sticky interfaces that are normally buried in the native states. The cooperative nature of the folding mechanisms of small, single–domain proteins protects them from misfolding and aggregation; larger proteins, whose folding is less cooperative, are therefore more at risk. Proteins need to be kinetically as well as thermodynamically resistant to local unfolding that exposes aggregation-prone regions (Maksym and Laura, 2013).

Three notable examples, for which the relationship between the folding and the aggregation energy landscapes has been characterized in detail, are lysozyme, transthyretin and $\beta 2-$microglobulin as reviewed in ref. (186-188) of (Maksym and Laura, 2013). For all three proteins, partly folded intermediates have been identified that are critical for aggregation. Other illustrations of the competition between the conversion of partly folded intermediates into the native state and into aggregation include tailspike protein, luciferace, stefin, ure2, prion protein, superoxide dismutase, serpin, $\gamma-$crystalline, and SH3. Although many of these proteins form amyloid-like aggregates that are rich in $\beta-$sheet structure, aggregate containing substantial native structure have also been observed. One mechanism by which native-like aggregate can form is 3D domain swapping, and aggregates of several disease-associated proteins have been shown to assemble via domain-swapped intermediates. A striking recurring feature of intermolecular assembly leading to both domain swapping and aggregation is proline residues (Maksym and Laura, 2013).

2.10 Protein misfolding and conformational diseases

Protein misfolding gives rise to the malfunctioning of living systems. When proteins misfold, they can clump together ("aggregate"). These clumps can often gather in the brain, where they are believed to cause the symptoms of the growing number of age-related diseases, including Alzheimer's and Parkinson's disease as well as other neurodegenerative disorders. With all the things proteins do to keep our bodies functioning and healthy, they can be involved in disease in many different ways. The more we know about how certain proteins fold the better new proteins we can design to combat the disease-related proteins and cure the diseases. Of all the ways that proteins can go bad, becoming an amyloid is surely one of the worst. In this state, sticky elements within proteins emerge and seed the growth of sometimes deadly fibrils. By the 1980s, researchers had come to understand that these artificially induced fibrils had the same peculiar structure seen in disease-linked amyloid, such as the amyloid-β deposits in the brains of people with Alzheimer's disease. Some mutations and toxins, and the cellular wear and tear associated with ageing, can result in proteins that are less well folded and less protected by chaperoning and disposal mechanisms and thus more liable to become amyloids (Jim, 2010).

In work reported in February, 2010 a team led by David Eisenberg at the University of California, Los Angeles, sifted through tens of thousands of proteins looking for segments with the peculiar stickness needed to form amyloid. According to Eisenberg, they found that "effectively all complex proteins have these short segments that, if exposed and flexible enough, are capable of triggering amyloid formation". They believed that not all proteins form amyloids. However, the amyloids are restricted because most proteins hide these sticky segments out of harm's way or otherwise keep their stickness under control (Jim, 2010).

They found that 95% of the predicted amyloid-prone segments within them are buried within the structures of their host proteins, and that those that are exposed are too twisted and inflexible to zip up with partner segments. It seems that most proteins have evolved to fold in a way that effectively conceals their amyloid-prone segments.

According to Chris Dobson, a structural biologist at the University of Cambridge, "The amyloid state is more like the default state of a protein and in the absence of specific

protective mechanisms; many of our proteins could fall into it". Amyloids have been found in some of the most common age-related disease, and there is evidence that ageing itself makes some amyloid accumulation inevitable (Jim, 2010).

Table 2.2. Some human conformational diseases caused by protein deposits

Disease	Disease protein	Site of Folding	Characteristic pathology
Alzheimer's disease	Amyloid $\beta-protein$ / presenilin	ER	Extracellular plaques; tangles in neuronal cytoplasm
Huntington's disease	Long glutamine stretches within certain proteins	Cytosol	Intranuclear inclusions and cytoplasmic aggregates
Parkinson's disease	$\alpha-synuclein$	Cytosol	Lovy body formation
Scrapie/Creutzfeldt-Jacob disease	Prion Protein	ER	Spongiform degeneration; extracellular plaques; amyloid inside and outside neurons
Sickle cells anaemia	haemoglobin	Cytosol	
$\alpha_1-antitryp\,\sin$	$\alpha_1-antitryp\,\sin$	ER	Inclusions in hepatocytes leading to emphysema, liver cirrhosis
Cystic fibrosis	cystic fibrosis trans-membrane regulator	ER	
Cancer	p53	Cytosol	Result of cell damage
Familial amyloidoses	transthyretin/lysozyme	ER	
Osteogenesis imperfect	procollagen	ER	
Scurvy	Collagen	ER	Teeth damage
Cataracts	Crystallins	Cytosol	Irritation of eye and disorderliness

ER means Endoplasmic Reticulum

Diverse diseases have been shown to arise from protein misfolding and can be grouped as conformational diseases (as shown in table 2.2). Typical examples are serpinopathies such as α1- antitrypsin deficiency leading to emphysema or liver cirrhosis and familial encephalopathy with neuroserpin inclusion bodies, and various neurodegenerative disorders such as Alzheimer's disease, Huntington's disease, Parkinson's disease and the prion diseases and many cancers or cancer related syndromes (Cheolju and Myeong-Hee, 2005).

In these disorders, specific peptides or proteins misfold, often as a result of mutations, and give rise to protein aggregates. In serpinopathies, serpin molecules such as α1-antitrypsin and neuroserpin form loopsheet polymers. In Alzheimer's disease, extracellular amyloid- β peptide deposition is thought to be intimately associated with the disease (as shown in figure 2.11 and 2.12). Several genetic loci related to Parkinson's disease were found and one of them is α-synuclein which forms intracellular aggregates. Huntington's disease is caused by a mutant version of the protein huntingtin. It has a longer expansion of amino-terminal polyglutamine domain and thus is more prone to aggregation. Prion diseases, such as Creutzfeldt-Jakob disease, are caused by deposition of prion protein aggregates in the brain and nervous system. Some of the diseases result from loss-of function. In the case of α1-antitrypsin deficiency, misfolded α1-antitrypsin is retained in hepatocyte and secretion to blood plasma is blocked. The lack of circulating α1- antitrypsin induces an imbalance between the anti-proteolytic activity and elastase, which causes the onset of emphysema due to failure to protect elastic tissues of the lung from proteolysis by elastase. In other cases like the neurodegenerative diseases, misfolded proteins escape the protective mechanisms and form intractable aggregates within cells or in the extracellular space. It is thought that either the aggregates of the disease protein themselves or the work done by the aggregates, or the process of their formation confers cellular toxicity. This idea supports the notion that misfolded disease proteins act through a gain-of-function (Cheolju and Myeong-Hee, 2005).

Since so many diseases are related to the misfolding of proteins, if we want to cure them, we have to know the mechanisms involved in the folding process. Besides this, the industry of medicine, food, environment and energy etc, are all looking forward to the knowledge of protein folding eagerly (Liu, 2009).

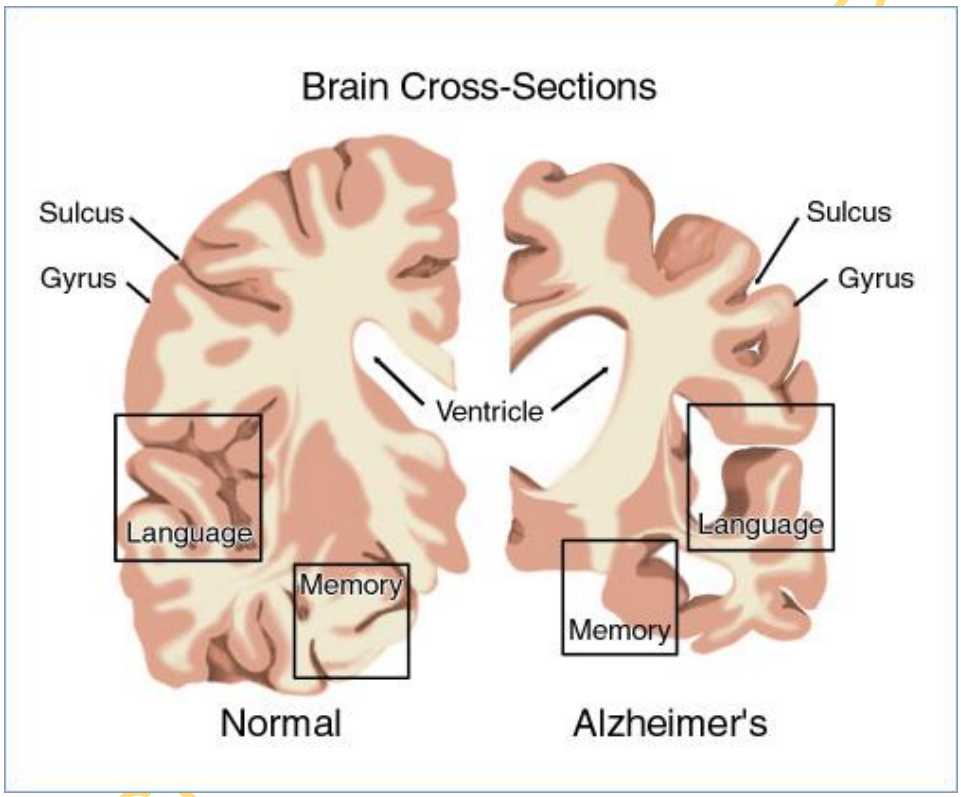


Figure 2.11. Cross sections of normal and Alzheimer's disease (AD) brain showing the dramatic atrophy in regions responsible for memory and language skills. (Source: Eckhard Mandelkow, Max Planck research group, Hamburg).

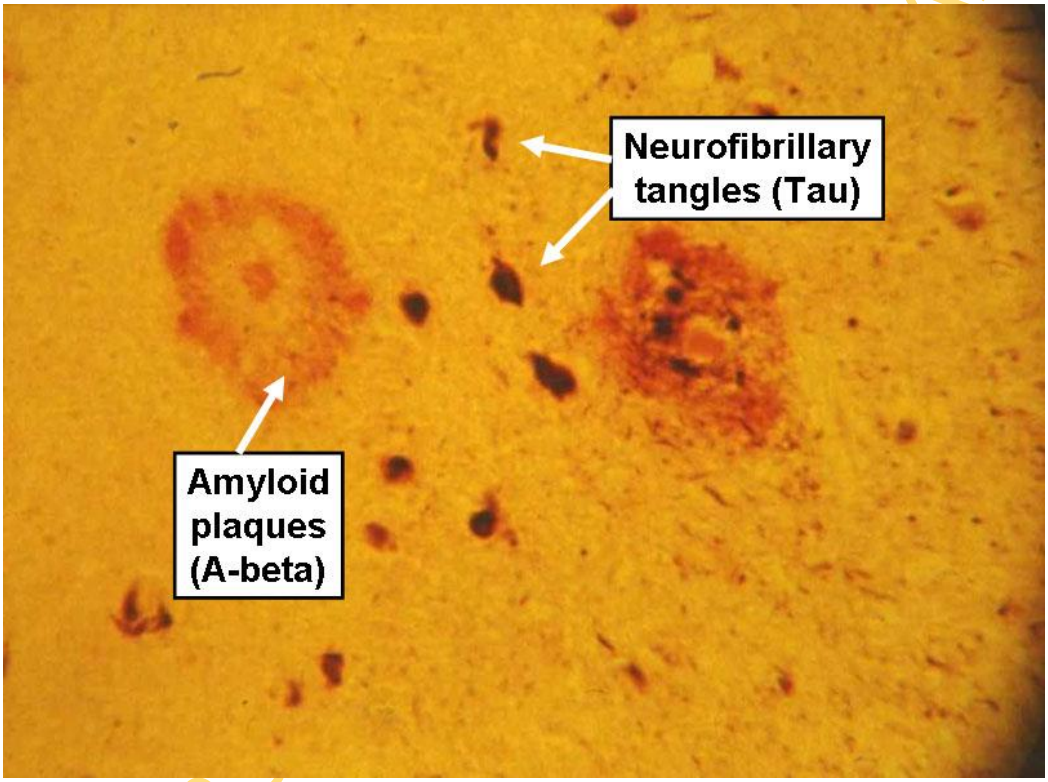


Figure 2.12. Microscopic image of brain tissue from an Alzheimer disease (AD) patient showing the typical AD deposits called plaques and tangles. (Source: Eckhard Mandelkow, Max Planck research group, Hamburg).

Chunmei et al., observed the folding co-operativity, which described how sharp and smooth the folding transition occurs, and how essential for proteins to fold quickly, effectively and to avoid misfolding and aggregation (Chunmei et al., 2012). With the fulfilment of the Human Genome project, the proteome project comes into a full-fledge stage. All these make the protein folding problem stay in an essential position in contemporary science (Liu, 2009). Yi Cao et al., observed that (>30%) proteins in living cells required cofactors, for electron transfer, metal ion transport and storage, to enzymatic catalysis. To reach their functional states, these proteins need to fold into their unique three-dimensional structure in the presence of cofactors, intermingle with each other, raising questions about the interplay between these two processes and whether the binding of cofactors occurs before or after the protein has acquired its three-dimensional structure. Addressing this question is not only of fundamental importance for understanding the roles of cofactors in the functions of these proteins, but, may also offer new insight into understanding of the protein folding problem (Yi and Hongbin, 2011).

2.11 The time scale in protein folding

Surprisingly protein folding time scales covers an extremely broad dynamic range: microseconds to milliseconds for small helical proteins, milliseconds to seconds for largely $\beta-$sheet proteins, minutes for proline cis/trans isomerization, minutes to hours for repairing misfolded proteins, and decades for prion or amyloid diseases to emerge. Plaxco and Baker (1998) very elegantly noted that the folding rate for two-state proteins vary directly with the average sequence distance between contacting residues in a protein or contact order. Thus, more local $\alpha-$helical proteins fold faster than $\beta-$sheet proteins, because the average distance between contacting residues is greater in the $\beta-$sheet protein which has non-local, distant contacts across $\beta-$strands. Experimentally, many of these time scales are readily accessible using NMR coupled with hydrogen exchange (HX) or line broadening measurements, continuous-flow or temperature jumps, stopped-flow techniques or manual dilution (Bryan, 2002).

2.12 The interaction energies and forces relevant to protein stability

Understanding the nature of the energies and forces relevant to protein stability is fundamental to the process of describing protein folding kinetics. There are different types of interaction between the atom and molecules; this interaction energy is usually given in units of kcal/mole, with the following equivalences (Kerson, 2005).

1Kcal/mole = 0.0433ev = 503K

This section briefly highlights the interaction energies that guide a protein to its native structure.

2.12.1 The hydrophobic effect

The hydrophobic effect is widely believed to be the main driving force behind the formation of the native structure (Wust et al., 2009; Ying et al., 2011; Hans-Joachim and Dirk, 2007; Meng et al., 2010; Erik, 2000). The hydrophobicity of an amino acid is a measure of the thermodynamic interaction between the side chain and water. It is also, the tendency of nonpolar (hydrophobic) molecules to associate with water. The hydrophobic effect can be viewed as an entropic effect and provides an important component of the driving force for protein folding. In the liquid phase, water molecules form a loosely connected network, which, among other things explains water's high heat capacity. This network is disturbed when a non-polar molecule is introduced into the solvent. Water molecules and nonpolar molecule interacts unfavourably, and the water molecule re-arranges their network to compensate for this. This leads to an increased ordering of the water and a decrease in entropy. If two or more nonpolar molecules are introduced, they will associate since the number of ordered water molecules will be reduced, and hence the entropy will increase, compared to if the molecules are separated (Erik, 2000). The 20 amino acids show varying degree of hydrophobicity (see table 2.1), and this has important consequences for protein structures. Due to the effective attraction between hydrophobic amino acids induced by the hydrophobic effect, protein folds into compact structures with highly hydrophobic cores and surfaces consisting mainly of polar amino acids. The hydrophobic core is important for the stability of the native structure. This can, for example, be seen in the evolution of globin proteins where the hydrophobic core is

conserved through large evolutionary distances (Erik, 2000). The hydrophobic force is the propensity of hydrophobic amino acids to cluster in buried regions, leaving polar ones exposed to a polar solvent which is water in living organisms (Luca, 2005). The hydrophobicity (h) of a protein is defined as the fraction of amino acids that are hydrophobic. If we adopt the kyte-Doolittle scale to define the hydrophobic amino acids (Liu, 2009), the amino acids with positive kyte-Dolittle value are regarded as hydrophobic; and those with negative values are hydrophilic. Accordingly, there are 8 hydrophobic amino acids (I,V,L,F,C,M,A,G), and 12 hydrophilic ones (T,S,W,Y,P,H,E,N,Q,D,K,R). According to the kyte-Doolittle scale, the hydrophobicity of natural proteins varies mainly from 0.20 to 0.75 and exhibits a Gaussian-like distribution N (0.5, 0.054) as shown in figure 2.13.

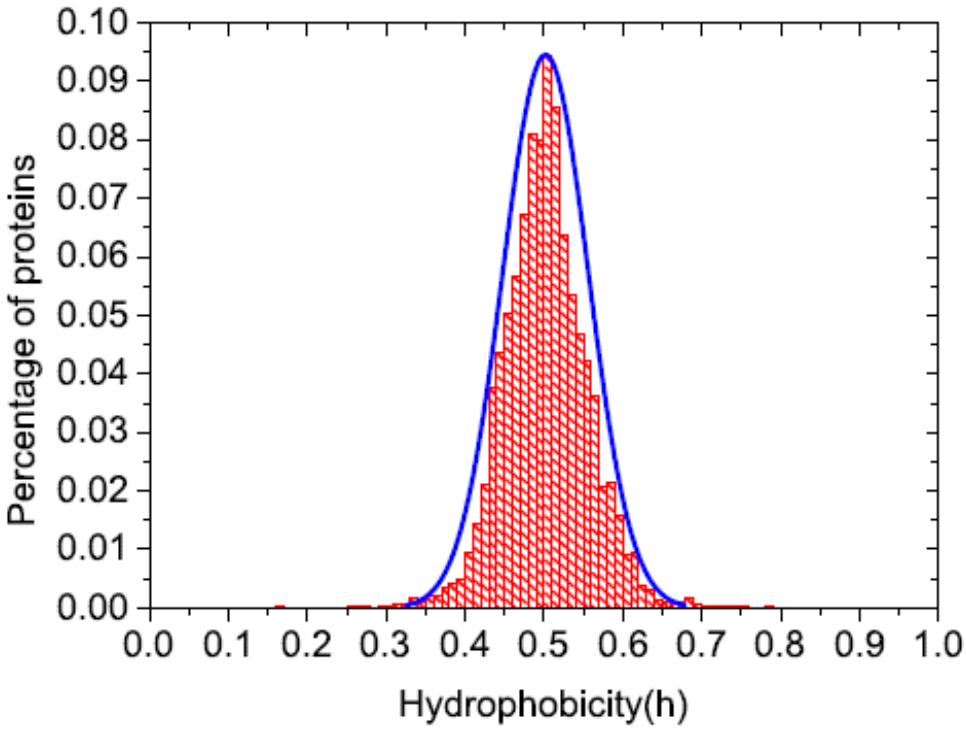


Figure 2.13.The distribution of hydrophobicity for natural proteins in PDB. The curve is fitted by Gaussian distribution N (0.5, 0.054) (Liu, 2009).

As discussed previously, the hydrophobic interaction is a main attraction force in protein folding. Thus the hydrophobicity is important for the compactness of a protein in water solution. If all amino acids are hydrophobic, corresponding to poor solvent conditions, the protein is highly compressed by solvent pressure and has a scaling exponent $\upsilon \approx 1/3$. On the other hand, if all amino acids are hydrophilic, which corresponds to good solvent conditions, the protein will be extended with $\upsilon \approx 3/5$. The detailed dependence of the scaling exponent with respect to the hydrophobicity is given in figure 2.14.

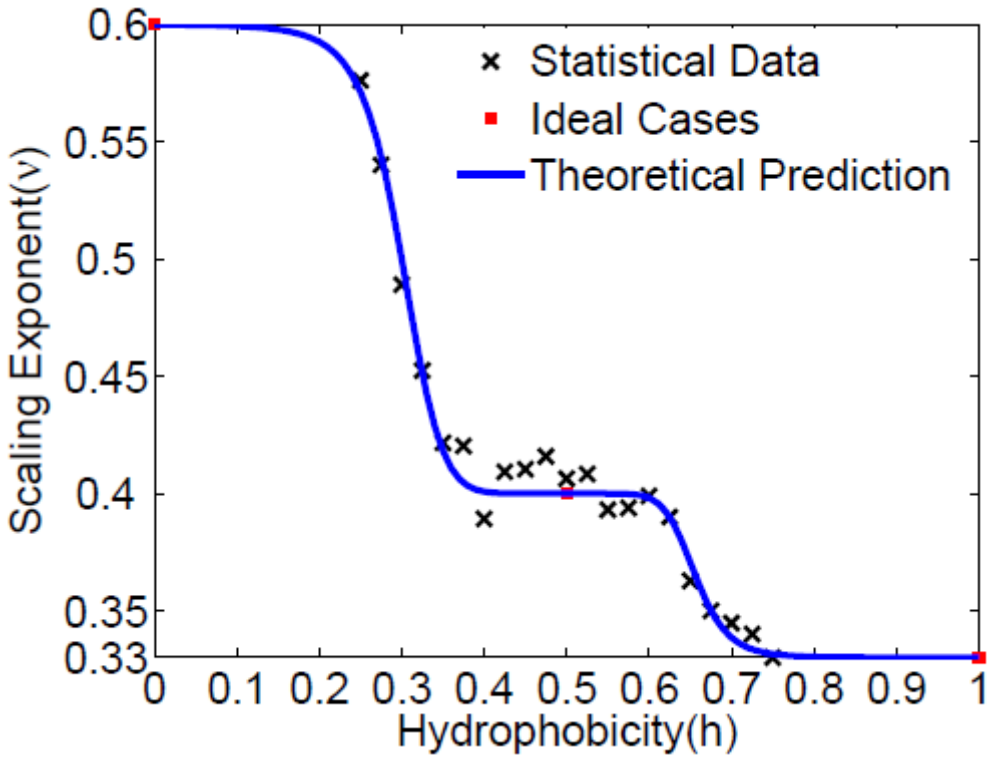


Figure 2.14. Dependence of scaling exponent on hydrophobicity (Liu, 2009).

It shows that the scaling exponent is almost unchanged with $\upsilon \approx 2/5$ with 0.4 < h < 0.6 for small value of h (h < 0.4), υ increases to $\sim 3/5$, as h approaches 0.2; while for large values of h (h > 0.6), υ decreases to ~1/3. Therefore, proteins with lower hydrophobicity have a larger scaling exponent than those with higher hydrophobicity. This is justified since higher hydrophobicity would result in more compact conformations due to stronger solvent pressure (Liu, 2009). In a theoretical study of the influence of hydrophobicity on the scaling exponent, an ensemble of proteins with the same hydrophobicity in certain solvent condition is considered. The state of this protein-solvent coupled system is denoted as X (h,p), where h is the hydrophobicity of the protein, and p is the polarity of the solvent. While the scaling exponent of a protein with hydrophobicity h in a solvent with polarity p is denoted by $\upsilon(h, p)$. Furthermore, it is assumed that each protein can only be in one of the three following states: the extended state X_1 (good solvent) with scaling exponent υ_1=3/5, the native state X_2 (water solution) with $\upsilon_2 = 2/5$, and the compressed state X_3 (poor solvent) with $\upsilon_3 = 1/3$. Then the average scaling exponent is assumed to be a linear combination of the exponents for proteins in the above three states.

$$\upsilon = \upsilon_1 \frac{[X_1]}{X_T} + \upsilon_2 \frac{[X_2]}{X_T} + \upsilon_3 \frac{[X_3]}{X_T} \tag{2.1}$$

Where $[X_i]$ stands for the concentration of proteins in state X_i, and $[X_T] = [X_1] + [X_2] + [X_3]$. In the real case, there may be many other intermediate states, but we omit them for simplicity. It is widely known that the hydrophobicity of a protein is closely related to the polarity of its surrounding solvent. Increasing the polarity of the solvent leads to higher hydrophobicity of the protein. Likewise, decreasing the polarity of solvent leads to an increase of the hydrophobicity; thus, the dependence of a protein's compactness on its hydrophobicity is equivalent to the dependence on the polarity of the solvent.

2.12.2 Electrostatic

Electrostatic energies occur between charges on the protein group like amino- and carboxy- termini and on many ionizable side chains; and they are often the result of the pair-wise application of Coulomb's law to charged functional groups located on a protein's surface (Bryan, 2002). The ion-ion interaction is given by Coulomb's law;

$$E_{i-i} = q_1 q_2 / 4\pi\varepsilon\varepsilon_o r \quad (2.2)$$

For the neutral atoms, they may form dipoles or multipoles. The general formula for the ion-dipole interaction is;

$$E_{i-d} = -p^2 q^2 / (4\pi\varepsilon_o)^2 3K_B T r^4 , \quad (2.3)$$

And for the dipole-dipole interaction is

$$E_{d-d} = -2p_1^2 p_2^2 / 3K_B T (4\pi\varepsilon_o) r^6 \quad (2.4)$$

Formally, electrostatics is a long range interaction, but in fact, its actual range is rather short due to the shielding of water solution. Besides, the signature of electrostatics is also dependent on the pH value and ionic strength (Liu, 2009).

Electrostatic interactions are important for the thermodynamic stability of proteins and have been exploited extensively in nature. Electrostatic interactions between side chains of amino acid residues can be modulated by changing the protonation/deprotonation state of charged residues via the change of environmental pH. Thus, thermodynamic stability of proteins often depends on pH values (Peng et al., 2010).

2.12.3 Hydrogen bonds

Hydrogen bond contains both positive (H-donor) and negative (H-acceptor) partial charges, which represents a combination of covalent and electrostatic interactions, particularly it contains the electrostatic interaction which arises from the partial sharing of hydrogen atom between a proton donor group, which is strongly polar, such as FH, OH, NH, SH etc. and a proton acceptor atom, which is strongly electronegative, such as F, O, N e.t.c. Usually, the hydrogen bond length varies in a narrow range (2.9 ± 0.1 Å), with their directions in a line. Hydrogen bonding is a weak interaction, which is about 2 - 10 kcal/mol depending on the electro negativity, bonding length and orientations. Its strength is usually stronger than normal dipole forces between residues, but is only about 1/10 as strong as normal covalent bonds within a residue. In $\alpha - helix$ and $\beta - sheet$, hydrogen bond exists abundantly, and is thought to be important in stabilizing the secondary structure (Liu, 2009). When closely examining

the details of protein H-bond formation, the unfolded amides and carbonyls are solvated with water and thus require backbone desolvated to form native H-bonds. It is expected that the free energy gained from forming a native protein amide-carbony H-bond is lost to the desolvation of these moieties in the unfolded state. An alternative manner of explaining the role of the protein H-bond energy is that while H-bonds do not contribute to the net stability of the protein, the H-bond is not readily broken once formed deep inside a protein core, because of the large penalty associated with burying an unsatisfied polar functional group within a non-polar environment. Thus H-bonds guide the formation of protein structure and encode structural specificity. Because once formed, the structure cannot readily slip into an alternate conformation e.g a $\beta - sheet$ may not simply shift out of register without breaking H-bonds in an unfavourable, non-polar environment ((Bryan, 2002). The H bond (E_{Hbond}) contribution involves backbone-backbone (bb) bond and backbone-side-chain (sc) bond

$$E_{Hbond} = \epsilon_{hb}^{\otimes} \sum_{bb-bb} \varsigma(x_{ij}) \tau(\alpha_{ij}, \beta_{ij}) + \epsilon_{hb}^{\oplus} \sum_{bb-sc} \varsigma(x_{ij}) \tau(\alpha_{ij}, \beta_{ij}) \tag{2.5}$$

Where x_{ij} represents OH distance, α_{ij} and β_{ij} represent the NHO and HOC angles respectively. The function $\varsigma(x)$ is given by

$$\varsigma(x) = 5\left(\frac{\partial_{hb}}{x}\right)^{12} - 6\left(\frac{\partial_{hb}}{x}\right)^{10} \tag{2.6}$$

And the angular dependence as

$$\tau(\alpha, \beta) = \begin{cases} (\cos\alpha \cos\beta)^{\frac{1}{2}} & if\ \alpha, \beta > 90^0 \\ 0 & otherwise \end{cases} \tag{2.7}$$

Also, the term ϵ_{hp} stands for effective hydrophobic attraction and is represented as

$$\epsilon_{hp} = \sum_{i<j} \omega_{ij} \eta_{ij} \tag{2.8}$$

Where the sum is over all pairs of non-polar amino acids; the parameters $\omega_{ij}\ (\geq 0)$ are constants that determine the strength of attraction between amino acid i and j. While η_{ij} is a geometric factor and a measure of the degree of contact between two side

2.12.4 Van der Waals interaction

The Van der Waals potential is another widely assumed local interaction, which comprises electron shell repulsion, dispersion forces, transitory forces and permanent forces. Van der Waals interaction shows attractive effect when two amino acid residues at moderate distances and repulsion effect when they come to each other close enough. Van der Waals energies are derived from classic Leonard-Jones potentials where the optimal distance between atoms is approximately the sum of the Van der Waal radii.

$$e_{L-J} = \varepsilon[2(\sigma/r)^2 - (\sigma/r)^6] \tag{2.9}$$

Van der Waals interaction is much weaker than chemical bonds; even random thermal motion around room temperature can easily overcome or disrupt them. However, it is regarded as necessary in forming the correct tertiary structure in protein folding (Liu, 2009; Bryan, 2002).

2.12.5 Configurational entropy

Configurational entropy behaves in a different ways when compared to other interaction; it tends to destabilize the native state of protein structure even when other tends to stabilize it. Albeit it tends to increase the degrees of freedom available to the protein chain in the unfolded state relative to the native state which comes from both the side chains and the backbone, even though the peptide backbone of most residues in a globular protein is relatively fixed (i.e. low entropy), and those residues that are most buried within the core of the protein have even fewer backbones degrees of freedom. Hence, the configurational entropy of the backbone is dependent on the side chain as is observed from Ramachandran plot. The amino acid compositions also play a role in the effect of the configurational entropy. In estimating the contributions of conformational entropy to protein stability, computational methods are the best

approach such as statistical survey of rotomer populations in proteins of known structure and Monte Carlo simulation (Kenneth, 2001).

2.13 Protein folding pathways

Protein folding is often considered a determinate process, whereby specific intermediates populate along a singular pathway. Most folding experiments, but not all are integrated in the context of a homogeneous transition state ensemble, while theoretical work, on the other hand, has led to a funnel picture in which folding occurs via structurally distinct, heterogeneous routes. This controversy, "the classical versus new view debate", has become central to the protein folding pathway discourse. The contention surrounding this issue is whether a protein traverses the same barrier each time it folds to the native state. Moreover, does the protein populate the same sequence of structure along a given pathway each time it folds? Or does a protein gradually build up structure from a variety of nucleation sites where the connectivity of the various intermediate structures is better described by a multi-dimensionally landscape rather than a linear determinant process? (Bryan, 2002). Since protein folding is an evolution of the folded structure in time, dynamics is an essential part of the folding process.

The complexity arises on one hand from the vast number of degrees of freedom available to the polypeptide chain, and on the other from the intricate network of weak, non-covalent interactions, which stabilize the native and intermediate structures. The folding pathways are controlled by two main factors: the thermodynamic stability of the partially folded intermediates, and second the dynamics of the polypeptide chain motions through which these structures are sampled. Both the thermodynamic and dynamic components contribute to the kinetics of folding. Kinetics in general refers to the macroscopic change in the population of protein conformational states in time. The energy landscape theory asserts that without much loss of kinetic information, protein folding can be captured by one or a small number of reaction coordinates, while such reaction coordinates are seldom accessible experimentally, their existence allows systematic conceptual treatment of folding kinetics based on reaction rate theories (Ginka et al., 2011).

Pioneering work has demonstrated two possible pathways for the folding of proteins in the presence of cofactors (Yi and Hongbin, 2011; Wilson et al., 2004; Wittung-

stafshede, 2002) as shown in the figure 2.15 below. In pathway I (binding-before-folding) cofactors can interact with and bind unfolded polypeptide in a specific manner to form an intermediate complex, which in turn significantly reduces the conformational entropy of the protein. Then, the cofactor-unfolded peptide complex serves as a nucleus for subsequent folding. Examples of proteins folding in this pathway include a Zurin (Yi and Hongbin, 2011) and Fe-s cluster proteins (Yi and Hongbin, 2011). In pathway II (binding after folding), unfolded polypeptides fold independently of cofactors to form apo-proteins (proteins without cofactors bound). Escherichria coli ribonuclease HI (12) and straphylococal nuclease A (13), which bind Mg^{2+} and Ca^{2+}, respectively, are two examples that follow this pathway (Yi and Hongbin, 2011).

