LEBANESE AMERICAN UNIVERSITY

FRAGMENT BASED PROTEIN STRUCTURE PREDICTION

By

MEGHRIG OHANES TERZIAN

A thesis
Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science

School of Arts and Sciences
July 2013

LAU
الجامعة اللبنانية الأميركية
Lebanese American University

Lebanese American University
School of Arts and Sciences - Beirut Campus

Thesis Proposal Form

Name of Student: **Meghrig Terzian** I.D.#: **201000119**

Program / Department: **Master of Science in Computer Science**

On (dd/mm/yy): **27/04/2012** has presented a Thesis proposal entitled:

Fragment Based Protein Structure Prediction

in the presence of the Committee Members and Thesis Advisor:

Advisor: Nashat Mansour [redacted] **27/04/2012**
(Name and Signature)

Committee Member: Ramzi Haraty [redacted] **27/04/2012**
(Name and Signature)

Committee Member: Hiyam El Sibai [redacted] **27/04/2012**
(Name and Signature)

Comments / Remarks / Conditions to Proposal Approval:

Date: May 2, 2012 Acknowledged by [redacted]
(Dean, School of Arts and Sciences)

cc: Department Chair
School Dean
Student
Thesis Advisor

LEBANESE AMERICAN UNIVERSITY
School of Arts and Sciences - Beirut Campus

Thesis Defense Result Form

Name of student Meghrig Terzian I.D: 201000119

Program / Department: Computer Science & Mathematics

Date of thesis defense: 18/07/2013

Thesis title: Fragment Based Protein Structure Prediction

Result of Thesis defense:

☐ Thesis was successfully defended. Passing grade is granted

☐ Thesis is approved pending corrections. Passing grade to be granted upon review and approval by thesis Advisor

☐ Thesis is not approved. Grade NP is recorded

Committee Members:

Advisor: Dr. Nashat Mansour [redacted] 18/7/2013
(Name and Signature)

Committee Member: Dr. Ramzi Haraty [redacted]
(Name and Signature)

Committee Member: Dr. Mirvat El-Sibai [redacted]
(Name and Signature)

Advisor's report on completion of corrections (if any):

Changes Approved by Thesis Advisor: N. M. Signature: [redacted]

Date: 18/07/2013 Acknowledge by [redacted]
(Dean, School of Arts and Sciences)

Cc: Registrar, Dean, Chair, Advisor, Student

LAU
اللبنانية الأميركية
Lebanese American University

LEBANESE AMERICAN UNIVERSITY

School of Arts and Sciences - Beirut Campus

Thesis Approval Form

Student Name: Meghrig Terzian I.D. #: 201000119

Thesis Title: Fragment Based Protein Structure Prediction

Program : Master of Science in Computer Science.

Department : Computer Science & Mathematics

School : School of Arts and Sciences

Approved by :

Thesis Advisor: Dr. Nashat Mansour Signature [redacted]

Member : Dr. Ramzi Haraty Signature :

Member : Dr. Mirvat El-Sibai Signature

Date : 18/07/2013

THESIS COPYRIGHT RELEASE FORM

LEBANESE AMERICAN UNIVERSITY NON-EXCLUSIVE DISTRIBUTION LICENSE

By signing and submitting this license, you (the author(s) or copyright owner) grants to Lebanese American University (LAU) the non-exclusive right to reproduce, translate (as defined below), and/or distribute your submission (including the abstract) worldwide in print and electronic format and in any medium, including but not limited to audio or video. You agree that LAU may, without changing the content, translate the submission to any medium or format for the purpose of preservation. You also agree that LAU may keep more than one copy of this submission for purposes of security, backup and preservation. You represent that the submission is your original work, and that you have the right to grant the rights contained in this license. You also represent that your submission does not, to the best of your knowledge, infringe upon anyone's copyright. If the submission contains material for which you do not hold copyright, you represent that you have obtained the unrestricted permission of the copyright owner to grant LAU the rights required by this license, and that such third-party owned material is clearly identified and acknowledged within the text or content of the submission. IF THE SUBMISSION IS BASED UPON WORK THAT HAS BEEN SPONSORED OR SUPPORTED BY AN AGENCY OR ORGANIZATION OTHER THAN LAU, YOU REPRESENT THAT YOU HAVE FULFILLED ANY RIGHT OF REVIEW OR OTHER OBLIGATIONS REQUIRED BY SUCH CONTRACT OR AGREEMENT. LAU will clearly identify your name(s) as the author(s) or owner(s) of the submission, and will not make any alteration, other than as allowed by this license, to your submission.

Name: Meghrig Ohanes Terzian

Signature: [redacted] Date: 18/07/2013

PLAGIARISM POLICY COMPLIANCE STATEMENT

I certify that I have read and understood LAU's Plagiarism Policy. I understand that failure to comply with this Policy can lead to academic and disciplinary actions against me.
This work is substantially my own, and to the extent that any part of this work is not my own I have indicated that by acknowledging its sources.

Name: Meghrig Ohanes Terzian

Signature: 

Date: 18/07/2013

ACKNOWLEDGMENTS

This dissertation would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

First, I would like to express my most sincere gratitude to my advisor, Dr. Nashat Mansour, and my committee members, Dr. Ramzi Haraty and Dr Mirvat El-Sibai, for their constant guidance throughout my research.

This thesis would have remained a dream had it not been for the financial support provided by the Lebanese American University during my graduate studies.

Finally, I would like to thank my family, specially my mother for her patience and support throughout the years of my graduate studies; I dedicate this thesis to my fiancé Shuntt Tanielian for being a great source of encouragement and for always believing in me.

Fragment Based Protein Structure Prediction

Meghrig Ohanes Terzian

Abstract

In recent years, the protein structure prediction problem has come under extensive investigation, and computational structure prediction methods are circumventing the time-consuming experimental methods by accelerating the prediction process. This work presents a fragment based protein tertiary structure prediction method that provides suboptimal structures. In addition, it demonstrates the advantage of using the CHARMM36 energy model. The method is based on a two-phase Scatter Search metaheuristic that minimizes the energy function. Backbone fragments selections extracted from the Robetta server are followed by side chain selections, extracted from the Dunbrack Library. The results, evaluated on three proteins, show that the algorithm produces tertiary structures with promising root mean square deviations, within reasonable times.

Keywords: Protein Structure Prediction, Scatter Search, CHARMM36, Fragments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

Introduction

1.1 – Protein Structure Prediction

Proteins, macromolecules found in all biological organisms, are composed of a linear sequence of up to 5,000 amino acids and involved in a wide variety of functions within cells including cell structure, cell motility, cell signaling, enzyme catalysis, and substance transport. Structural proteins, such as actin and tubulin, are responsible for cell structure. Motor proteins, such as myosin, are responsible for muscle contraction and cell motility. Proteins responsible for cell signaling, such as insulin, bind to a signaling molecule and provoke a biochemical reaction. Enzymatic proteins, such as pepsin, are fundamental for metabolism and accelerate the rates of biochemical reactions. Transport Proteins, such as hemoglobin, transport molecules and ions within the cell or across the cell membrane and can be engaged in vesicular transport. These various functions mentioned are determined by the structure attained by the protein after proper folding via its unique sequence of amino acids, driven by non-covalent interactions between amino acids such as hydrogen bonding, ionic interactions, Van der Waals forces, hydrophobic packing, and environmental effects such as presence of water or lipids and the pH of the surrounding.

Predicting protein tertiary structure provides information about the functionality, localization and interactions between proteins and consequently contributes in drug design and disease prevention associated with protein misfold. The experimental methods for protein structure prediction, mainly X-ray crystallography and Nuclear Magnetic Resonance not only consume time but are expensive, computational methods if managed to be coded properly have proven to be quite inexpensive with respect to the experimental methods on one hand and much less time consuming on another note.

1.2 – Methods for Protein Structure Prediction

1.2.1 – Experimental Methods

The laboratory techniques used to determine the tertiary protein structure include X-ray crystallography, Cryo-electron microscopy, Nuclear Magnetic Resonance spectroscopy, and Small-angle X-ray and neutron scattering. These methods consume time, money, and are not applicable on all proteins.

X-ray crystallography is a form of very high resolution microscopy, which enables protein structure visualization at atomic level and contributes in better understanding of protein function. It mainly involves five steps: protein purification, crystallization, data collection and processing, phase determination, model building, and refinement and analysis. After protein purification and crystallization, the directions and intensities of x-ray rays diffracted from the crystals are calculated, thus the protein structure is predicted (Ilari and Savino, 2008).

Cryo-electron microscopy is a microscopy method where beams of electrons are transmitted through samples under very low temperatures. Unlike X-ray crystallography it allows the observation of samples instantaneously, without the need to pass through a series of steps that might lead to inappropriate conformational changes. But because of the low resolution of the maps, it is not possible to determine structures based on Cryo-electron microscopy maps only, models from protein crystallography are used to understand the Cryo-electron microscopy maps (Bartesaghi and Subramaniam, 2009).

Nuclear Magnetic Resonance manages the rotating states of the nuclei using radio frequency pulses and high magnetic fields. The locations and intensities of the crest on the spectrum imitate the nucleic positions and the chemical environment of the molecule (Wüthrich, 1990).

Small-angle X-ray and neutron scattering provides information about the structure, domain organization, and contacts of macromolecules in a solution. Joint with deuterium labeling it ascertains the positions of explicit components within a complex (Neylon, 2008).

1.2.2 – Computational Methods

Computational approaches for protein structure prediction lie in two groups. The first group, comparative modeling, predicts structures using proteins of known structures as templates. The second, *ab initio,* predicts structures using the amino acid sequence of the structure to be predicted.

1.2.2.1 – Comparative Modeling Methods

Methods lying in this group are further divided into two subgroups, homology modeling and fold recognition or threading.

Homology Modeling

Homology Modeling is based on the supposition that new proteins evolve from previous proteins after undergoing a series of alterations by amino acid substitution, addition, or deletion that do not affect their tertiary structures. Strong sequence similarity often indicates strong function similarity (Kopp and Schwede, 2004).

Homology modeling based predictions start with a database search to select homologous sequences to the query sequence from protein data banks. This selection is performed by comparing protein sequence profiles, Hidden Markov Models (HMM), or by BLAST pairwise alignments. In the latter, selection of sequences with less than 30% similarity, limits their significance. Profile methods, on the contrary, identify proteins that do not share significant homology. Using sophisticated methods like comparing sequences to profiles (PSI-BLAST), profiles to profiles (FFAS), or HMMs to HMMs, generates more precise homologues (Chen et al., 2009). PSI-BLAST generates homologous with twice the accuracy of BLAST. Moreover, FFAS identifies more than twice as many cases as PSI-Blast (Jaroszewski et al., 2011).

From the identified homologues, either multiple templates each with average similarity or a single best template is chosen. Multi-template modeling produces structures including the best fragments of multiple templates, resulting in an average model. Alternatively, using a single best template does not result in an average model, but leads to the dilemma of determining the best template (Kopp and Schwede, 2004).

The extracted templates and the query sequence are aligned using single or multiple sequence alignment algorithms to identify structurally conserved regions (SCR), active site residues, disulphide bridges, and salt bridges. When performing an alignment, selecting the proper aligning algorithm and scoring method are crucial (Doong, 2007).

After the alignment phase, the modeling phase starts where the main chain, loops and side chains are modeled successively, then the energy is refined bringing the conformation to the nearest local minimum. Finally, the model is evaluated by assessing the values of bond angles, bond lengths, and dihedral angles. Swiss-model and Modeler are examples of homology modeling servers (Kopp and Schwede, 2004).

Homology modeling faces a number of challenges which primary include the precise sequence-structure alignment, best template selection in cases where sequences do not share significant homology; on another note the model quality might decline as a result of refinement, and finally there sometimes is a lack of sufficient amount of structural information available to build models.

Fold Recognition or Threading.

Fold Recognition is based on the assumption that although proteins evolve due to mutations by nucleotide substitution, addition, or deletion, they maintain their role and structure. The best sequence-structure alignment is generated from the amino acid sequence and a database of probable protein conformations (Abual-Rub and Abdullah, 2008).