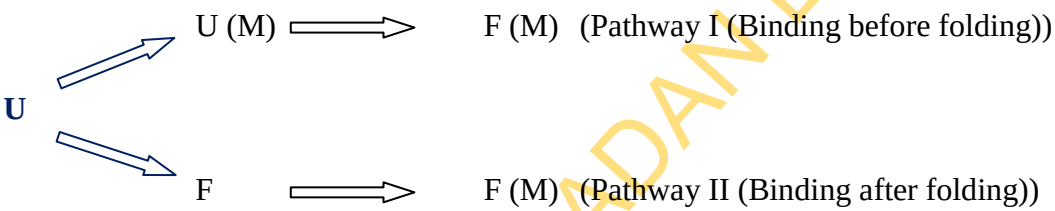


Figure 2.15. Pathways for the folding of proteins

Allan Chris et al., observed that the proper folding of proteins is critically important in cellular activities, and involves substantial interesting physics and complexity. Hence, this physical chemical process has generated tremendous interest in the multidisciplinary fields of protein science (Allan and Ashok, 2011). The protein folding field has advanced greatly over the past 50 years due to knowledge acquired through a multitude of experimental (Ginka et al., 2011; Benedetti et al., 2011) and computational approaches. Experimental strategies have included ensemble methodologies such as Nuclear Magnetic Resonance (NMR) Spectroscopy, Fluorescence and CD spectroscopy, small- angle X-ray scattering, X-ray crystallography and protein engineering (Peng et al., 2010), among many others (Allan and Ashok, 2011). Greta and co (Greta et al., 2012), observed that the numerous protein-folding studies have demonstrated that protein domains have evolved to fold without accumulation of low-energy intermediates. Such low energy intermediates might aggregate through exposure of hydrophobic patches, which would explain these intermediates are rare (Greta et al., 2012).

2.14 Folding energy landscapes

An energy landscape is a surface defined over conformation space indicating the potential energy of each and every possible conformation of the molecule. Protein folding as well as other chemical processes are best understood using the idea of their underlying energy landscape, an approach with its basis in the statistical mechanics of glasses and phase transition. This approach is the theoretical manifestations of the interactions that contribute to the chemical processes. In an energy landscape as shown in figure 2.16, valleys indicate stable low energy conformations and mountains indicate unstable high energy conformations. Foldable sequences and unfoldable sequences determined by the amino acid sequences should be manifested in their underlying energy landscape. The energy of the conformations of the folding sequence and the reaction coordinates Q is expected to be proportional with some roughness that is introduced by non-native contacts. This correlation of energy and structure introduces a bias in favour of the native conformation as well as a bias against the non-

native structure. Such a correlation is responsible for the funnel shape of the landscape (Oren et al., 2001).

Folding requires that there exists a temperature high enough for the process to occur (i.e. the protein is not frozen in one of the minima) yet low enough so that the ground state is stable. The folding temperature (T_f) is the temperature below which the native state is stable, also as a result of the roughness of the landscape there is another temperature in glasses (T_g) which the kinetics are controlled by non-native traps. For a sequence to fold, it is necessary that the folding temperature be higher than the glass temperature, $T_f > T_g$. It can be deduced that the contest between the energetic bias toward the native state and the landscape roughness plays a central role in the folding process. Consequently, the energy landscape theory offers a solution to many of the kinetic and thermodynamic perplexities of protein folding (Oren, 2001; Andrej et al., 1995; Hans, 2010).

In general, folding can be viewed as the motion of the polypeptide chain on a complex energy landscape (Ginka et al., 2011; Wust et al., 2009; Ying et al., 2011). The energy landscape theory declares that without much loss of kinetic information protein folding can be captured by one or a small number of reaction coordinates. This reaction co-ordinate is rarely obtainable experimentally. Down-hill trajectories to the folded state are opposed primarily by chain entropy. While the landscapes for polymers with a randomly chosen order of amino acids are predicted to be rugged, the landscapes of natural proteins have been smoothed to resemble a funnel, which means that many conformations have high energy and few have low energy. This funnel topology makes predicting the mechanism of folding easy once the structure is known.

The mathematical basis for understanding folding focuses on the statistics of the energy landscapes for finite size systems. Purely random heteropolymers with defined sequences done in the laboratory posed some difficulties in studying them; because they are not soluble and will crash out of the solution. On the other hand, this can be well established by computer simulation which actually confirmed the basic qualitative ideas of energy landscape theory (Broglia et al., 2007).

The energy landscape of a random heteropolymer like the landscape of structural glasses ultimately resembles the most extreme case of energetic ruggedness, the so-

called random energy model introduced by Dorrinda to model spin glasses. From figure 2.16, the radial coordinates (Q) express the entropy of an ensemble of states with a fixed value of the fraction of the native structure, which is correlated with the energy E. These variables are expressed in the vertical axis. The near linear relation between energy and entropy means that free energy barriers in the profile are small on the scale of total energy. Notice that the transition state ensemble at Q = 0.6 occurs at a lower Q and therefore higher entropy than the glass transition at Q = 0.71 from the ruggedness.

Natural proteins simply do not seem to be as highly frustrated as typical random heteropolymer would be. Proteins do not fold by gradual loss of entropy until you run out of state. Instead, there are many local themes of consistency and symmetry between a given sequence and the structure it adopts. In contrast to a glass, the energy landscape of a minimally frustrated system can be naturally divided into layers with common energetic properties. The state within each layer having a similar low energy also has a specified degree of geometrical similarity to the ground state. In this case as the energy decreases, i.e. the deeper the layer is, the more the structure of a configuration in that layer resembles the native structure. This energy decreases faster as the global minimum is approached than would be expected for a random heteropolymer.

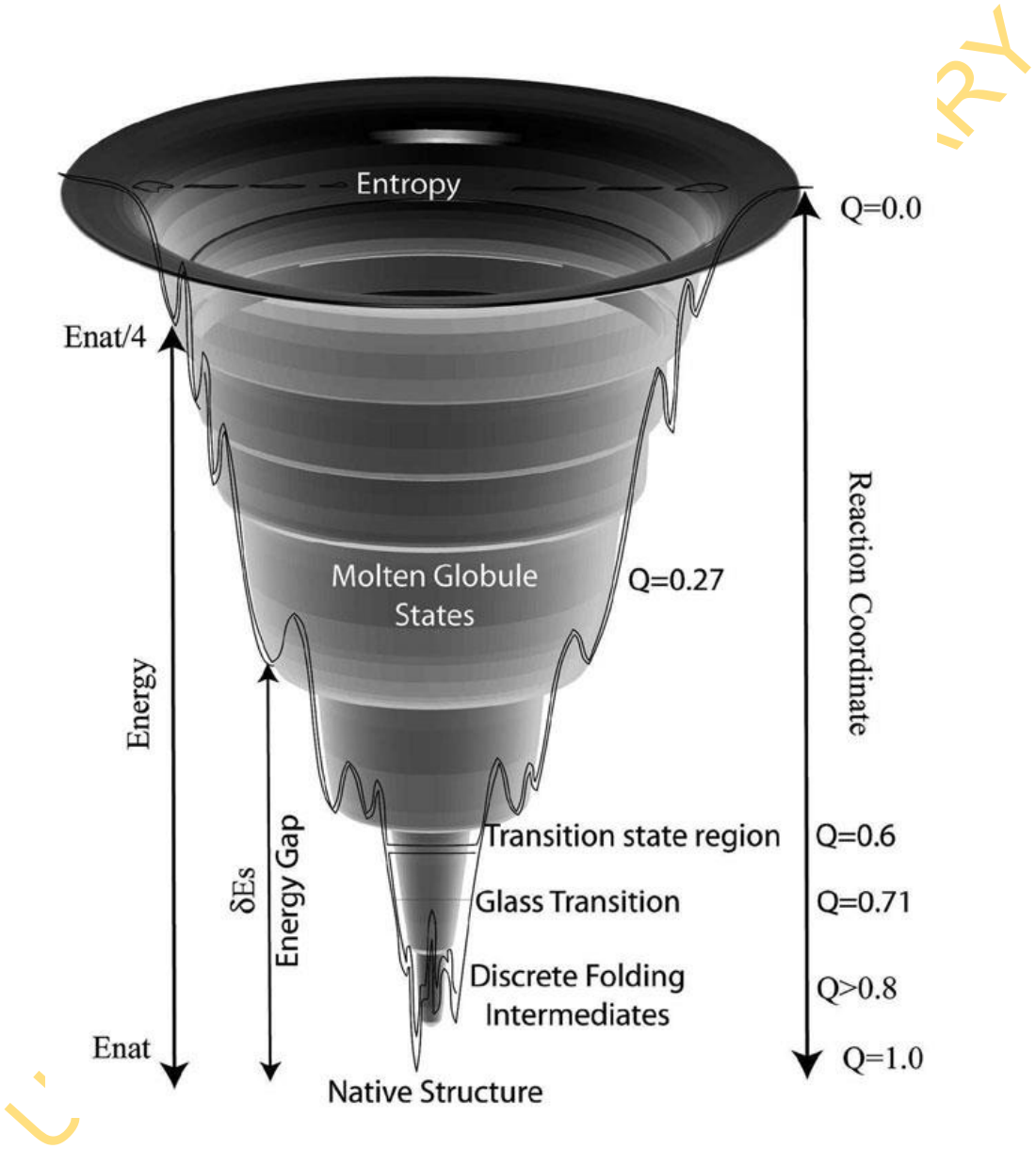


Figure 2.16. A schematic illustration of a typical small protein folding funnels with its major landmarks (Broglia et al., 2007)

2.15 Proteins and frustration

Frustration occurs when a system is unable to simultaneously achieve a minimum energy for each entity involved (Carlo et al., 2009). The protein folding phenomenon was largely an experimental endeavour until the formulation of an energy landscape theory of proteins by Joseph Bryngelson and Peter Wolynes in the late 1980s and early 1990s. This approach introduced the principle of minimal frustration. The principle of minimal frustration underlies the thermodynamic picture of protein folding. According to this picture, protein negotiates a rough, funnel-shaped energy landscape during the folding process and eventually settle in a state that, as much as possible, satisfies the energetical constraints arising from the multitude of interatomic covalent, electrostatic and Van der Waals interactions (Edwards et al., 2012; Scott et al., 2012; Carlo et al., 2009). The notion of minimal frustration has been made quantitatively precise by using the statistical mechanics of spin glasses. Protein is regarded as a system with minimal frustration in its native structure, that is, the protein folding mechanism is also governed by the minimal frustration principle (Takeshi, 2009).

Therefore, it is important to study a general behaviour of a system with minimal or no frustration, in order to understand the protein folding problem (Takeshi, 2009). A global criterion for the landscape to be funnelled to the native state emerges from this theory. Minimal frustration implies protein structure is also robust to mutation. However, neither the proteins' kinetic foldability nor their mutational robustness denies the possibility that some frustration from conflicting signals may be present locally in some proteins. Such local frustration, being tolerable, might naturally arise from random neutral evolution. Local frustration also could be a functionally useful adaptation. The possible adaptive value for a molecule to have spatially localized frustration arises from the way such frustration may sculpt protein dynamics for specific functions. In a monomeric protein the alternate configurations caused by locally frustrating an otherwise largely un-frustrated structure could provide specific control of the thermal motions, so the protein can function much like a macroscopic machine having only a few moving parts. Alternatively, a site frustrated in a monomeric protein may become less frustrated in the final larger assembly containing that protein, thus guiding specific association. Thermodynamic folding studies of enzymes also show that catalytic sites exhibit signs of frustration. These arguments

suggest that quantitative methods for localizing frustration in proteins can give insights into the functional constraints on the evolution of protein energy landscapes. The only feature of frustrated systems which survives in the case of proteins is the difficulty of predicting the ground state conformation of the system. This prediction is the essence of the protein folding problem (Broglia and Tiana, 2003; Carlo et al., 2009).

2.16 Thermodynamic view of protein folding

In order to understand and have a better knowledge of protein folding and unfolding it is expedient to have a clear understanding of thermodynamics and its properties such as entropy, enthalpy and free energy which are useful in comprehending the protein stability. In 1960s, C.B. Anfinsen proposed his most famous "Thermodynamic hypothesis" (Anfinsen, 1973) according to several experimental discoveries, that for a single-domain proteins, the three-dimensional structure of the protein is completely determined by the information embedded in the amino acid sequence and the native state is the one in which the Gibb's free energy is the minimum.

Anfinsen's famous hypothesis have been convinced by a lot of experiments, especially for many small proteins, their folding and unfolding apparently reach thermal equilibrium. This gives rise to the thermodynamic view of protein folding. According to thermodynamics, the structural changes from denatured state to folded state in protein folding are usually called conformational transition, whose characteristics can be described through thermodynamics. The denatured state has considerable conformational freedom. It is not a rigid structure, but individual segments of a polypeptide chain that can move relative to one another. The denatured state has an inherently high configurational entropy from the simple Boltzmann formula $S = k \ln W$, where W is the number of accessible states and K is the Boltzmann constant i.e $k = R/N$. On the other hand; the native state is very conformationally restricted and has a low entropy. Thus, as a protein folds, it loses considerable entropy, which must be balanced by a gain of enthalpy for the free energy to favour folding. The enthalpy of packing of side chains in the native state is favourable and compensates, just barely, for its low entropy. The classic thermodynamic description of protein folding is that it has large negative values of ΔS and ΔH. Also the thermodynamics of solvent water also contribute to the value of ΔH and ΔS. The

entropy and enthalpy of water must be added to the entropy and enthalpy of the protein to give the gross thermodynamic properties of the denatured or native state. The conformational transition of large proteins usually adopts three-state transition and are irreversible denatured with molten globule as an intermediate state; while for small proteins, two-state transition is often seen (Liu, 2009; Allan and Ashok, 2011; Bryan, 2002).

2.16.1 Two-state transitions

Transitions of protein can be from coil-globular or coil-helix transitions which can either be a two-state or three-state transitions. For small single-domain proteins with less than 100 amino acids, a two-state folding transition from 1μs to 1ms is usually observed. Small proteins are reversibly denatured and they regain their native structure spontaneously when normal conditions are restored. It can be simply explained as equilibrium between an unfolded state and a folded state, with no stable intermediates. The reaction coordinate of such a process will consist of two energy minimum separated by a single energy barrier

$$\mathbf{U} \rightleftharpoons \mathbf{F} \tag{2.10}$$

The equilibrium of the reaction is determined by the flows from the unfolded state (U) to the native (folded) state (F) and vice versa

$$C_f[U]_{eq} = C_u[F]_{eq} \tag{2.11}$$

Where C_f and C_u are the microscopic rate constants for the folding and unfolding reactions, respectively. The equilibrium constant (C) for folding is

$$C = \frac{[F]_{eq}}{[U]_{eq}} = \frac{C_f}{C_u} \tag{2.12}$$

The equilibrium constant is connected with the free energy of folding (ΔG^0) by the Van't Holf relation

$$\Delta G^o = -RT \ln C = -RT \ln\left(\frac{[F]_{eq}}{[U]_{eq}}\right) = -RT \ln\left(\frac{C_f}{C_u}\right) \quad (2.13)$$

The stability of protein molecule (and any other conformational state) is dictated by the magnitude of Gibbs' free energy and is usually described by the difference in free energy between the unfolded and the native state, ΔG^o. From the standard convention a positive value of ΔG^o indicates that the native state is energetically favoured over the unfolded state. The free energy difference, ΔG is defined in term of the enthalpy ΔH^0, and entropy, ΔS^o, differences as

$$\Delta G^o = \Delta H^o - \Delta S^o \quad (2.14)$$

For the case of protein denaturation, ΔH^o and ΔS^o are dependent on temperature through the heat capacity difference ΔC_p as

$$\Delta H^o = \Delta H_R + \Delta C_P (T - T_R) \text{ , and} \quad (2.15)$$

$$\Delta S^o = \Delta S_R + \Delta C_P \ln(\frac{T}{T_R}), \quad (2.16)$$

$$\Delta G^o = \Delta H_R^o - \Delta S_R^o + \Delta C_P \left\{(T - T_R) - T \ln(\frac{T}{T_R})\right\} \quad (2.17)$$

where the subscript R indicates the value of ΔH_R and ΔS_R are the enthalpy and entropy changes at the reference temperature, T_R, T is the absolute temperature in K and ΔC_P is the heat capacity change.

For two-state folders, a strong correlation is found between folding rates and contact order. And its strength suggests that topology is a major dominant of the folding rate. This correlation makes sense in the local contacts should be faster to form than nonlocal ones. Since helical structures are more local than $\beta-$sheets, helical proteins tend to fold faster than $\beta-$sheet proteins (Liu, 2009).

2.16.2 Three-state transition

In three-state transition, large proteins tends to fold with stable intermediates, and shows three-state transition: there is an initial fast collapse from unfolding disordered structures to partially regular structures-molten globule known as the intermediate and then a slow process which rearranges the structures and searches for global free energy minimum into the native state

(2.18)

The first folding from part is usually from $10^{-6}s$ to $10^{-3}s$ accompany by a dramatic reduction of the radius of gyration. This stage, the hydrophobic collapse appears to be the primary driving force. The hydrophobic residue moves to the surface and form hydrogen bonds with the outer water solution. At the end of this stage, a partially regular intermediate structure-molten globule is formed. The last folding part is a relatively slow process lasting from $10^{-3}s$ to 1s, with no obvious collapse of radius of gyration. This stage is similar to the crystal nucleation process. The side-chains to form tight and specific Van der Waal's interaction of a native state are the essential final step. This structure rearrangement is considered to be energetically more difficult than the formation of folding intermediates. So it's time scale is also much longer than the latter one (Liu, 2009).

2.17 Kinetic view of protein folding

From the 1980s, the kinetic view of protein folding became popular. It states that the free energy surface of natural proteins may not contain an absolute global minimum, but rather a group of local minima. During protein folding, only a tiny fraction of the total possible conformations can be explored. And this subset of conformations can be viewed through a kinetic pathway. Thus, in the kinetic view, the native state is not only determined by the initial conditions. Different folding pathways may also lead to different final conformations. For example, *in vitro*, some large proteins may easily get trapped into some misfolded states. These states are higher in free energy than the native state. But since there exist large energy barriers, the misfolded state will not change into the native state automatically. While *in vivo*, folding of these kind of large proteins usually needs the help of chaperon, which can greatly reduce the energy

barrier between the misfolded state and the native state, as well as prevent the aggregation of folding intermediate. As a result, in the kinetic view, not all protein chains, that satisfy the requirement of thermodynamic, will fold. And of those do, not all will fold in a biologically reasonable time. Only those proteins that can fold in a short time will be chosen by evolution to function in living cells. (Liu, 2009; Allan and Ashok, 2011; Bryan, 2002). Characterizing the folding and unfolding kinetics of proteins is important for understanding the energetic landscape, leading to the active native conformations of these molecules (Ashlee & Hongbin, 2010). The question as to whether the protein structure is under kinetic or thermodynamic control is not a speculative question. It is raised again and again when one faces practical problems of protein physics and engineering. For example: when trying to predict protein structure from its sequence, what have we to look for? The most stable or the most rapidly folding structure; when designing a de novo protein, what have we to do? To maximize stability of the desired structure or to create a rapid folding pathway to this structure (Finkelstein and Galzitskaya, 2004).

2.18 Protein Structure Prediction (PSP)

One of the main goals of PSP is to model the free energy of the given amino acid chain and then to find minimum energy conformations. Experiment and theory show folding proceeds fairly directly to the native structure which is energetically very stable. The aim of predicting the three-dimensional (native) structure of proteins from the sequences of their amino acid alongside their folding pathways has been one of the most important tasks in computational structural biology. The predicted structures are very crucial to pharmacology and medical science. There are some experimental methods to find the native structure of a protein; the foremost of them are x-ray crystalography and nuclear magnetic resonance (NMR) spectroscopy. These two methods are very complicated, capitally intensive, time consuming and laborious such that the number of known sequences outweighs the structures. Hence, computer simulation (CS) has played an important role in the protein structure prediction (PSP) problem. Although, PSP has been proven to be NP complete (i.e problem considered cannot be solved optimally within a reasonable time ‘polynomial time’) even from the application of the simplest hydrophobic-hydrophilic (HP) lattice model like 2D-square and 3D-cubic lattices (Liu et al., 2013).

The Motivation for this structure prediction stems from vastly different areas such as: Medicine (which help to understand biological functions, since binding of proteins with ligands and with other proteins, nucleic acids, carbohydrates and lipids constitute much of the cellular activity of living organisms), Drug Design, Agriculture (genetic engineering of richer and more resistant crops), Industry (Synthesis of enzymes e.g. those that can be incorporated in a mixture with detergents) (Marian, 2006).

2.18.1 Experimental methods

The prediction of protein structure is one of the most important in protein sciences. For the past two decades, experimental method has played a vital role in protein structure prediction. In this case, the atomic level resolution requires that the position of atoms is known with precision and certainly with respect to each other, hence the three-dimensional structural pictures of proteins will be able to establish. There are varieties of methods available for this purpose.

The foremost of them is X-ray crystallography by (Kendrew et al., 1953) which is the most powerful method for determining the structure of protein, because of its capability to provide an atomic structure of the whole assembly. The method adopted by X-ray crystallography relies on the diffraction of X-rays by electron dense atoms constrained in a crystal according to Bragg's law. The crystals are subjected to X-ray radiation and the resulting diffraction pattern can be interpreted as a reaction of the primary beam source from sets of parallel planes in the crystal. The amplitudes and phases of the diffraction data are used to calculate electron density maps. The corresponding protein structure can then be obtained by fitting the amino acid sequence to the electron density maps. Over 80% of the protein structures deposited in the PDB is determined by X-ray crystallography.

Next to this method is Nuclear magnetic resonance (NMR) spectroscopy (Wuthrich, 1990). On the other hand, it does not require a protein crystal, but treats the protein in solution and is more suitable to study their dynamics and interaction. The solution is subjected under an external magnetic field and high frequency radiation results in the splitting of the degenerate energy levels of the nuclear spin states mostly for spin ½ nuclei. The environment of the component atoms of the proteins determines the

magnitude of the energy level splitting and can be used to identify resonance frequencies with particular atoms in the protein. The result is a network of distances involving pairs of spatially-proximate hydrogen atoms. The distances are derived from the Nuclear Overhauser Effects (NOEs) between neighbouring atoms. The resulting distances together with other experimental information are converted to a 3D structure with a computational procedure in which an energy function is minimized and structure coordinates which conform to the experimental data are found. NMR allows determination of atomic structures of increasingly large subunits and even their complexes. Recently, an extra step has been added to the NMR procedure, in which the resulting models are used again to calculate the spectra, and by matching of the calculated spectra to the experimental one, an iterative procedure for improvement is pursued. NMR constitutes about 9% of the structures in protein data bank.

Also, we have Electron Microscopy (EM) which relies on electron diffraction by particles immersed in frozen lattice. This method is gaining reputation as a result of requiring little sample preparation and is particularly suitable to very large protein complexes. We also have Mass Spectroscopy (MS) in which a charged particle passing through a magnetic field will be deflected along a circular path on a radius, which is proportional to the mass to charge ratio (m/e). MS is often used to identify the primary structure of an unknown polypeptide and more insight has been gained about MS for determining tertiary structure and functions.

Fluorescence Spectroscopy (FS) is another experimental method which is based on the electronic transition which provides information on the absorbance and fluorescence of chromophores found in proteins, usually aromatic residues. Circular Dichroism (CD) is another method that requires the measurement of differential absorption of right and left circularly polarized light as a function of wavelength. The CD allows the prediction of secondary structure with reasonable accuracy. We also have hydrogen exchange (HE) which is good to check the local stability of proteins. Lastly, we have Single–Molecule Experiments (SMS) mostly we have atomic force microscopy (AFM) and single molecule force spectroscopy (SMFS). SM adopts the method of investigating the properties of a single individual molecule that can be isolated for the purpose of an experiment or analysis. Other methods are Transmission Electron Microscopy (TEM), Scanning Electron Microscopy (SEM), Infrared Spectroscopy

(IRS), Electron Spin Resonance (ESR), Optical Rotatory Dispersion (ORD) e.t.c. (Marian, 2006; Liu, 2009).

2.18.2 Computational methods

The computational approaches to the protein structure prediction (PSP) have some wonderful advantages comparing with analytical and experimental methods. They can be classified into two main categories: Comparative modelling and Ab initio approach (Kihara et al., 2001; Hyun-suk, 2006; Hoque et al., 2007; Mahmood et al., 2013).

(a) Comparative modelling uses the existing database of experimentally determined protein structures as starting points. This class can be further split into two main subclasses:

(i) Homology modelling: Homology modelling is based on the assumption that two homologous proteins (proteins that share similar amino acid sequences) will presumably contain similar 3D structures in which their functions are strongly conserved during the process. Thus a strong sequence similarity usually indicates strong structure similarity, although the converse may not necessarily be true. The sequence of the solved structure is modified to that of the unknown structure and the resulting optimized conformation is the predicted three-dimensional model of the unknown structure.

(ii) Threading method: This method scans the amino acid sequence of the unknown structure against a database of experimental structures, a scoring function is evaluated for each comparison to assess the compatibility of the sequence to the structure, thereby producing plausible three-dimensional models. These methods depend on the database of protein sequence and the respective structure. However, since they depend on the sequence samples in the database, their results may become unrealistic for unlike sequences and become less accurate for longer sequences.

(b) The Ab initio or De novo (i.e. from the origin) approach is based on the physical principles governing the interactions of amino acids in a polypeptide chain and the surrounding solvent. First, an accurate model of the physical interaction within the polypeptide chain is necessary. Ab-initio prediction is the only choice to infer the protein structure from primary sequence information when no suitable templates can be found based on the intrinsic properties (hydrophilic and polar) of amino acids. This

is captured in a potential energy function which describes the interatomic physical interactions. The potential energy function must be accurate enough to capture the important interactions, yet simple enough so that calculations can be performed with today's computational power in real time.

The computational models of protein folding are typically formulated to find the global minimum of a potential energy function. Force fields of different resolutions (from all-atom to highly simplified coarse grained models) have been developed. Second, the concept of ab initio folding is based on the Anfinsen's thermodynamic hypothesis (Anfinsen, 1973) which assumes that the native fold of the protein populates its global energy minimum; as well as Levinthal paradox (Levinthal, 1968)which state that protein fold into their specific 3D conformations in a time-span far shorter than it would be possible for protein molecules to actually search the entire conformation space for the lowest energy state. However, in contrast, protein cannot be a random process which concludes that folding pathways must exit which motivate the ab initio based computation. Ab initio among the three approaches is the most computationally demanding and in contrary, is also the most promising in providing reliability, accuracy, usability and flexibility in checking the functional divergence of a protein by modifying its structure and sequence. This task is carried out by a variety of global optimization techniques such as energy minimization (Hoque et al., 2007), Monte Carlo-based methods (Mahmood et al., 2013) and molecular dynamic procedures (Marian, 2006; Liu, 2009).

2.19 The state-of-the-art approaches for ab initio PSP

In order to search for the ground state conformation, there are numerous existing search methods such as molecular dynamic, statistical mechanical model and stochastic search methods that attempt to solve the PSP problem by exploring the feasible structures known as conformation either on 2D, 3D and FCC HP lattice model. Over the years, Molecular dynamics have been used for PSP, but due to its NP completeness, the methods are easily trapped by local energy minimum and involve too much time for a protein of reasonable size. By and large, the stochastic search

methods, which are heuristic, like the Monte Carlo, genetic algorithm, tabu search, ant colony algorithm, and simulated annealing have been prominent for the PSP problem.

The iterative Monte Carlo methods based on local search approach have been in the forefront in the search for the lowest energy conformation. Thachuk (Thachuk et al., 2007), presents a replica exchange Monte Carlo (REMC) algorithm which is a classical Monte Carlo search method coupled with random walk at the same time for the PSP. This method, sample conformations according to the Boltzmann distribution in the energy space and employs VSHD moves, a combination of three moves and pull move neighbourhood search for both 2D and 3D HP lattice model to the benchmark instances which gave them the ground state structure when compared with the previous state-of-the-art results. REMC is also known as parallel tempering, exchange Monte Carlo, and multiple Markov chain Monte Carlo. (Ron and John, 1993) used a genetic algorithm (GA) for 2D lattice model. They believed that GA is an extension of MC by including information exchange between a set of parallel simulations. In comparison with MC method, they concluded that GA is more superior to conventional MC in term of searching effectiveness in a model of protein folding. However, Huang (Huang et al., 2010) used a genetic algorithm on optimal secondary structure (GAOSS) by ameliorating the evolutionary Monte Carlo algorithm for PSP in the 2D HP model. Their results showed that GAOSS obtains the conformation faster and pave way for more ground state conformation.

Besides MC search, Jacek (Jacek et al., 2004), used the tabu search strategy (TSS) by using conformational motif as a problem domain knowledge to find the optimal conformations of the 2D benchmark sequences. (Mahmood et al., 2013) used tabu based spiral search local method on 3D FCC, their algorithm employs a novel H-core directed guidance that squeezes the structure around a dynamic hydrophobic-core centre with the application of random work which employs pull moves coupled with relay-restart technique to enhance the H-core and prevent it from early convergence. (Alena and Holger, 2005) used ant colony optimization algorithm (ACO) for both 2D and 3D HP model to obtain the lowest energy when compared with the previous state of the art algorithm. Moreover, Guo (Guo et al., 2006) designed a hybrid elastic net algorithm (ENL) coupled with local search strategies which ameliorate the

multimapping problem of the original elastic net algorithm to produce the minimal energy for benchmark instances.

Presently, none of the aforementioned heuristic algorithms appears to completely dominate the others in terms of solution quality and run-time when applied to both the 2D and 3D lattice HP model.

CHAPTER 3

METHODOLOGY

3.1 Coarse-Grained (CG) models

Generally, folding of protein has been a big problem in the sciences; the processes involved are very complex as a result of the large number of the degrees of freedom, since the complexity of proteins depends solely on the different physical and chemical features of their monomers that is, the 20 types of amino acids. As a result, Coarse–Grained (CG) representations of the polypeptide chain have been a veritable tool in the simulation of protein folding; this method has been used by many authors to understand the physical principle of folding and serves as an essential tool in theoretical studies of protein folding. In this model, instead of representing each atom in the protein, definite groups of atoms can be treated as a single coarse-grained site. In most cases, each residue corresponds to a single coarse-grained site placed at the alpha carbon, although models with multiple sites per residue have been employed (Ronald and Charles, 2009). The constriction of the success recorded by the atomistic protein simulations by available computer power was not as a result of the volume of the atoms involved but rather the long equilibration time associated with a system that in a highly nontrivial phase space can so easily get stuck.

Sequel to this, CG simulations plays a role of addressing this problem by lowering the level of resolution. A smaller number of beads or atoms enhance the speed of MC simulations by reducing the computational requirements and the molecular friction which smoothens out the free energy landscape. From the wide point of view of many authors, one of the main research methodologies of protein folding has been computer simulation. However, any computer simulation based upon all-atom details consume a lot of CPU time to perform based on contemporary computers for long chains. It is strongly desirable to improve the modelling efficiency by circumventing the atomic details to some extent. In that case, the interactions are approximated to capture the

important physical properties and the amino acid residues are coarse-grained by single monomers (Tristan and Markus, 2009; Yantao et al., 2004).

During the last decade, off-lattice models using traditional simulation like molecular dynamics have been successful in studying protein folding. Hence, performing MD at atomistic detailed is a big challenge to explore the folding landscape of a small protein. Sequel to this, lattice Monte Carlo simulation based on a coarse-grained model circumventing the atomic detailed is a better approach to model the protein folding process with very high efficiency, albeit off-lattice MC simulations are also available for the coil-helix transition (Yantao et al., 2004)

Important insight has been gained by studying CG models which are not meant for specific structure predictions, but rather to elucidate generic physical principles of protein folding and is also significant for revealing the universal behaviour of protein. It is simple but non-trivial model from which exact solutions exist, it also provide the thermodynamic studies that requires a proper sampling of the conformation space. The CG model is of two types, one is lattice-based and the other is off-lattice based model, both of which have two types of amino acids (hydrophobic and polar). The amino acids are represented as point-like interaction centres, and the solvent is only implicitly included through effective hydrophobicity forces. In this research work we will use this model both to learn about the forces required for the formation of unique native structures (Erik, 2000).