A database of protein conformations, an energy function to measure the correctness of an alignment, a search algorithm to find the best alignment, and a way of determining the best fold among the best scoring alignments is needed in fold recognition. Once a template has been identified, the remaining of the process is the same as comparative modeling (Abual-Rub and Abdullah, 2008).

The complexity depends on whether variable length gaps are allowed in alignments and whether the energy function includes amino acids interactions. Various Methods have been proposed to tackle computational difficulty such as Molecular Dynamics,

Neural Networks, Monte Carlo simulations, and Genetic Algorithms (Abual-Rub and Abdullah, 2008).

The efficiency of Fold Recognition is limited by the size of the Protein Data Bank (PDB). However, if sequence similarities are less than 25%, it outperforms comparative modeling (Sikder and Zomaya, 2005).

1.2.2.2 – *Ab Initio* Methods

Ab initio is based on Anfinsen's theory stating that the lowest energy value protein conformation is the most stable one (Anfinsen, 1973). Given the particular amino acid sequence of a protein, *ab initio* methods predict its tertiary structure.

The *ab initio* method is divided into two classes. The first being fragment based and the second biophysics based. While fragment based methods employ database information, biophysics based methods don't (Floudas, 2007).

A typical *ab initio* method starts with random conformations, generates substitute conformations using heuristics, calculates their energies, and keeps on generating substitute conformations until the ending criterion is reached, where the solution is the conformation with the lowest energy. The efficiency of *ab initio* methods depends on the utilized energy function accuracy and the search algorithm efficiency (Bradley et al., 2005).

The main challenge of this method is the search space vastness. To limit the search space a number of models such as the Hydrophobic-Polar model, UNRES model, dihedral angles model, and Force Field model have been developed described in section 2.2. Detailed models include the interactions between all atoms of the protein sequence (Unger and Moult, 1993).

1.3 – Thesis Objectives, Contribution, and Methodology

The protein structure prediction problem being NP-Complete, the need for approximation algorithms is inevitable to tackle it (Unger and Moult, 1993). The employment of the fairly recent Scatter Search (SS) metaheuristic makes the model computationally tractable.

This work presents a fragment based protein tertiary structure prediction method that provides suboptimal structures. In addition, it demonstrates the advantage of using the CHARMM36 energy model. The method is based on a two-phase Scatter Search metaheuristic that minimizes the energy function. Backbone fragments selections extracted from the Robetta server are followed by side chain selections, extracted from the Dunbrack Library. The results, evaluated on three proteins, are assessed by calculating their energy and root mean square deviation (RMSD) values and by visualizing them. The best structures generated are compared with structures generated by ROSETTA, I-TASSER, and previous work performed by Mansour et al. (2011).

1.4 – Enhancements to previous work

A number of enhancements are made to the Scatter Search algorithm proposed by Mansour et al. (2011).

First, the most recent version of the CHARMM force field, CHARMM36, is employed. While the functional form is the same as the previous CHARMM version, some of the parameter values have changed, especially for backbone and side chain torsions.

Moreover, VSWITCH and FSHIFT methods to calculate interactions between non-bonded atom pairs, which shift and switch interatomic forces at long distances making large simulations computationally feasible without destabilizing the macromolecule are implemented.

Lastly, a two phase fragment based Scatter Search algorithm that assembles the side chains after assembling the backbone is employed.

1.5 – Thesis Organization

The rest of the thesis is organized as follows. Chapter 2 explains the structure of proteins and protein structure prediction models. It also describes the evaluation methods and the energy function used, and the assumptions made in this work. Chapter 3 summarizes previous research performed in the fragment based protein structure prediction area and how advancement in this area is assessed by CASP. Chapter 4 explains the basic Scatter Search algorithm and the tailored Scatter Search

algorithm to predict protein structures. In addition, it illustrates how a solution is represented in this work. It also explains how side chains are assembled and how transformations from Cartesian to Torsion representations are performed. Chapter 5 discusses the experiments performed and the results obtained. Chapter 6 concludes the research work and provides suggestions for possible future work.

CHAPTER TWO

Proteins

Computational approaches for protein structure prediction use protein structure details to predict structures correctly. This section explains the structure of proteins and protein structure prediction models. It also describes the evaluation methods, the energy function used, and the assumptions made in this work.

2.1 – Protein Structure

Proteins are assembled from the set of twenty amino acids, made up of carbon, hydrogen, nitrogen, oxygen, and sulfur atoms. Amino acids are comprised of a central Cα atom connected to an amino group (NH_2), a carboxyl group (COOH), and a side chain (R) (figure 2.1) (Setubal & Meidanis, 1997). The twenty different amino acids result from the variations in side chains. Table 2.1 lists the names, the three and one letter abbreviations and structures of the twenty amino acids

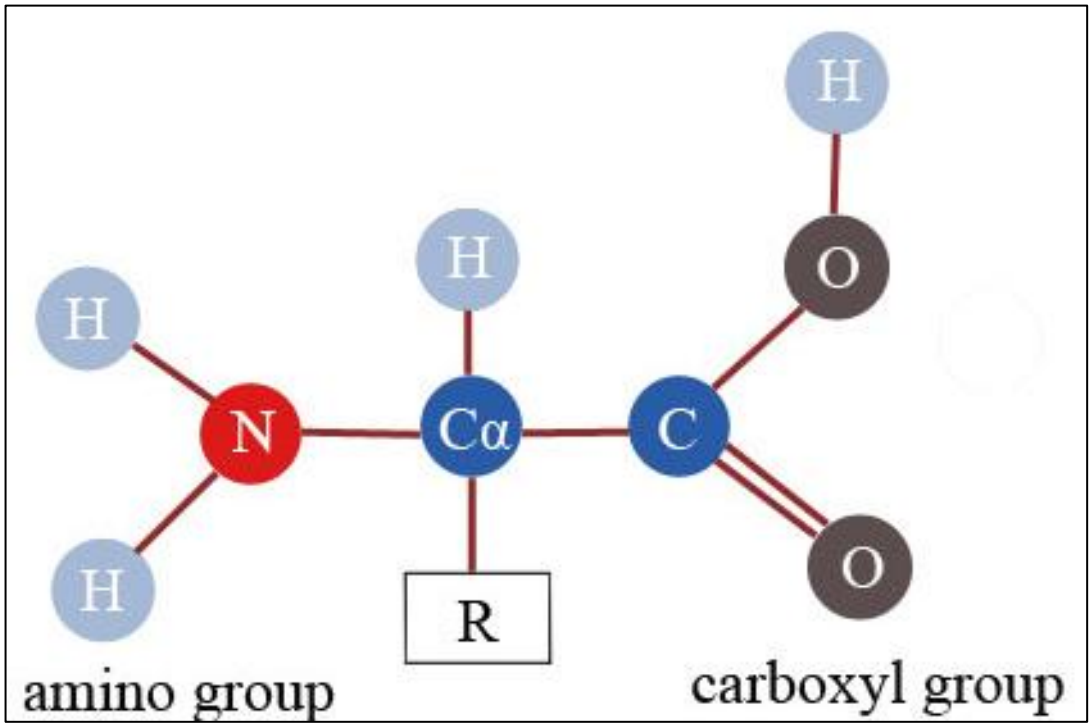
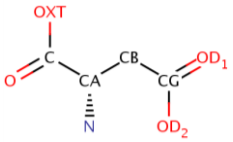
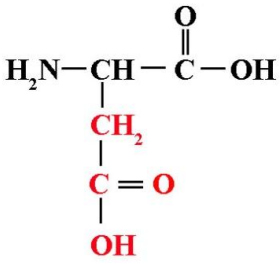
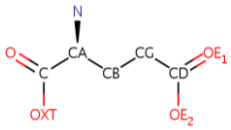
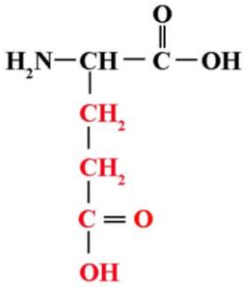
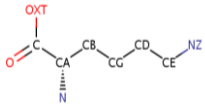
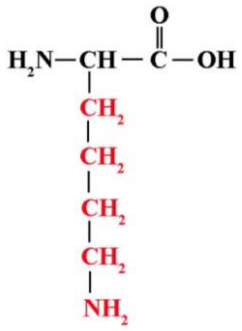
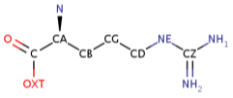
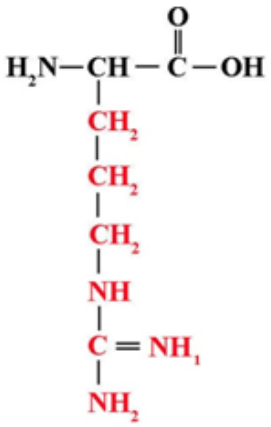
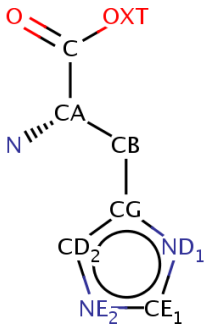
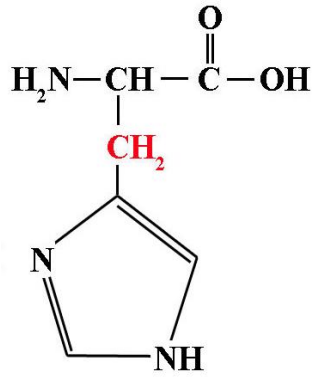
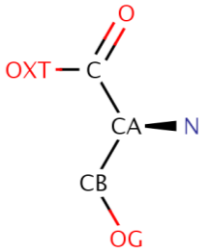
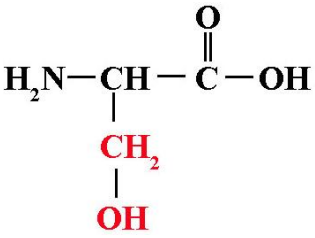
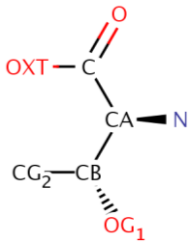
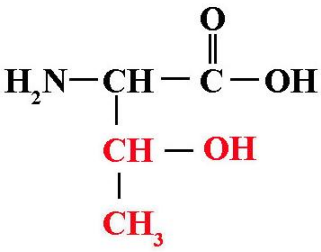
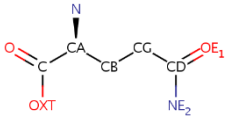
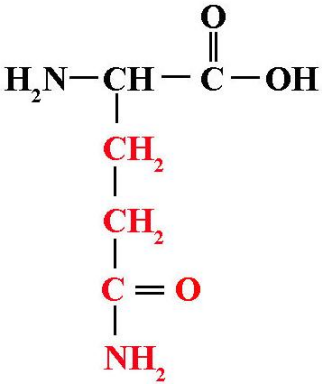
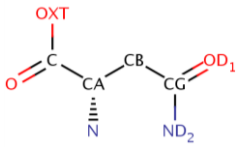
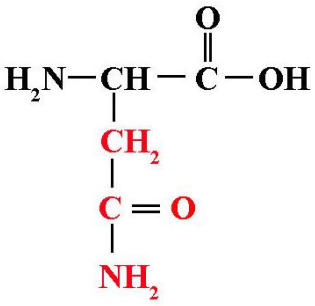
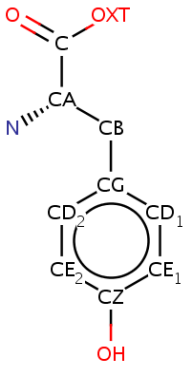
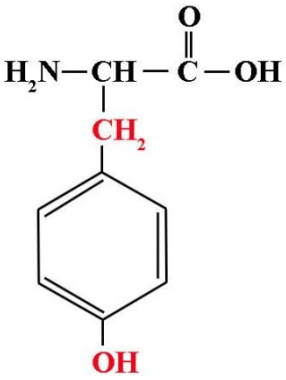
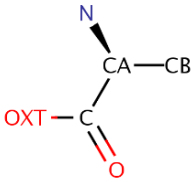
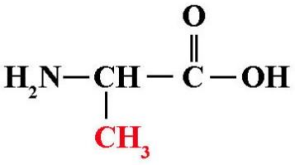
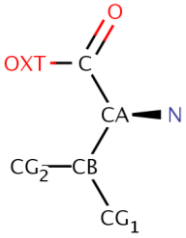
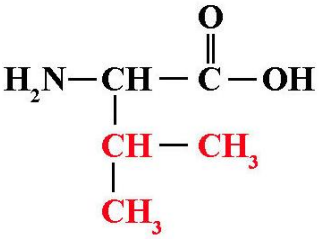
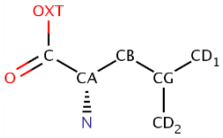
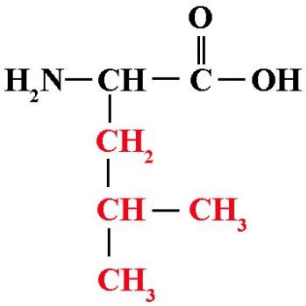


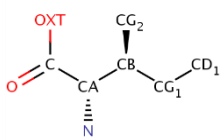
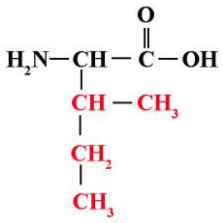
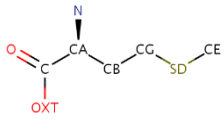
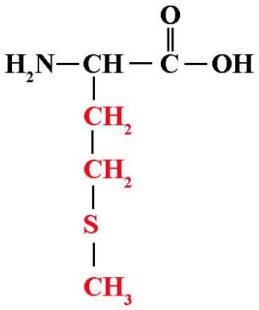
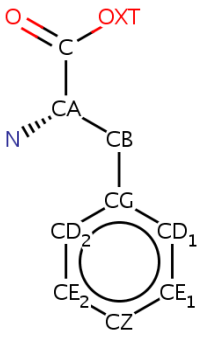
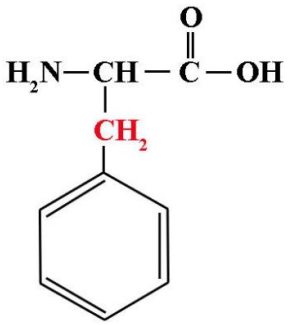
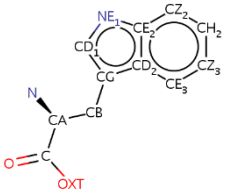
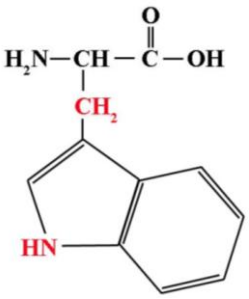
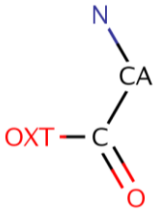
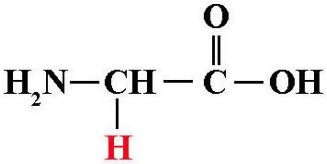
Figure 2.1. Structure of Amino Acids.

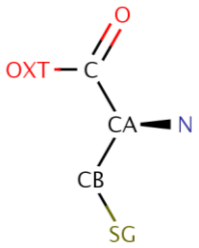
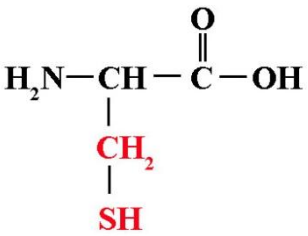
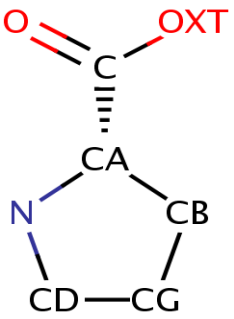
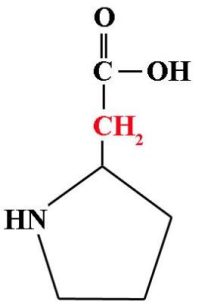
Table 2.1. Amino Acid Details.

Name	Three Letter Code	One Letter Code	Atomic Structure	Atomic Detailed Structure
ASPARTIC ACID	ASP	D		
GLUTAMIC ACID	GLU	E		
LYSINE	LYS	K		
ARGININE	ARG	R		

HISTIDINE	HIS	H		
SERINE	SER	S		
THREONINE	THR	T		
GLUTAMINE	GLN	Q		

ASPARAGINE	ASN	N		
TYROSINE	TYR	Y		
ALANINE	ALA	A		
VALINE	VAL	V		
LEUCINE	LEU	L		

ISOLEUCINE	ILE	I		
METHIONINE	MET	M		
PHENYL-ALANINE	PHE	F		
TRYPTO-PHAN	TRP	W		
GLYCINE	GLY	G		

CYSTEINE	CYS	C		
PROLINE	PRO	P		

Side chains can have a linear or a ring shaped structure. Chi angles, formed as a result of the planes created by the side chain atoms, can at most be four in number. They are represented by four dihedral angles Chi1 (χ_1), Ch2 (χ_2), Chi3 (χ_3), and Chi4 (χ_4), denoted in Greek alphabet as Cbeta (CB), Xgamma (XG), Xdelta (XD), and Xepsilon (XE) (Setubal & Meidanis, 1997).

Two Amino acids bond together by a condensation reaction and form a covalent bond called a peptide bond by losing a water molecule (Figure 2.2). Many amino acids bonded together form a polypeptide chain. This polypeptide chain is referred to as the backbone of a protein. Sketching peptide chains starts with the amino group or the N-terminal which is not bonded to another amino acid and ends with the free carboxyl group on the other hand or the C-terminal (Setubal & Meidanis, 1997).

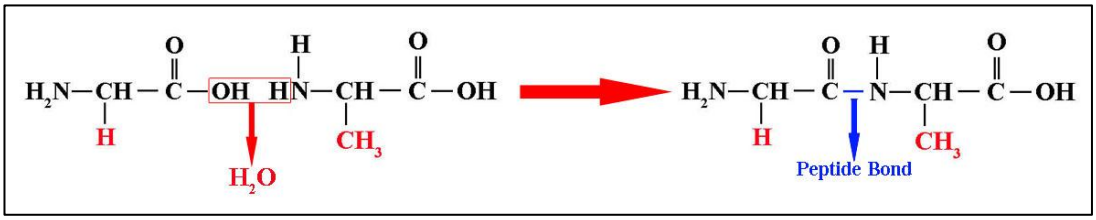


Figure 2.2. Formation of a Peptide Bond.

Amino acid side chain interactions are essential for the protein structure. Ionic bonds can be formed between charged amino acid side chains, and hydrogen bonds can be formed between polar amino acids, in addition to Van der Waals interactions between them. All bonds formed by side chain atoms are not covalent, except bonds between cysteine S atoms that form disulfide bridges. Thus the sequence and the position of amino acids in a protein chain controls the twists and folds of a protein (Setubal & Meidanis, 1997).

Protein structure is described on four levels: primary, secondary, tertiary and quaternary.

The primary structure is the sequence of amino acids associated covalently. The primary structure can be represented as a sequence of three or one letter abbreviations of amino acids, or as a linear sequence of atomic structures of amino acids connected to each other in a polypeptide chain (Figure 2.3) (Setubal & Meidanis, 1997).

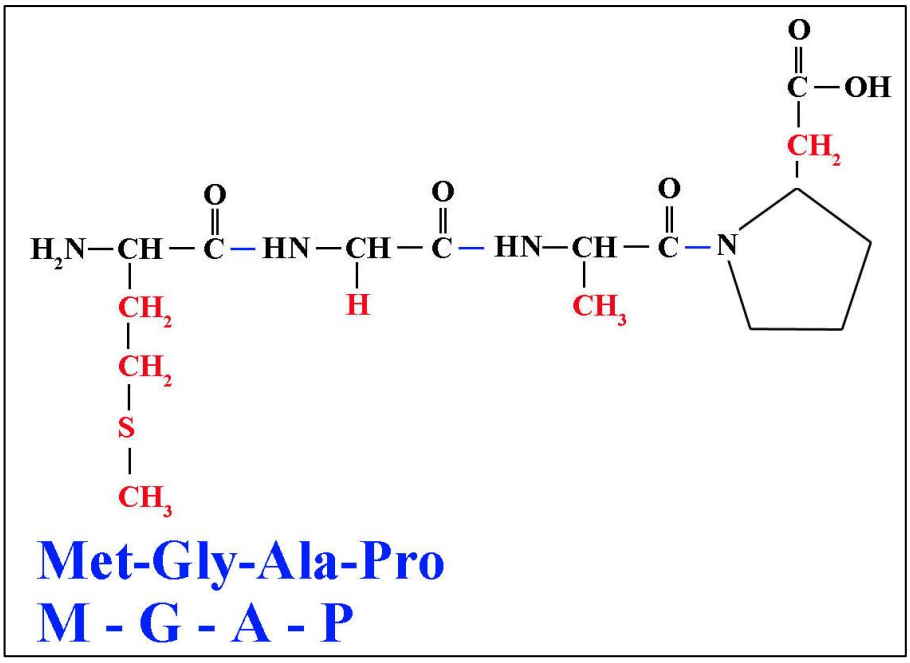


Figure 2.3. Primary Structure of an Amino Acid Chain.

The secondary structure of a protein results from hydrogen bonding between backbone atoms causing a fold in the protein sequence. Hydrogen bonds occur when an electropositive hydrogen atom is attracted to an electronegative atom such as nitrogen, oxygen, fluorine. There are three forms of secondary structure: Alpha Helix, Beta Pleated Sheet, and Loop (Setubal & Meidanis, 1997).

Alpha helices occur when N-H and C=O backbone atoms of two amino acids 3.6 residues away get connected by a hydrogen bond, forming a turn. The helix formed is right-handed and the side chains stick out of it (Figure 2.4) (Setubal & Meidanis, 1997).

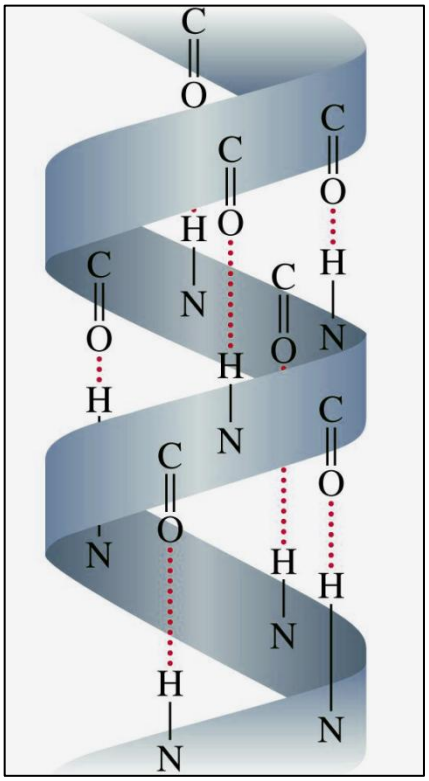


Figure 2.4. Alpha Helix.

Beta pleated sheets, which are less common than alpha helices, occur when adjacent strands of proteins typically 3 to 10 amino acids long interact laterally through hydrogen bonds between backbone N-H and C=O atoms. The strands may be parallel with the amino groups of both strands at the same side or antiparallel. Side chains of amino acids of a strand point to either sides of a strand (figure 2.5) (Setubal & Meidanis, 1997).

Attaching alpha helices to alpha helices, alpha helices to beta pleated sheets, or beta pleated sheets to beta pleated sheets are done by loops. They are short chains of amino acids not forming beta pleated sheets or alpha helices, but simply connecting them (figure 2.5) (Setubal & Meidanis, 1997).