Proteins are represented as a linear and self-avoiding chains which contain two types of monomers; H (hydrophobic) and P (polar) monomers on a 2D lattice model for which self-avoiding walk (SAW) is used to represent the protein chain such that covalently linked residues occupy neighbour lattice sites. The energy of a conformation is the sum of the energies of pair-wise contacts between monomers. Two monomers are defined to be in contact if they are neighbours on the lattice and not connected by a covalent bond. The energy of the contact depends only on the identity of the two amino acids residues involved. The interaction energies for residue pairs are determined from the statistical distribution of contacts in real proteins. There are two types of non-covalent interaction in proteins; local (short range) and nonlocal (long range), so defined as the distance along the chain, $|i-j|$, (that is, the distance in the

sequence between the interacting residues) rather than spatial distances, $|r_i - r_j|$. They are geometrically different. Local interactions, which are sequenced independent are those formed by monomers that are neighbours along the sequence and they participate in defining the secondary structure. While the non-local interactions which are sequence dependent facilitate the hydrophobic attraction and participate in defining the tertiary structure of the protein. In order to attain the native structure and stability of a protein, the sequence dependent non-local is more relevant to the protein folding while local interaction is less important in the classical point of view (Abkevich et al., 1995). However, local interaction has an important role in driving the conformational search during various models of proteins (Anders and Sandelin, 1998; Victor and Serrano, 1996; Anders et al.,1997).

In this model, focusing on the hydrophobic effect, protein is described by its sequence of N amino acids, $\{\sigma_i\}$, which takes the values H and P with i as a monomer index, their positions being $\{q_i\}$. The bond vectors $\{b_i\} = \{r_{i+1} - r_i\}$, have fixed unit length. The energy of a structure is given by sequence–independent local interactions and sequence-dependent nearest-neighbour contact interactions

$$E = \kappa E_L + E_{Non} \tag{3.1}$$

$$E_{Non} = \sum_{1 \le i < j \le N} \in (\sigma_i, \sigma_j) \Delta(q_i - q_j) \tag{3.2}$$

$$E_L = 2 \sum_{i=2}^{N-1} (1 - b_i \cdot b_{i-1}) \tag{3.3}$$

Where $\Delta(q_i - q_j) = 1$, if monomers i and j are lattice neighbours and $\Delta_{ij} = 0$ otherwise. q_i defines the type of amino acid residue at position i. The energy depends on three parameters which determine the strength of non-local interaction $\in_{HH,}$ $\in_{HP,}$ and $\in_{PP}$ according to Lau and Dill (1989) see ref. (7) of (Anders and Sandelin, 1998) $\in_{HH} = -1$, $\in_{HP} = \in_{PP} = 0$. While κ determines the strength of the local interactions. For $\kappa = 0$ the model is identical to the HP model. The hydrophobic effect is modelled by having a stronger attraction between HH pairs than between HP and PP pairs of amino acids.

3.1.1 **Lattice model**

Currently, the investigation of the folding process of real proteins via full simulation or by calculating their native structure directly is not feasible. Consequently, lattice protein model which abstracts from real protein, has come to a full-fledged stage to address this complexity since the real protein sequences are usually not apt due to model restrictions in sequence or structure space.

It is often computationally intractable to undertake protein structure studies using full atom representation. The challenge is to reduce complexity while maintaining detail. In order to achieve this, lattice protein models are often used, but in general only the protein backbone or the amino acid centre of mass is represented. Lattice models have proven to be extremely useful tools to address the complexity of the protein structure prediction problem (PSP). This model can be used to extract essential principles, make predictions and harmonize our understanding of many different properties of proteins. The discretisation of the space of conformations serves as one of the hallmark approximations made by lattice model. Even though, this discretisation prevents a completely accurate model of protein structures, it preserves important features of the problem of computing minimum energy conformations. The discretisation also provides the mathematical structure that can be used to analyse the computational complexity of PSP problems. Among the lattice model, face-centred cubic (fcc) HP lattice model which is shown to yield very good approximations of real protein structures has been distinct in capturing the main features of protein structure (Martin., 2012). Folding is an intrinsically statistical phenomenon and no conclusion can be derived from a single folding or unfolding trajectory. Lattice and other simplified analytical models are the statistical mechanician's contribution to the protein folding. Their intimate connection with statistical mechanics is very important as it often allows us to compare the simulation with statistical-mechanical analytical theories. It finds its applications in the folding process, native structure properties, sequence evolution, co-translational folding and cooperative folding and so on.

The pioneering work of Levitt and Wharshel in the 1970s, creating the first detailed energy-minimization lattice model for studying the folding of bovine pancreatic trypsin inhibitor (BPTI) marked the beginning of the physically based models in the

study of protein folding. The fact that most current structure prediction methods used a similar representation to that of Levitt and Wharshel is a strong testament of the power of such an approach. Lattice models are of two types. The first type is "designed to understand the basic physics governing the protein folding process," The second type aims at "realistic folding of real proteins and are therefore parameterized using real proteins as templates by statistical sampling of the available structures and are often referred to as statistical potentials" (Istrail and Lam, 2009).

In the first basic physics category, the major models that provided deep insights into the physical principles of folding are: Go {simplicity, 1983}, Wolynes {funnel-like energy landscape, 1995}, Dill {hydrophobic interactions, hydrophobic-hydrophilic pattern, 1990}, Shakhnovich {statistical mechanics, 1994}, Karplus {diffusion-collision, 1999}. The leading models in the second category are due to Skolnick, 1990, Miyazawa and Jernigan, 1985, Crippen, 1996, Eisenberg, 1991, Sippl, 1990, Scheraga, 1997.

According to (Istrail and Lam, 2009), lattice protein models are a common abstraction of proteins and is often used to investigate the folding process and the native structure properties of proteins. The high level of abstraction facilitated the possibility of large-scale studies which is impossible in more realistic full atom protein representations as a result of the large size of sequence and structure space.

The lattice used plays a vital role in modelling a real protein conformation from specific lattice protein structure. The energy function is typically contact based, i.e. it sums sequence specific pair-wise potentials for amino acids that are in close distance within the conformation, and even more complex energy functions incorporate distance-based potentials. Hence, the applied energy function directly depends on the modelled sequence space which is the amino acid letter code. Figure 3.1 gives examples of different lattice protein models used in literature and such contacts are highlighted in red in the 2D-models' drawings. They are typically defined by the minimal distance between neighboured positions within the lattice.

A lattice-based PSP model represents conformations of proteins as non-overlapping embedding of the amino-acid sequence in the lattice. According to (William and Alantha, 2001), lattice models can be classified based on the following properties:

1. The physical structure, which specifies the level of detail at which the protein sequences are represented. The structure of the protein is treated as a graph whose vertices represent components of the protein.
2. The alphabet of types of amino acids that are modelled either the letter of the 20 naturally occurring type of amino acids or the binary alphabet of hydrophobic (H) and polar (P).
3. The set of protein sequences that are considered by the model.
4. The energy formula used, which specified how pairs of amino acid residues are used to compute the energy of a conformation.
5. The lattice in which protein conformations are expressed; this determines the space of possible conformations for a given protein e.g cubic and diamond lattices.

3.1.1.1 The dimensions of lattice model

The 2D square lattice and the 3D cubic lattice are the most thoroughly studied lattices and consequently have extensive literature on exact computational methods, approximation algorithms, and complexity results. In three dimensions, a lattice of major importance is the face-centred-cubic (FCC) lattice. It has been shown that the neighbourhood of amino acids in proteins closely resembles an FCC lattice, providing evidence for the importance of the FCC lattice in modelling protein folds. Furthermore, the kissing number of a sphere in 3D space is known to be 12, the same as the degree of each vertex in the FCC structure. Therefore, the number of degrees of freedom for placing adjacent spheres in three dimensions is achieved by the vertices of the FCC lattice. This is intimately tied to Kepler's conjecture, recently proved by Thomas Hales, which states that the face-centred-cubic lattice is the densest packing of identical spheres in three dimensions and therefore provides the densest possible hydrophobic core for any lattice-based protein folding model.

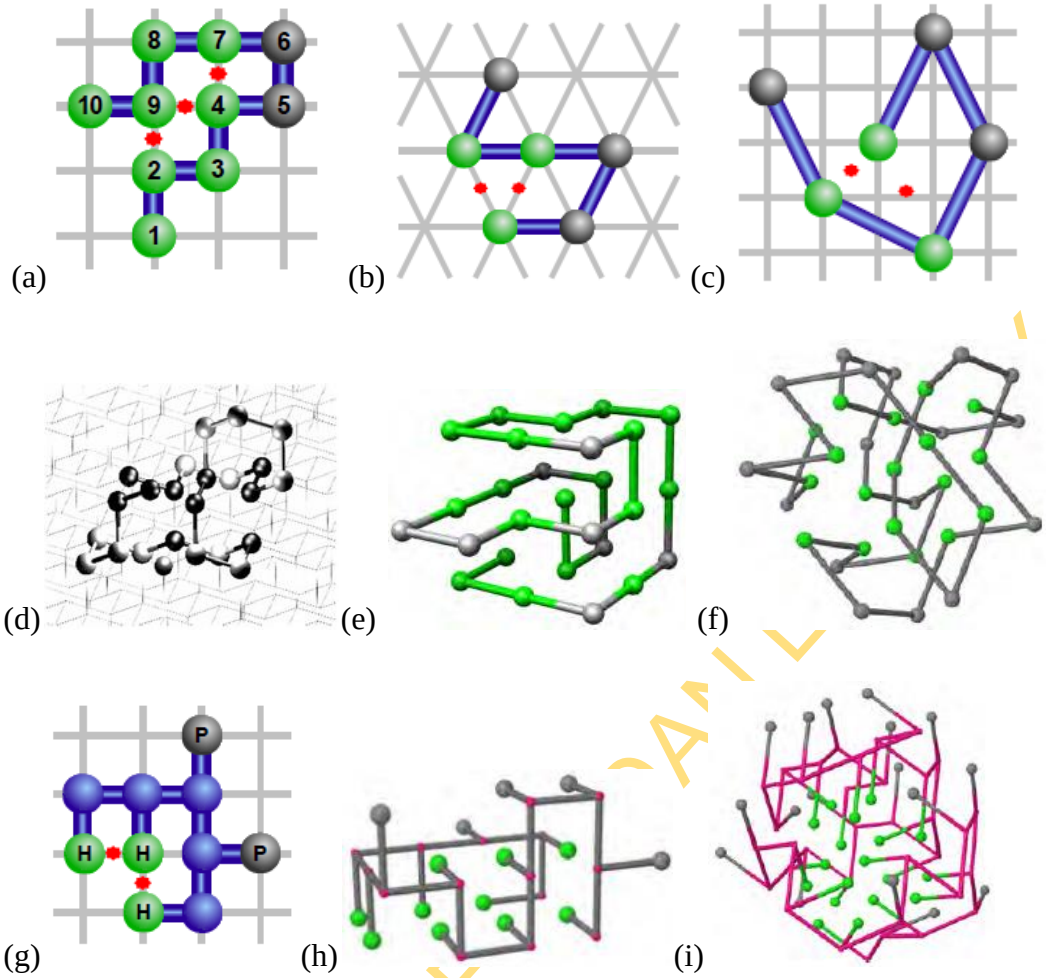


Figure. 3.1. Examples of lattice protein models used in the literature adapted from (Martin M. , 2011): Backbone-only models in (a) 2D-square lattice (Lau and Dill , 1989), (b) 2D-triangular lattice (Bockenhauer et al., 2008), (c) 2D view of 3D-210 "chess knight" lattice (Sun et al., 1999), (d) 3D-diamond lattice (Krasnogor et al., 2002), (e) 3D-cubic lattice (Thachuk et al., 2007), and (f) 3DFCC lattice (Mann et al., 2008b). Side chain models in (g) 2D-square lattice (Bromberg and Dill, 1994), (h) 3D-cubic lattice (Hart and Istrail, 1997), and (i) 3D-FCC lattice (Mann et al., 2009c). In 2D model figures (a,b,c,g), favourable contacts are highlighted (red stars).

In lattice models, a fold of a protein sequence is defined by placing the amino acids on lattice nodes and the protein chain as a self-avoiding path on the lattice; in off-lattice models, the placement of the protein is in 3D space, with the only restriction being the self-avoidance of the backbone and of the branching side-chains (Dill et al., 1995).

3.1.1.2 **Lattices**

A lattice is a finite set of regularly spaced points or lattice vector (also called lattice points) in a space of dimension d = 1, 2 or 3. In dimension 1 we have a string of points on a line, which we can enumerate from 1 to N ("N" denote the number of lattice sites, regardless of dimension). Each line segment between lattice sites is called a bond, and a lattice site is called nearest neighbours if there is a bond connecting them. In general, except for the lattice site on the "boundary" of the lattice, each lattice site in a d-dimensional lattice has 2d nearest neighbours (as shown in figure 3.2).

In order to discretise the structure space of proteins, two or three-dimensional lattices can be used. Such a lattice L is a set of 2D or 3D coordinates (also named nodes, points, or vectors) that contains the zero coordinate $\vec{0}$ and forms an additive group with operator + and its inverse operator - for any two points $\vec{x}, \vec{y} \in L$. Consequently, two nodes $\vec{x}; \vec{y} \in L$ are neighbours within the lattice if their distance vector is a neighbourhood vector, i.e.

$$(\vec{x} - \vec{y}) \in N_L \text{; where } \vec{x} \text{ and } \vec{y} \text{ are neighbours.} \tag{3.4}$$

To obtain a regular lattice, all vectors of the neighbourhood N_L have to be of equal length:

$$\forall \vec{r}_1, \vec{r}_2 \in N_L : \left|\vec{r}_1\right| = \left|\vec{r}_2\right| \tag{3.5}$$

where $\left|\vec{y}\right|$ denotes the length of the vector $\vec{y}$. This property makes the formalized lattices a subgroup of the Bravais lattices where the spanning neighbourhood vectors can be of different lengths. The number of lattice neighbourhood vectors $\left|N_L\right|$ is an

important property of a lattice L and is called its coordination number which is a measure of the lattices complexity (Martin, 2011).

Table 3.1. The three most common lattices with their co-ordination number; The Visualization is given in the figure. 3.2.

Lattice Name	**ID**	**Neighborhood NL**	**$\|N_L\|$**
Square	SQR	$\{\pm(1,0,0),\pm(0,1,0)\}$	4
Cubic	CUB	$\{\pm(1,0,0),\pm(0,1,0),\pm(0,0,1)\}$	6
Face Centered Cubic	FCC	$\left\{\begin{matrix}(1,1,0),(1,0,1),(0,1,1),\\(1,-1,0),(1,0,-1),(0,1,-1)\end{matrix}\right\}$	12

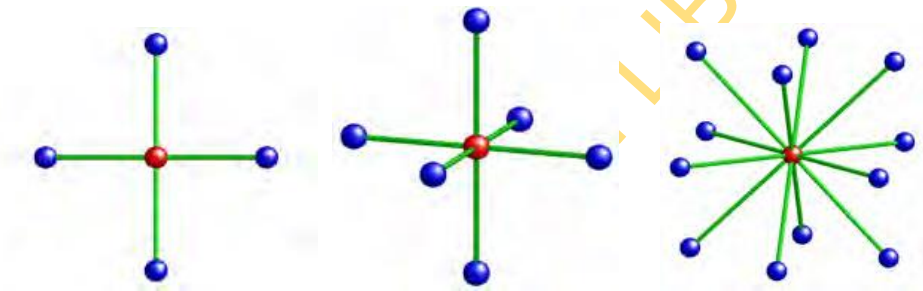


Figure. 3.2. Visualization of the lattice neighborhood NL in different lattices (from left to right: 2D-square, 3D-cubic, and 3D-FCC lattice). The reference point is given in red, the set of neighboring vectors are depicted in green, and the reached neighbored points are colored in blue (Martin, 2011).

3.1.1.3 The square lattice

The square lattice is the easiest and most used lattices without parity problem. It is defined as the set of lattice points $\mathcal{Z}^2$ and generated by its basis; it has a coordination number of four, i.e each point has 4 neighbours

$$\left\{\begin{pmatrix}1\\0\\0\end{pmatrix},\begin{pmatrix}0\\1\\0\end{pmatrix}\right\}. \tag{3.6}$$

The set of minimal vectors is given by

$$\left\{\begin{pmatrix}\pm 1\\0\\0\end{pmatrix},\begin{pmatrix}0\\\pm 1\\0\end{pmatrix}\right\}. \tag{3.7}$$

3.1.1.4 The cubic lattice

The cubic lattice (as shown in figure 3.3) is the simplest and probably the most prominent three dimensional lattices. It is defined as the set of lattice points Z^3 and generated by its basis; it has a coordination number of six that is each point has 6 neighbours

$$\left\{\begin{pmatrix}1\\0\\0\end{pmatrix},\begin{pmatrix}0\\1\\0\end{pmatrix},\begin{pmatrix}0\\0\\1\end{pmatrix}\right\}. \tag{3.8}$$

The set of minimal vectors is given by

$$\left\{\begin{pmatrix}\pm 1\\0\\0\end{pmatrix},\begin{pmatrix}0\\\pm 1\\0\end{pmatrix},\begin{pmatrix}0\\0\\\pm 1\end{pmatrix}\right\}. \tag{3.9}$$

The cubic lattice has a parity problem from its property which artificially restricts possible contacts in the cubic lattice. In this property the lattice points are naturally partitioned into two disjoint classes by the neighbour relation (Sebastian, 2005). The first one is point with even sum

$$Z^3_{|even} = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in Z^3 \mid x + y + z \right\} \text{ is even} \tag{3.10}$$

And point with odd sum

$$Z^3_{|odd} = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in Z^3 \mid x + y + z \right\} \text{ is odd} \tag{3.11}$$

Every point in one of the two classes has only neighbours in the other class, since adding any neighbour vector to a point changes the parity of the sum of its coordinates

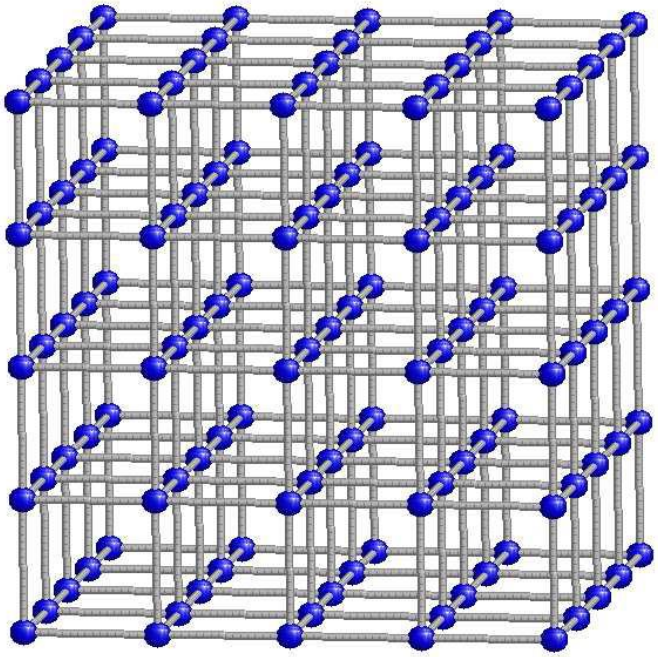


Figure 3.3. The cubic lattice (Sebastian, 2005)

3.1.1.5 The faced-centered cubic lattice (FCC)

FCC has been proved by many authors to be the densest packing of spheres in three dimensions. FCC is defined as the set of points

$$D_3 = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in Z^3 \mid x + y + z \right\} \text{is even} \tag{3.12}$$

The length of the minimal vector which is the minimal distance between two lattice points is $\sqrt{2}$ with twelve minimal vectors

$$\begin{pmatrix} \pm 1 \\ \pm 1 \\ 0 \end{pmatrix}, \begin{pmatrix} \pm 1 \\ 0 \\ \pm 1 \end{pmatrix}, or \begin{pmatrix} 0 \\ \pm 1 \\ \pm 1 \end{pmatrix}. \tag{3.13}$$

The unit cell of FCC (as shown in figure 3.4) is constructed by placing points at the corners of a cube and at the centre of its faces; this principle actually explains the origin of the name of the lattice "face-centered cubic"

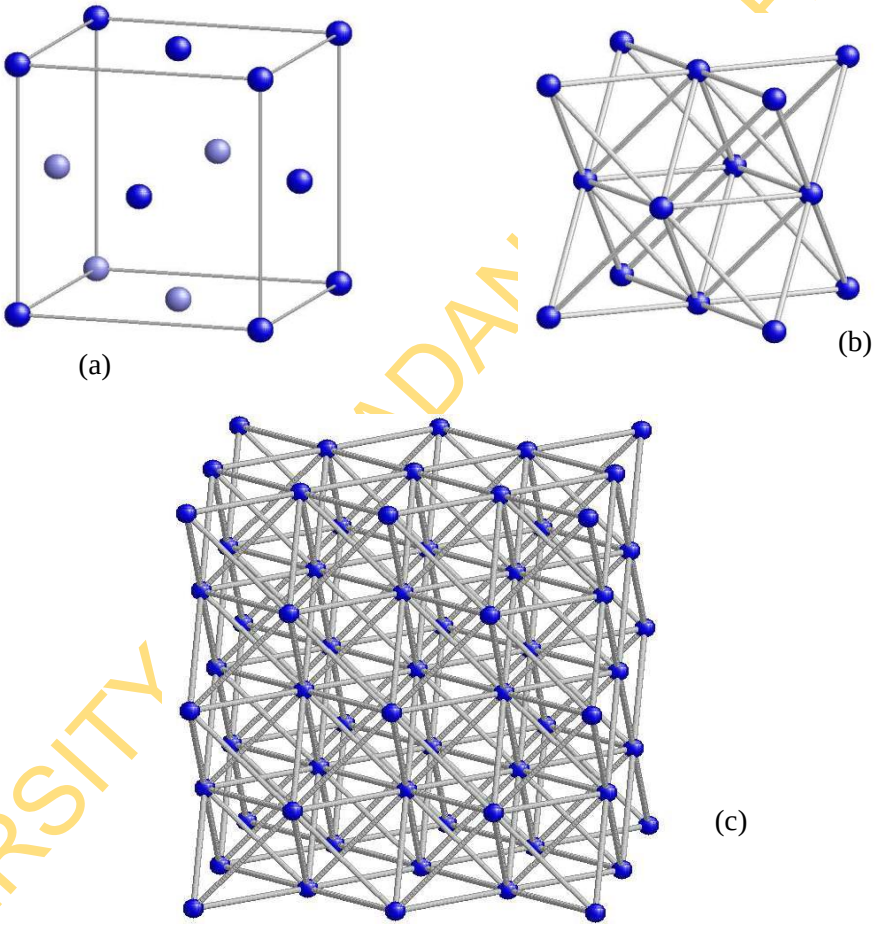


Figure3.4. Unit cell of the face-centered cubic lattice. (a) Cube with lattice points at corners and centers of faces. (b) edges between neighbors. (c) the Larger cutout of Face-centered cubic (Sebastian, 2005)

3.1.1.6 Classes of lattice protein structure

Lattice protein models are a coarse abstraction of real proteins where structures are discretised using a lattice. One can distinguish two major classes of lattice protein structures: backbone-only (main chain) and side chain models. Within backbone-only models, only the protein's backbone is represented with one monomer per amino acid. In side chain models, the abstraction is extended by a second monomer for each amino acid to model its side chain. All monomers are confined to the underlying lattice, such that connected monomers in sequence are neighboured within the lattice. A first graphical sketch is given in Fig. 3.2. In the following, both structure abstractions are discussed in detail.

3.1.1.6.1 Backbone-only models

Backbone-only models of proteins are usually represented by the C_α – positions of amino acids. Occasionally their centroid/centre of mass is represented. The representative monomer has no volume nor mass, i.e. is independent of the amino acid side chain size. Given a lattice L, a backbone-only lattice protein structure ζ of length x is a sequence of lattice nodes $\zeta = (\zeta_1, \ldots\ldots \zeta_x)$ such that all occupied nodes are different and successive monomer nodes are neighboured within the lattice, i.e.

$$\zeta_i \in L: \qquad \forall 1 \le i \le n \qquad (3.14)$$

$$\zeta_i \neq \zeta_j: \text{ i.e. } \zeta \text{ is self-avoiding} \qquad \forall 1 \le i < j \le n \qquad (3.15)$$

$$\zeta_i - \zeta_{i-1} \in N_L: \text{ i.e. } \zeta_i \text{ is neighboured to predecessor } \zeta_{i-1} \qquad \forall 1 < i \le n \qquad (3.16)$$

Examples of Backbone-only model structures in different lattices are in figure 3.1 and on the left in figure 3.5. The complexity of the lattice, i.e. its coordination number, directly influences the growth of the resulting structure space.

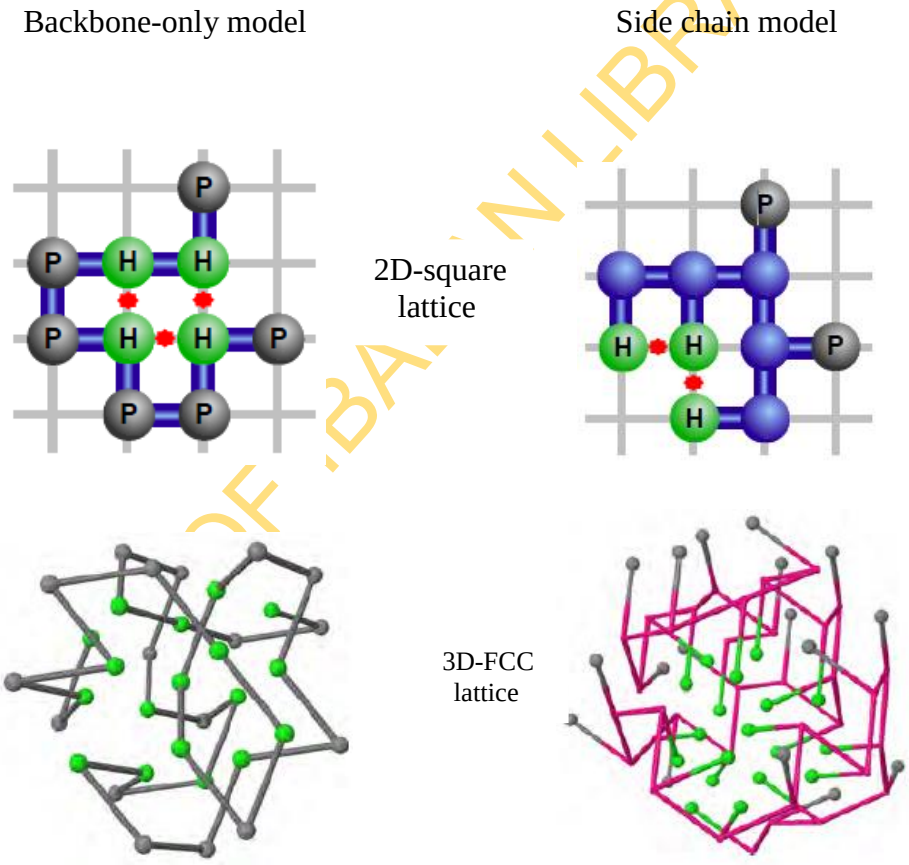


Figure 3.5. Comparison of backbone-only and side chain lattice protein models in different lattices (Martin, 2011).

Backbone-only lattice protein models are often used to perform large scale protein studies, the monomers are placed side by side and no space is retained for side chains than as it would be possible in real structures. This usually results in two compact structures when compared to real proteins. A possible solution is the application of sophisticated energy functions that try to compensate for the effect or the use of side chain models (Martin, 2011).

3.1.1.6.2 Side chain models

In side chain models a monomer is used to represent both the backbone and side chain of each amino acid within the lattice. The backbone monomer represents the C_α – backbone-atom while the side chain monomer abstracts the side chain's centroid, i.e. its geometric centres, or the side chain's centres of mass. Again, both monomers are without volume or mass and have to be neighboured within the lattice. This simple representation approximates the frequency weighted ratio of side-chain to main-chain volumes of the amino acid residues found in proteins. Side chain lattice proteins are more realistic due to the explicit representation of the amino acid side chain. Side chain enables the reconstruction of full atom protein data and disallowed unnatural collapse during structural studies this is as a result of increased complexity, since the side chain structural space grow much faster than for backbone-only models (Martin, 2011). Examples of side chain models in different lattices are given in figure 3.1 and on the right of figure 3.5. Given a lattice L, a side chain lattice protein structure ζ of length x is a sequence of coordinate pairs

$$\zeta = \left(\left(\zeta_1^b, \zeta_1^s\right), \ldots\ldots, \left(\zeta_x^b, \zeta_x^s\right)\right) \tag{3.17}$$

Where ζ_i^b and ζ_i^s denote the backbone and side chain monomer position of *ith* amino acid, respectively (Martin, 2011). Backbone and side chain monomers as well as successive backbone monomers have to be neighbored in the lattice, i.e.

$$\zeta_i^b, \zeta_i^s \in L: \qquad \forall 1 \le i \le n \tag{3.18}$$

$$\zeta_i^b \neq \zeta_j^b \wedge \zeta_i^s \neq \zeta_j^s: \text{ i.e. } \zeta^b \textit{and } \zeta^s \text{ are self-avoiding} \qquad \forall 1 \le i < j \le n \tag{3.19}$$

$$\zeta_i^b \neq \zeta_j^s: \text{ i.e. } \zeta^b \text{ vs. } \zeta^s \text{ is self-avoiding} \qquad \forall 1 \le i, j \le n \tag{3.20}$$

$$(\zeta_i^s - \zeta_i^b) \in N_L: \text{ i.e. side chain } \zeta_i^s \text{ is neighboured to backbone } \zeta_i^b \quad \forall 1 \le i \le n \tag{3.21}$$

$(\zeta_i^b - \zeta_{i-1}^b) \in N_L$: i.e backbone ζ_i^b is neighboured to predecessor ζ_{i-1}^b $\forall 1 < i \leq n$ (3.22)

3.1.2 HP energy lattice models

Since deterministic approaches are not helpful in identifying minimum energy conformations, to find a nondeterministic (stochastic) heuristic approach that can extract minimal energy conformations efficiently is of great importance. The greatest challenge lies in the huge search space, as well as the complexity of the energy surface which contain a lot of local minima and a few global minima (Jingfu et al., 2013). In order to simplify many of the required calculations, we choose the lattice model which captures the main features of the PSP. In this thesis, our research is limited only to the backbone-only HP model in 2D-square lattice; which is shown to yield very good approximations of real protein structures and has been studied widely with a conviction that sequences with non-degenerate ground state are not too scarce (Jingfu et al., 2013; Erik, 2000; Irback et al., 1998; Seno et al., 1996).

To describe a simplified model of the protein folding, some parameters need to be considered such as: (1) How to obtain the primary sequences of the protein? (2) How is the space modelled in which the protein folds? (3) How is the energy of the unique conformation of the protein modelled?

The HP lattice model is a protein model that abstracts from real proteins in two important ways. Firstly, the model is restricted to only the backbone structure of the protein rather than modelling the positions of all atoms of the protein. Hence, these positions are constrained to the points of a lattice. Secondly, only the hydrophobic interaction between the amino acids is modelled, therefore the model distinguishes only two kinds of amino acids namely hydrophobic (H) and polar (P).

The HP lattice model is a standard model from the perspective of statistical mechanics that shows rich thermodynamic behaviours. Since the HP chain is constructed specifically to represent an individual protein. Hence, thermodynamic properties of the chain depend on its chain length and the sequence of H and P monomers uniquely. This model is the most frequently used model, which is based on the observation that the hydrophobic interaction between amino acids is the main driving force for protein

folding. In HP model, amino acids are represented as a reduced set of H and P according to the hydrophobicity of a single amino acid. The Hs form the protein core, while the Ps have an affinity with the solvent and so tend to remain on the outer surface. The H-core formation is the main objective of HP based PSP. A folding of a protein in this model means that amino acids are embedded in the lattice such that adjacent amino acids in the sequence occupy adjacent grid points in the lattice and no grid point in the lattice is occupied by more than one amino acid, a process known as the self avoiding walk.