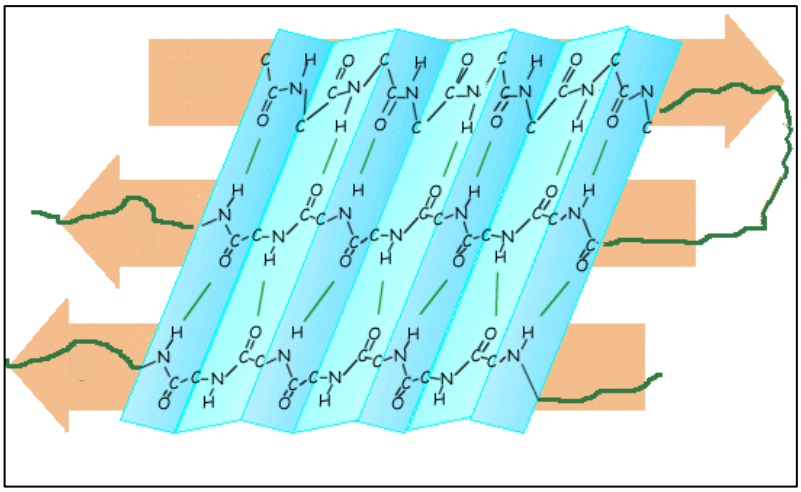


Figure 2.5. Beta Pleated Sheets and Loops.

The tertiary structure is the compact three dimensional structure of a protein, resulting from the different interactions mentioned above between amino acid side chain atoms (Figure 2.6).

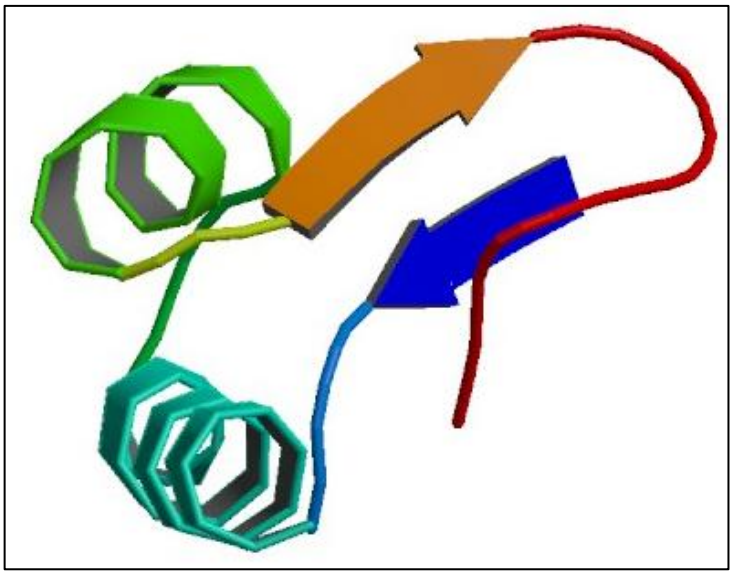


Figure 2.6. Tertiary Structure of Crambin.

And finally, protein quaternary structure exists when several polypeptide chains are present in a protein complex. Each polypeptide chain is called a subunit. Polypeptide chains in a quaternary structure can be identical or different. The quaternary structure is maintained the same way as a tertiary structure is maintained. Figure 2.7 represents the quaternary structure of hemoglobin made up of four chains colored differently (Setubal & Meidanis, 1997).

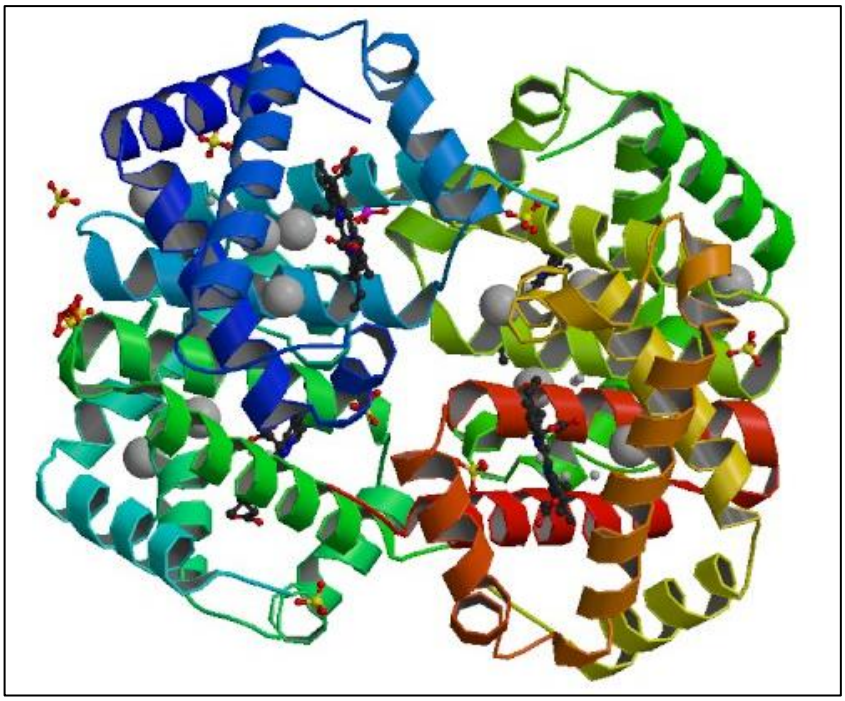


Figure 2.7. Quaternary Structure of Hemoglobin.

2.2 – Protein Models

2.2.1 – Hydrophobic-Polar model

Proteins in the Hydrophobic-Polar model are placed in a 2D or 3D lattice where their amino acids occupy cells in the lattice designated by H or P standing for hydrophobic and polar respectively. In this model the lattice where the hydrophobic amino acids are placed in the center is the most stable structure, given that hydrophobic amino acids do not react with solvents and move to the inside while the polar amino acids stay on the outside. Two neighboring hydrophobic amino acids in the lattice, that are not successive in the sequence, carry a -1 energy value. There is adjacency when two H amino acids are connected by a bond. The goal is finding the protein fold with the maximum number of H-H bonds, i.e. to find a path by visiting each amino acid once. What differentiates the 3D lattice from the 2D lattice is the representation of amino acids using the x, y and z plane (Dill, 1985).

Even though this representation minimizes the search space, predicting protein structures in the Hydrophobic-Polar model remains complex. Many methods have been developed to find good suboptimal solutions. Methods based on evolutionary Monte Carlo approaches were proposed (Liang and Wong, 2001). A particle swarm

optimization algorithm in the 2D lattice was proposed to tackle the problem (Zhang and Li, 2007). An Ant Colony Optimization algorithm for 2D and 3D lattices was also proposed (Shmygelska and Hoos, 2005). A genetic algorithm with the Monte Carlo method in 2D and later in 3D lattices was furthermore proposed (Unger and Moult, 1993). Later, another genetic algorithm approach was developed that outperformed the latter approach by attaining a higher number of H-H bonds with a lower number of energy values (Patton et al., 1995). Afterwards, a genetic algorithm accompanied with a backtracking technique to determine collision problems was proposed (Johnson and Katikireddy, 2006). A genetic algorithm that repairs the unreasonable candidates generated by crossover and guarantees that mutation will lead to suitable and possible candidate solutions in 3D lattices was moreover proposed (Mansour et al., 2010). Recently, a particle swarm optimization based algorithm in the 3D lattice model was proposed (Mansour et al., 2013).

Table 2.2 shows an illustration of protein in a 2D lattice, where L represents the sequence length, E represents the energy value, and H and P represent hydrophobic and polar amino acids respectively. Five conformations of the same protein of length 48 in a 2D lattice are represented in Figure 2.8. White squares denote polar amino acids, and black squares hydrophobic amino acids. The best conformation, having the maximum number of H-H bonds, is the first conformation in the figure having an energy value of -23. Even though a lattice does not completely show the exact folding, some show secondary structures like α helices and β Sheets (Figure 2.9) (Hu et al., 2008).

The Hydrophobic-Polar model generates good conformations on small proteins, it is not feasible on long protein chains, since the lattice becomes very intricate.

Table 2.2. Protein Sequences in a 2D Lattice.

L	**E**	**Sequence**
18	-9	PHPPHPHHHPHHPHHHHH
18	-8	HPHPHHHPPPHHHHPPHH
48	-23	PPHPPHHPPHHPPPPPHHHHHHHHHHPPPPPPPHHPPHHPPHPH HHHH

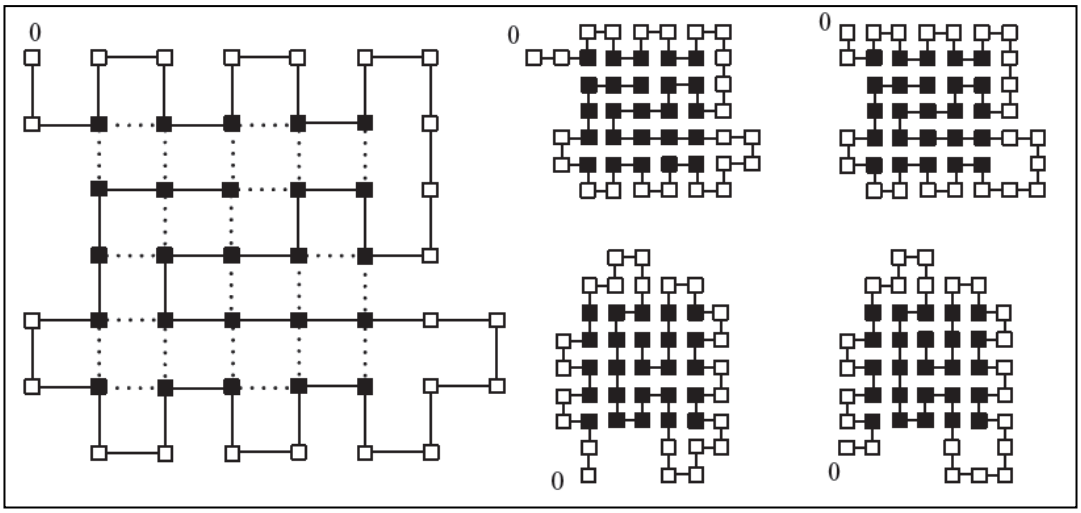


Figure 2.8. Different Conformations of a Protein in a 2D Lattice (Hu et al., 2008).

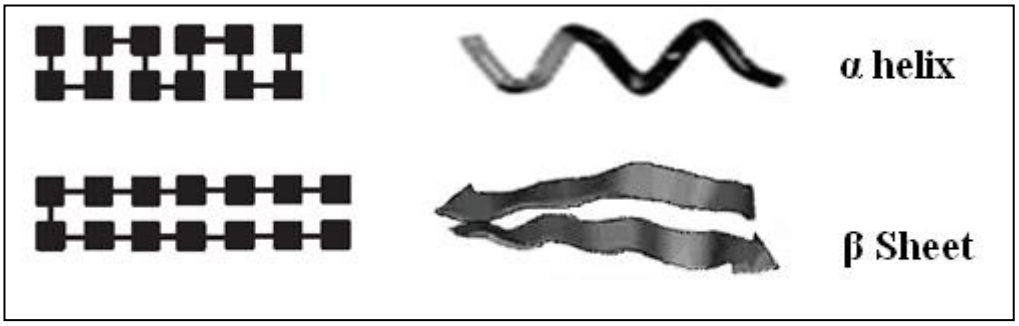


Figure 2.9. Secondary Structures of a Protein Sequence in a 2D Lattice (Hu et al., 2008).

2.2.2 – UNRES Model

The UNRES model is a simplified protein model where the polypeptide chain is represented as a series of α carbon (Cα) atoms connected to each other via virtual bonds (dC), and connected to united side chains (SC) via virtual bonds (dX) and united peptide groups (p) positioned at the center of successive α-carbon atoms. This representation allows performing protein conformational searches in real time. The interaction positions are the united side chains and united peptide group. The virtual bond lengths (Cα-Cα and Cα-SC) have constant values. The side chain angles, α_{SC} and β_{SC}, and the virtual bond angles, θ and γ, have variable values (Figure 2.10) (Liwo et al., 2000). The physics based UNRES force field is represented in equation 2.1.

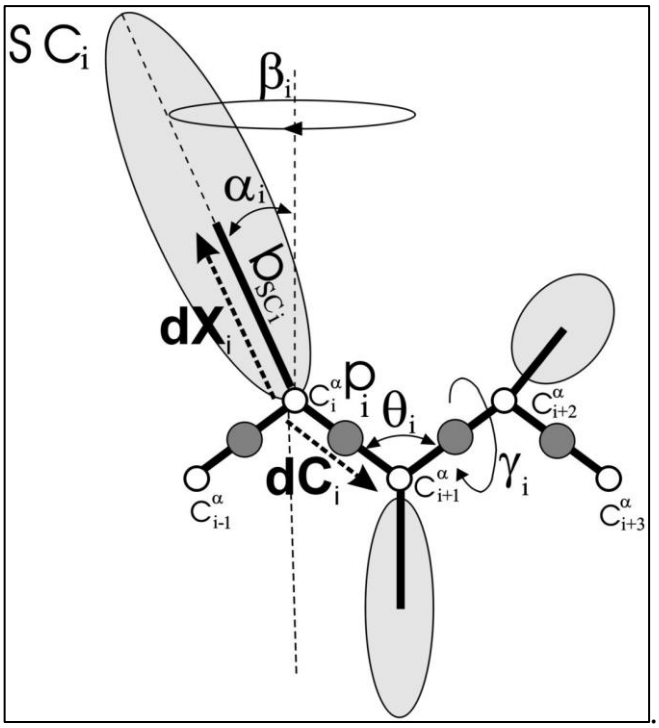


Figure 2.10. UNRES Model (Liwo et al., 2000).

$$U = \sum_j \sum_{i<j} U_{SCiSCj} + w_{SCp} \sum_j \sum_{i \neq j} U_{SCipj} + w_{pp}^{el} f_2(T) \sum_j \sum_{i<j-1} U_{pipj}^{el} + w_{pp}^{vdW} \sum_j \sum_{i<j-1} U_{pipj}^{vdW} +$$
$$w_{tor} f_2(T) \sum_i U_{tor}(\gamma_i) + w_{tord} f_3(T) \sum_i U_{tord}(\gamma_i + \gamma_{i+1}) + w_b \sum_i U_b(\theta_i, \gamma_{i-1}, \gamma_{i+1}) + w_{rot} \sum_i U_{rot,i}$$
$$+ \sum_{m=2}^{N_{corr}} w_{corr}^{(m)} f_m(T) U_{corr}^{(m)} + w_{turn}^{(3)} f_3(T) U_{turn}^{(3)} + w_{turn}^{(4)} f_4(T) U_{turn}^{(4)} + w_{turn}^{(6)} f_6(T) U_{turn}^{(6)} + w_{bond} U_{bond}(d_i)$$
$$+ w_{ss} \sum_{disulfidebonds} U_{ssi} + n_{ss} E_{ss}$$

(2.1.)

T_0 is set to 300K. $f_n(T)$ is the temperature scaling multiplier. U_{SCiSCj} is the average free energy of hydrophobic or hydrophilic side chain solvent interactions. U_{SCiPj} is the discarded volume energy of side chain peptide group interactions. U_{pipj}^{el} is the mean electrostatic interaction between backbone peptide groups. w_{pp}^{vdW} is the mean van der Waals interaction between backbone peptide groups. The torsion and double torsion energies of the rotation around a virtual bond or two consecutive virtual bonds are represented by U_{tor} and U_{tord}. U_b is the virtual bond angle bending energy and U_{rot} is the side chain energy. $U_{corr}^{(m)}$ and $U_{turn}^{(m)}$ represent the relationship between peptide group electrostatic and backbone local interactions. $U_{bond}(d_i)$ denotes the multiple minima in virtual-bond stretching potentials, where di is the measure of the

i^{th} virtual bond. U_{SSi} is the energy resulting from the distortion of disulfide bonds from their balanced configuration. E_{SS} is the energy resulting from the formation of disulfide bonds in the chain, and n_{SS} is the number of disulfide bonds. The w terms are the weights of the energy terms derived by optimizing the energy landscape that aims to lower the energy value (Maisuradze et al., 2010).

2.2.3 – Dihedral Angles Model

In the dihedral angles model of protein presentation, the backbone conformation is determined by three torsion angles, Phi φ, Psi ψ and Omega ω, and the conformation of side chains is determined by the Chi χ angles. Phi, determining how far C atoms of successive amino acids are, is the angle formed by the C-N-Cα and N-Cα-C planes and rotating around the N-Cα bond. Psi, determining how far N atoms of successive amino acids are, is the angle formed by the N-Cα-C and Cα-C-N planes and rotating around the Cα-C bond. Omega, determining the distance between Cα-Cα atoms of successive amino acids, is the angle formed by the Cα-C-N and C-N-Cα planes and rotating around the C-N bond. Chi angles, formed as a result of the planes created by the side chain atoms, can at most be four in number. They are denoted by Chi1 (χ_1), Ch2 (χ_2), Chi3 (χ_3), and Chi4 (χ_4) (figure 2.11).

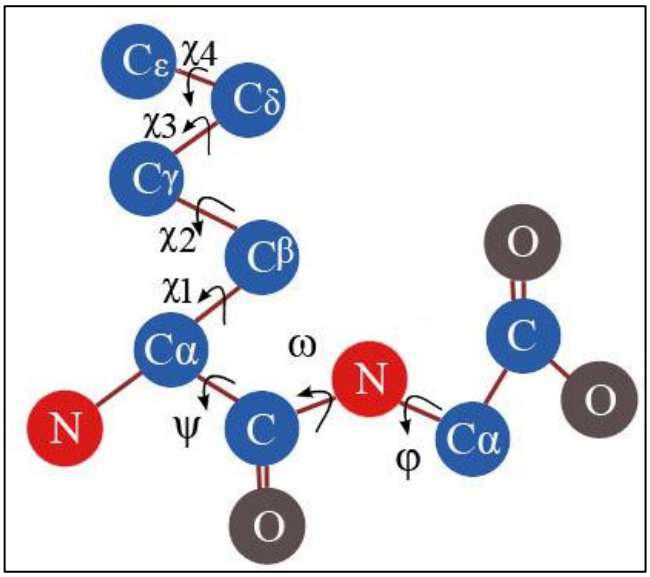


Figure 2.11. Protein Dihedral Angles Model.

2.2.4 – Force Field Models

When simulating a protein tertiary structure it is essential to represent the energy of a protein as a function of its atomic coordinates, including forces on individual atoms, referred to as force fields. These force field models implicate protein physical forces and chemical reactions (Ponder & Case, 2003).

The most commonly used protein force fields include a potential energy function describing the bonded and non-bounded interactions between atoms. While bonded interactions include bonds, angles and torsion terms, non-bonded ones include van der Waals and electrostatic interactions (Ponder & Case, 2003).

Force fields can be classified into three types: All-atom, United-atom, and Coarse-grained. In All-atom force fields parameters are supplied for every single atom. In United-atom force fields parameters are supplied for all atoms except non-polar hydrogen atoms. In Coarse-grained force fields parameters are supplied for atom groups. These parameters are derived experimentally and through quantum mechanical calculations. Many of the more recent developments in protein force fields address effects of solvation (Lee et al., 2003).

Force fields can further be classified into two major groups physics-based and knowledge-based. Physics-based relies on basic physical theories and knowledge-based reply on information extracted from known protein structures. CHARMM , AMBER, and UNRES are physics-based force fields, TASSER is a knowledge-based force field and ROSETTA's force field is a combination of both types (Lee et al., 2003).

2.3 – Protein Structure Prediction Problem

We acquire our current understanding of the protein structure prediction problem, as a result of successive discoveries made in the field. In 1894, Emil Fischer proposed that the tertiary structure of a protein determines its function (Cramer, 2007). In 1931, Hsien Wu introduced that protein misfold leads to function loss (Wu, 1931). In the early 1950s, Frederick Sanger experimentally sequenced Insulin (Sanger and Thompson, 1953). In 1958, John Kendrew was the first to determine the tertiary structure of Myoglobin, using X-ray crystallography (Takano, 1977). In 1961, Christian Anfinsen proposed that a protein’s correct conformation has the lowest

potential energy (Anfinsen, 1973). The Ramachandran plot, revealing the possible local conformations in protein structures that lead to their secondary structure, was first established in 1963 and further elaborated in 1968 (Ramachandran et al., 1963; Ramachandran and Sasisekharan, 1963). In 1968, Levinthal's Paradox stated that proteins fold very quickly (Levinthal, 1968). In 1973, Anfinsen demonstrated that the conformation of a protein can be inferred only from its sequence of amino acids. Later he introduced his thermodynamics hypothesis, known as Anfinsen's dogma, which states that protein folding depends on the physical processes, the particular amino acid sequence and the surrounding environment (Anfinsen, 1973).

Even though proteins fold very quickly, this period is long for computers to imitate. Computational protein structure prediction methods face two difficulties: First, the possible conformations a protein chain can adopt are infinite and enormous to be analyzed computationally. Second, the assumptions made in interaction simulations would not produce accurate prediction.

The protein folding problem is given a protein sequence, the list of amino acid codes, predicting the tertiary structure of that protein, where the native state has the lowest energy.

Protein structure prediction is a complex problem that requires a variety of techniques to solve different portions of the problem. To tackle this complexity, different computational methods have been developed discussed earlier in section 1.2.2.

2.4 – Evaluation Methods and Proteins Used for Evaluation

2.4.1 – CHARMM36 Energy Function

The All-atom CHARMM36 (Chemistry at HARvard Macromolecular Mechanics) protein force field function, developed at Harvard, computes the potential energy of a protein structure. The potential energy is the sum of individual terms representing the internal and non-bonded contributions. Internal terms include bond, angle, Urey-Bradley, improper torsion, torsion, and backbone torsional correction energy values. The non-bonded terms include electrostatic, Van der Waals, and solvation values. Equation 2.2 represents the 7 terms of the CHARMM36 energy function E as a function of the conformation c (Brooks et al., 1983).

$$E(c) = \sum_{bonds} K_b (b - b_o)^2 + \sum_{angles} K_\theta (\theta - \theta_o)^2 + \sum_{impropers} K_{imp} (\varphi - \varphi_o)^2 +$$

$$\sum_{torsions} K_x (1 + \cos(n\chi - \delta)) + \sum_{solvation} \sigma_i A_i + \sum_{electrostatic} \frac{q_i q_j}{r_{ij}} +$$

$$\sum_{vanderWaals} \varepsilon_{ij} \left(\left(\frac{R\min_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R\min_{ij}}{r_{ij}} \right)^{6} \right) + \sum_{Urey-Bradley} Ku(u - u_o)^2 + \sum_{\alpha carbons} CMAP(\phi, \psi)$$

(2.2.)

The energy resulting from bonds is the summation of bond stretches of the target protein taken from the ideal bond lengths, calculated by term 1 of the equation, where b is the bond length in the predicted conformation, b_o is the ideal bond length between the bonded atoms and K_b is the bond force constant that determines the strength of the bond (Brooks et al., 1983).

The angles energy is the summation of bond bending of the target protein from the real bond angles, where θ is the measured angle in the predicted conformation, θ_o is the ideal bond angle between the bonded atoms, and K_θ is the angle force constant that determines the elasticity of the bending angle (Brooks et al., 1983).

The improper torsion energy is the summation of out of plane bending, calculated by term 3 of the equation, where φ_o is the best torsion angle, φ is the torsion angle of the predicted structure, and K_{imp} is the improper torsion force constant. The earlier three terms present the difference in geometry of the predicted conformation from the ideal conformation of the target protein hence if the values of these three terms are near zero, the predicted conformation of a protein is favorable (Brooks et al., 1983).

The torsion energy, representing the rotation energy around a covalent bond is calculated by term 4 of the energy function where χ is the torsion angle between the planes containing the first and last 3 atoms, n is the multiplicity typically 1, 2, or 3, K_χ is the torsion force constant, and δ is the phase angle (Brooks et al., 1983).

Non-bonded interactions are calculated between all atom pairs within a user specified cutoff distance, except for pairs that are covalently bonded or separated by two covalent bonds (Brooks et al., 2010).

The electrostatic energy is calculated using the Coulomb potential, that computes the force exerted from the interaction of the charges of two non bonded atoms (Brooks et

al., 1983). The interaction of two charged atoms i and j, where q_i and q_j are the charges of the two atoms, ε_r is the dielectric constant, and r_{ij} is the Euclidean distance between the two atoms, is calculated by term 6 of the equation (Brooks et al., 1983).