In this model the space in which the proteins fold is discretised in terms of a 2-D grid lattice. Any conformation of the protein corresponds to an embedding of the amino acid sequence into the grid must obey the following rules:

1. Every position in the amino acid sequence is assigned to one point of the lattice
2. Adjacent positions in the sequence must be assigned to adjacent points of the lattice
3. No two positions of the sequence are assigned to the same lattice point (self-avoiding walk)

This lattice model simplifies a protein‘s primary structure, i.e. the sequence of amino acid to a linear chain of beads. Each bead represents an amino acid which can either be: H (hydrophobic, i.e nonpolar) or P (hydrophilic, i.e polar). The HP lattice model has been described by many authors as the Ising model of protein folding (Michael and Adam, 2012; Ying et al., 2011).

In 1985, Ken Dill proposed the hydrophobic-hydrophilic (HP) model which has been subjected to the huge amount of literatures due to its fundamental role in protein folding modeling (Backofen et al., 2000; Lau and Dill, 1989). The Hydrophobic-Polar (HP) model was first introduced by (Lau and Dill, 1989) to describe globular proteins that have affinity for water. Despite its simplicity, the model is powerful enough to capture a variety of properties of actual proteins. Its energy function focuses on hydrophobic interactions that are known to have a large influence on protein folding and structure. Thus, it abstracts from all the possible attracting and repulsing forces by considering the 20 types of amino acids into two classes based on their hydrophobicity; hydrophobic (H) and hydrophilic/polar (P). The model captures the fact that native

protein folds tend to form very compact cores driven by dominant hydrophobic interactions. Each amino acid both in sequence and structure is referred to as monomer which can be classified either as hydrophobic (H-monomers) or hydrophilic/polar (P-monomers) and two hydrophobic amino acids are said to be in contact if they are adjacent in the fold but nonadjacent in the primary sequence. Since the goal of HP model is the formation of highly compact hydrophobic cores, the optimization function is to maximize the number of contacts between hydrophobic atoms (H-H contacts) (Martin, 2011). With lattice models, an energy value is associated with every conformation taking into account particular neighbourhood relationships of the amino acids on the lattice. The resulting energy function is then given by:

$$E(\sigma,\zeta) = \sum_{1\le i<j\le N} \in (\sigma_i,\sigma_j)\Delta(q_i - q_j) \tag{3.23a}$$

Where the interaction energy between monomers i and j located at positions q_i and q_j, respectively is defined as

$$\Delta(q_i - q_j) = \begin{cases} 1 & if \quad |q| = 1 \quad without \quad \text{cov}\, alent \quad bond \\ 0 & otherwise \end{cases} \tag{3.23b}$$

and,

$$\in(\sigma_i,\sigma_j) = \begin{cases} -1 & if \quad \sigma_i \quad and \quad \sigma_j \quad are \quad both \quad hydrophobic(H) \\ 0 & otherwise \end{cases} \tag{3.23c}$$

This equation shows that only interactions between hydrophobic residues, (i.e given a sequence the energy of a conformation is the number of hydrophobic-hydrophobic contacts) so called HH-contacts, are relevant for the energy calculation. All other interaction types (HP- or PP-contacts) result in no energy contribution. Structures with low energy show therefore a close packing of hydrophobic residues, usually resulting in a globular structure where hydrophobic monomers are gathered in its centres. This phenomenon is intended and does reproduce the hydrophobic cores observed in real globular protein structures (Martin, 2011). Lau and Dill (Lau and Dill, 1989) described the HP lattice model of protein by maximizing the number of H-H contacts subject to the following:

1. Allocation (Assignment); Each amino acid must occupy one lattice point

2. Self-avoiding walk (Non-overlapping); No two amino acids may share the same lattice point, i.e. a given lattice site can be occupied only once

3. Contact (Connectivity); every two amino acids that are consecutive in the protein's sequence must also occupy adjacent lattice points.

The objective function of this model is to maximize the number of H-H contacts that is, the number of adjacencies in the lattice between hydrophobic amino acids. (Greeberg et al., 2004), shows that these three requirements couple with the above objective function give a correct solution to the HP model.

3.1.2.1 The 2D HP lattice model

One of the potent approaches to PSP is to model the free energy of the given amino acid sequence and then to find the minimum energy conformations. The HP model of protein folding is a free energy model where the low energy conformation is favoured with a hydrophobic core by allowing the hydrophobic residues which are less ionic and low affinity for water cluster inside while the hydrophilic/polar residues which are ionic and bond well with water are at the surface. In this model, the conformations of a sequence are restricted to self-avoiding paths on a two-dimensional lattice. A folding of a protein in the 2D HP model is that amino acids are embedded in the lattice such that the adjacent residue in sequence occupy adjacent grid points in the lattice, and no grid point in the lattice is occupied by more than one residues. The problem of PSP is addressed by placing the amino acids to lattice points such that every vertex is visited only once and the local neighbour of amino acid sequence must be maintained in the whole process.

For a given sequence, a structure is called native if and only if its E (HP) is minimal among all structures of sequence (Abkevich et al., 1995). Also, the energy of a given conformation (ζ) is defined as the number of topological neighbouring (TN) contacts between those of Hs (i.e the adjacent hydrophobic amino acids that are not neighbours in sequence) in the lattice, denoted by H-H with $E_{HH} = -1$ and $E_{HP} = E_{PP} = 0$ as shown in Figure 3.6. If a conformation is denoted as $\zeta = \zeta_1, \zeta_2, ..., \zeta_n$, $\zeta_i \in [H, P]$ and

$i \in \{1,2,.....,n\}$ where ζ_i is H if *ith* amino acid is hydrophobic and P if it is polar. Therefore, if we have λ such H-H TN contacts, its energy $E(\zeta) = \lambda(-1)$, the energy evaluation focuses on hydrophobicity only, the used sequence alphabet can be reduced to two based on the amino acids as hydrophobic (H) or polar (P), i.e. $E(\sigma, \zeta) = (H, P)$. Hence, a conformation with the highest number of H-H contacts indicates a conformation with the lowest free energy.

The Hydrophobicity can be classified according to (Ullah et al., 2009) as seen in table 3.2. Other hydrophobicity classification and assignment schemes are by (Sandelin, 2004; Jingfu et al., 2013) e.t.c. In this thesis, we followed the classification by (Ullah et al., 2009) throughout our work. The sequence and structure in HP model have been shown to be protein-like with respect to several properties such as: volume exclusion among residues (Xia and Levitt, 2004b), the structures of HP show protein-like secondary structures (Helling et al., 2001), also, the hydrophobic cores and polar exteriors in HP model are in agreement with real protein structures (Sandelin, 2004).

	H	P
H	-1	0
P	0	0

Figure 3.6. HP energy model (Lau and Dill, 1989)

Table 3.2. The classification of Hydrophobic-Polar of amino acids by (Ullah et al., 2009); for converting an amino acid sequence into an HP sequence and calculating the energy function.

Hydrophobic (H)	Hydrophilic/Polar (P)	
C: Cysteine	A: Alanine	Q: Glutamine
F: Phenylalanine	D: Aspartic Acid	R: Arginine
I: Isoleucine	E: Glutamic Acid	S: Serine
L: Leucine	G: Glycine	T: Threonine
M: Methionine	H: Histidine	
V: Valine	K: Lysine	
W: Trytophan	N: Asparagine	
Y: Tyrosine	P: Prolin	

3.2 General techniques

In this section, time discrete Markov Chains and Markov Chain Monte Carlo methods are introduced. These are essential for folding simulations and local search methods within the discrete structure space of lattice proteins. They are applied to find energy minimal structures or to simulate the folding process. Furthermore, they can be applied to traverse the sequence space when searching for sequences with low-degenerated ground states

3.2.1 Monte Carlo method (MCM) for protein folding

Monte Carlo method is a stochastic technique driven by random numbers and probability statistic to sample conformational space when it is infeasible or impossible to compute an exact result with a deterministic algorithm. It applies the theories of statistical physics to the study of macroscopic systems (disordered system, fluids, and cellular structures) as a result of their large degree of freedom and probabilistic nature

The name "Monte Carlo" (a computer simulation of random numbers, i.e. using random numbers as a tool to compute something that is not random) was originally coined by Metropolis and Ulam during the Manhattan project of World War II as a result of the simulation technique to the game of chance. Monte Carlo simulation (a series of random steps in conformation space, each perturbing some degrees of freedom of the molecule) is a standard method often used to compute several pathways in understanding thermodynamic and kinetics mechanisms of long polypeptide chains in the context of a lattice model. Protein from its unfolding to the native state can be viewed analogously as a phase change problem from the thermodynamic point of view（Oyewande, 2012; David and Kurt, 2000; Oren et al., 2001）. The Lattice Monte Carlo method based on coarse-grained model is effectively useful to model protein folding by circumventing the atomic details.

MCM consists of two steps:

- Generating a new "trial conformation"
- Deciding whether the new conformation will be accepted or rejected

In protein conformation via polypeptide torsion moves, choosing a new trial conformation, we let the computer select an amino acid position along the polypeptide

backbone and randomly select which of the several rotatable bonds in that amino acid will be modified (e.g the ϕ, ψ torsion angle). Finally, a new value is randomly selected for this torsion angle from a predefined set of values.

Once a new "trial conformation" is created, it is necessary to determine whether this conformation will be accepted or rejected. If rejected, the above procedure will be repeated, randomly creating new trial conformations until one of them is accepted. If accepted, the new conformation becomes the "current" conformation, and the search process continues from it. The trial conformation is usually accepted or rejected according to a temperature-dependent probability of the metropolis type

$$\varpi_{m,n}(T) = \begin{cases} 1 & if \quad G(m) \le G(n) \\ \exp^{-(G(m)-G(n))/kT} & otherwise, \quad i.e. \quad G(m) > G(n) \end{cases} \tag{3.24}$$

$$= \min\left\{1, \quad \exp^{-(G(m)-G(n))/kT}\right\} \tag{3.25}$$

Where $\beta = \frac{1}{KT}$ and ΔG is the change in the potential energy. This means that if the energy of the new trial conformation is lower than that of the current conformation, $\Delta G < 0$, it is always accepted. But even if the energy of the trial conformation is higher than the current energy, $\Delta G > 0$, there is a certain probability, proportional to the Boltzmann factor, that it will be accepted. To detect whether a higher energy trial conformation is accepted, a random number $\Re$ in the range [0, 1] is selected and compared to the metropolis probability in the above equation. If $\Re < P$, the conformation is accepted; otherwise it is rejected. This acceptance probability satisfies the principle of detailed balance, ensuring that if the process continues for a long enough time, then a stationary solution will be achieved (Oren et al., 2001).

Temperature plays a very important role in MC just as in molecular dynamic (MD) simulation. At high temperature there is a significant probability of climbing up energy slopes, allowing the search process to cross high energy barriers, although MC simulations tend to move toward low energy states. This probability becomes significantly smaller at low temperatures, and it vanishes altogether in the limit of $T \to 0$, where the method becomes equivalent to a minimization process. Thus, high

temperature MC is often to sample broad regions of conformational space. The popularity of MC simulations is as a result of its ease of use and their good convergence properties. Nevertheless, straightforward application of MC methods to biomolecules is often challenging due to very low acceptance ratios which significantly reduce the efficiency of the method. The reason for the low acceptance ratio, which is the ratio between accepted MC moves and total MC trial moves is the compact character of most biomolecules. This means that many move attempts end up rejected as a result of clashes between the molecules. Torsion move-sets are recommended to partially ameliorate the problem (Oren et al., 2001).

MC methods have the ability to analyse thermodynamic equilibrium but not suitable for investigating dynamic phenomena, since dynamic properties of a system depend on time; Unlike the MD method which is useful for thermodynamic equilibrium but are more advantageous for investigating the dynamic properties of a system in a nonequilibrium situation. The MC method generates a series of microscopic states under a certain stochastic law, regardless of the equation of motion of particles. Since the MC method does not use the equations of motion, it cannot include the concept of explicit time.

3.2.2 Classes of Monte Carlo method

MCM are classified into three categories; (i) Static (ii) Quasi-Static (iii) Dynamic

The static methods are those that generate a sequence of statistically independent samples from the desired probability distribution (ξ).

Quasi-static methods are those that generate a sequence of statistically independent batches of samples from the desired probability distribution (ξ).

Dynamic methods are those that generate a sequence of correlated samples from some stochastic process (usually a Markov process) having the desired probability distribution (ξ) as its unique equilibrium distribution

Several types of moves are common for Monte Carlo based on the lattice model. These are:

- Pull move
- Corner moves (a flip of the residue across the diagonal of a square formed by neighbouring bonds)
- Crankshaft rotations (rotation of the beads i+1 and i+2, while keeping the adjacent beads i and i+3 fixed).
- Rotation of the end beads.

Albeit, the particular choice of moves or their probabilities will alter the local structural dynamics, but the thermodynamic and the kinetic properties remain unchanged as long as the moves are ergodic.

3.2.3 Markov chains

A Markov process is a nondeterministic (stochastic) process that has the Markov property. That is, Markov property exists for stochastic process if the conditional probability distribution of future states of the process given all the past states depends only upon the present/current state.

$$P(\sigma_{k+1} \mid \sigma_{0,}\sigma_{1,}.....,\sigma_k) = P(\sigma_{k+1} \mid \sigma_k) \tag{3.26}$$

A Markov chain is a sequence of random variables $\{\sigma_k\}$ with no or very restricted history, producing a series of discrete states generated by a Markov process. Where σ_k is called the state of the process at time k

Within this thesis only 1st-order time-homogeneous Markov Chains with no history are used, i.e. "the conditional probability distribution of future states of the process depends only upon the present/current state". The latter is known as the Markov Property. Furthermore, we will focus on time-discrete Markov Chains where we consider only discrete time points and enforce the random process to be in a distinct state at each time point. The transition probability density (transition karnel) $P(n \mid m)$ is the probability density that the process moves from m to n. The transition kernel must satisfy:

$$\int P(n \mid m)dn = 1 \tag{3.27}$$

Hence, Markov chain is said to be homogeneous or stationary if the transition karnels do not depend on the time. A condition called ergodicity is satisfied in our Markov process to reach any state of the system from any other state if we run it for long enough. If a Markov chain is ergodic, then there exists a unique steady state distribution Ω independent of the initial state.

$$\Omega(n) = \int P(n \mid m)\Omega(m)dm \tag{3.28}$$

Another condition on Markov chain is the condition of detailed balance (reversible condition). This condition ensures the generation of Botlzmann probability distribution at equilibrium in the sense that the rate at which the system makes transitions into and out of any state must be equal. This condition guarantees the invariance of Ω under the transition kernel

$$P(m \mid n)\Omega(n) = P(n \mid m)\Omega(m), \forall m, n \tag{3.29}$$

i.e the unconditional probability of moving n to m is equal to the unconditional probability of moving m to n, where m and n are both generate from Ω. The conversion of Markov Chains into a distribution where detailed balance is fulfilled the so called steady state, can be used to sample from that distribution. This is done by Markov Chain Monte Carlo methods.

3.2.4 **Markov chain Monte Carlo methods** (MCMC)

MCMC is a very general method to sample from probability distributions $P(\sigma)$ that is difficult to directly sample from by mean of simulation. The approach is to use a sequence of samples (σ_k) from a Markov chain, such that the sequence converges on $P(\sigma)$ as $k \to \infty$. We construct a reversible Markov chain that has the desired distribution as its equilibrium distribution. As discussed above, any start probability distribution for a reversible Markov Chain will converge to its stationary distribution. Thus, the state of the Markov Chain reached after a large number of steps can be assumed to be seen according to the equilibrium distribution. Since the latter equals the desired distribution, the reached state is then used as a sample from the distribution. The quality of the sample improves as a function of the number of steps (Clote and Backofen, 2000). In the following, some Markov Chain Monte Carlo algorithms that

are used within the thesis are introduced (Oyewande, 2012; Newman and Barkema, 1999).

3.2.4.1 Metropolis-coupled Markov chain Monte Carlo (MCMCMC)

Metropolis (Metropolis et al., 1953) introduced a method to obtain a sequence of random samples from a Boltzmann distribution. Subsequently, the algorithm was generalized by Hastings (1970). The Metropolis Monte Carlo Method is a general scheme for generating a sequence of conformations beyond short chains weighted according to the desired Boltzmann distribution. Given an energy function E that assigns energy to each state of a Markov Chain the stationary distribution Ω^* is given by the Boltzmann distribution:

$$\Omega^* = \frac{1}{Z(T)} \exp[-\Delta G(\sigma)/k_B T], \tag{3.30}$$

$$Z = \sum \exp[-\Delta G(\sigma)/k_B T] \tag{3.31}$$

Where Z is the canonical partition function at constant temperature T using the Boltzmann constant k. In lattice models, the conformation state is discrete, as a result of this all conformations can be enumerated and thermodynamically averages can be calculated exactly.

Given a system Φ and its neighbouring system Θ, where $m \in \Phi$ and $n \in \Theta$. The acceptance probability of the transition from state $m \in \Phi$ to a neighbour $n \in \Theta$ defined at temperature T by the Metropolis criterion $\varpi(T)$ as follows

$$\varpi_{m,n}(T) = \begin{cases} 1 & if \quad G(m) \le G(n) \\ \exp^{-(G(m)-G(n))/k_B T} & otherwise, \quad i.e. \quad G(m) > G(n) \end{cases} \tag{3.32}$$

$$= \min\left\{1, \quad \exp^{-(G(m)-G(n))/k_B T}\right\} \tag{3.33}$$

Thus, a transition to a state with lower energy is always accepted, while the transition to a state with higher energy is only accepted according to the probability defined by

the temperature dependent Boltzmann weight of the energy difference. If the transition is rejected, the current state is maintained.

3.2.5 Self avoiding walk (SAW)

A self-avoiding walk is one that requires only one monomer per lattice site, i.e. two different monomers cannot occupy the same space (Andrej, Eugene, & Martin, 1995). SAW is a random walk that is prohibited from revisiting an old site previously visited. SAW has various contexts of applications in physical and biological Sciences e.g Physics, Chemistry and Biology. SAW was first proposed about 5 decades ago as a model of a linear polymer molecule in a good solvent such as DNA, RNA and proteins (Madras and Sokal, 1988; Madras and Slade, 1993; Alan, 1996).

Since this time a lot of progress has been made in the development of new and more efficient algorithms for simulating the self-avoiding walk which involves firstly by generating a random step, and accept it if an old site is not revisited, and generate another step otherwise. The process is repeated until N acceptable steps are obtained which is one state of a polymer chain. In order to calculate the statistical averages, we need to have an ensemble of polymer chains by generating a sufficient number of chains independently. These new algorithms reduce the CPU time for generating an "effectively independent" N-step SAW from $N^{\approx 3.2}$ to $N^{\approx 2}$ or even N. SAW lives on a discrete lattice and has non-tetrahedral bond angles (e.g 90^0 and 180^0 on the simple cubic lattice) with energy independent of the bond rotation angles, and a repulsive hard-core monomer-monomer potential. The SAW has been studied extensively by a variety of methods. Thus, considerable work has been devoted to developing numerical methods for the study of long SAWs (Madras and Sokal, 1988; Madras and Slade, 1993).

3.2.5.1 Numerical methods for the Self-avoiding walk

This method is basically of two categories: exact enumeration and Monte Carlo.

3.2.5.2 Exact enumeration

In an exact-enumeration study, one first generates a complete list of all SAWs up to a certain length (usually $N \approx 15-35$ steps), keeping track of the properties of interest

such as the number of such walks or their squared end-to-end distances. An extrapolation is performed to the limit $N \to \infty$, using techniques such as the ratio method or differential approximats (Alan, 1996).

3.2.5.3 Monte Carlo method

In contrast, the Monte Carlo method aims to probe directly the regime of fairly long SAWs (usually $N \approx 10^2 - 10^5$ steps). It generates random numbers instead of enumeration. An extrapolation to the regime of extremely long SAWs is still required, but this extrapolation is much less severe than in the case of exact-enumeration studies, because the point of departure is already much closer to the asymptotic regime. Monte Carlo studies of the self-avoiding walk go back to the early 1950's and indeed these simulations were among the first applications of a new invention, like, the "high-speed electronic digital computer to pure science". These studies continued throughout the 1960's and 1970's and benefitted from the increasing powerful computers that became available. From the beginning of 1980's, vast progress has been made in the development of new and more efficient algorithms for simulating the self-avoiding walk. These new algorithms reduce the CPU time for generating an "effectively independent" $N - step$ SAW from $\sim N^{\approx 3.2}$ to $\sim N^{\approx 2}$ or even $\sim N$ (Alan, 1996).

In this thesis, we are concerned with self-avoiding walks in lattice model, particularly the d-dimensional rectangular lattice Z^d with origin -1. SAW is very important as a model for the spatial arrangement of linear polymer molecules in chemical physics in which the walk represents a molecule composed of many (perhaps 10^5 or more) monomers linked in a chain, with the self-avoidance constraint reflects the fact that no two monomers may occupy the same position in space. From heuristic arguments and empirical studies, almost nothing is known in rigorous terms about the above problems for the most interesting cases of low-dimensional lattices with $2 \le d \le 4$. SAWs in one dimension square lattice are trivial; it has only two different self-avoiding walks with fixed step size if the first move is to the left, every subsequent step is to the left; if the first move is to the right, the walk continues to the right. While in SAW 2-D square lattice, the walker must step North, South, East or West with equal probability and at the same time the walker must avoid previously visited locations. The first step has 4 allowed directions, NSEW. Every subsequent step has 3, 2, 1, or 0 allowed directions.

In higher dimensions rather more is known, essentially because the self-avoidance constraint becomes less significant and the behaviour resembles that of simple (non-self-avoiding walks, which are well understood) (Dana and Alistair, 2000). We present Monte Carlo algorithms for approximating the number of self-avoiding walks of a given length for a given dimension d, and for generating self-avoiding walks of a given length almost uniformly at random.

3.3 This work: Optimized searching procedure

We use local search based on self-avoiding move-biased Monte Carlo (MBMC) simulation methods by finding all sequences which satisfy the thermodynamic requirement with unique ground state energy minimal in a reasonable time through all the sets of contacts for each sequence and then calculate the energy of each of the sets. In the Monte Carlo step, a move is selected at random until a move is found that conserves the unit bond lengths and does not result in more than one monomer per lattice site a process known as self-avoiding walk. Immediately the move is found, the corresponding energy change in the system ΔE, is evaluated as in equation 3.23a. A sequence folds in a given Monte Carlo simulation if it finds the native conformation (the compact self-avoiding chain with the lowest energy) within a reasonable small number of Monte Carlo steps. Each MC must update the current conformation using the coupled diagonal- pull move search strategy. The MC search is based on the idea of iteratively improving a given candidate solution by exploring its local neighbourhood. In PSP the neighbourhood of a conformation can be thought to consist of slight perturbations of the respective conformation. The neighbourhood (move sets) for PSP specifies a perturbation as a feasible change from a current conformation ζ at time t to a valid conformation at time $t+1$. Hence, the neighbourhood of a conformation ζ is a set of valid conformation ζ' that is obtained by applying a specific set of perturbations to ζ. In this thesis, we consider two such neighbourhoods; the pull moves and the diagonal moves (corner-flip) for 2D model.

The pull moves introduced unprecedented by (Lesh et al., 2003) to update the conformation. In this work, we choose this move for our simulation. The pull move has been proven to be very proficient, complete, reversible, and satisfy ergodicity for

square and cubic HP lattice model under the variability of local search methods. Also, Successive pull move never generates infeasible conformation.

The basic idea of the pull move on 2D square lattice which is feasible only when there is at least one free vacancy of its neighbours. We describe this process by choosing randomly a vertex from the chain with length n to ensure a free lattice point in the grid adjacent with either the predecessor or successor of the vertex in the chain and then move it to this free lattice. This move might alter the chain, so we need to repair the chain by pulling the chain, i.e. the old position of the moved vertex will be occupied by its successor (or predecessor), again leaving a free position where the next vertex of the chain is moved, and so on, until a valid conformation is reached (see figure 3.7). During the process of the pull move, if the ith amino acid is moved first, we define the pull move as the pull move of the ith amino acid. Consider the pull move of the ith amino acid whose position is (x_i, y_i), if there exist an index $i \in \{2,.....,n-1\}$ and a vertex $\Re$ which is empty and adjacent to both the ith and $(i-1)$ amino acid, then forward pull move is feasible. Also, if $\Re$ is adjacent to the $(i+1)$ amino acid, we move directly amino acid i to $\Re$ which gives a new conformation this is called direct move. Otherwise, we move amino acid i to $\Re$, $(i+1)$ to i, $(i+2)$ to $(i+1)$, and so on, until we have a valid conformation.

Diagonal moves (corner flip) can potentially be performed on any residue excluding the end residues. For this to be possible, the residue i must be mutually adjacent to $i+1$, and $i-1$ lattice. A diagonal move occurs between $i+1, i$, and $i-1$ If the mutually adjacent position is empty, residue i can be moved to it (see figure 3.7)

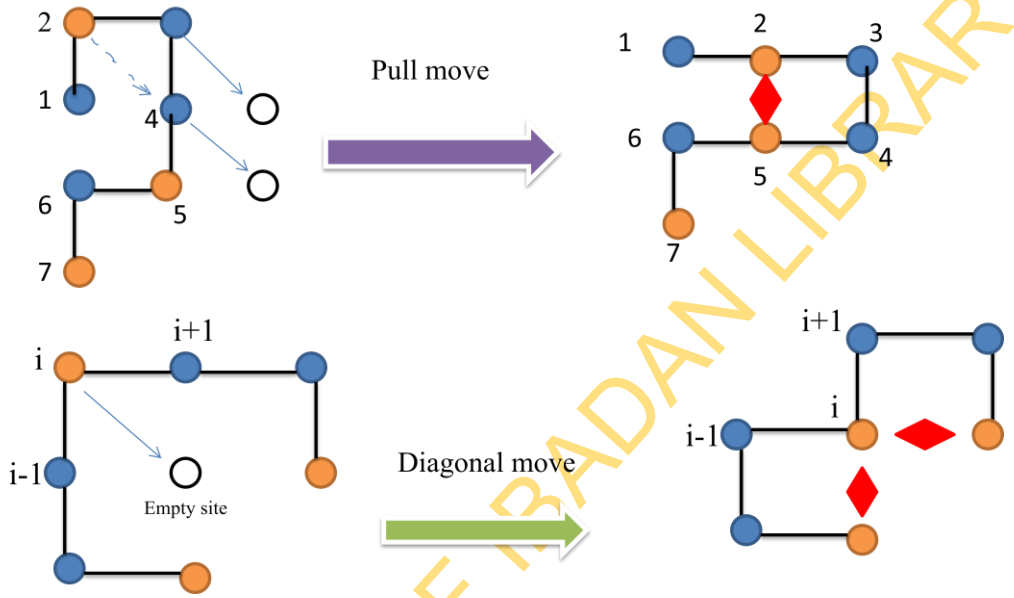


Figure 3.7. Example of pull move and diagonal move in 2D space. The blue atoms are Ps, the Orange atoms are Hs while the **black** lines are the peptide bond and the red diamonds are the H_H contact

3.4 Procedure of the algorithm

We put forward an improved MC method called move-biased Monte Carlo simulation (MBMC) based on self-avoiding walk and the neighbourhood search strategy in our algorithm. The improved method is developed for the protein folding problem in the HP lattice model. The calculating procedure is presented as follows.

1. The adopted method generates an initial conformation 'ζ' following a self-avoiding walk on square lattice points. It places the first amino acid at (0, 0) followed by a random selection of a basis vector to place the amino acid at a neighbouring free lattice point. The mapping proceeds until a self-avoiding walk is found for the whole protein sequence.
2. We compute the energy $E(\zeta)$ as a self-avoiding walk on square lattice point for each conformation.
3. Let $i = 1$
4. We execute coupled (diagonal-pull) moves for all legal move positions of the ith amino acid of the current conformation ζ. If the coupled move is executed successfully, we compute the energies of the corresponding legal conformations obtained by coupled moves and pick out the conformation with the lowest energy as a newly updated conformation of ζ, expressed as $\zeta^{\otimes}$
5. We compute $E(\zeta^{\otimes})$
6. If the $E(\zeta^{\otimes}) < E(\zeta)$, then let $\zeta = \zeta^{\otimes}$, $E(\zeta) = E(\zeta^{\otimes})$ and go to the last procedure; otherwise go to (7)
7. If $r(0,1) < \exp\{[E(\zeta) - E(\zeta^{\otimes})]/k_B T\}$, where $r(0,1)$ denotes a random number between 0 and 1, then let $\zeta = \zeta^{\otimes}$, $E(\zeta) = E(\zeta^{\otimes})$, and go to (9); otherwise go to (8)
8. From the current conformation ζ, we produce the new conformation $\zeta^{\otimes}$ by coupled move search strategy. If $\zeta^{\otimes}$ is a legal conformation, then we update the current conformation ζ with $\zeta^{\otimes}$, i.e we let $\zeta = \zeta^{\otimes}$ and $E(\zeta) = E(\zeta^{\otimes})$
9. Stop if the move is ergodic; otherwise we go step (2)

3.4.1 Periodic boundary conditions

A periodic boundary condition (as shown in figure 3.8) was applied since our simulations are performed on finite systems. We set the boundary conditions to treat the edges or boundaries of the lattice; these boundaries can be effectively eliminated by wrapping the d-dimensional lattice on a (d+1)-dimensional torus on the row and the column. This boundary condition often known as 'periodic boundary condition' (PBC) is the process of identically repeating the basic cell in an infinite number of times so that the first monomer in a row/column sees the last monomer in the row/column as a nearest neighbour and vice versa. In writing the algorithm, we ensured that we applied the periodic boundary conditions to the array to guarantee the monomers on one edge (uppermost/leftmost) of the lattice are neighbours of the corresponding monomers on the other edge (lowermost/rightmost). This ensures that all monomers have the same number of neighbours and local geometry. Other boundary conditions like skew-periodic, free-periodic, anti-periodic, mean-field and free-edge are also available for studying properties or geometrical nature of systems rather than their bulk properties (Kurt and Dieter, 2010; Oyewande, 2012).

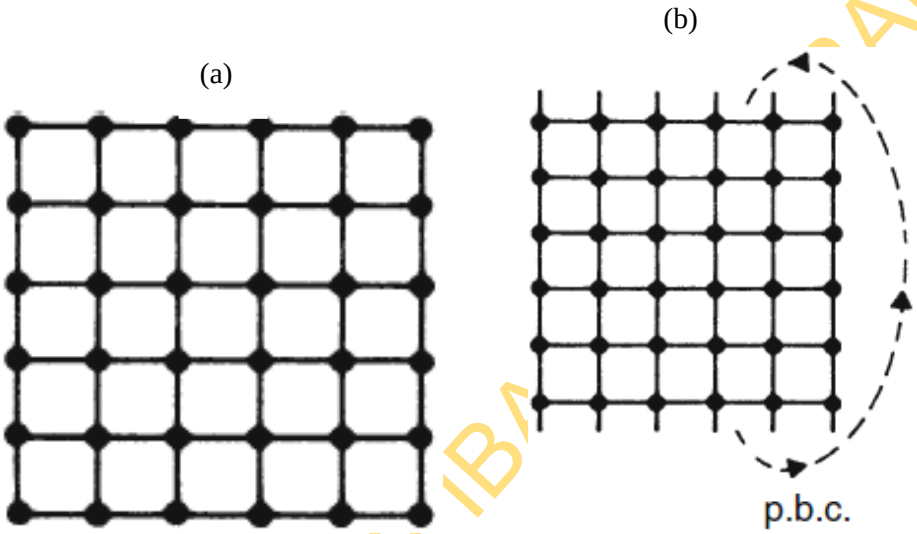


Figure 3.8. Boundary conditions for (a) a free boundary and (b) a periodic boundary

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Protein-like sequence in HP model

The folding of a protein in the HP lattice model means that amino acids are embedded in the lattice such that adjacent amino acid in the sequence occupy adjacent grid points in the lattice and no grid point in the lattice is occupied by more than one amino acid a process called self-avoiding walk (as shown in figures 4.14 – 4.22). We demonstrate our strategy of protein-like sequences in simplified lattice protein model.

Currently the investigation of folding process of real protein neither via full simulation nor to calculate their native structure of protein directly is not feasible (Martin, 2011). To reduce this complexity lattice protein model using computational methods is adopted with the following conditions:

1. A model-specific classification of protein-like sequence have to be calculated for the model
2. Identification of protein-like sequence
3. The identified protein-like sequences must satisfy thermodynamic requirements, i.e. possess a (unique) stable native structure
4. The identified protein-like must be able to satisfy kinetic requirement, i.e. must be able to fold to this structure within a short time interval

A sequence that fulfils the condition three and four above is called protein-like sequence.

For example, the biological sequence of real protein 2.5D (PD ID: 2m7t) with residues 33 obtained by NMR is given as "$\sigma_{orig.}$:

GCPQGRGDWAPTSCSQDSDCLAGCVCGPNGFCG",

and the converted HP protein model sequence is given as "σ_{HP}: PHPPPPPPHPPPPHPPPPPHHPPHHHPPPPHHP" .

Table 4.1 shows the benchmark sequences used for the experiments, their PDB IDs and their corresponding derived HP sequences taken from (Jacek et al., 2004). A sequence is a protein-like if it can adopt only one optimal structure i.e when the degeneracy $(\Phi(\sigma)=1)$. The degeneracy (Φ) of a protein, which is a measure of its thermodynamic stability is a veritable tool in the classification of sequences as protein-like or not. A protein sequence σ is degenerate if their exit more than one optimal structure with minimal energy (the same topological contact). This degeneracy can be used to filter the large number of conformations to a smaller number and as a guide to obtain the stability of a protein; that is, its thermodynamics status. Our interest is in non-degenerated sequences, i.e. sequences with very low degeneracy. The approach adopted is designed in the 2D-HP backbone model using a move biased Monte Carlo simulation (MBMC) based on self-avoiding walk. This program enables folding studies in the field of lattice proteins. The outcome of our method is a data set consisting of protein-like sequences mainly classified as good folder, bad folder and unclassified as shown in table 4.2.

Table 4.1.The standard benchmark sequences for 2D HP lattice model taken from (Jacek et al., 2004; Unger and Moult, 1993; Toma and Toma, 1996).

IN	N	PDB ID	Protein Sequence (H-hydrophobic, P-polar)	E[a]
1	20	SI-1	$(HP)^2PH^2PHP^2HPH^2P^2HPH$	-9
2	24	SI-2	$H^2P^2HP^2HP^2HP^2HP^2HP^2HP^2H^2$	-9
3	25	SI-3	$P^2HP^2H^2P^4H^2P^4H^2P^4H^2$	-8
4	36	SI-4	$P^3H^2P^2H^2P^5H^7P^2H^2P^4H^2P^2HP^2$	-14
5	48	SI-5	$P^2HP^2H^2P^2H^2P^5H^{10}P^6H^2P^2H^2P^2HP^2H^5$	-23
6	60	SI-7	$P^2H^3PH^8P^3H^9(HP)^2P^2H^{12}P^4H^6PH(HP)^2$	-36
7	64	SI-8	$H^{11}(HP)^3P(H^2P^2)^2HP^2(H^2P^2)^2HP^2(H^2P^2)^2(HP)^2H^{12}$	-42
8	85	SI-9	$H^4P^4H^{12}P^6(H^{12}P^3)^3HP^2(H^2P^2)^2HPH$	-53

[a]putative energy value; IN means instances and N means sequence number

Table 4.2. Classification of protein like sequences for the benchmark instances in an HP lattice model.

	N = 20 (474)	N = 24 (230)	N = 25 (200)	N = 36 (230)	N = 48 (200)	N = 60 (100)	N = 64 (100)	N=85 (100)
2^N	$104x10^4$	$167x10^5$	$335x10^5$	$687x10^8$	$281x10^{12}$	$115x10^{16}$	$184x10^{17}$	$386x10^{26}$
#good	35	22	11	39	83	18	6	37
#Bad	377	172	135	121	137	21	32	42
#UC	58	36	54	70	80	61	62	21

UC: Unclassified; values in bracket are the number of runs scaned per sequence

Good folders are those sequences that reach the ground state either sequentially or through global folding. The bad folders cannot adopt the native structure in short time interval, while the unclassified ones are those that are in between the good and the bad structures.

The main goal is to separate the non-degenerated sequences into the two sets good and bad folders. According to (Mazzoni and Casetti, 2006), good folders are presumed to be the more protein-like sequences as a result of the ability to fold into their native conformation in a quick succession i.e "a given conformation "ζ" is said to be good or designable if there is at least one sequence "σ" out of the possible 2^N that has the "ζ" as its nondegenerate ground state" while the bad folders represent random protein sequences that are able to form a random coil but no stable functional native structure (Irback and Sandelin, 1999; Seno et al., 1996).

The properties of good and bad folders are related to the attribute of the energy landscape, often called folding funnel (as shown in figure 4.1 - 4.9). In the case of good folders, the folding funnels dominate the landscape and engineer the folding process downwards to the native fold (Klemm, 2008). Over the years, many researchers are interested in the final structure of the folding and not how it got there. Generally it is believed that the final structure does not cause disease, but rather the intermediate steps along the way (as shown in figure 4.1 – 4.9) caused the toxicity found in the disease, hence the path along the folding play a much more role than the final structure. An intermediate state is an essential stepping-stone that guides a protein through the folding process to the native state and a critical species in misfolding process that lead to aggregation and diseases.

We are going to do a good and the bad folding classification for the large set of non-degenerated HP sequences in table 4.1. For each protein sequence, we perform a series of maximum of 500 very short folding simulations with 100000 MC steps at the given value of the folding temperature $k_B T_f$ and adopted the process for each protein model and length in table 4.1 to find the global energy minimum.

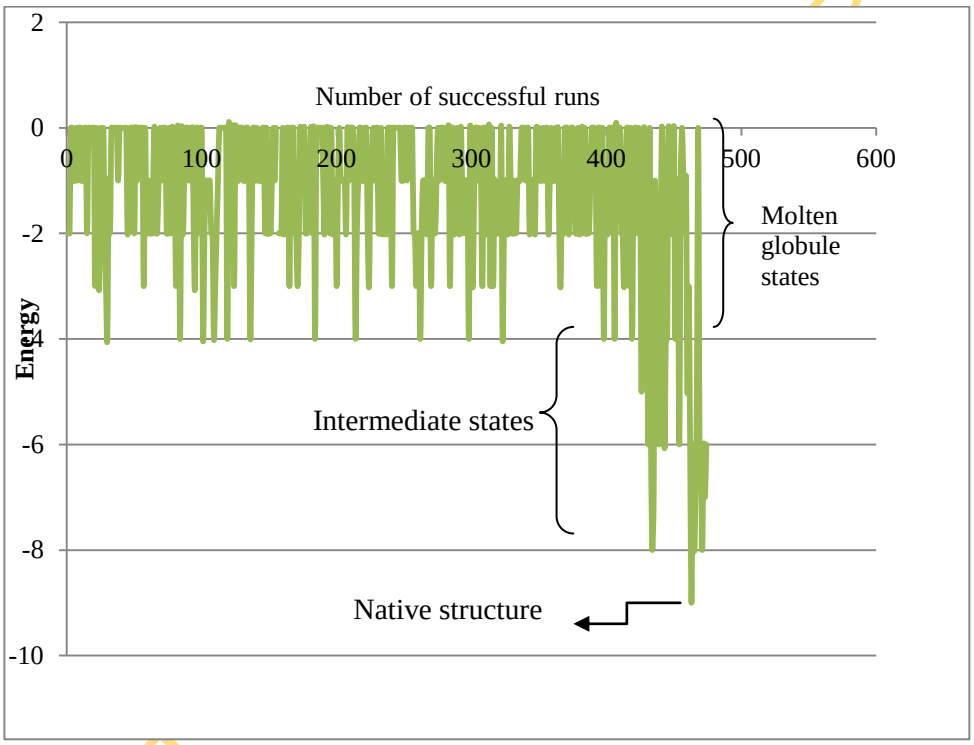


Figure 4.1. The energy landscape for N = 20. Each funnel represents a conformation of energy against the number of iterations

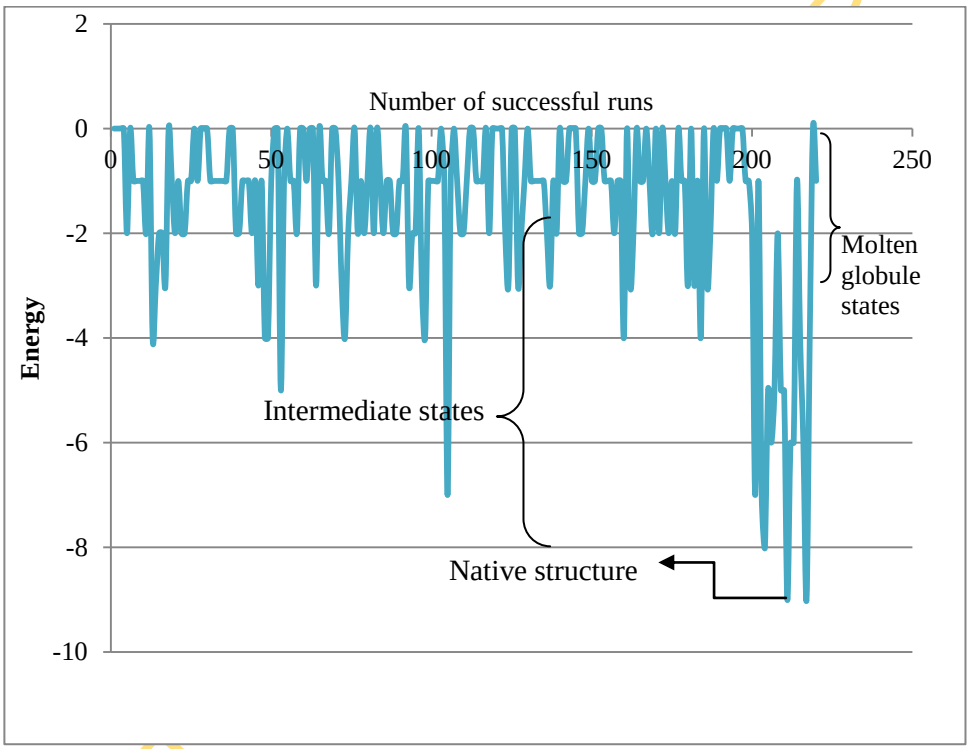


Figure 4.2. The energy landscape for N = 24. Each funnel represents a conformation of energy against the number of iterations

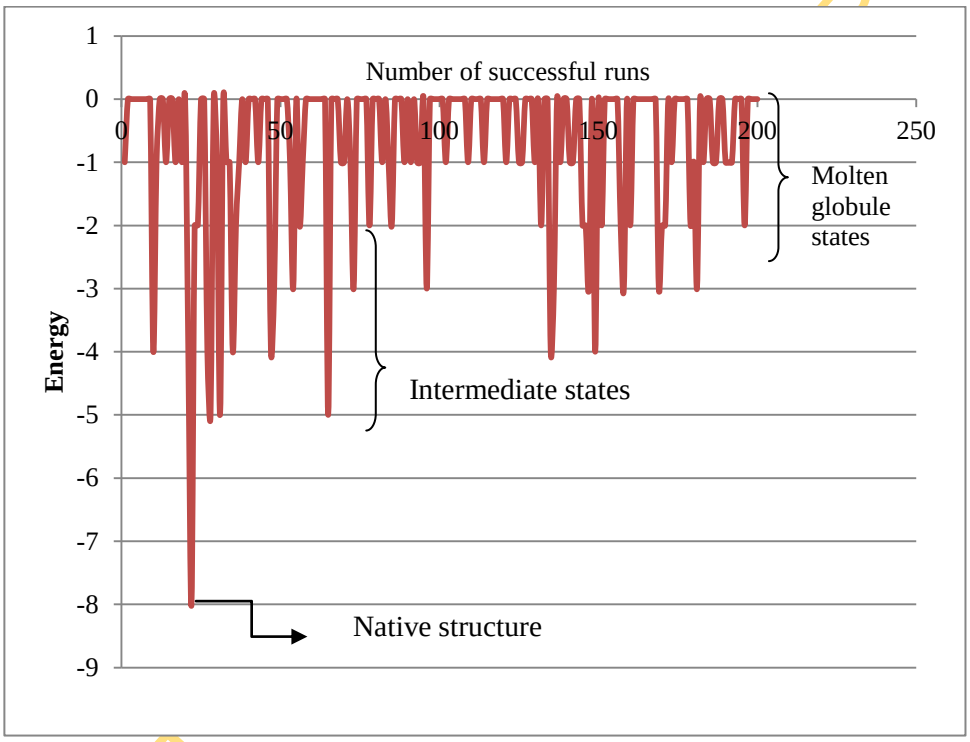


Figure 4.3. The energy landscape for N = 25. Each funnel represents a conformation of energy against the number of iterations

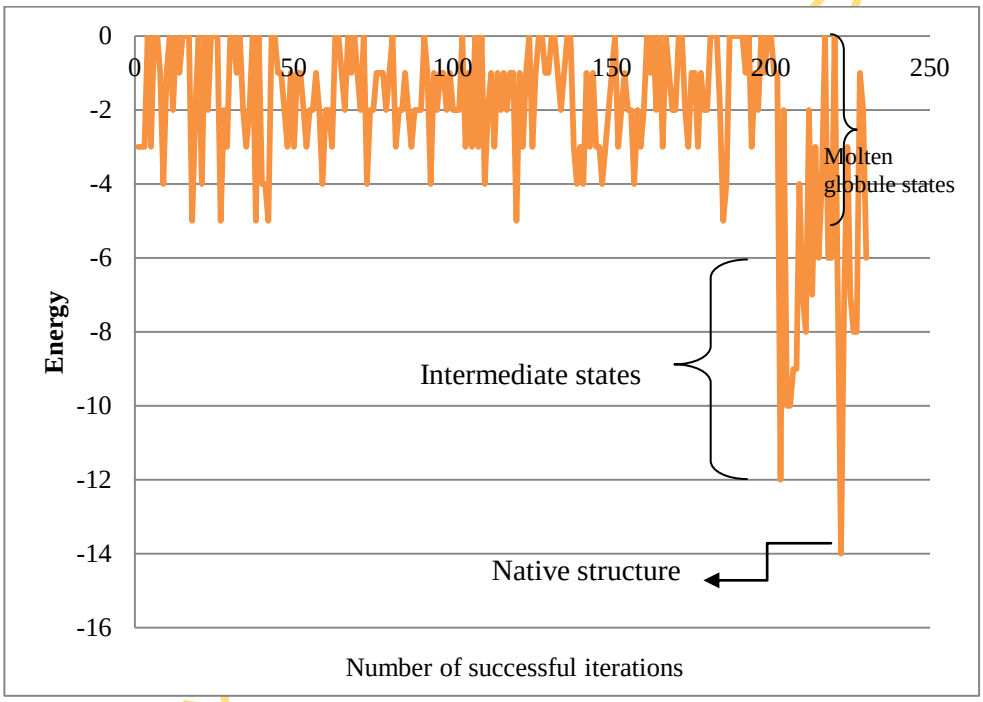


Figure 4.4. The energy landscape for N = 36. Each funnel represents a conformation of energy against the number of iterations

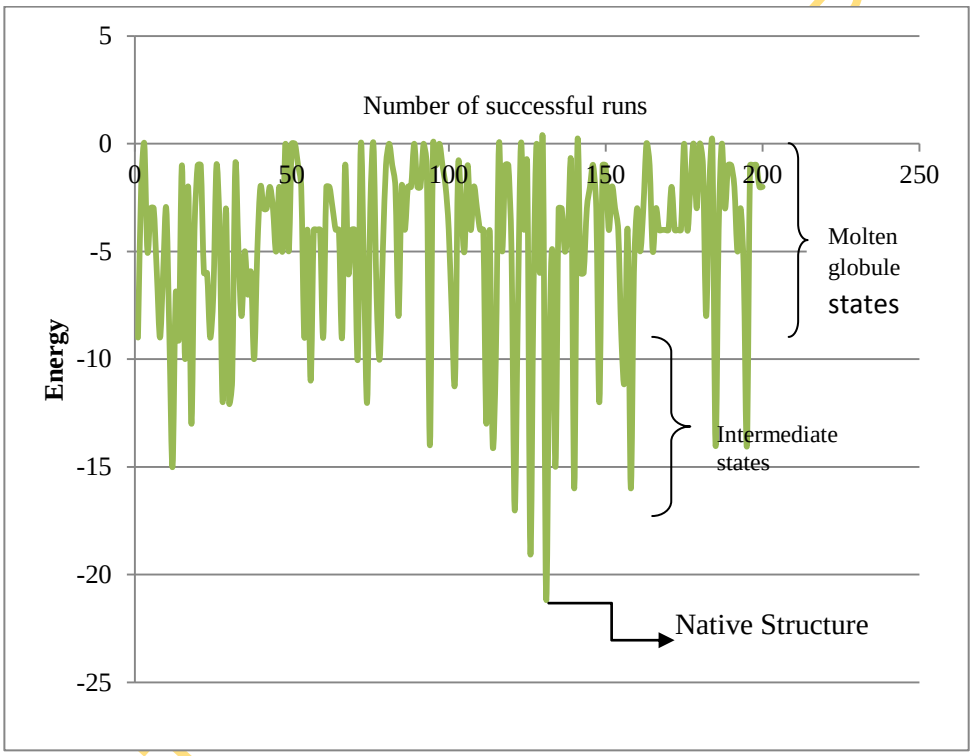


Figure 4.5. The energy landscape for N = 48. Each funnel represents a conformation of energy against the number of iterations

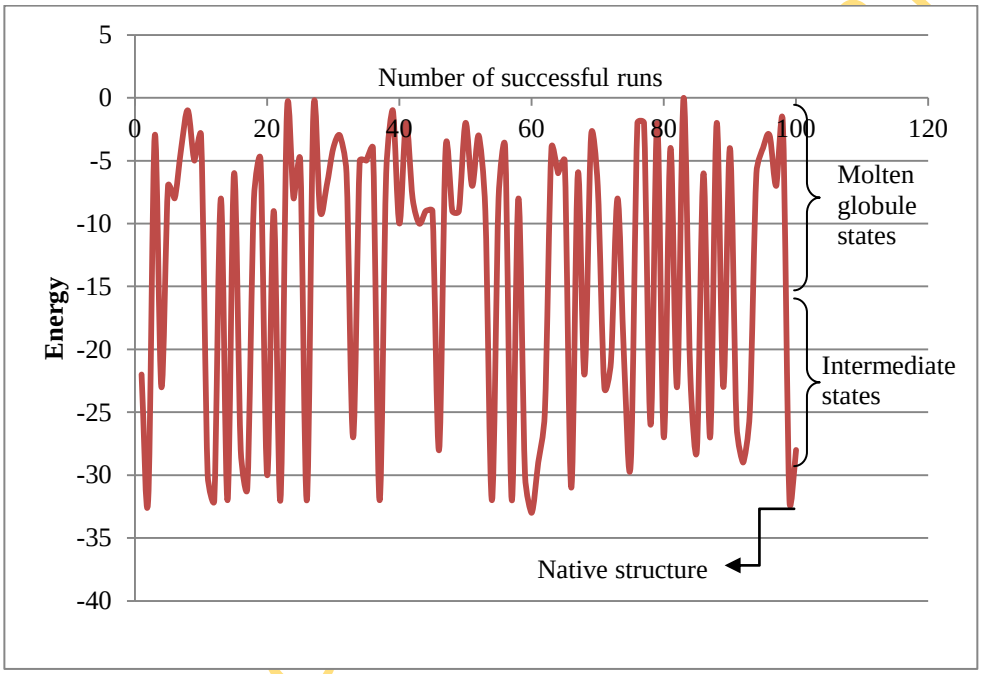


Figure 4.6. The energy landscape for N = 60. Each funnel represents a conformation of energy against the number of iterations

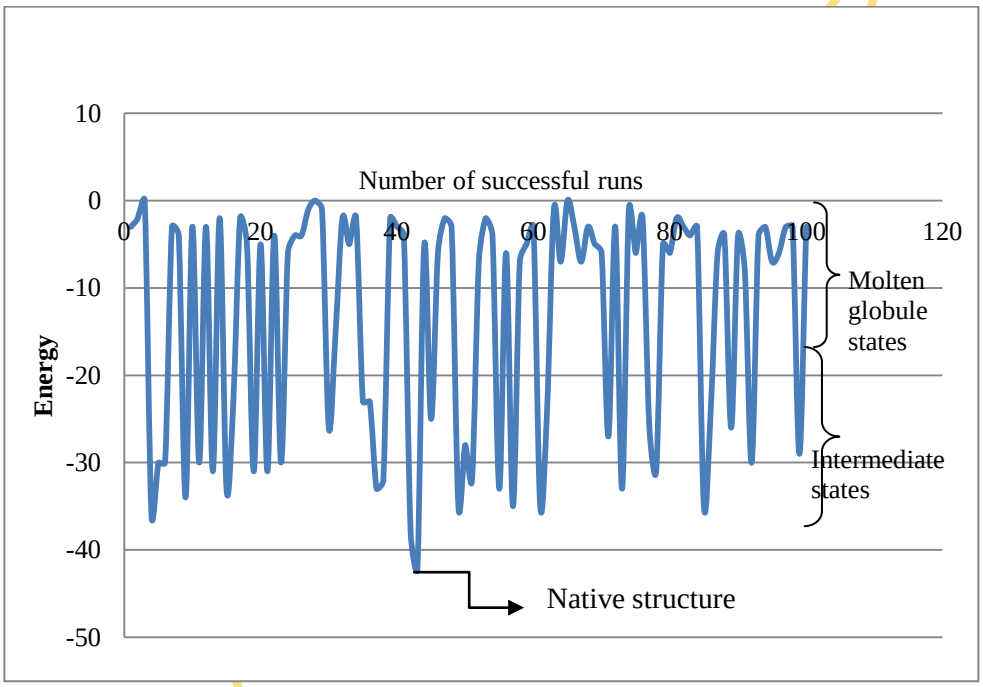


Figure 4.7. The energy landscape for N = 64. Each funnel represents a conformation of energy against the number of iterations

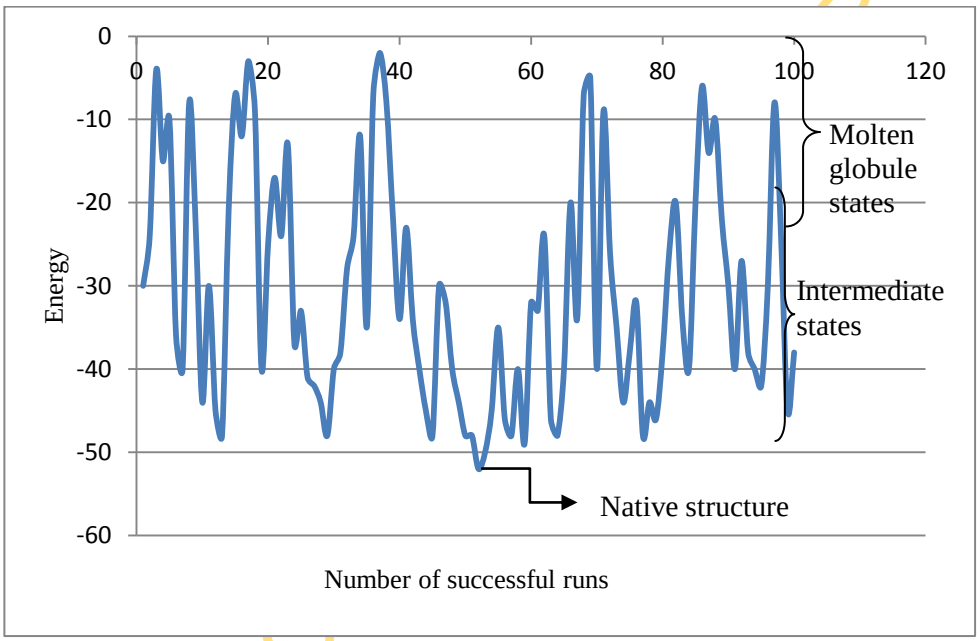


Figure 4.8. The energy landscape for N = 85. Each funnel represents a conformation of energy against the number of iterations

4.2 **Numerical results**

Protein folding problem in the HP lattice model is NP-hard (NP complete) i.e the problem considered cannot be solved optimally within a reasonable time (polynomial time) (Berger and Leighton, 1998; Crescenzi et al., 1998; Unger and Moult, 1993). Consequently, heuristic methods may solve the problem better and quicker, but don't guarantee finding a global optimal solution because their solution is locally optimal. In this thesis, we implement move biased Monte Carlo (MBMC) on 2D-square HP lattice models. This method incorporates MC in addition to the coupled (diagonal-pull) neighbourhood, move search strategy for the protein structure prediction. The MBMC is a class of heuristic global optimization and a generation of Monte Carlo (MC) method. This method is tested with short chain residues of lengths $N=16$ taking from (Seno et al., 1996) and $N=18$ from (Irback et al., 1998); where (i) exact enumeration are feasible (ii) Suitable structures are trivial (iii) authentication of the folding properties are trivial and (iv) the residues have been used extensively as the benchmark for algorithm performances. For large chains $N\geq 18$ it is infeasible to explore the entire sequence space with any method.

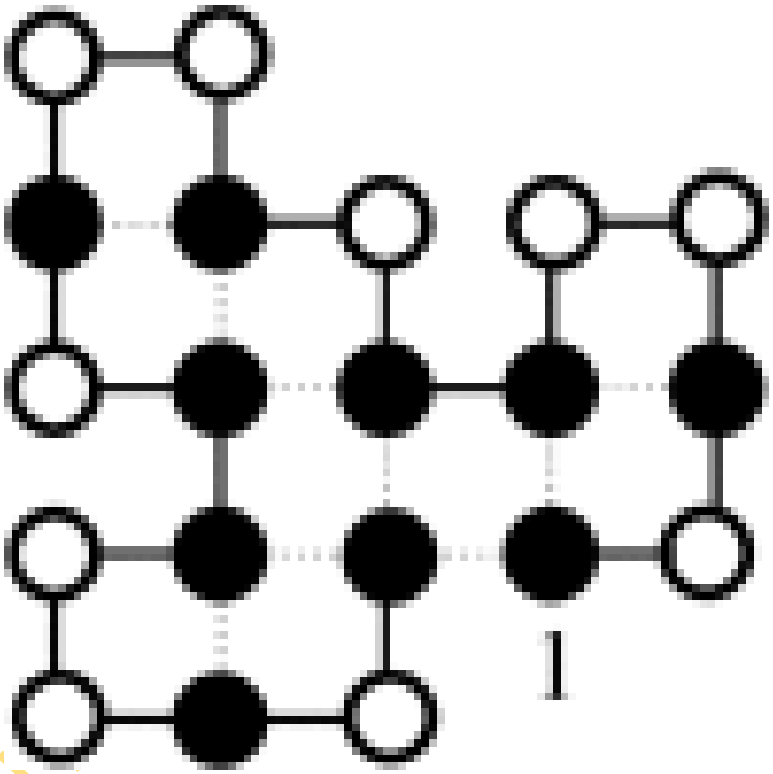


Figure 4.9. Example of protein conformation in the 2D HP model, The sequence is from instance 1of table 4.1 HPHPPHHPHPPHPHHPPHPH; the black and the white circles represent hydrophobic and polar amino acids respectively, The dotted lines represents the H-H contacts underlying the energy calculation. The energy of this conformation is -9, which is optimal for the given sequence (Alena and Holger, 2005).

The implementation of MBMC as a tool to fold up given sequences on a square lattice model includes the eight (8) instances listed in table 4.1 which have been partly used in literature (Jacek et al., 2004; Ron and John, 1993; Alena and Holger, 2005; Thachuk et al., 2007). For proper porting in the HP model, the original protein sequences $\sigma_{orig.}$ are converted to HP sequences σ_{HP} based on their hydrophobicities as shown in table 3.2. Using MBMC, we enumerate for a given sequence a set of optimal structures showing a compact hydrophobic core and shape. Since our algorithm is a stochastic algorithm, it searches for the lowest energy in all runs for each option since the optimal result in each run within given step numbers cannot be ascertained. We implement the algorithm in Silverfrost FTN 95 compiler and run it on a laptop computer with an intel Pentium Dual-core CPU, 2.30 GHz processor and 4.00GB of RAM.

4.2.1 The physical mechanism on protein conformation

Folding of protein is driven by nonbonded interactions, which are represented as "contact energies" in the lattice model, i.e. interactions between residues that are situated on adjacent (or nearest-neighbour) lattice site but are not covalently bonded together to each other. It is generally accepted that protein folding is driven mainly by the hydrophobic effect which is the tendency of protein monomers to be repelled by water molecules. But, what is really lacking is an understanding of the specific physical mechanism or principles governing the folding process. The physical mechanism we considered in this thesis is the influence of directional probabilities which to our knowledge is a new approach. The mechanism may not uniquely obtain the folded structure, but it will drive it towards a sort of basin of attraction, which will give the basis for convergent evolution of the conformation.

Our method for maximizing the directional probability $P(\sigma)$ is based on two-dimensional hydrophobic-polar (HP) lattice backbone-only model (Dill, 1985). To obtain a conformation that is stable with unique ground state energy minimum in this model, the directional probability $P(\sigma)$ plays a vital role.

We vary the probabilities of the four possible directions along which a self-avoiding walker may move and analyse the influence of the variation of these probabilities on the sequence length (i.e length of simulated protein)

We let, $\alpha = \frac{P_u}{P_d}$, where $P_u = 0.005 \cdots 0.5$, $P_d = 0.25$ so that $0.02 \le \alpha \le 2.0$ (4.1)

$$\beta = \frac{P_r}{P_l}, \quad \text{where } P_r = 0.5 \cdots 0.005, \; P_l = 0.25 \text{ so that } \; 2.0 \ge \beta \ge 0.02 \quad (4.2)$$

$$d = \frac{\alpha}{\beta} \quad (4.3)$$

Where P_u, P_d, P_r, and P_l are the probabilities of up, down right and left step respectively.

From figure (4.10 a & b), large d represents a very low beta relative to alpha (since the maximum of alpha is 2.0) d increases as beta tends to 0 this means that two directions are formed for high d, probability of up and that of down. For low beta, the probability of right is lower than left; that is why the sequence length may be short because two directions are favoured.

The fluctuation in the sequence length as a function of alpha in figure 4.11 is due to the other factors (directions). If all other directional probabilities are fixed; then the fluctuations will give way to a more steady variation. A run for such parameter revealed that some variations still which can be explained is of a competition between the three directions. While the probabilities of the walker in figure 4.12 moving right with respect to left vary with the sequence length. Figure 4.13 shows a gradual decrease in the sequence length as alpha increases, which confirms that the variation in the sequence length stabilizes with reduction in the competition between the directions (for this figure we used $P_d = 0.02$, $P_r = 0.02$, and $P_l = 0.25$ to simulate a situation in which only the left direction is dominant).