The Van der Waals energy is calculated using the Lennard-Jones 6-12 potential that models the attractive and repulsive forces between atoms. Electron collision results in repulsive forces and variation of charges in the electron clouds of one atom results in attractive forces. Attractive forces exist for extended distances between atoms, whereas it is absent for short distances where repulsive forces exist. Both energies are calculated by term 7 of the equation where the first part depicts the repulsive forces and the second the attractive ones. $Rmin_{ij}$ is the distance at which the Lennard-Jones term is the minimum, r_{ij} is the interatomic distance, and ε_{ij} is the Van der Waals depth (Brooks et al., 1983).

It is computationally extensive, around 75% of SS execution time, to calculate the non-bonded forces for all atom pairs except those that are covalently bonded or separated by two covalent bonds. While attraction between pairs of atoms is almost null at considerable distances and at smaller distances, it increases until reaches an equilibrium position, of minimum energy, where there is neither attraction nor repulsion. If this distance continues decreasing, the atom clouds overlap and repulsion increases exponentially (Bonetti et al., 2010). This is tackled by using VSWITCH and VSHIFT algorithms for van der Waals and electrostatic interactions respectiverly. All interactions beyond 12 Å distance and less than 1.5 Å are ignored (Brooks et al., 1983).

For van der Waals interactions, the VSWITCH method uses two cutoff values, ctofnb=12 Å and ctonnb=10 Å. Atoms below ctonnb value posses energy calculated by the Lennard-Jones potential, atoms within the two cutoff values posses energy calculated by the Lennard-Jones potential multiplied by the sw(d, ctonnb, ctofnb) function value, where d is the interatomic distance (equation 2.3) (Brooks et al., 1983).

$$\mathrm{sw}(\mathrm{d},\mathrm{ctonnb},\mathrm{ctofnb}) = \frac{(\mathrm{ctofnb}-\mathrm{d})^2(\mathrm{ctofnb}+2\mathrm{d}-3\mathrm{ctonnb})}{(\mathrm{ctofnb}-\mathrm{ctonnb})^3} \qquad (2.3.)$$

To tackle the same problem in electrostatic interactions, the VSHIFT method is used that multiplies the coulomb interaction value by sh(d, ctonnb) when the pairs are separated by a distance less than the cutoff value ctofnb=12 Å, where d is the interatomic distance (equation 2.4). The idea behind this is shifting the energy smoothly before setting it to zero (Steinbach et al., 1994).

$$\mathrm{sh(d, ctofnb)} = (1 - d / \mathrm{ctofnb})^2 \qquad (2.4.)$$

The solvation energy, presenting protein-water interaction energies, is based on the solvation parameter σ_i and solvent accessible area A_i of each atom i (Wesson & Eisenberg, 1992) is calculated by term 5 of the equation. The values of σ_i represent the hydrophobicity of each atom type i (Wesson & Eisenberg, 1992). The solvent accessible areas are calculated using the methodology proposed by Hasel et al., which is an extension to Wodak and Janin's methodology by equation 2.5, where Si is the solvent accessible surface area of an atom i having radius r_i and a probe radius r_s=1.4 Å, b_{ij} is the solvent accessible surface area removed from atom i when overlapped with atom j, p_i corrects systematic errors caused due to hybridization and substitution of atom i, and P_{ij} is the connectivity parameter set according to the connection between the two atoms i and j (Wodak & Janin, 1980; Hasei et al., 1988).

$$A_i = S_i \Pi(1 - p_i p_{ij} b_{ij} / S_i)$$

$$S_i = 4\Pi(r_i + r_s)^2$$

$$b_{ij} = \Pi(r_i + r_s)(r_j + r_i + 2r_s - d)[1 + (r_j + r_i)/d] \qquad (2.5.)$$

The Urey-Bradley component represents angle bending using 1,3 non bonded interactions where K_u is the respective force constant and d the distance between the 1,3 atoms in the harmonic potential. Finally, the CMAP term corrects small errors in protein backbone, by grid based energy correction maps or dihedral crossterms. Fussy treatment of protein backbone is vital for successful predictions (Brooks et al., 1983).

Parameter values needed to compute torsion, electrostatic, torsion, solvation and Van Der Waals energy values are extracted from CHARMM36 topology and parameter files (MacKerell et al., 1998; Alexander et al., 2004; Robert et al.).

2.4.2 – RMSD

The generated structures are evaluated by computing the root mean square deviation (RMSD) that measures the average distance between the Cartesian coordinates of predicted and reference protein atoms. RMSD is expressed in Å and calculated by equation 2.6, where *ai* is the *ith* amino acid in the predicted protein, *bi* is the *ith* amino acid in the reference protein, and n is the number of Cα atoms. The RMSD value of two identical structures is 0 and increases with the increase of structural difference between the two proteins (Carugo and Pongor, 2001). In this work we calculate the RMSD in two ways. The Cα-RMSD, to calculate Cα atoms average distance, and all-atom RMSD, to calculate all atoms average distance. Cα-RMSD is utilized in phase one and all atom-RMSD in phase two.

$$RMSD = \frac{\sqrt{\sum_{i} (x_{ai} - x_{bi})^2 + (y_{ai} - y_{bi})^2 + (z_{ai} - z_{bi})^2}}{\sqrt{n}} \qquad (2.6.)$$

RMSD is a global measure, a small perturbation in one part of a protein results in a large RMSD value. For instance, if we assess a prediction where the first residues (accounting for the 80% of total residue number) align correctly with the target and the last 30% don't using RMSD, RMSD returns a large value reflecting a bad prediction. Such a prediction is considered to be good prediction. Thus, it is necessary to consider local regions of the proteins when evaluating their structure.

2.4.3 – Solution Visualization

The solutions are visualized using an open source molecular visualization tool, PyMOL. The tool takes as input the generated structure in a PDB file format, and sketches the tertiary structure of the protein (Schrödinger, 2013).

2.4.4 – Evaluated Proteins

The algorithm is tested on three proteins Crambin (1CRN), Repressor of Primer (1ROP), and Uteroglobin (1UTG).

A relatively short protein, having only 46 amino acid residues, Crambin is a storage protein found mostly in plants especially in the seeds of the Abyssinian cabbage. It is not known for causing any human disease which has rendered it rather uninteresting, yet having 6 cysteine residues in its relative short sequence and ability to show excellent diffraction have been the main reasons behind its extensive theoretical and experimental research. A crambin has along its 2 beta pleated sheets, another 2 alpha helices which yield its anti-parallel 3D structure (Teeter, 1984).

Repressor of Primer protein more commonly known as 1ROP is a 63 amino acid residue homodimer. Each of the two dimmers of the 1ROP consists of two alpha helices making it a total package of four almost entirely stable helices connected by a loop. This globular coiled-coil protein's function relies in the regulation of plasmid replication via intricate RNA-RNA interactions. The simplicity in the folding of the 1ROP has caused it to become a paradigm in the studying of sequence-structure relationships (Banner et al., 1987).

Uteroglobin, also known as 1UTG or Blastokinin, is a mammalian protein initially discovered in pregnant rabbits. Uteroglobin plays a major role during embryo implantation which is also a reason why it is a steroid or progesterone induced protein. Human uteroglobin is a 91 amino acid residue protein, about three quarters of which form alpha helices. In Rabbits uteroglobin is made up 70 amino acids. This variability in the number of amino acid residues across the mammalian species forms a superfamily of proteins called Secretoglobin (Scgb) whose members may be different in the number of amino acid residues yet are similar in structure and function. Uteroglobins are homodimeric proteins made up of identical amino acid sequences bonded together with disulphide bridges via cysteine residues along the chain. The dimeric structure of 1UTG creates a cavity in which water molecules could be fitted (Mukherjee et al., 2007).

2.5 – Assumptions

A couple of assumptions have been made in this thesis to simplify the representation of a solution including:

Bond lengths and bond angles are set to constant values, thus bond, angle, and improper torsion components are removed from the CHARMM36 energy function equation. This supposes that the structure is static, and bond lengths and angles do not change as a result of variations in the torsion angles. The CMAP term is also ignored for simplicity.

Hydrogen atoms are combined with neighboring heavy atoms referred to as the "extended-atom representation", which considerably reduces the size of the problem.

CHAPTER THREE

Literature Review on Fragment Based Protein Structure Prediction

In this chapter an overview of the leading fragment based PSP algorithms are presented and how PSP methods are assessed through CASP.

3.1 – Fragment Based Protein Structure Prediction Methods

The most successful method in *Ab initio* category, fragment based, employs peptide fragments, secondary structures and statistical information from PDB structures to predict protein tertiary structures. The basic principle behind this method is the presence of a strong relationship between an amino acid sequence and structure (Kopp and Schwede, 2004).

A typical *ab initio* fragment based method starts with generating fragments from the PDB. Then heuristics are used to optimize conformations and generate native like structures by using energy functions and evaluation methods.

3.1.1 – I-TASSER

Iterative Threading Assembly Refinement (I-TASSER) is a unified meta server for protein structure and function prediction (Roy at al., 2012).

Starting from the query sequence, I-TASSER uses BLAST to identify sequence homologs. Then, the homologs are aligned using multiple sequence alignment to form a sequence profile and are utilized for secondary structure prediction by PSIPRED. Using the sequence profile and the secondary structure, the query sequence is threaded using LOMETS (Local Meta-Threading-Server) which is a web service that generates tertiary models by gathering top template hits from ten threading programs (Roy et al., 2010).

Fragments are utilized to assemble well aligned structural regions of the segments with unaligned regions. A simplified representation of residues is used to improve the efficiency of the search, where a residue is made up of a Cα atom and side-chain center of mass. The fragment assembly is conducted using a Monte Carlo search with

multiple parallel simulations at different temperatures. The force field utilized is knowledge based and includes probabilistic values extracted from the PDB, threading constraints, and SVMSEQ contact predictions. The conformations produced in the low-temperature simulations during the refinement simulation are clustered by SPICKER to determine low free-energy states. By averaging the coordinates of all the clustered structural decoys centroids are obtained (Roy et al., 2010).

Then, the fragment assembly search is executed for another time on the selected cluster centroids to eliminate steric clashes and improve the overall conformation. The final models are built by REM from Cα traces by optimizing hydrogen bonding networks and using the lowest energy clustered decoys generated in the refinement stage (Roy et al., 2010).

Finally, the function of the query sequence is deduced by matching the predicted structures with proteins in the PDB (Roy et al., 2010).

I-TASSER uses four evaluation matrices to score its predictions, C-score, RMSD, TM-score and Cluster density (Zhang, 2008).

The C-score estimates the quality of predictions depending on the threading quality the structure assembly searches correctness. It is classically in the range [-5,2] where a higher value indicates a high confidence (Zhang, 2008).

The TM-score and RMSD, measuring the structural similarity between two structures, are based on the C-score value. TM-score is utilized since RMSD is a global measure and a small disorientation in one part of a protein results in a large RMSD value even if the overall conformation is accurate. A TM-score less that 0.5 implies an accurate model and a TM-score less than 0.17 implies a random similarity model (Zhang, 2008).

Cluster density, used by SPICKER, is the frequency of structure decoys in the SPICKER cluster. A higher cluster density indicates a more frequent structure in the simulation course and thus a better quality (Zhang, 2008).

I-TASSER maintained position one for protein structure prediction in the last four CASP experiments. I-TASSER also seized position one for function prediction

in CASP9. The server is available online (http://zhanglab.ccmb.med.umich.edu/I-TASSER) to the academic community (Roy at al., 2012).

3.1.2 – UCL Group Methods

The UCL (University College London) bioinformatics group developed several algorithms to tackle protein structure prediction and function annotation including FRAGFOLD for prediction of tertiary structure, PSIPRED for prediction of secondary structure, pGenTHREADER and pDomTHREADER for folding recognition and many others (Buchan et al., 2010).

FRAGFOLD starts the folding simulation with supersecondary fragment selection for each position in the query sequence. Supersecondary structures include α-Hairpins, α-Corners, β Hairpins and β-α-β Units, and Split β-α-β units. The energy function utilized includes terms for short-range, long-range, solvation, steric clashes, and hydrogen bonds with their corresponding weights. The energy minimization phase is conducted using a Simulated Annealing approach (Jones & McGuffi, 2003).