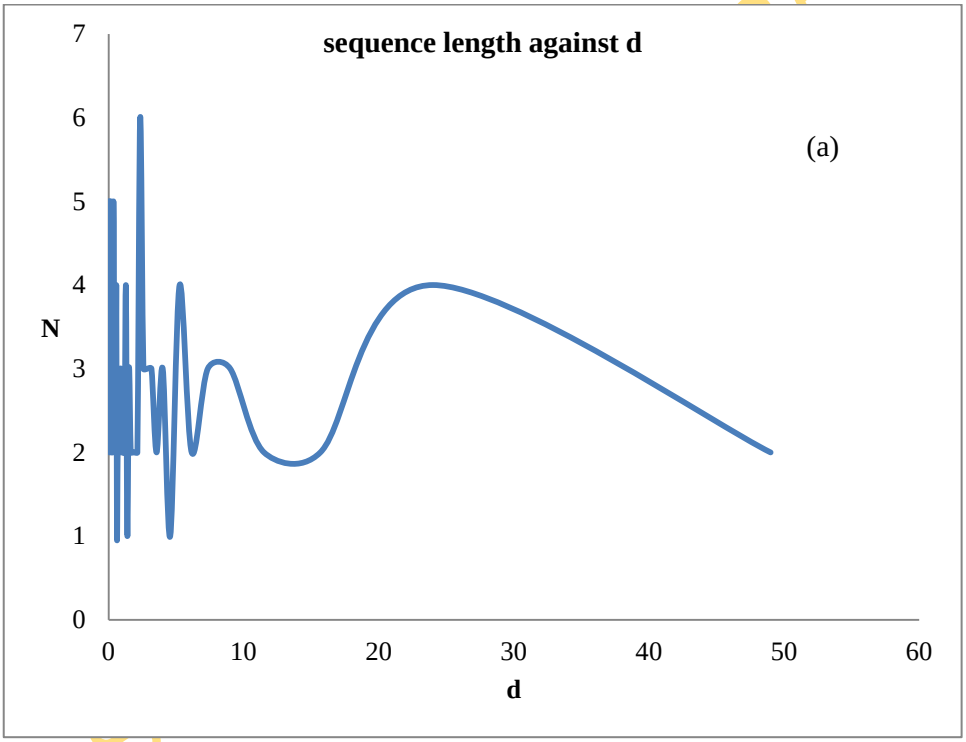


Figure 4.10a. The plot of sequence length (N) against d when the realisation is 5

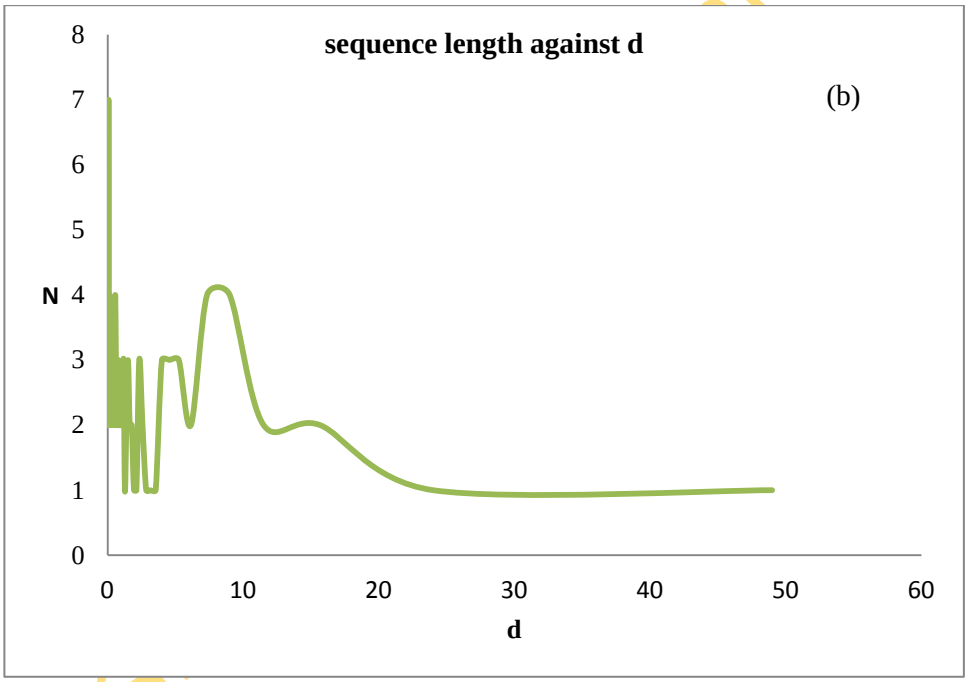


Figure 4.10b. The plot of sequence length (N) against d when the realisation is 10

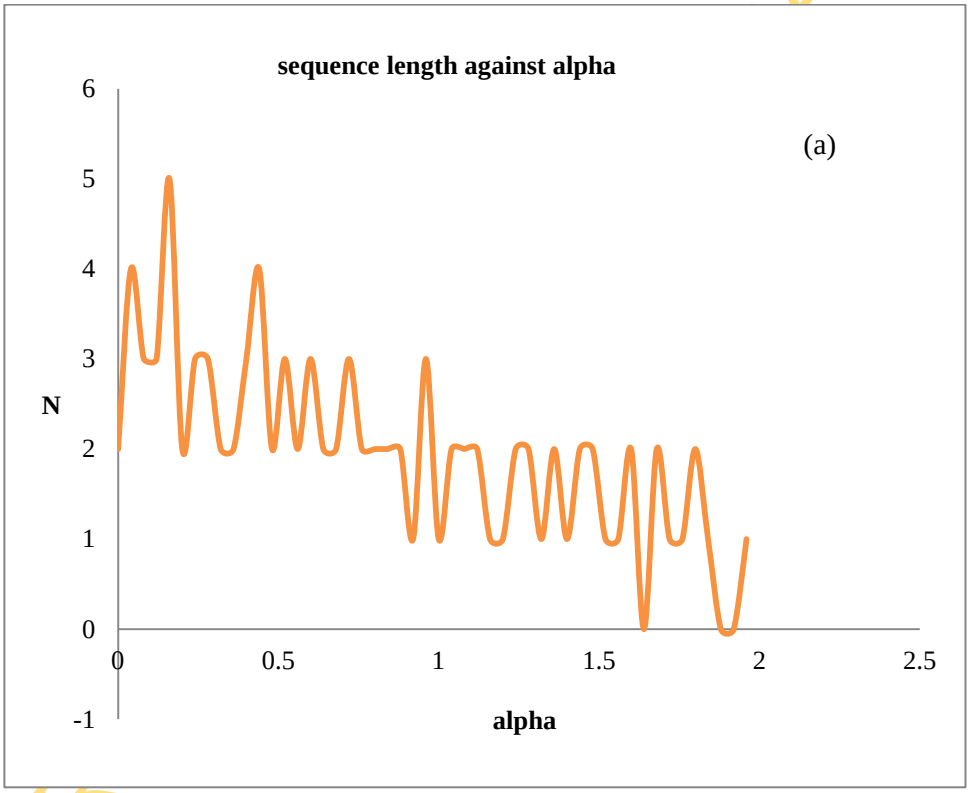


Figure 4.11. The plot of sequence length (N) against alpha

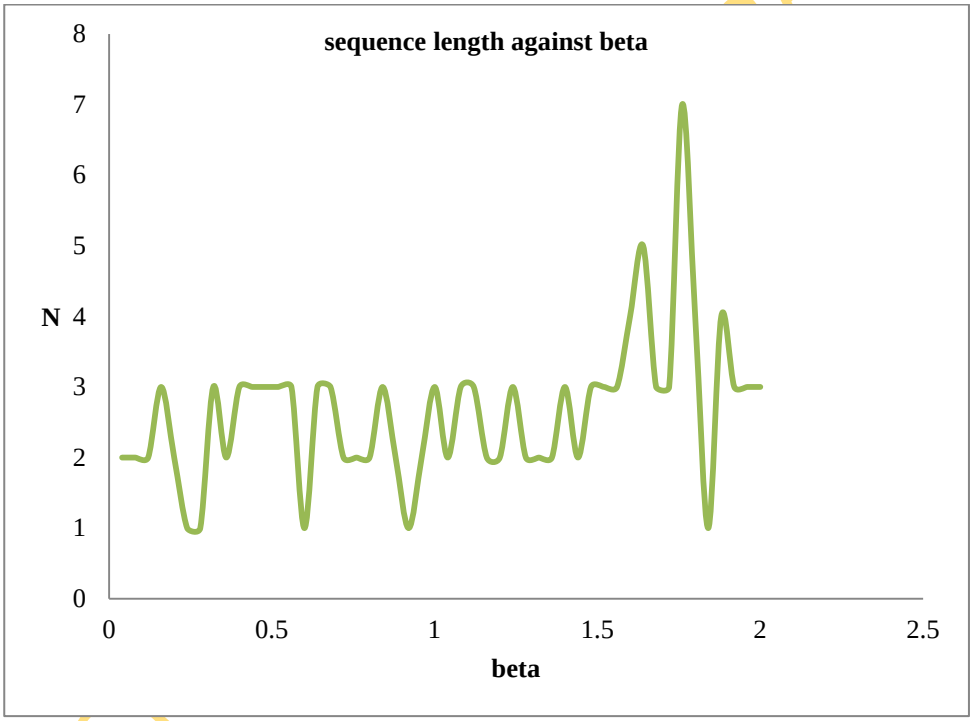


Figure 4.12. The plot of sequence length (N) against beta

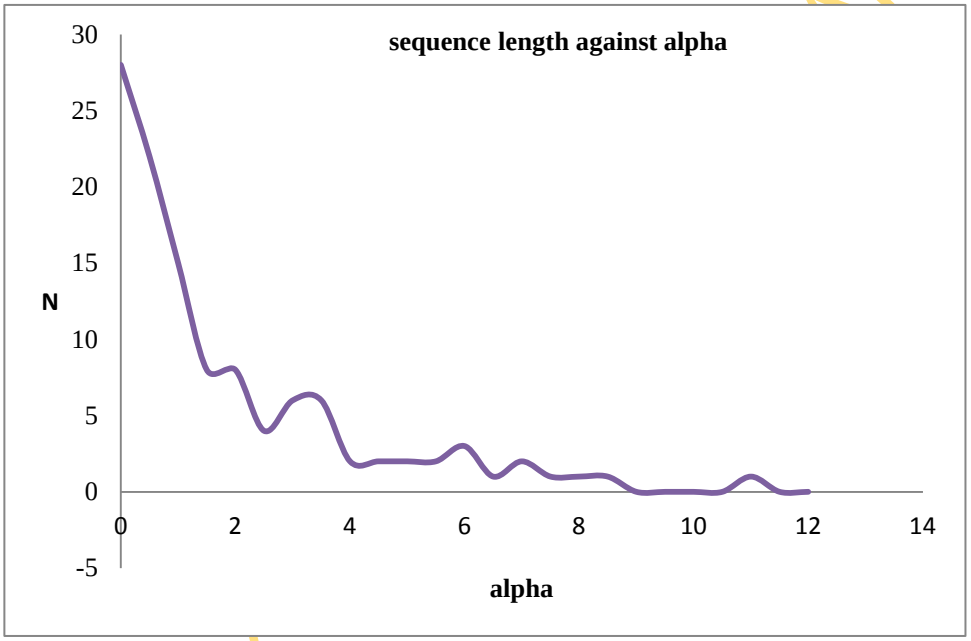


Figure 4.13. The plot of sequence length (N) against alpha

4.3 Evaluation of H-H contact

After each iteration, the conformation is evaluated by counting the H-H contacts (topological neighbour) where the two amino acids are non-consecutive (i.e the amino acids where there is no peptide bond). When a contact takes place between two non-consecutive hydrophobic amino acids it leads to a reduction of free energy of the resulting molecules. Hence, when the number of non-consecutive hydrophobic amino acids contact is maximize, protein reaches its ground state. This is done by checking for the number of amino acid present in the sequence, the checking is done from the first amino acid to the last amino acid for hydrophobic or polar. Hence, the energy value is the negation of the H-H contact count.

Table 4.3. This table shows the comparison of MBMC energy with the putative energy values of the benchmark instances for the 2D HP lattice model.

IN	N	|H|	|P|	Protein Sequence (H-hydrophobic, P-polar)	E^a	E^b	#
1	20	10	10	$(HP)^2PH^2PHP^2HPH^2P^2HPH$	-9	-9	474
2	24	10	14	$H^2P^2HP^2HP^2HP^2HP^2HP^2$ HP^2H^2	-9	-9	220
3	25	9	16	$P^2HP^2H^2P^4H^2P^4H^2P^4H^2$	-8	-8	200
4	36	16	20	$P^3H^2P^2H^2P^5H^7P^2H^2P^4H^2$ P^2HP^2	-14	-14	200
5	48	25	23	$P^2HP^2H^2P^2H^2P^5H^{10}P^6H^2$ $P^2H^2P^2HP^2H^5$	-23	-23	200
6	60	43	17	$P^2H^3PH^8P^3H^{10}PHP^3H^{12}$ $P^4H^6P\ H^2PHP$	-36	-35	100
7	64	42	22	$H^{12}(PH)^2(P^2H^2)^2P^2HP^2H^2$ $PPH^2P^2\ HP^2\ (H^2P^2)^2(HP)^2H^{12}$	-42	-42	100
8	85	59	26	$H^4P^4H^{12}P^6(H^{12}P^3)^3HP^2$ $(H^2P^2)^2HPH$	-52	-52	100

[a] the putative energy value
[b] the energy obtained by MBMC
[#] the number of energy evaluations required by the best run to achieved the optimum or a sub-optimum conformation

4.4 Optimal structure prediction

The protein structure prediction (PSP) problem is to determine the optimal protein structure from a given protein sequence within the model (as seen in figure 4.14). This problem is actually how to embed a sequence of H or P abstracted from real protein in a lattice such as square, cubic, triangular e.t.c. A structure is HP-optimal if it minimizes the energy function based on $E(H,P)$. The structure prediction has been limited to X-ray crystallography by Kendrew et al., (1953) or nuclear magnetic resonance (NMR) spectroscopy by Wuthrich, (1990). The two methods are very complicated and expensive such that the number of known sequences outweighs the structures. Consequently, in order to increase the number of identified protein structures has called for the computational methods where the state-of-the-art algorithms are benchmarked every two years by CASP experiment (Critical Assessment of Techniques for protein structure prediction). Currently, the overall accuracy of computational approaches is limited, but with significant progress in the quality of the predictions. Mathematically, the PSP is expressed as follows, for an HP sequence $\sigma = \sigma_1, \sigma_2, ..., \sigma_x$, we try to find a conformation with minimum energy of σ i.e to find $\zeta^{\otimes} \in \zeta(\sigma) \ni E(\zeta^{\otimes}) = \min\{E(\zeta)/\zeta \in \zeta(\sigma)\}$ where $\zeta(\sigma)$ is the set of all valid conformations i.e the self-avoiding walk of the sequence σ. PSP has been proven to be NP-complete even with the simplest lattice model, e.g. 2D-square and 3D-cubic lattice model (Crescenzi et al., 1998; Berger and Leighton, 1998).

4.5 Protein encoding

There are two isomorphic encoding strategies for HP models which complement the direct, coordinate presentation in the lattice model: relative encoding and absolute encoding. In relative encoding (Hoque et al., 2009), the move direction is defined relative to the direction of the previous move, while in absolute encoding (Backofen et al., 1999), the direct coordinate presentation is replaced by letters or numbers representing directions with respect to the lattice structure. In this thesis, following the absolute encoding in the 2D square lattice, (as shown in figure 4.14) the permitted moves are: right $\rightarrow$, left $\leftarrow$, up $\uparrow$ and down $\downarrow$.

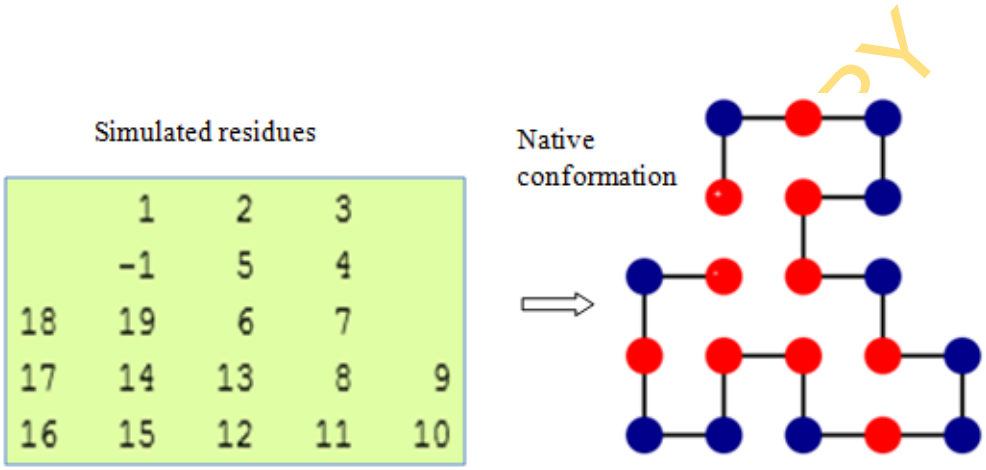


Figure 4.14. The Protein Structure Prediction for the backbone-only HP model in 2D-square lattice. The simulated residues (left) are embedded in the square lattice to produce the native conformation (right). The numbers -1 to 19 are the amino acid sequences given in instance 1 of table 4.1. H-monomers are given in red, P-monomers in blue and the black lines are the backbone connections. The plus and the minus signs are the starting and the ending points respectively.

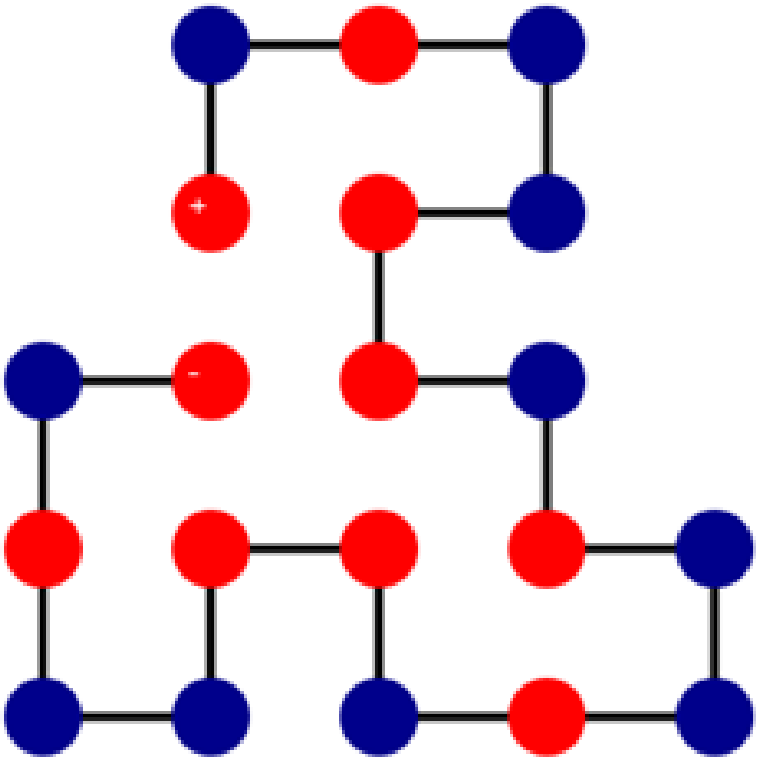


Figure 4.15. Optimal conformation with an energy of -9 for instance 1 (the length-20 sequence) found by the MBMC method. H-monomers are given in red, P-monomers in blue and the black lines are the backbone connections. The plus and the minus signs are the starting and the ending points respectively.

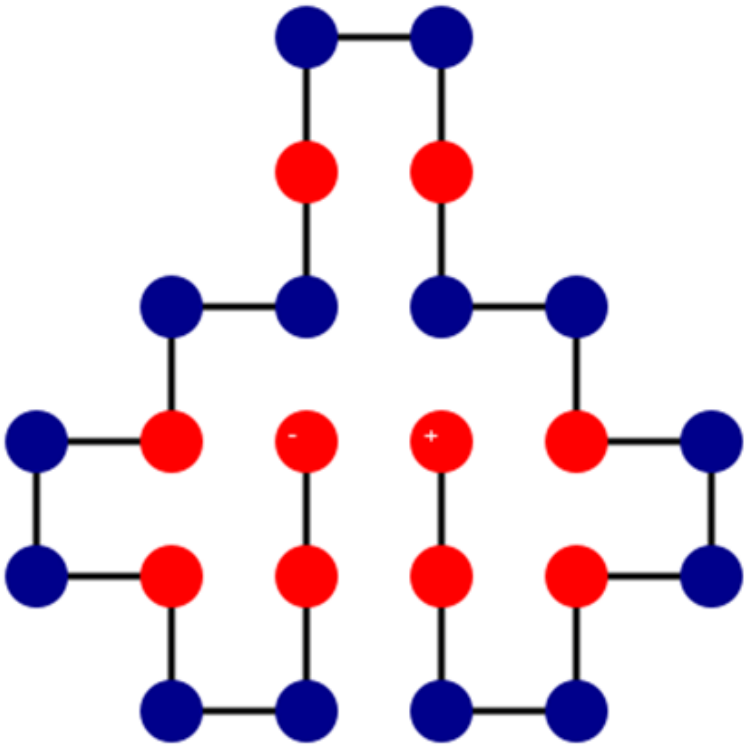


Figure 4.16. Optimal conformation with an energy of -9 for instance 2 (the length-24 sequence) found by the MBMC method. H-monomers are given in red, P-monomers in blue and the black lines are the backbone connections. The plus and the minus signs are the starting and the ending points respectively.

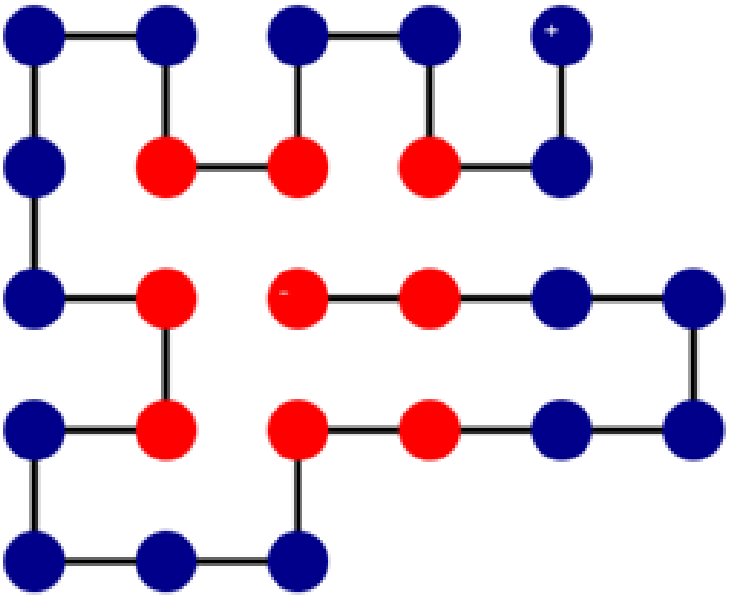


Figure 4.17. Optimal conformation with an energy of -8 for instance 3 (the length-25 sequence) found by the MBMC method. H-monomers are given in red, P-monomers in blue and the black lines are the backbone connections. The plus and the minus signs are the starting and the ending points respectively.

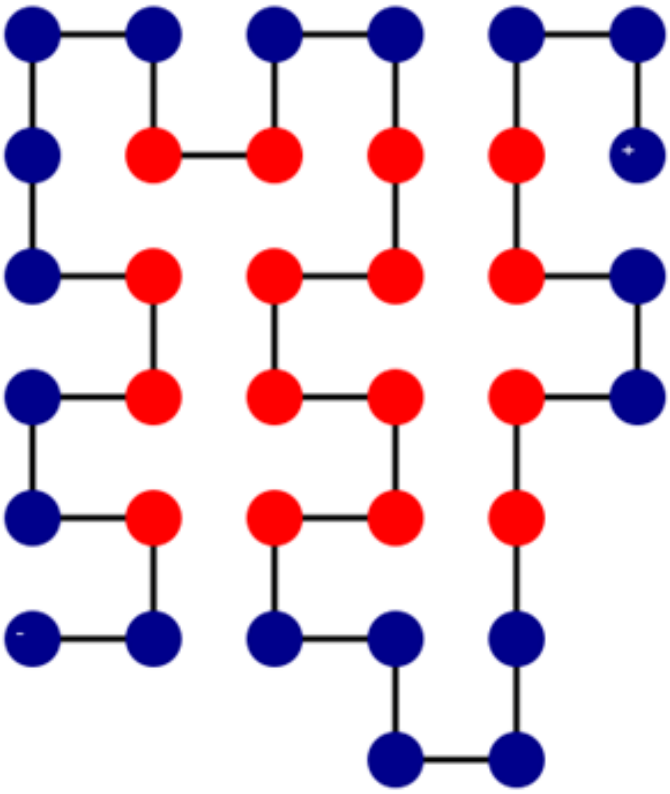


Figure 4.18. Optimal conformation with an energy of -14 for instance 4 (the length-36 sequence) found by the MBMC method. H-monomers are given in red, P-monomers in blue and the black lines are the backbone connections. The plus and the minus signs are the starting and the ending points respectively.

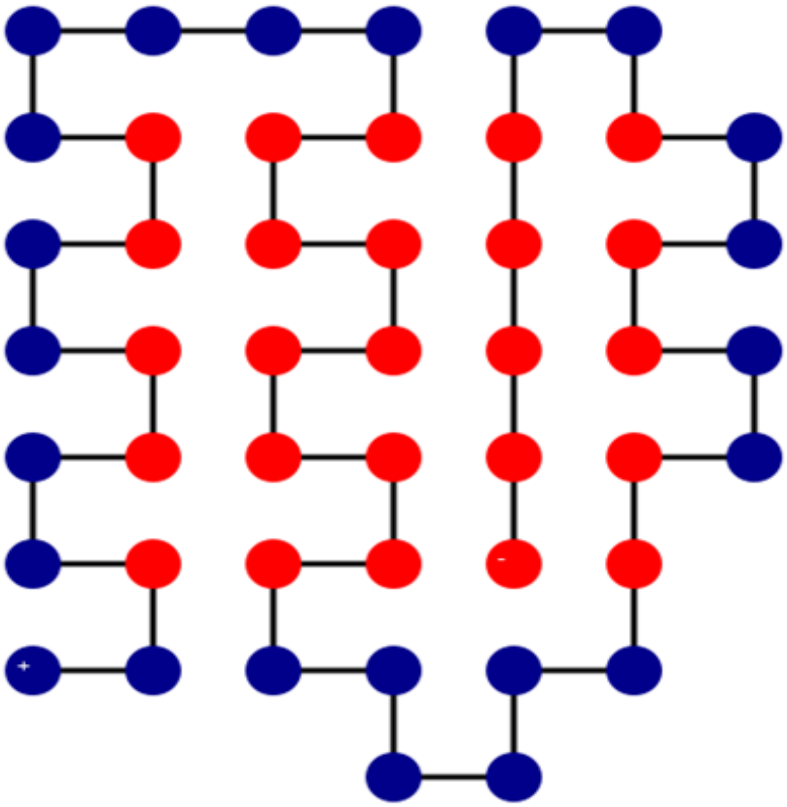


Figure 4.19. Optimal conformation with an energy of -23 for instance 5 (the length-48 sequence) found by the MBMC method. H-monomers are given in red, P-monomers in blue and the black lines are the backbone connections. The plus and the minus signs are the starting and the ending points respectively.

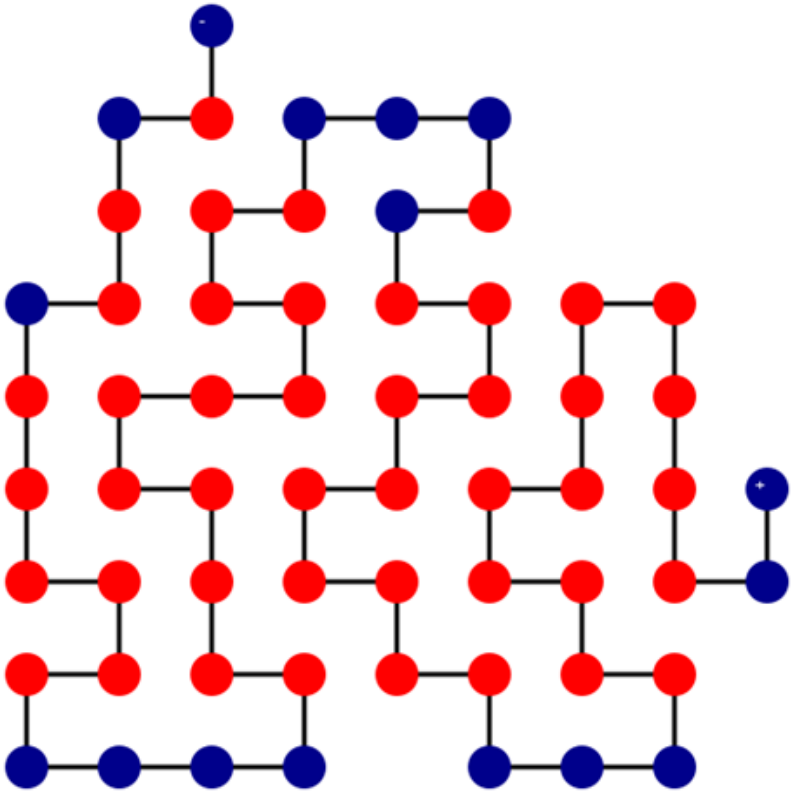


Figure 4.20. Optimal conformation with an energy of -35 for instance 7 (the length-60 sequence) found by the MBMC method. H-monomers are given in red, P-monomers in blue and the black lines are the backbone connections. The plus and the minus signs are the starting and the ending points respectively.

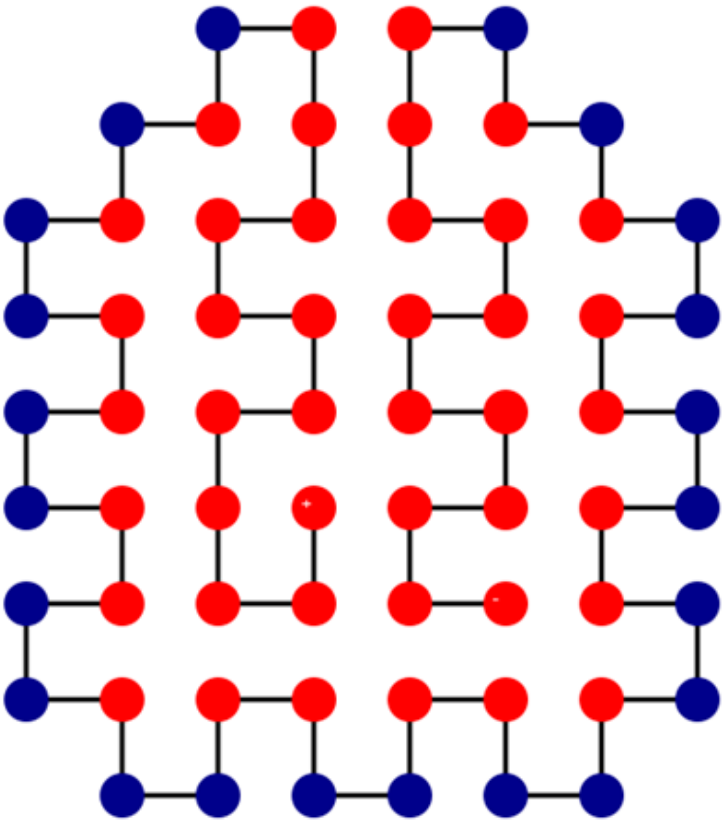


Figure 4.21. Optimal conformation with an energy of -42 for instance 8 (the length-64 sequence) found by the MBMC method. H-monomers are given in red, P-monomers in blue and the black lines are the backbone connections. The plus and the minus signs are the starting and the ending points respectively.

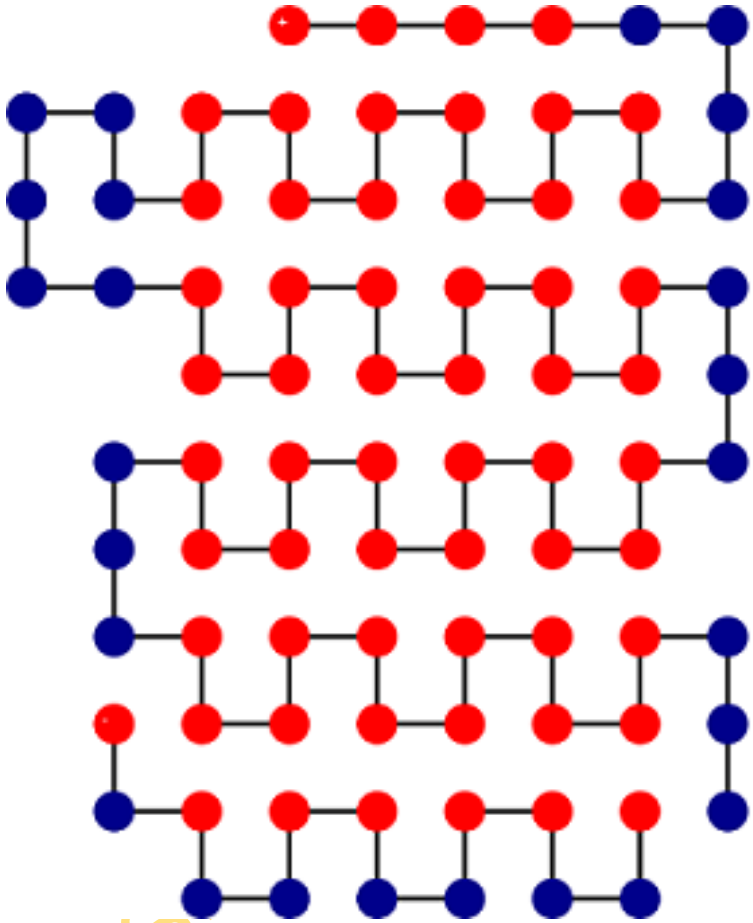


Figure 4.22. Optimal conformation with an energy of -52 for instance 9 (the length-85 sequence) found by the MBMC method. H-monomers are given in red, P-monomers in blue and the black lines are the backbone connections. The plus and the minus signs are the starting and the ending points respectively.