PSIPRED implements a two-stage neural network approach using PSI-BLAST output to predict protein secondary structure (Buchan et al., 2010).

pGenTHREADER and pDomTHREADER are two improved versions of the GenTHREADER algorithm that use linear regression Support Vector Machine (SVM) to predict protein folds after combining profile-profile alignments with secondary structure gap penalties, and pairwise and solvation based potentials. What differentiates the two methods is the representation and combinations of the above mentioned features on one hand, and the scoring and confidence values on another. pGenTHREADER is used for sensitive fold recognition and pDomTHREADER for domain superfamily discrimination. Both methods outperform sequence profile based methods (Lobley, 2012).

UCL achieved rank seven in the *ab initio* category in CASP9, but in CASP10 the lack of homologous sequences for the targets in the PDB made successful predictions unachievable. The servers are available online (http://bioinf.cs.ucl.ac.uk) to the academic community.

3.1.3 – ROSETTA

ROSETTA, an integrated package for protein structure prediction and functional design, is one of the leading ab initio performers in CASP (Kaufmann et al., 2009).

ROSETTA uses fragments of length three and nine to model the protein backbone. Then the model is refined and rotamers from the Dunbrack library are assembled to model the side chains (Rohl et al., 2004).

Since ROSETTA utilizes fragments of length three and nine, for each query sequence to be predicted, a customized library of fragments defining the conformational space to be searched is generated. All windows of length three and nine of the query sequence are compared with windows from the PDB and scored using Psipred, SAM-Tninenine and JUFO. The final fragment list for a query sequence is composed of 200 three and 200 nine residue fragments (Gront et al., 2011).

The fragment assembly phase is guided by Monte Carlo Simulated Annealing search, where a nine residue window and a fragment from the top 25 fragments of that particular window are selected randomly for 28000 times. After placing the torsion angles of a fragment in the solution and calculating the energy, the fragment insertion move is accepted if it decreases the energy value. If no moves are accepted in 150 successive fragment selections, the probability of accepting a move of increased energy is increased and shifted to its original value as soon as a move is accepted. Throughout the fragment assembly phase the usage of the terms in the energy function, and the SS score are fluctuated. The model is refined by attempting 8000 three residue fragment insertions by using the complete scoring and energy functions. The same algorithm is used to assemble the side chains (Kaufmann et al., 2009).

Two different energy functions are used in ROSETTA, Knowledge-Based Centroid Energy Function and Knowledge-Based All Atom Energy Function. In the latter, side chains are represented by their centroids. The low-resolution energy function includes solvation, electrostatic, van der Waal, hydrogen bonding, and steric clashes. Even though this simplified representation softens the energy representation, the energy function is no more exact. The purpose behind this is accepting structures

with defects that hold energy values close to the global minima. The all atom high-resolution energy function includes van der Waal, solvation, orientation-dependent hydrogen bonding, knowledge-based electrostatic, and knowledge-based conformation-dependent amino acid internal free energy terms. The probability values used in the energy function are collected using Bayesian statistics from the PDB. Generated structures are assessed by RMSD calculation (Kaufmann et al., 2009).

The software suite is available online on Rosetta Commons (www.rosettacommons.org), where joint research is conducted to further improve the suite. Fragments can be generated from the Robetta Server (http://robetta.bakerlab.org/).

3.1.4 – Mansour et al. Previous Work

Mansour et al. presented a scatter search algorithm for predicting all-atoms protein structures using CHARMM22 energy model. Their algorithm uses a randomization-based Diversification Generation Method. Phi and psi values are extracted from the Ramachandran plot. The algorithm checks if structures are valid and feasible (Mansour et al., 2011).

The algorithm's best produced structures for the three evaluated proteins where for 1crn 9.01 Å, for 1ROP 12.14 Å and for 1UTG 14.78 Å (Mansour et al., 2011).

3.2 – CASP

Individual methods assess presented methods against others on a particular data set depending on specific criteria and in specific ways. It is extremely important to compare methods based on some standard criteria including sensitivity of fold recognition, specificity of fold recognition, prediction reliability assessment, alignment accuracy, scope of applicability for structure/function prediction, and computational efficiency.

CASP, Critical Assessment of Techniques for Protein Structure Prediction, is a biannual worldwide experiment that evaluates the computational methods for protein structure prediction. The goal is to identify the progress made in the field and emphasize future focus areas.

Testing different prediction methods in a blind manner that enables direct comparison of protein models to its real structure was the basis of CASP experiments initiated by John Moult in 1994. Proteins whose sequences are known but not yet publicly available are used as prediction targets. Participants predict the structure of protein using different algorithms and different methods. Once the tertiary structure is released, their accuracy is assessed.

In CASP10 none of the teams outstandingly predicted a target in the *ab initio* category, while in CASP9 Rosetta predicted the structure of the T0581 target outstandingly. The results of CASP10 are available on the communities' website (www.predictioncenter.org).

CHAPTER FOUR

Scatter Search Metaheuristic

This chapter thoroughly describes the Scatter Search Algorithm. After providing a brief overview on the basic Scatter Search algorithm, the Scatter Search algorithm adapted to provide good suboptimal solutions for protein structure prediction problem is detailed. In addition, it illustrates how a solution is represented in this work. It also explains how side chains are assembled and how transformations from Cartesian to Torsion representations are performed.

4.1 – Scatter Search Basic Algorithm

Scatter Search (SS) is a population based, evolutionary and stochastic meta-heuristic that generates and maintains high quality solutions by controlling the search space through randomization, recombination and diversification. Scatter Search was introduced by Glover in 1977 in a heuristic study for integer programming and remained idle until it was combined with Tabu Search in 1993. The final approach used today was proposed by Glover in 1998 (Glover, 1998). Scatter Search generates a random set of candidate solutions, improves them and selects 20% of these solutions and places them in the reference set. Half of the selected solutions are high quality and the other half diverse. Then it iterates through a subset generation, solution combination, improvement and reference set update methods, where new subsets are generated, combined, improved and included in the reference set according to a certain criteria. Figure 4.1 describes the major methods in Scatter Search. In the following sections, we describe the design of the various methods that adapt scatter search to provide good suboptimal solutions for the protein structure prediction problem.

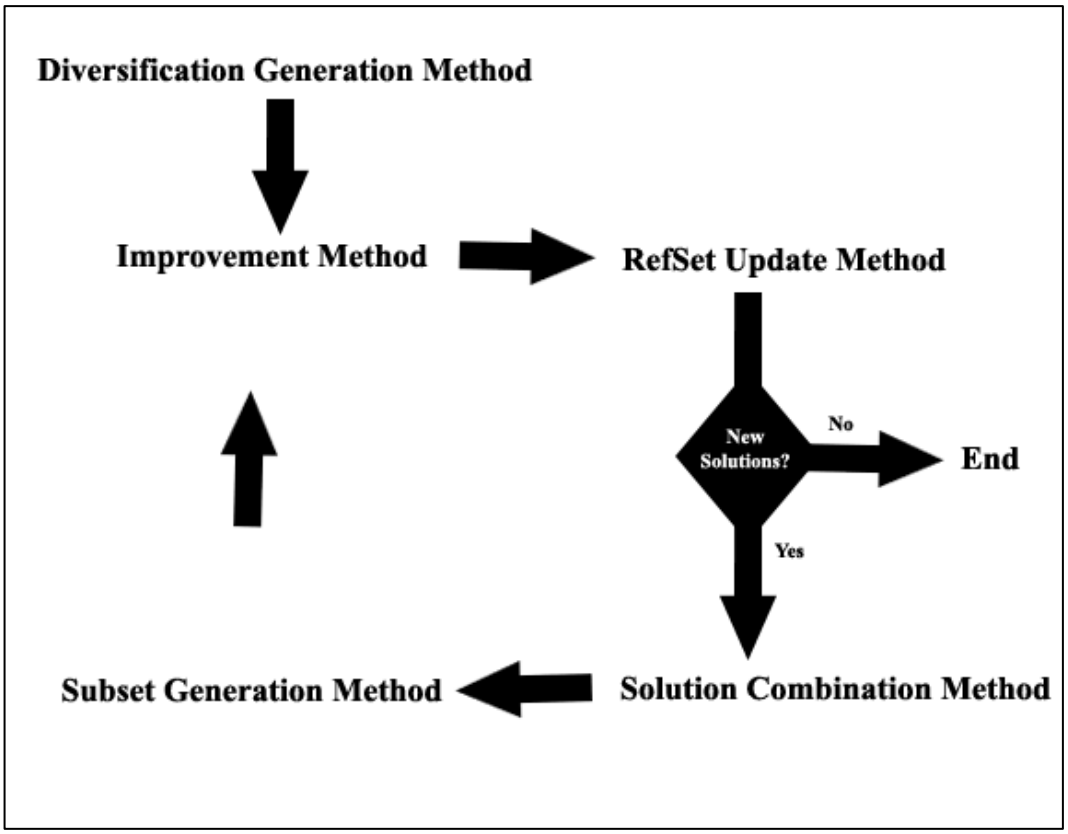


Figure 4.1. Diagram of Scatter Search.

4.2 – Solution Illustration

A PROTEIN in a solution is represented as a list of consecutive objects, AMINO ACIDs, in addition to van der Waals, electrostatic, torsion, and solvation energy values. The position of an Amino Acid in a PROTEIN object list is consistent with its position in the protein chain. Consequently the size of the PROTEIN object is equal to the number of amino acids of the protein. Each AMINO ACID consists of a name, Phi, Psi, omega, Chi1 to Chi4 angle values (if present), van der Waals, electrostatic, torsion, and ASA energy values, and a list of ATOM objects representing the atoms of that particular AMINO ACID. An ATOM has a name and a POSITION object, representing the Cartesian coordinates of that atom.

A protein consisting of three Amino acids is represented as follows:

- PROTEIN1 {AMINOACID1, AMINOACID2, AMINOACID3, VdwEnergy, EsEnergy, TorEnergy, ASA}.

- AMINOACID1: { LYS, 0, 151.5, 195.3, 175.3, 166.3, 48.9, 136.5, 0, VdwEnergy, EsEnergy, TorEnergy, ASA, ATOM1, ATOM2, ATOM3, ATOM4, ATOM5,

ATOM6,ATOM7, ATOM8, ATOM9}. LYS is the three letter code for the amino acid Lysine, following are the phi, psi, omega, chi1 to chi5 angle values, energy values and the atoms of this Amino Acid.

- ATOM1: {N, POSITION} where N stands for nitrogen and the POSITION object represents the Cartesian coordinates of the N atom calculated from its dihedral angle values.

- POSITION: {16.608, 27.235, - 0.943}

4.3 – Diversification Generation Method

The Diversification Generation Method (DGM) generates random, diverse and valid initial solutions. These solutions are formed by randomly selecting a nine width window of consecutive amino acids from the protein chain, then randomly selecting a 9 width fragment (containing phi, psi and omega values) for this particular position and placing the torsion angles in their corresponding spots. Next the Cartesian coordinates of the atoms of each amino acid are calculated and the energy of the solution is computed. These steps are repeated until the chain is full and the generated structure is valid, that is, each amino acid in the chain has phi, psi and omega values and there are no collisions between the atoms.

The first amino acid of the generated solutions have the phi, psi, chi and omega values of the corresponding reference protein and its atoms have the Cartesian coordinates of the corresponding reference protein atoms, this enables the torsion to Cartesian conversion process explained in section 4.9. The Diversification Generation Method Algorithm is presented in Figure 4.2.

```
Given : protein size PS, Population Size (PopSize=100)
For all solution[P], P=0,1.....PopSize
        While solution[P] is not valid
                Set the torsion angles of solution[P] to zero
                While the protein chain of solution[P]  is not full
                        Randomly select a 9 width window of amino acids from P
                        Randomly select a fragment from the list for this window
                        Place the torsion angles into solution [P]
                End While
                For each a in solution [P]
                        TorsionToCartesian(a-1, a, solution [P])
                End For
                CalculateEnergy(solution[P])
        End While
End For
```

Figure 4.2. Diversification Generation Method Algorithm.