Table 4.4. This table shows the performance comparison of MBMC with various methods for the 2D HP lattice model. The number in each cell is the minimum energy obtained by the corresponding method for the respective HP sequence. The numbers in parentheses are the numbers of valid conformations scanned before the lowest-energy values are found. The values are taken from (Guo et al., 2006; Jingfa et al., 2011).

N	E[a]	MBMC	CMC	GA	EMC	ACO	ENLS
20	-9	-9 (474)	-9 (292443)	-9 (30,492)	-9 (9374)	-9 (500)	-9 (800)
24	-9	-9 (220)	-9 (2492221)	-9 (30,491)	-9 (6929)	-9 (500)	-9 (800)
25	-8	-8 (200)	-7 (2694572)	-8 (20,400)	-8 (7202)	-8 (500)	-8 (800)
36	-14	-14 (200)	-12 (6557189)	-12 (301,339)	-14 (12447)	-14 (500)	-14 (800)
48	-23	-23 (200)	-20 (9201755)	-22 (126,547)	-23 (165791)	-23 (500)	-23 (800)
60	-36	-35 (100)	-33 (8262338)	-34 (208,781)	-35 (203729)	-34 (100)	-36 (800)
64	-42	-42 (100)	-35 (7848952)	-37 (187,393)	-39 (564809)	-32 (100)	-39 (800)
85	-53	-52 (100)	N/A	N/A	-52 (44 029)	-53	N/A

N/A: not available
[a] the putative energy value

Table 4.5. This table shows the comparison of performances run time of various methods on the eight 2D HP sequences listed in table 4.1.

N	MBMC t(s)	CMC t(s)	GA t(s)	EMC t(s)	ACO t(s)	ENLS t(s)
20	8.90	?	5.60	?	(<1)	?
24	8.51	?	6.00	?	(<1)	?
25	8.37	?	3.66	?	(<1)	?
36	9.14	?	54.60	?	(4sec.)	?
48	9.45	?	?	?	(1min)	?
60	9.46	?	?	?	(20min.)	?
64	9.52	?	?	?	(1.5hrs)	6
85	12.85	?	?	?	?	38

4.6 Relative Improvement (RI)

The relative improvement explains how close our predicted results to the lower bound of free energy with respect to the energy obtained from the state-of-the-art approaches. We compare MBMC (new) and CMC as the (reference). We calculate for the eight (8) instances, the RI of the new method (n) with respect to the reference (ref.) using the formula in equation 4.4

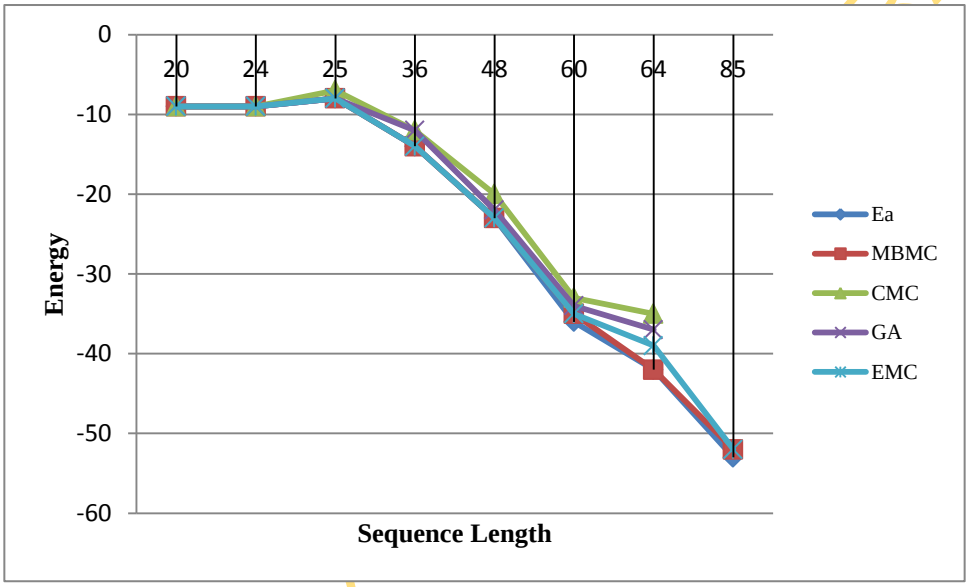
$$RI = \frac{E_n - E_{ref.}}{E_{l.b} - E_{ref.}} \times 100\% \quad (4.4)$$

Where E_n and $E_{ref.}$ represent the energy values obtained by new methods and reference respectively, and $E_{l.b}$ is the lower bound of free energy for the protein in the HP model. RI is presented for each protein of the eight (8) instances having known their lower bound free energy values. The results in table 4.6 show a comparison of improvement (%) on the conformation quality in terms of the ground state conformation energy level. MBMC and CMC show very good progress in the first two instances, but MCMB was able to improve the search quality in terms of minimizing the free energy level for higher instances. The bold-faced values are the minimum and maximum improvements for the same column. The relative improvements with respect to CMC range from 33.3% to 100%; while with respect to GA range from 50% to 100%.

Table 4.6. Relative improvement by MBMC with respect to CMC

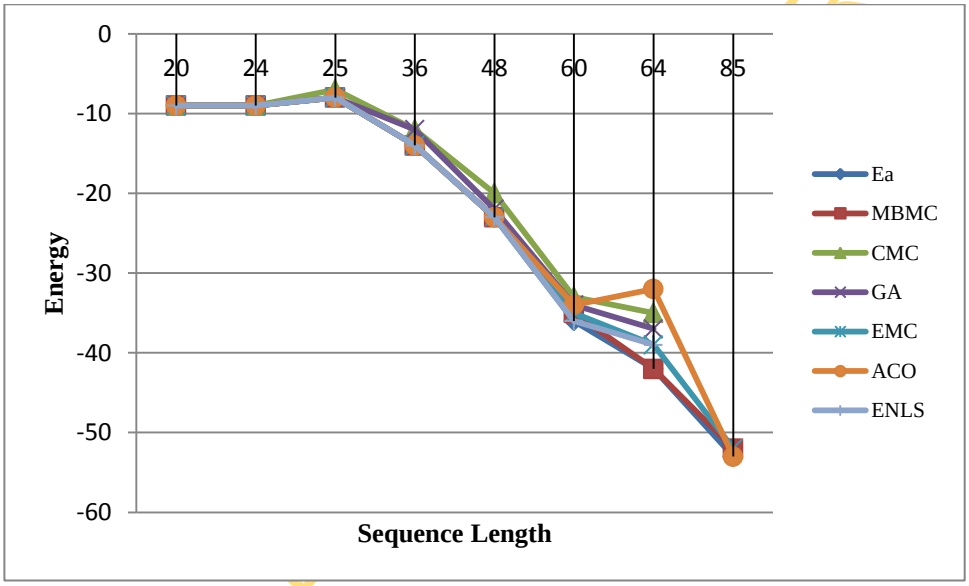
N	E[a]	MBMC	CMC	RI (%)	GA	RI (%)
20	-9	-9	-9	Nil	-9	Nil
24	-9	-9	-9	Nil	-9	Nil
25	-8	-8	-7	**100%**	**-8**	Nil
36	-14	-14	-12	**100%**	**-12**	**100%**
48	-23	-23	-20	**100%**	**-22**	**100%**
60	-36	-35	-33	**66.7%**	**-34**	**50%**
64	-42	-42	-35	**100%**	**-37**	**100%**
85	-53	-52	N/A		N/A	

The relative improvements (RI) of MBMC to CMC and GA are presented in the table 4.6. The values of RI are calculated using the formula in equation 4.4. The bold face values are the minimum and maximum values of RI for the respective columns. N/A means not available.



Ea: Putative energy value; MBMC: move biased Monte Carlo; CMC: Conventional Monte Carlo; GA: Genetic algorithm; EMC: Evolutionary Monte Carlo

Figure 4.23. Relative improvement of MBMC with three other Monte Carlo based methods (CMC, EMC and GA), Ea is the lowest bound energy. The results are calculated for at least 100 iterations.



Ea: Putative energy value; MBMC: move biased Monte Carlo; CMC: Conventional Monte Carlo; GA: Genetic algorithm; EMC: Evolutionary Monte Carlo; ACO: Ant colony Optimization; ENLS: Hybrid elastic net.

Figure 4.24. Relative improvement of MBMC with three others that are Monte Carlo based (CMC, EMC and GA) and two others that are not Monte Carlo based (ACO and ENLS), Ea is the lowest bound energy. The results are calculated for at least 100 iterations.

4.7 Discussion

Table 4.4 summarizes the results obtained by MC-coupled move (MBMC) algorithm's performance as well as other methods' reported in the literature on eight different sequences ranging from 20 to 85 residues in length. From Table 4.4, one can see that the lowest free energies for the three shortest protein sequences SI-1 (20-mer), SI-2(24-mer) and SI-3(25-mer) obtained by the methods in table 4.4 are all the same except SI-3(25-mer) for CMC were the free energy is -7. For the other sequences SI-4 (36-mer), SI-5(48-mer), SI-6(50-mer), SI-7(60-mer) and SI-8(64-mer) CMC and GA methods have the least energy for SI-4 (36-mer), SI-5(48-mer), SI-8(64-mer) and SI-9(85-mer). For SI-8(64-mer), only MBMC finds the ground-state conformations (GSC) with the lowest free energy of -42. Also, for SI-9(85-mer) sequence, MBMC, EMC and ACO methods find the GSCs with the lowest free energy of -52 while the result of CMC and GA are unknown. The ground state conformations by MBMC for all the instances in Table 4.4 are shown in figure (4.14 – 4.22); It is obvious that each of these conformations possesses a compact hydrophobic core. Moreover, the MC-coupled move method scans fewer valid conformations (as shown in Table 4.4) than the CMC, GA and EMC methods to obtain the lowest free energy for every sequence. From Table 4.5, it is seen that MC-coupled move method requires much less time to obtain the ground state energies. Hence, the MC-coupled move method explores the conformation surfaces more efficiently than the CMC, GA and EMC methods.

In general, the MBMC optimization approach outperforms the CMC and GA since it reaches final conformations in less iteration for all the benchmark instances.

CHAPTER 5
CONCLUSION AND RECOMMENDATION

5.1 Conclusion

Protein folding is a very fundamental problem in protein sciences. The heuristic HP lattice protein model has been auspicious for the protein folding problem. In this thesis, we focused on lattice protein models that discretise the possible structure space of proteins. We investigated backbone-only lattice protein models in 2D lattices and demonstrated the strength of the move-biased Monte Carlo (MBMC) method in searching for the unique ground state energy conformation (GEC) of some standard benchmark protein sequences. This new method is a class of heuristic and a generation of stochastic Monte Carlo method which can be applied in a rather straightforward way to 2D for Protein structure prediction. Our method incorporates a coupled neighbourhood search strategy (diagonal-pull moves) on premature conformations to obtain the ground state energy conformations. This is sequel to the new local search procedures based on the long range move for effective search neighbourhood.

We compared our results with other heuristic search methods which achieved the state-of-the-art results in terms of CPU run time and the number of conformations scanned. Our results show that MC method with the coupled (diagonal-pull) moves explores the conformation surface more efficiently by finding a different ensemble of native conformations with the lowest known energy within a comparable computational time compared to the CMC, GA, EMC and ACO in the benchmark sequences studied.

In general, the results presented in this work indicate that with the coupled moves local search procedures, the MBMC optimization approach outperforms the CMC, GA, EMC and ACO since it reaches the native conformations in less iteration and CPU time for higher benchmark instances in contrast to CMC, GA and EMC which have no records of CPU time. We also found that the effect of directional probabilities is

crucial for the performance of the algorithm; particularly it enhanced the formation of high-quality conformations for the benchmark sequences studied.

Finally, we have demonstrated that MBMC performs better than CMC on long sequences in 2D lattice model. In addition, we intend to develop and study MBMC algorithm for other types of protein folding problems such as three-dimensional (3D) lattice model and 2D, 3D triangular lattice model. Also, to look at the influence of the directional probabilities of the four possible directions which, a SAW may take on the compactness of the resulting structure and on the compact quality.

5.2 Our contribution

In this thesis, we developed a move-biased Monte Carlo (MBMC) simulation method for ab initio protein native structure prediction using an HP energy model on the two-dimensional square lattice. The introduced method combines the advantages of Monte Carlo and local neighbourhood search moves (diagonal-pull) for protein structure prediction.

In our approach, we considered the effect of directional probabilities for protein conformations. This effect is a new idea which is responsible for the physical mechanism governing the protein folding. This mechanism, albeit may not uniquely obtain the folded structure, but it will drive it towards a better conformation, which will give the basis for convergent evolution of the conformation.

On a set of benchmark proteins sequences, this effect on the optimization approach (MBMC) makes it a significant method competing with other heuristic approach and particularly outperforms the conventional Monte Carlo (CMC), genetic algorithm (GA), evolutionary Monte Carlo (EMC) and ant colony optimization (ACO) methods since it reaches the native conformations in less iteration and CPU time for higher benchmarks protein sequences

REFERENCES

Abkevich, V., Gutin, A. and Shakhnovich, E. 1995. Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *Journal of Molecular Biology* 252: 460-471.

Alan, S. D. 1996. Monte Carlo methods for the self-avoiding w alk. *Nuclear Physics B (Proc. Suppl.)* 47:172-179.

Alena, S. and Holger, H. H. 2005. An ant colony optimization algorithm for the 2D and 3D hydrophobic polar protein folding problem. *BMC Bioinformatics* *6*: 30.

Allan, C. M. and Ashok, D. A. 2011. Protein folding at single-molecule resolusion. *Biochimica et Biophysica Acta* 1814:1021-1029.

Anderea, R., Cristian, M., Flavio, S. and Amos, M. 2001. A self-consistent knowledge-based approach to protein design. *Biophysical Journal* 80: 480-490.

Anders, I. and Sandelin, E. 1998. Local interactions and protein folding: A model study on the square and triangular lattices. *Journal of Chemical Physics* 108: 2245-2250.

Anders, I., Carsten, P., Frank, P. and Ola, S. 1997. Local interactions and protein folding: A three-dimensional off-lattice approach. *Journal of Chemical Physics* 107: 273-282.

Andrej, S., Eugene, S. and Martin, K. 1995. Kinetics of protein folding, A lattice model study of the requirements for folding to the native state. *Journal of*

Molecular Biology 235: 1614-1636.

Anfinsen, C. 1973. Principle that govern the folding of protein chains. *Science* 181: 223-239.

Ashlee, J. and Hongbin, L. 2010. Measuring "unmeasurable" folding kinetics of proteins by single-molecule force spectroscopy. *Journal of molecular Biology* 402: 610-617.

Backofen, R., Will, S. and Clote, P. 2000. Algorithmic approach to quantifying the hydrophobic force contribution in protein folding. *Proceedings of the Pacific Symposium on Biocomputing (PSB 2000)* 5: 92-103.

Benedetti, F., Micheletti, C., Bussi, G., Sekatskii, S. and Dietler, G. 2011. Nonkinetic modeling of the mechanical unfolding of multimodular proteins. *Biophysical Journal* 101: 1504-1512.

Berger, B. and Leighton, T. 1998. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *Journal of Computional Biology* 5.*1*: 27-40.

Blum, C. 2005. Ant colony optimization: introduction and recent trends. *Physics of Life Review* 2: 353-373.

Bockenhauer, H., Ullah, A., L., L. K. and Steinhofel, K. 2008. A local move set for protein folding in triangular lattice models. *Proceedings of the 8th International Workshop, Karlsruhe.* K. Crandall and J. Lagergren. Eds. Berlin: Springer. 5251: 369-381.

Broglia, R. A. and Tiana, G. 2003. Physical models for protein folding and drug

design. *Proceeding Idea-Finding Symposium*, Germany: Frankfurt Institute for Advanced Studies. 23-33.

Broglia, R. A., Serrano, L. and Tiana, G. 2007. Protein folding and drug design. *Proceedings of the international school of Physics " Enrico Fermi" , IOS Press and Societa Italiana DiFisica, Italy.*

Bruno, R. and Valerie, D. 2013. Using simulations to provide the framework for experimental protein folding studies. *Archives of Biochemistry and Biophysics* 531: 128-135.

Bryan, A. K. 2002. Protein folding: new methods unveil rate-limiting structures. Chigago: PhD Thesis. Dept. of Biochemistry and Molecular Biology. University of Chicago. 1- 255.

Branden, C. and Tozer, J. 1999. *Introduction to protein structure.* 2nd ed. New York, USA: Garland.

Carlo, T.-A., Ylva, I., Per, J. and Stefano, G. 2009. Folding and stability of globular proteins and implications for function. *Current opinion in structural Biology* 19: 3-7.

Cebrian, M., Dotu, I., Van Hentenryck, P. and Clote, P. 2008. Protein structure prediction on the face centered cubic lattice by local search. *Proceedings of the Conference on Artificial Intelligence* 1: 241-246.

Cheolju, L. and Myeong-Hee, Y. 2005. Protein folding and diseases. *Journal of Biochemistryand Molecular Biology* 38: 275-280.

Chikenji, G., Kikuchi, M. and Iba, Y. 1999. Multi-self-overlap ensemble for protein

folding: ground state search and thermodynamics. *Condensed Materials Archive* , 27.

Chunmei, L., Chang, T., Meng, Q., Dawie, Z., Yi, C. and Wei, W. 2012. Low folding cooperativity pf HP 35 Revealed by single-molecule force spectroscopy and molecular dynamic simulation. *Biophysical Journal* 102: 1944-1951.

Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A. and Yannakakis, M. 1998. On the complexity of protein folding. *Journal of Computational Biology* 5.3: 423-465.

Cutello, V., Nicosia, G., Pavone, M. and Timmis, J. 2007. An immune algorithm for protein structure prediction on lattice models. *IEEE Transaction on Evolutionary Computing* 11: 101-117.

Dana, R. and Alistair, S. 2000. Self-testing algorithms for self-avoiding walks. *Journal of mathematical Physics* 41, 3.

David, L. P. and Kurt, B. 2000. *Monte Carlo simulations in statistical physics.* United Kingdom: Cambridge University Press.

Deutsch, J. and Kurosky, T. 1996. New algorithm for protein design. *Phys. Rev. Lett.* 76: 323-326.

Diego, F. U., Joseph, H. A., Elizabeth, K. A. and Peter, W. G. 2007. Localizing frustration in native proteins and protein assemblies. *PNAS* .

Dill, K. A. 1985. Theory for the folding and stability of globular proteins. *Biochemistry* 24.*6*: 1501-1509.

Dill, K. A., Ozkan, S., Shell, M. and Weikl, T. 2008. The protein folding prblem. *Biophys* 37: 289-326.

Dotu, I., Cebrian, M., P., V. H. and Clote, P. 2011. On lattice protein structure prediction revisited. *IEEE/ACM Trans. Comput. Biol. and Bioinform* 8: 1620.

Dotu, I., Cebrian, M., Van Hentenryck, P. and Clote, P. 2011. On lattice protein structure predicted revisited. *IEEE Transactions on Computational Biology and Bioinformatics.* IEEE.

Erik, S. 2000. *Thermodynamics of protein folding and design.* PhD. Thesis. Dept. of Theoretical Physics. Lund University. 1-131.

Finkelstein, A. and Galzitskaya. 2004. Physics of protein folding. *Physics of Life Reviews* 1, 23-56.

Ginka, B. S., Ronan, M. D., Nicolae-Viorel, B. and Jon, K. 2011. Dynamic of protein folding: Probing the kinetic network of folding-unfolding transitions with experiment and theory. *Biochimical et Biophysical Acta* 1814: 1001-1020.

Greeberg, H., Hart, W. and Lancia, G. 2004. Opportunity for combinatorial optimization in computational biology. *INFORMS Journal of Computing* 16: 211-231.

Greta, H., Soren, P. W., Celestine, C. N., Kristian, S. and Stefano, G. 2012. An expanded view of the protein folding landscape of PDZ domains. *Biochemical and Biophysical Research Communication* 42: 550-553.

Guo, J., Mi, D. and Sun, Y. 2010. Folding kinetics of two- state proteins based on the model of general random walk in native contact number space. *Physica A 389*: 761-766.

Guo, Y.-Z., Feng, E.-M. and Wang, Y. 2006. Exploration of two-dimensional hydrophobic-polar lattice model by combining local search with elastic net algorithm. *The Journal of Chemical Physics* 125: 154102.

Hans, F. 2010. *An Introduction to Biological Physics and Molecular Biophysics.* C. S. Shirley and C. S. Winnie. Eds. New York: Springer.

Hans-Joachim, B. and Dirk, B. 2007. Protein folding in the HP model on grid lattices with diagonals. *Discrete applied mathematics 155*: 230-256.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their Applications. *Biometrika,* 57 (1): 97

Helling, R., Li, H., Melin, R., Miller, J., Wingreen, N. and Zeng, C. T. 2001. The Designability of protein structures. *J. Mol. Graph Model* 19.1: 157-167.

Hoque, M. T., Chetty, M., Lewis, A. and Sattar, A. 2011. Twin removal in genetic algorithms for protein structure prediction using low-resolution model. *IEEE Transactions on Computational Biology and Bioinformatics* 8: 234-245.

Hoque, M., Chetty, M. and Sattar, A. 2009. *Biomedical data and applications. Studies in computational intelligence. A. Sidhu, T. Dillon. Eds.* Berlin Heidelberg: Springer. Vol. 224

Hoque, M., Chetty, M. and Sattar, A. 2007. Protein folding prediction in 3D FCC HP lattice model using genetic algorithm. *Proceedings of the IEEE Congress Evolutionary Computation:* Singapore. 4138-4145.

Hsu, H., Mehra, V., Nadler, W. and Grassberger, P. 2003. Growth algorithm for lattice heteropolymers at low temperatures. *Journal of Chemical Physics* 118: 444-451.

Hughes, B. 1995. *Random walks and random enviroments, Volume 1: Random walks.* Oxford:Clarendon Press.

Hyun-suk, Y. 2006. *Optimization approaches to protein folding.* PhD.Thesis. School

of industrial and systems engineering. Georgia Institute of Technology, Georgia. 1-114.

Irback, A. and Sandelin, E. 1999. Monte Carlo study of the phase structure of compact polymer chains. *Journal of Chemical Physics* 110: 12256-12262.

Irback, A., Peterson, C., Potthast, F. and Sandelin, E. 1998. Monte Carlo procedure for protein design. *Phys. Rev. E* 58: R5249-R5252.

Istrail, S. and Lam, F. 2009. Combinatorial algorithms for protein folding in lattice models: A survey of mathematical results,. *Commun. Inf. Syst.* 9.4: 303-346.

Jacek, B., Ken, D., Piotr, L. and Micostan, M. 2004. A tabu search strategy for finding low energy structures of proteins in HP-model. *Computer Methods in Science* , 10: 7-19.

Jim, S. 2010. Protein folding : The dark side of proteins. *Nature* 464: 828-829.

Jingfa, L., Gang, L. and Jun, Y. 2011. Protein-folding simulations of the hydrophobic-hydrophylic model by combining pull moves with energy landscape paving. *Physica Review E 84:* 031934.

Jingfu, l. L., Beibei, S., Zhaoxia, L., Weibo, H., Yuanyuan, S. and Wenjie, L. 2013. Energy-landscape paving for prediction of face-centered-cubic hydrophobic-hydrophilic lattice model proteins. *Physica Review E* 88: 052704.

Kendrew, J., Bodo, G., Dintzis, H., R.G. Parrish, W. H. and Phillips, D. 1953. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* 181: 662-666.

Kenneth, M. P. 2001. Protein structure, stability and folding. *Methods in Molecular Biology , 168*: 1-16.

Kerson, H. 2005. *Lectures on statistical physics and protein folding.* Singapore:

World Scientific Publishing Co. Pte. Ltd.

Kihara, D., Lu, H., Kolinski, A. and Skolnick, J. 2001. Touchstone: An ab initio protein structure prediction method that uses threading-based tertiary restrints. *PNAS* 98.18: 10125-10130.

Klau, G., Lesh, N., Marks, J. and Mitzenmacher, M. 2002. Human guided tabu search. Proceeding of the conference on artificial intelligence.

Klemm, K., Flamm, C. and Stadler, P. F. 2008. Funnels in energy landscapes. *The European Physical Journal B* 63: 387-391.

Kurosky, T. Deutsch, J. 1995. Design of copolymeric materials. *J.Phys* A27: L387-L393.

Kurt, B. and Dieter, H. W. 2010. *Monte Carlo simulation is statistical physics. 5th ed.* Berlin, Heidelberg: Springer-Verlag.

Lau, K. and Dill, K. 1989. A lattice statistical model for the conformational and sequence spaces of proteins. *Macromolecules* 22: 3986-3997.

Lesh, N., Mitzenmacher, M. and Whitesides, S. 2003. A complete and effective move set for simplified protein folding. *Proceedings of the seventh annual international conference on research in Computational molecular biology (RECOMB'03)*. 188-195.

Levinthal, C. 1968. Are there pathways to protein folding? *Journal of Chemical Physics* 65: 44-45.

Liang, F. and Wong, W. 2001. Evolutionary Monte Carlo for protein folding simulations. *Journal of Chemical Physics* 115.7: 3374-3380.

Liu, H. 2009. *The statistical models for globular proteins folding in water solution.*

PhD. Thesis. Dept. Applied Mathematics. Tsinghua University China. 1-200.

Luca, M. 2005. *Coarse-grained models for protein folding and function.* PhD. Thesis. International school for advanced studies (ISAS) ICTP, Italy. 5-105

Madras, N. and Slade, G. 1993. *The self-avoiding walk, probability and its applications.* Boston: Birkhauser.

Madras, N. and Sokal, A. 1988. The pivot algorithm: A highly efficient Monte Carlo method for the self-avoiding walk. *Journal of Statistical Physics* 50: 109-186.

Mahmood, R. A., Hakim, N. M., Tamjidul, H. M., Swakkhar, S., Nghia, P. D. and Abdul, S. 2013. Spiral search: a hydrophobic-core directed local search for simplified PSP on 3D FCC lattice. *BMC Bioinformatics* 14: S16.

Maksym, T. and Laura, I. S. 2013. The how's and why's of protein folding intermediates. *Archives of Biochemistry and Biophysics* 531: 14-23.

Mann, M., Will, S. and Backofen, R. 2008b. CPSP-tools-exact and complete algorithms for high-throughput 3D lattice protein studies. *BMC Bioinformatics* 9: 230.

Marian, N. 2006. *Protein folding with coarse-Grained off-lattice models of the polypeptide chain.*: PhD. Thesis. Cornell University, New York. 1-146.

Martin, M. 2011. *Computational methods for lattice protein models.* PhD. Thesis. der Albert-Ludwigs-Universitat Freiburg, Germany. 1-137.

Martin, M., Rhodri, S., Cameron, S., Rolf, B. and Charlotte, D. M. 2012. Producing High-Accuracy Lattice Models from protein atomic coordinates including side Chains. *Hindawi Publishing Corporation Advances in Bioinformatics* 2012*:* 6.

Mazzoni, L. N. and Casetti, L. 2006. Curvature of the energy landscape and folding of

model proteins. *Physical Review Letters* 97(21) : 218, 104.

Meng, G., Huaiqiu, Z., Xin-Qiu, Y. and Zhen-su-she. 2010. Water dynamics clue to key residues in protein folding. *Biochemical and Biophysical Research Communication* 392: 95-99.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* 21:1087-1092.

Michael, P. A. and Adam, D. 2012. Wang-Landau simulations of adsorbed and confined lattice polymers. *Physics Procedia* 34: 6-13.

Michael, T. 2003. *Protein Folding, HIV and Drug Design.* Department of Physics. Michigan State University, Michigan: APS New. Feb.10.

Moret, M. 2011. Self-organized critical model for protein folding. *Physica A* 390: 3055-3059.

Morrissey, M. P. and Shakhnovich, E. 1996. Design of proteins with selected thermal properties. *Fold. Des.* 1: 391-405.

Newman, M. and Barkema, G. 1999. *Monte Carlo methods in statistical physics.* New York: Oxford University Press.

Nikolas, B. S., Csilla, V., Stephen, W. A. and David, W. L. 2012. Exploring the energy landscapes of protein folding simulations with Bayesian computation. *Biophysical Journal* 102: 878-886.

Nobuhiro, G. 2007. Physics and biology of protein. *Progress of theoretical physics supplement* 170.

Olav, Z. and Ulrich, H. H. 2008. Uderstanding protein folding: small proteins in

silico. *Biochimica et Biophysica Acta* , 1784: 252-258.

Oren, B. M., Alexander, M. D., RouX, B. and Masaktsu, W. 2001. *Computational Biochemistry and Biophysics.1st ed.* New York: Eastern Hemisphere.

Oyewande, E. O. 2012. Surface science and Markov chain Monte Carlo simulation of disordered systems. *arXiv:1207, 0744v1 [cond-mat.stat-mech] 3 Jul 2012* , 1-20.

Pain, R. 2000. *Mechanism of protein folding.* Oxford: Oxford University.

Peng, Z., Yi, C., Tianjia, B., Suzana, S. K. and Hongbin, L. 2010. Single molecule force spectroscopy reveals that electrostatic interactions affect the mechanical stability of proteins. *Biophysical Journal* 100: 1534-1541.

Petsko, G. A. 2001. Size doesn't matter. *Genome biology* 2: 1003.1-1003.2.

Plaxco, K. W. and Baker, D. 1998. Limited internal friction in the rate-limiting step of a two-state protein folding reaction. *PNAS* 95: 13591-13596.

Prusiner, S. 1991. Molucular biology of prior diseases. *Science* 252: 1515-1522.

Ramachandran, G. and Sasisekharan, V. 1968. Conformation of polypeptides and proteins. *Advances in protein Chemistry* 23: 49-85.

Ramakrishnan, R., Ramachandran, B. and Pekny, J. 1997. A dynamic Monte Carlo algorithm for exploration of dence conformational spaces in heteropolymers. *Journal of Chemical Physics* 106.6: 2418-2424.

Rashid, M., Newton, M. A., Hoque, M. T. and Shatabda, S. 2013. Spiral search: a hydrophobic-core directed local search for simplified PSP on 3D FCC lattice. *BMC Bioinfo.* 14: S16.

Ron, U. and John, M. 1993. Genetic algorithms for protein folding simulation. *Jounal Molecular Biology* 231: 75-81.

Ronald, D. H. and Charles, L. B. 2009. Insights from Coarse-Grained Go models for protein folding and dynamics. *Int. J. Mol. Sci.* 10: 889-905.

Sandelin, E. 2004. On hydrophobicity and conformational specificity in proteins. *Biophysical Journal*. 86: 23-30.

Scott, E. A., Johannes, W. and Frauke, G. 2012. Dynamic prestress in a globular protein. *Plos Computational Biology* 8.5: e1002509.

Sebastian, W. 2005. *Exact, Contraint- based structure prediction in simple protein models.* Jena: der Fakultat fur Mathematik und Informatik der Friedrich-Schiller Universitat Jena. 1-144

Seno, F., Vendruscolo, M., Maritan, A. and Banaver, J. R. 1996. An optimal protein design procedure. *Phys. Rev. Lett.* , 77: 1901-1904.

Slade, G. 2011. *The Self-avoiding walk: a brief survey. Selected papers based on the presentations at the 33rd conference on stochastic processes and thier applications. Surveys in stochastic processes. Blath, Jochem, et al. Eds.* Berlin, Germany: European Mathematical Society (EMS).

Sooyoung, C. and Faming, L. 2011. Folding small proteins via annealing stochastic approximation Monte Carlo. *Biosystem* 105: 243-249.

Takeshi, K. 2009. Spin correlations in a non-frustrated one-dimensional spin system and formation of the ground state as a model of protein folding. *Physica A* 388: 129-136.

Thachuk, C., Shmygelska, A. and Hoos, H. 2007. A replica exchange monte carlo algorithm for protein folding in the HP model. *BMC Bioinformatics* 8: 342.

Tristan, B. and Markus, D. 2009. Generic coarse-grained model for protein folding and aggregation. *Journal of Chemical Physics* 130: 235106.

Ullah, A., Kapsokalivas, L., Mann, M. and Steinhofel, K. 2009. Protein folding simulation by two-stage optimization. *Proceeding of ISICA'09, CCIS. 51*. Wuhan, China: Springer. 138-145.

Unger, R. Moult, J. 1993. A genetic algorithm for 3D protein folding simulations. In soft computing-A Fusion of foundations, Methodologies and Applications. *The 5th International Conference on Genetic Algorithms*. Morgan Kaufmann Publishers. 581.

Victor, M. and Serrano, L. 1996. Local versus nonlocal interactions in protein folding and stability- an experimentalist's point of view. *Folding and Design* 1: 71-77.

William, H. E. and Alantha, N. 2001. *Protein structure prediction with lattice models.* Sandia National Laboratories & Massachusetts Institute of Technology. USA

Wilson, C., Apiyo, D. and Wittung-stafshede, P. 2004. Role of cofactors in metalloprotein folding. *Q. Rev.Biophys* 37: 285-314.

Wittung-stafshede, P. 2002. Role of cofactors in protein folding. *Acc.Chem.Res*. 35: 201-208.

Wust, T., Landau, D., Gervais, C. and Xu, Y. 2009. Monte Carlo simulations of systems with complex energy landscapes. *Computer Physics Communication* 180: 475-479.

Wuthrich, K. 1990. Protein structure determination in solution by NMR spectrscopy. *J. Biol. Chem. 265.36*: 22,059-22,062.

Xia, Y. and Levitt, M. 2004b. Funnel-like organization in sequnce space determines the distributions of protein stability and folding rate preferred by evolution. *Proteins* 55.1: 107-114.

Yantao, C., Qi, Z. and Jiandong, D. 2004. A Coarse-Grained and associated lattice Monte Carlo simulation of the coil-helix transition of a homopolypeptide. *Journal of Chemical Physics* 120: 7.

Yi, C. and Hongbin, L. 2011. Dynamic of protein folding and cofactor binding monitored by single-molecule for spectroscopy. *Biophysical Journal* 101: 2009-2017.

Ying, W. L., Thomas, W. and David, L. P. 2011. Monte Carlo simulation of the HP model (The "Ising model" of protein folding). *Computer Physics Communication* 182: 1896-1899.

Zhao, X. 2008. Advances on protein folding simulations based on the lattice HP models with natural computing. *Applied Soft Computing* 8: 1029-1040.

Zia, R., Dong, J. and Schmittman, B. 2011. Modeling translation in protein synthesis with TASEP: a tutorial and recent development. *Cond-mat.stat-mech arxive 1108.3312v1* , 20-25.

http://en.wikipedia.org/wiki/Protein, retrieved April 20, 2013

http://en.wikipedia.org/wiki/Protein_folding , retrieved April 20, 2013

http://en.wikipedia.org/w/index.php?title=Protein_structure&oldid=551224293, retrieved April 20, 2013

http://www.rcsb.org/, retrieved August 18, 2014

Appendix A program document

A.1 A FORTRAN code for calculating the physical mechanism

```
Program saw_fixed

        Implicit none

 !!! declaration
   integer :: step, row, col, prow, pcol, Nsteps!prow=previous row
   real, parameter :: pd=0.02, pl=0.25, pr=0.02, pu_incr=0.01!directional probabilities
!comment out
   !for simulating conformational bias towards different(4) possible directions.
   !Should all add up to 1 (including the last, not included, one -wgt_left).
   real :: pu, alpha, beta, dee!pu=prob_up,pr=prob_right,alpha=prob_up/prob_down,
beta=prob_right/prob_left,dee=alpha/beta
   integer :: pu_loop, Nloop, rz
   integer, parameter :: start_pt=-1, Nrz=50 !The different value "start_pt", instead of
"1"
   !for the first lattice site stepped on, is distinguish it as a starting point on an excel
plot
   integer :: seq_len, length(Nrz) !lattice site occupation number = ith step of sawalker
after starting point.
        integer, parameter :: Msteps=10000 !Mstep=maximum number of steps in the
random walk
   integer, parameter :: Lrow = 100, Lcol = 100 !length/size of each dimension.
        integer, dimension(1:Lrow, 1:Lcol) :: site! 2-dim lattice of (discrete) position
sites.
        !integer :: size=0, get(8)=0 !variables required in subroutine "random_seed".
        !we let "get" have the same size as "values"
        !integer, parameter :: size = 8
        !integer :: put(size), get(size)
        integer :: put(20), get(20)
```

```
        character(len=10) :: date, time, zone !required by subroutine "date_and_time"
        real :: rn !rn=random number; output of subroutine "random_number".
0<=rn<1
        character(LEN=12):: rcfmt, stfmt, analfmt !see below
        parameter (rcfmt='(i5), 2x', stfmt='(i5), 2x', analfmt='(f6.2), (f6.2)')!output
format for writing row and col(rcfmt), and site(stfmt)

        call date_and_time(date, time, zone, put) !we use this subroutine to generate
varying initial values for
          !"random_seed"; i.e. since this program will be run at different times in the
year of our lord, it is not possible
          !to get the same "values" at these different times.

!  open files for output
      open(unit=10, file='saw_fixed.dat', status='Unknown')
          open(unit=20, file='saw_dee.dat', status='Unknown')
          open(unit=30, file='saw_alpha.dat', status='Unknown')
          open(unit=40, file='saw_beta.dat', status='Unknown')

        !initialize the random number generator by using "random_seed"
        !call random_seed(put)
        !call random_seed(get) ! generate seed for random_number generator
          call random_seed()

! initialize 2-dim square lattice of (discrete) position sites.
   site = 0 !none of the lattice sites has been stepped on
   !Nsteps = 35
   Nloop = (0.245-0)/pu_incr + 1
   pu = 0.0 - pu_incr
   do pu_loop = 1, Nloop
   !set parameters
      pu = pu + pu_incr
      !pr = 0.5 - pu
```

```
      alpha = pu/pd
      beta = pr/pl !beta>=0.02

      dee = alpha/beta

  ! Loop over realizations

  length=0

  do rz=1, Nrz
! SAW steps
   do step = 1, Msteps
      if(step /= 1) then !if not the SAWalker's first (or starting) step then
      !choose a nearest neighbour site for next step of SAWalker

         !call date_and_time(date, time, zone, put)
         !call random_seed(put)
         !write(*,*) put(1),put(2),put(3),put(4),put(5),put(6),put(7),put(8)
         !read(*,*)
         call random_number(rn)
         !write(*,*) 'rn=', rn
         !read(*,*)
         prow=row
         pcol=col
         if(rn<pd) then
            row=row+1 !walker goes down
            if(row==Lrow+1) then  !apply periodic boundary condition
               row=1
            endif
         else if(rn<(pd+pu)) then
            row=row-1 !walker goes up
            if(row==0) then  !apply periodic boundary condition
```

```
        row=Lrow
      endif
    else if(rn<(pd+pu+pr)) then
      col=col+1 !walker goes right
      if(col==Lcol+1) then  !apply periodic boundary condition
        col=1
      endif
    else
      col=col-1 !walker goes left
      if(col==0) then  !apply periodic boundary condition
        col=Lcol
      endif
    endif


    !if(seq_len == Nsteps) then !this cuts short any occurrence of
    !unusual long chains
     !   exit
    !else if(site(row,col) == 0) then !site is unstepped, walker steps on it
    if(site(row,col) == 0) then !site is unstepped, walker steps on it
      !site(row,col)=1
      length(rz)=length(rz)+1
      site(row,col)=length(rz) !introduced here to keep track of sawalker's steps
    else
      exit !walk terminates
    endif

  else !choose a random site for starting position of SAW

      call random_number(rn)
      !row=int(1+rn*(Lrow-1)) !the "int" function truncates 0.x to zero!
      row=1+rn*(Lrow-1)
      call random_number(rn)!choose a random site for starting position of SAW
```

```
            !col=int(1+rn*(Lcol-1))
            col=1+rn*(Lcol-1)
            site(row,col)=start_pt !starting point of saw. site_ij has been stepped on.
            length(rz)=0 !start tracking sawalker's steps
            !write(*,*) "row=", row, "col=", col, "site=", site(row,col)
        endif

    enddo

    enddo
! write(20, analfmt) dee, seq_len
! write(30, analfmt) alpha, seq_len
! write(40, analfmt) beta, seq_len
seq_len=0
do rz = 1, Nrz
    seq_len=seq_len+length(rz)
enddo
seq_len=seq_len/Nrz  !average sequence length


    write(20, *) dee, seq_len
    write(30, *) alpha, seq_len
    write(40, *) beta, seq_len

enddo


! output data
    write(10, "(t6)", advance="no")

        do col=1, Lcol
            if(col<Lcol) then
                write(10, rcfmt, advance="no") col
```

```
      else
        write(10, rcfmt) col
      endif
    enddo

    do row=1, Lrow
      write(10, rcfmt, advance="no") row

      do col=1, Lcol
        if(col<Lcol) then
          write(10, stfmt, advance="no") site(row,col)
          !write(*,*) "row=", row, "col=", col, "site=", site(row,col)
        else
          write(10, stfmt) site(row,col)
          !write(*,*) "row=", row, "col=", col, "site=", site(row,col)
        endif
      enddo

    enddo

      close(10)

!     stop 'data saved in sawfx.dat'
      End
```

A.2 a FORTRAN code for MCSAW for each conformation

```
Program saw_fixed
        Implicit none
 !!! declaration
   integer :: step, row, col, prow, pcol, Nsteps!prow=previous row
   real, parameter :: wgt_down=0.25, wgt_up=0.25, wgt_rgt=0.25!directional weights
!comment out
   !for simulating conformational bias towards different(4) possible directions.
   !Should all add up to 1 (including the last, not included, one -wgt_left).
   integer, parameter :: start_pt=-1 !The different value "start_pt", instead of "1"
   !for the first lattice site stepped on, is distinguish it as a starting point on an excel
plot
   integer :: occ_num !lattice site occupation number = ith step of sawalker after
starting point.
        integer, parameter :: Msteps=50000 !Mstep=maximum number of steps in the
random walk
   integer, parameter :: Lrow = 100, Lcol = 100 !length/size of each dimension.
        integer, dimension(1:Lrow, 1:Lcol) :: site! 2-dim lattice of (discrete) position
sites.
        !integer :: size=0, get(8)=0 !variables required in subroutine "random_seed".
        !we let "get" have the same size as "values"
        !integer, parameter :: size = 8
        !integer :: put(size), get(size)
        integer :: put(20), get(20)
        character(len=10) :: date, time, zone !required by subroutine "date_and_time"
        real :: rn !rn=random number; output of subroutine "random_number".
0<=rn<1
        character(LEN=12):: rcfmt, stfmt !see below
        parameter (rcfmt='(i5), 2x', stfmt='(i5), 2x')!output format for writing row and
col(rcfmt), and site(stfmt)
        call date_and_time(date, time, zone, put) !we use this subroutine to generate
varying initial values for
```

```
!"random_seed"; i.e. since this program will be run at different times in the year of our lord, it is not possible
!to get the same "values" at these different times.
!  open files for output
      open(unit=10, file='sawfx.dat', status='Unknown')
    !initialize the random number generator by using "random_seed"
    !call random_seed(put)
    !call random_seed(get) ! generate seed for random_number generator
      call random_seed()
! initialize 2-dim square lattice of (discrete) position sites.
  site = 0 !none of the lattice sites has been stepped on
  Nsteps = 59
! SAW steps
  do step = 1, Msteps
    if(step /= 1) then !if not the SAWalker's first (or starting) step then
    !choose a nearest neighbour site for next step of SAWalker
      !call date_and_time(date, time, zone, put)
      !call random_seed(put)
      !write(*,*) put(1),put(2),put(3),put(4),put(5),put(6),put(7),put(8)
      !read(*,*)
      call random_number(rn)
      !write(*,*) 'rn=', rn
      !read(*,*)
      prow=row
      pcol=col
      if(rn<wgt_down) then
        row=row+1 !walker goes down
        if(row==Lrow+1) then  !apply periodic boundary condition
          row=1
        endif
      else if(rn<(wgt_down+wgt_up)) then
        row=row-1 !walker goes up
        if(row==0) then  !apply periodic boundary condition
```

```
            row=Lrow
          endif
        else if(rn<(wgt_down+wgt_up+wgt_rgt)) then
          col=col+1 !walker goes right
          if(col==Lcol+1) then  !apply periodic boundary condition
            col=1
          endif
        else
          col=col-1 !walker goes left
          if(col==0) then  !apply periodic boundary condition
            col=Lcol
          endif
        endif
        if(occ_num == Nsteps) then !this cuts short any occurrence of
        !unusual long chains
          exit
        else if(site(row,col) == 0) then !site is unstepped, walker steps on it
          !site(row,col)=1
          occ_num=occ_num+1
          site(row,col)=occ_num !introduced here to keep track of sawalker's steps
        else
          !ordinarily the saw should terminate here but we now put
          !the condition that if the protein chain is not up to the specified length
(msteps)
          !then the saw should not terminate but instead start afresh until msteps
attained.
          !write(10, *) "row=", row, "col=", col, "site=", site(row, col), "ocno",
occ_num
          row=prow  !give back the 'eyin'
          col=pcol
        endif
      else !choose a random site for starting position of SAW
        call random_number(rn)
```

```
            !row=int(1+rn*(Lrow-1)) !the "int" function truncates 0.x to zero!
            row=1+rn*(Lrow-1)
            call random_number(rn)!choose a random site for starting position of SAW
            !col=int(1+rn*(Lcol-1))
            col=1+rn*(Lcol-1)
            site(row,col)=start_pt !starting point of saw. site_ij has been stepped on.
            occ_num=0 !start tracking sawalker's steps
            !write(*,*) "row=", row, "col=", col, "site=", site(row,col)
        endif
    enddo
! output data
    write(10, "(t6)", advance="no")
        do col=1, Lcol
            if(col<Lcol) then
                write(10, rcfmt, advance="no") col
            else
                write(10, rcfmt) col
            endif
        enddo
        do row=1, Lrow
            write(10, rcfmt, advance="no") row
            do col=1, Lcol
                if(col<Lcol) then
                    write(10, stfmt, advance="no") site(row,col)
                    !write(*,*) "row=", row, "col=", col, "site=", site(row,col)
                else
                    write(10, stfmt) site(row,col)
                    !write(*,*) "row=", row, "col=", col, "site=", site(row,col)
                endif
            enddo
        enddo
            close(10)
            stop 'data saved in sawfx.dat'
```

End

A.3 C# code for the mapping of the conformation on square lattice

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Windows;
using System.Windows.Controls;
using System.Windows.Data;
using System.Windows.Documents;
using System.Windows.Input;
using System.Windows.Media;
using System.Windows.Media.Imaging;
using System.Windows.Navigation;
using System.Windows.Shapes;
using System.IO;

namespace Mr_Aisida
{
   /// <summary>
   /// Interaction logic for MainWindow.xaml
   /// </summary>
   public partial class MainWindow : Window
   {
      string dir = "Mr Aisida's Data 764756522"; string path; string dir_data = @"Mr
Aisida's Data 764756522\Data";
      string docpath =
Environment.GetFolderPath(Environment.SpecialFolder.Desktop); string temppath1;
string temppath2; //int count_file = 0;
      string[] storage_file = new string[10];
      // Create points that define curve.
```

```
        Line y0 = new Line(); Line y1 = new Line(); Line y2 = new Line(); Line y3 =
new Line();
        Line y4 = new Line(); Line y5 = new Line(); Line y6 = new Line(); Line y7 =
new Line();
        Line y8 = new Line(); Line y9 = new Line(); Line y10 = new Line(); Line y11 =
new Line();

        Line x0 = new Line(); Line x1 = new Line(); Line x2 = new Line(); Line x3 =
new Line();
        Line x4 = new Line(); Line x5 = new Line(); Line x6 = new Line(); Line x7 =
new Line();
        Line x8 = new Line(); Line x9 = new Line(); Line x10 = new Line(); Line x11 =
new Line();

        SolidColorBrush gridBrush = new SolidColorBrush(); SolidColorBrush
elipsBrush_red = new SolidColorBrush();
        SolidColorBrush elips_labelBrush = new SolidColorBrush(); SolidColorBrush
elipsBrush_green = new SolidColorBrush();
        SolidColorBrush elipsBrush_blue = new SolidColorBrush(); SolidColorBrush
elipsBrush_start = new SolidColorBrush();
        SolidColorBrush lineBrush_black = new SolidColorBrush(); char[] sep_atoms = {
'|' }; char[] sep_coords = { ',' };

        string atomic_h = null; string[] in_atomic_h = new string[1000]; string[]
in_atomic_h_inn = new string[1000]; int atom_number_h;
        int[] keep_atoms_h_x = new int[600]; int[] keep_atoms_h_y = new int[600]; int
count_atoms_h = 0; int count_atoms_h_ = 0; int[] out_atoms_h = new int[50];

        string atomic_p = null; string[] in_atomic_p = new string[1000]; string[]
in_atomic_p_inn = new string[1000]; int atom_number_p;
        int[] keep_atoms_p_x = new int[600]; int[] keep_atoms_p_y = new int[600]; int
count_atoms_p = 0; int count_atoms_p_ = 0; int[] out_atoms_p = new int[50];
```

```
        string atomic_o = null; string[] in_atomic_o = new string[1000]; string[]
in_atomic_o_inn = new string[1000]; int atom_number_o;
        int[] keep_atoms_o_x = new int[600]; int[] keep_atoms_o_y = new int[600]; int
count_atoms_o = 0; int count_atoms_o_ = 0; int[] out_atoms_o = new int[50];

        string atomic_a = null; string[] in_atomic_a = new string[1000]; string[]
in_atomic_a_inn = new string[1000]; int atom_number_a;
        int[] keep_atoms_a_x = new int[600]; int[] keep_atoms_a_y = new int[600]; int
count_atoms_a = 0; int count_atoms_a_ = 0; int[] out_atoms_a = new int[50];

        string username = null; int u = 0; int uu = 0;string combo_sel;

        public MainWindow()
        {
            InitializeComponent();
            //set_brush_for_plotting();
            path = System.IO.Path.Combine(docpath, dir);
System.IO.Directory.CreateDirectory(path);
            temppath1 = System.IO.Path.Combine(docpath, dir_data);
System.IO.Directory.CreateDirectory(temppath1);
            gridBrush.Color = Colors.Blue; elipsBrush_red.Color = Colors.Red;
elipsBrush_green.Color = Colors.Green;
            elipsBrush_blue.Color = Colors.DarkBlue; elips_labelBrush.Color =
Colors.White;
            elipsBrush_start.Color = Colors.Yellow; lineBrush_black.Color =
Colors.Black;
            atoms_h.IsEnabled = false; atoms_all.IsEnabled = false; atoms_p.IsEnabled =
false;
            //scaling();
            loaddatafile();
        }

        private void loaddata(string d)
```

```
        {
            temppath2 = System.IO.Path.Combine(temppath1, d);
            System.IO.StreamReader objReader = new System.IO.StreamReader(d);
            for (int y = 0; y < 4; y++) { storage_file[y] = objReader.ReadLine(); }
            objReader.Close();
            atoms_o.Text = storage_file[0]; atoms_p.Text = storage_file[1]; atoms_h.Text
= storage_file[2]; atoms_all.Text = storage_file[3];
        }

        private void loaddatafile()
        {
            foreach (string f in Directory.GetFiles(temppath1, "*.txt"))
            {
                comboBox1.Items.Add(f);
            }
        }

        private void comboBox1_SelectionChanged(object sender,
SelectionChangedEventArgs e)
        {
            combo_sel = comboBox1.SelectedItem.ToString(); loaddata(combo_sel);
        }

        private void Save_Click(object sender, RoutedEventArgs e)
        {//button Save
            if (user.Text == "") { MessageBox.Show("Please input a name for this current
project.)
             MessageBoxButton.OK); return; }
            else
            {
                username = user.Text;
                save_atoms(username);
            }
```

```
}

private void Set_Click(object sender, RoutedEventArgs e)
{//button Set
   canva.Children.Clear();
   for (int y = 0; y < 4; y++) { storage_file[y] = ""; }
   perform_Task();
}
private void atoms_p_TextChanged(object sender, TextChangedEventArgs e)
{//once you type into the Set_P textbox
   atoms_h.IsEnabled = true;
}
private void atoms_o_TextChanged(object sender, TextChangedEventArgs e)
{//no need
   atoms_p.IsEnabled = true;
}
private void atoms_h_TextChanged(object sender, TextChangedEventArgs e)
{
   atoms_all.IsEnabled = true;
}

private void atoms_all_TextChanged(object sender, TextChangedEventArgs e)
{
   Set.IsEnabled = true; Save.Content = "Save";
}

private void save_atoms(string myname)
{
   Rect bounds = VisualTreeHelper.GetDescendantBounds(canva);
   RenderTargetBitmap rtb = new RenderTargetBitmap((Int32)bounds.Width,
(Int32)bounds.Height, 96, 96, PixelFormats.Pbgra32);
   DrawingVisual dv = new DrawingVisual();
   using (DrawingContext dc = dv.RenderOpen())
```

```
            {
                VisualBrush vb = new VisualBrush(canva);
                dc.DrawRectangle(vb, null, new Rect(new Point(), bounds.Size));
            }
            rtb.Render(dv);
            PngBitmapEncoder png = new PngBitmapEncoder();
            png.Frames.Add(BitmapFrame.Create(rtb));
            using (Stream stm = File.Create(path + "\\" + username + ".jpg"))
            {
                png.Save(stm); Save.Content = "Saved";
            }
        }

        private void perform_Task()
        {
            count_atoms_o = 0; count_atoms_p = 0; count_atoms_h = 0;//reset
count_atoms to 0
            //checking all textbox details
            if ((atoms_o.Text == "") && (atoms_p.Text == "") && (atoms_h.Text == ""))
{ MessageBox.Show("Please input a numeric value into all spaces.
MessageBoxButton.OK); return; }
            else
            {
                //for atoms H
                atomic_h = atoms_h.Text; atom_number_h = atomic_h.Length;
                for (int y = 0; y < atom_number_h; y++)
                {
                    in_atomic_h = atomic_h.Split('|');//check for |:
                }
                for (int y = 0; y < in_atomic_h.Length; y++)
                {
                    in_atomic_h_inn = in_atomic_h[y].Split(',');//check for ,:
                    foreach (string new_atom in in_atomic_h_inn)
```

```
                {
                    int dOutput = 0;//check for user input's correctness
                    if ((int.TryParse(new_atom, out dOutput)))
                    {
                        out_atoms_h[count_atoms_h] =
Convert.ToInt16(in_atomic_h_inn[count_atoms_h]);
                        count_atoms_h++; count_atoms_h_++;
                    }
                    else { MessageBox.Show("The input number " + (count_atoms_h_ + 1)
+ " is wrongly typed. Please input the numeric value for H correctly,
MessageBoxButton.OK); return; }
                }
                keep_atoms_h_x[y] = Convert.ToInt16(out_atoms_h[0]);
                keep_atoms_h_y[y] = Convert.ToInt16(out_atoms_h[1]);
                count_atoms_h = 0;
            }

            //for atoms P
            atomic_p = atoms_p.Text; atom_number_p = atomic_p.Length;
            for (int y = 0; y < atom_number_p; y++)
            {
                in_atomic_p = atomic_p.Split('|');//check for |:
            }
            for (int y = 0; y < in_atomic_p.Length; y++)
            {
                in_atomic_p_inn = in_atomic_p[y].Split(',');//check for ,:
                foreach (string new_atom in in_atomic_p_inn)
                {
                    int dOutput = 0;//check for user input's correctness
                    if ((int.TryParse(new_atom, out dOutput)))
                    {
                        out_atoms_p[count_atoms_p] =
Convert.ToInt16(in_atomic_p_inn[count_atoms_p]);
```

```
                    count_atoms_p++; count_atoms_p_++;
                }
                else { MessageBox.Show("The input number " + (count_atoms_p_ + 1)
+ " is wrongly typed. Please input the numeric value for P correctly,
MessageBoxButton.OK); return; }
            }
            keep_atoms_p_x[y] = Convert.ToInt16(out_atoms_p[0]);
            keep_atoms_p_y[y] = Convert.ToInt16(out_atoms_p[1]);
            count_atoms_p = 0;
        }

        //for the order before now.
        //Now its for detecting the starting and ending particles particle
        atomic_o = atoms_o.Text; atom_number_o = atomic_o.Length;
count_atoms_o = 0;
        for (int y = 0; y < atom_number_o; y++)
        {
            in_atomic_o = atomic_o.Split('|');//check for |:
        }
        for (int y = 0; y < in_atomic_o.Length; y++)
        {
            in_atomic_o_inn = in_atomic_o[y].Split(',');//check for ,:
            foreach (string new_atom in in_atomic_o_inn)
            {
                int dOutput = 0;//check for user input's correctness
                if ((int.TryParse(new_atom, out dOutput)))
                {
                    out_atoms_o[count_atoms_o] =
Convert.ToInt16(in_atomic_o_inn[count_atoms_o]);
                    count_atoms_o++; count_atoms_o_++;
                }
```

```
                else { MessageBox.Show("The input number " + (count_atoms_o_ + 1)
+ " is wrongly typed. Please input the numeric value for the starting and ending points
correctly, MessageBoxButton.OK); return; }
                }
                keep_atoms_o_x[y] = Convert.ToInt16(out_atoms_o[0]);
                keep_atoms_o_y[y] = Convert.ToInt16(out_atoms_o[1]);
                count_atoms_o = 0;
            }
        }

        atomic_a = atoms_all.Text; atom_number_a = atomic_a.Length;
        for (int y = 0; y < atom_number_a; y++)
        {
            in_atomic_a = atomic_a.Split('|');//check for |:
        }
        for (int y = 0; y < in_atomic_a.Length; y++)
        {
            in_atomic_a_inn = in_atomic_a[y].Split(',');//check for ,:
            foreach (string new_atom in in_atomic_a_inn)
            {
                int dOutput = 0;//check for user input's correctness
                if ((int.TryParse(new_atom, out dOutput)))
                {
                    out_atoms_a[count_atoms_a] =
Convert.ToInt16(in_atomic_a_inn[count_atoms_a]);
                    count_atoms_a++; count_atoms_a_++;
                }
                else { MessageBox.Show("The input number " + (count_atoms_a_ + 1) +
" is wrongly typed. Please input the numeric value for all the points correctly,
MessageBoxButton.OK); return; }
                }
                keep_atoms_a_x[y] = Convert.ToInt16(out_atoms_a[0]);
                keep_atoms_a_y[y] = Convert.ToInt16(out_atoms_a[1]);
```

```
            count_atoms_a = 0;
        }

        //draw ellipse
        canva.Children.Clear(); //set_brush_for_plotting();

        for (int j = 0; j <= 600; j += 60)
        {
            for (int i = 0; i <= 600; i += 60)
            {
                Line line = new Line();
                /*** correct and working fine now. Thanks to Jesus. ***/

                /*******FOR SHOWING SCALING******
                TextBlock label = new TextBlock(); label.Height = 28; label.Width = 48;
                Canvas.SetLeft(label, j - 12); Canvas.SetTop(label, i - 12);
                label.Foreground = elipsBrush_red; label.FontSize = 12;
                label.Text = i.ToString() + "," + j.ToString();
                canva.Children.Add(label);
                **************/

                /*** correct and working fine now. Thanks to Jesus. ***/

                //Draw horizontal and vertical lines first
                if ((keep_atoms_a_y[0] == 0) && (keep_atoms_a_x[0] == 0)) { }
                else if ((i == keep_atoms_a_y[0]) && (j == keep_atoms_a_x[0]))
                {
                    //line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[0 + 1]; line.Y2 =
keep_atoms_a_x[0 + 1];//0 to 1
                    line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[0]; line.Y2 =
keep_atoms_a_x[0];//0 to 0
                    line.Stroke = lineBrush_black; line.StrokeThickness = 3;
                }
```

```
                if ((keep_atoms_a_y[1] == 0) && (keep_atoms_a_x[1] == 0)) { }
                else if ((i == keep_atoms_a_y[1]) && (j == keep_atoms_a_x[1]))
                {
                    //line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[1 + 1]; line.Y2 =
keep_atoms_a_x[1 + 1];//1 to 2
                    line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[1 - 1]; line.Y2 =
keep_atoms_a_x[1 - 1];//1 to 0
                    line.Stroke = lineBrush_black; line.StrokeThickness = 3;
                }

                if ((keep_atoms_a_y[2] == 0) && (keep_atoms_a_x[2] == 0)) { }
                else if ((i == keep_atoms_a_y[2]) && (j == keep_atoms_a_x[2]))
                {
                    //line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[2 + 1]; line.Y2 =
keep_atoms_a_x[2 + 1];//2 to 3
                    line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[2 - 1]; line.Y2 =
keep_atoms_a_x[2 - 1];//2 to 1
                    line.Stroke = lineBrush_black; line.StrokeThickness = 3;
                }

                if ((keep_atoms_a_y[3] == 0) && (keep_atoms_a_x[3] == 0)) { }
                else if ((i == keep_atoms_a_y[3]) && (j == keep_atoms_a_x[3]))
                {
                    //line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[3 + 1]; line.Y2 =
keep_atoms_a_x[3 + 1];//3 to 4
                    line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[3 - 1]; line.Y2 =
keep_atoms_a_x[3 - 1];//3 to 2
                    line.Stroke = lineBrush_black; line.StrokeThickness = 3;
                }

                if ((keep_atoms_a_y[4] == 0) && (keep_atoms_a_x[4] == 0)) { }
                else if ((i == keep_atoms_a_y[4]) && (j == keep_atoms_a_x[4]))
```

```
                {
                    //line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[4 + 1]; line.Y2 =
keep_atoms_a_x[4 + 1];//4 to 5
                    line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[4 - 1]; line.Y2 =
keep_atoms_a_x[4 - 1];//4 to 3
                    line.Stroke = lineBrush_black; line.StrokeThickness = 3;
                }

                if ((keep_atoms_a_y[5] == 0) && (keep_atoms_a_x[5] == 0)) { }
                else if ((i == keep_atoms_a_y[5]) && (j == keep_atoms_a_x[5]))
                {
                    //line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[5 + 1]; line.Y2 =
keep_atoms_a_x[5 + 1];//5 to 6
                    line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[5 - 1]; line.Y2 =
keep_atoms_a_x[5 - 1];//5 to 4
                    line.Stroke = lineBrush_black; line.StrokeThickness = 3;
                }

                if ((keep_atoms_a_y[6] == 0) && (keep_atoms_a_x[6] == 0)) { }
                else if ((i == keep_atoms_a_y[6]) && (j == keep_atoms_a_x[6]))
                {
                    //line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[6 + 1]; line.Y2 =
keep_atoms_a_x[6 + 1];//6 to 7
                    line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[6 - 1]; line.Y2 =
keep_atoms_a_x[6 - 1];//6 to 5
                    line.Stroke = lineBrush_black; line.StrokeThickness = 3;
                }

                if ((keep_atoms_a_y[7] == 0) && (keep_atoms_a_x[7] == 0)) { }
                else if ((i == keep_atoms_a_y[7]) && (j == keep_atoms_a_x[7]))
                {
                    //line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[7 + 1]; line.Y2 =
keep_atoms_a_x[7 + 1];//7 to 8
```

```
                line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[7 - 1]; line.Y2 =
keep_atoms_a_x[7 - 1];//7 to 6
                line.Stroke = lineBrush_black; line.StrokeThickness = 3;
            }

            if ((keep_atoms_a_y[8] == 0) && (keep_atoms_a_x[8] == 0)) { }
            else if ((i == keep_atoms_a_y[8]) && (j == keep_atoms_a_x[8]))
            {
                //line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[8 + 1]; line.Y2 =
keep_atoms_a_x[8 + 1];//8 to 9
                line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[8 - 1]; line.Y2 =
keep_atoms_a_x[8 - 1];//8 to 7
                line.Stroke = lineBrush_black; line.StrokeThickness = 3;
            }

            if ((keep_atoms_a_y[9] == 0) && (keep_atoms_a_x[9] == 0)) { }
            else if ((i == keep_atoms_a_y[9]) && (j == keep_atoms_a_x[9]))
            {
                //line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[9 + 1]; line.Y2 =
keep_atoms_a_x[9 + 1];//9 to 10
                line.X1 = i; line.Y1 = j; line.X2 = keep_atoms_a_y[9 - 1]; line.Y2 =
keep_atoms_a_x[9 - 1];//9 to 8
                line.Stroke = lineBrush_black; line.StrokeThickness = 3;
            }

            if ((keep_atoms_a_y[10] == 0) && (keep_atoms_a_x[10] == 0)) { }
            else if ((i == keep_atoms_a_y[10]) && (j == keep_atoms_a_x[10]))
            {
            Else { }
```