4.4 – Improvement Method

The Improvement Method (IM) enhances the solutions generated by the DGM. After saving the existing values of the torsion angles, for every amino acid position in the chain a fragment is randomly selected for that position and torsion angle values are inserted into the solution. If the newly generated solution is feasible and its potential energy value is lower than the old solution, the move is accepted. The improvement method is run 25 x for a selected protein size. Then the same procedure is repeated with length 3 fragments, with number of moves being 50 x protein size. Next, the Cartesian coordinates of the atoms are calculated and the potential energy is computed. The Algorithm for the Improvement Method is presented in Figure 4.3.

```
For all solution[P], P=0,1..... PopSize
        n9=0
        n3=0
        While n9< 25 * PS
                For each a in solution[P]
                        Randomly select a fragment for position a
                        Place the torsion angles into solution[P]
                        For each aa in solution [P]
                                TorsionToCartesian(aa-1, aa, solution [P])
                        End For
                        CalculateEnergy(solution[P])
                        Accept move if valid and E decreases else Undo Move
                        n9++
                End For
        End While
        While n3< 50 * PS
                For each a in solution[P]
                        Randomly select a fragment for position a
                        Place the torsion angles into solution[P]
                        For each aa in solution [P]
                                TorsionToCartesian(aa-1, aa, solution [P])
                        End For
                        CalculateEnergy(solution[P])
                        Accept move if valid and E decreases else Undo Move
                        n3++
                End For
        End While
```

Figure 4.3. Improvement Method Algorithm.

4.5 – Reference Set Update Method

The Reference Set Update Method (RSUM) constructs two reference sets (RefSet), high-quality and diverse solutions. The RefSet b contains b1 high-quality solutions, and b2 diverse solutions. Since b=20% of the population's size and PopSize=100, RefSet has 20 solutions. The b1 solutions are the top 10 minimum energy valued solutions generated from IM. The b2 solutions are the solutions having diverse energy values from the b1 high-quality solutions. After selecting the top 10 solutions of minimum energy and placing them in the RefSet (HQRefSet), for every solution not in the HQRefSet, the minimum distance between this solution and all solutions in the HQRefSet is computed and sorted in decreasing order of minimum distances. The first b2 (most diverse) solutions having the highest energy values are inserted into the RefSet (DivRefSet). The algorithm terminates when no new solutions are found to be inserted into the RefSet. The Algorithm for the Reference Set Update Method is presented in Figure 4.4.

```
Sort the solutions in increasing order of energy values.
Pick the first 10 lowest energy valued solutions and place them in b1
Repeat PopSize - 10% (PopSize ) times
        For each solution, solution[k] not in HQRefSet
                Count how many torsion angles are different (greater than 5%
                difference) from the angles in the 10 HQ solutions.
                Compute the minimum of the 10 distance values and record it for
                solution[k]
        End For
End repeat
Sort the PopSize - 10% (PopSize ) solutions in decreasing order of their minimum
distance and classify the first b2 solutions having the highest minima as the
DivRefSet
```

Figure 4.4. Reference Set Update Method Algorithm.

4.6 – Subset Generation Method

In the Subset Generation Method (SGM) subsets of the reference set are generated by using subset type 1 method, consisting of two elements in each set. (b!/2!(b – 2)!) subsets are generated, where b is the size of the RefSet. The Algorithm for the Subset Generation Method is presented in Figure 4.5.

```
copy RefSet
CombinedSolution array of size =PopSize
        For i=0 to (b!/2!(b-2)!)-1
                Select 2 solutions not yet combined S1 and S2
                Combine S1 and S2 (S1 + S2--> S3)
```

Figure 4.5. Subset Generation Method Algorithm.

4.7 – Solution Combination Method

In this Method for every amino acid in the combined solution, the dihedral angles from either candidate solutions are added and the energy function up to this amino acid is calculated. The ones that lower the energy value of the structure are chosen. The Algorithm for the Solution Combination Method is represented in Figure 4.5.

```
For each AA in the S3
        For each A in AA
                Place the dihedral angles of S1.AA  in S3
                E1=CalculateEnergy(S3, AA)
                Place the dihedral angles of S2.AA  in S3
                E2=CalculateEnergy(S3, AA)
                        If E1<E2
                                Set the torsion angle values of S3 equal to the values
                                of S1
                        Else
                                Set the torsion angle values of S3  equal to the values
                                of S2
                        End If
        End For
End For
```

Figure 4.6. Solution Combination Method Algorithm.

4.8 – Side chain Assembly

After the termination of phase one, the solution with the lowest Cα-RMSD value in the final Reference Set is chosen to go through the side chain assembly phase.

In this phase fragments from the Dunbrack library are chosen and inserted into the solution. The method utilized is the same method utilized in the Improvement method of the Scatter Search algorithm in phase one, with 100 * protein size attempted moves.

In this phase the energy function includes the energy values produced by the side chain atoms and all-atom RMSD value is calculated. The algorithm for the side chain assembly is presented in Figure 4.7.

```
For all solution[P], P=0,1..... PopSize
        While n< 100 * PS
                For each a in solution[P]
                        Randomly select a fragment for position a
                        Place the side chain angles into solution[P]
                        For each aa in solution [P]
                                TorsionToCartesian(aa-1, aa, solution [P])
                        End For
                        CalculateEnergy(solution[P])
                        Accept move if valid and E decreases else Undo Move
                        n9++
                End For
        End While
End for
```

Figure 4.7. Side Chain Assembly Algorithm.

4.9 – Dihedral to Cartesian Transformation

The backbones of proteins consisting of a sequence of connected atoms can be represented either by their atoms' Cartesian coordinates or by their bond lengths and angles. The Cartesian coordinates represent the exact position of atoms in 3D space while bond lengths and angles represent the internal coordinates of atoms, explicitly phi, psi, omega, chi1, chi2, chi3, and chi4. In PDB, proteins are represented in Cartesian coordinates, whereas in computational biology, proteins are represented in dihedral angles.

Both representations are utilized in this work. Cartesian Coordinates are used to calculate protein properties such as RMSD, energy, distance between atoms, etc. and torsion angle values are selected from fragments and placed in candidate solutions.

The method of translation from torsion representation to Cartesian used in this work is proposed by Parsons et al. (2005). The method calculates the Cartesian coordinates of atoms successively from one end to the other, by having the Cartesian coordinates of the three previous atoms in the chain, the bond length of the new bond, the bond angle relative to the previous bond, and the torsion angle about the previous bond. Having the coordinates of three atoms A B C, the bond length of the bond C-D, the bond angle between the points B-C-D, and the torsion angle according to the bond B-C, the coordinates of the point D can be calculated. To exemplify, in the Chain "CA - C - N - CA - C - N", to calculate the Cartesian coordinates of CA we need the bond length N-CA, the bond angle C - N - CA and the torsion angle according to the bond C-N, which is the omega in this case.

However, when we fed our algorithm the Cartesian coordinates of the PDB proteins, the algorithm failed to produce exact structures. For 1CRN, it generated the structure with 3.17 Å Cα-RMSD and 2.94 Å all-atom RMSD. For 1ROP, it generated the structure with 4.80 Å Cα-RMSD and 6.98 Å all-atom RMSD. For 1UTG, it generated the structure with 7.03 Å Cα-RMSD and 6.60 Å all-atom RMSD. This is mainly due to the parameter values utilized for bond lengths and bond angles that are not 100% accurate. Figure 4.7, 4.8 and 4.9 represent the PDB and edited by the code structures for 1CRN, 1ROP AND 1UTG respectively.

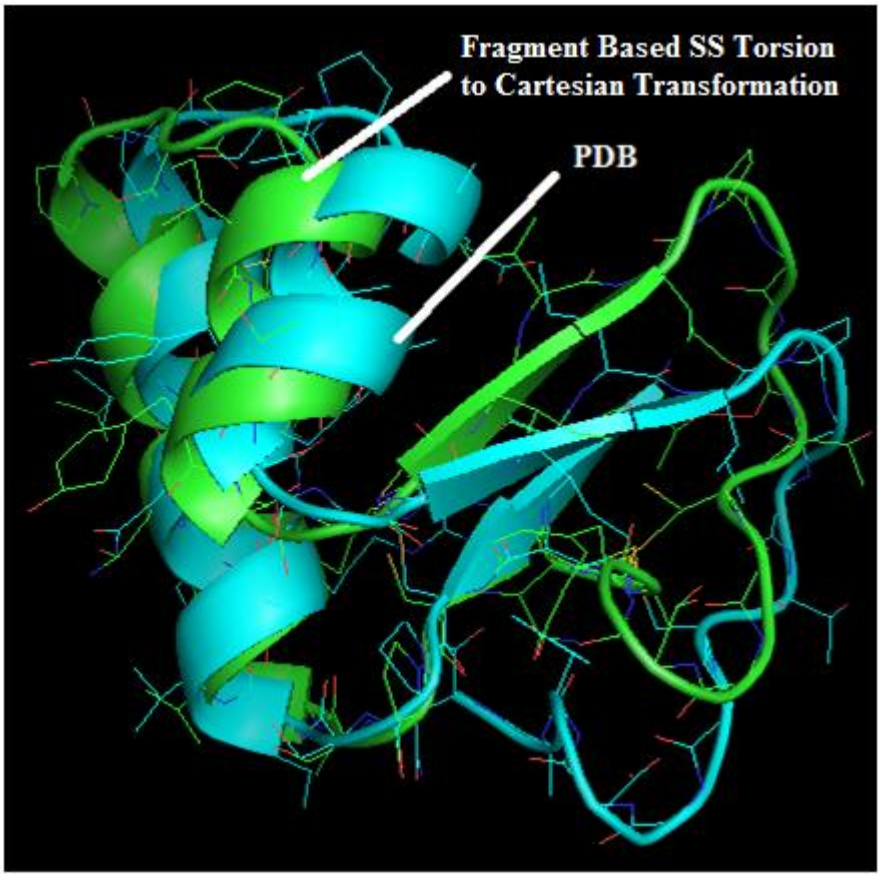


Figure 4.8. 1CRN Generated by Dihedral to Cartesian Transformation Algorithm.

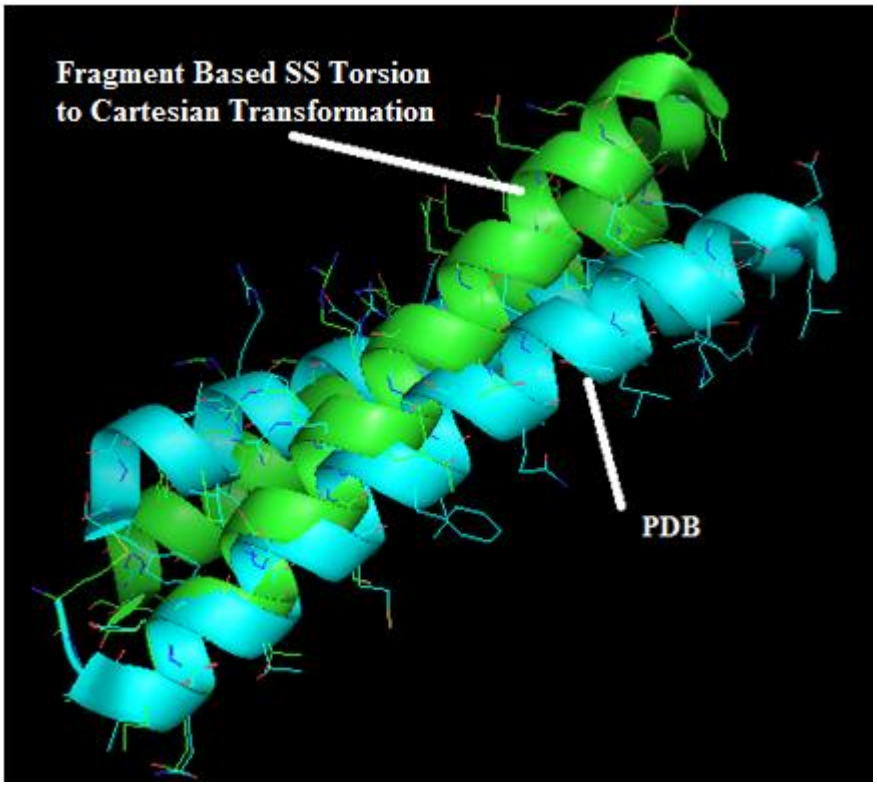


Figure 4.9. 1ROP Generated by Dihedral to Cartesian Transformation Algorithm.

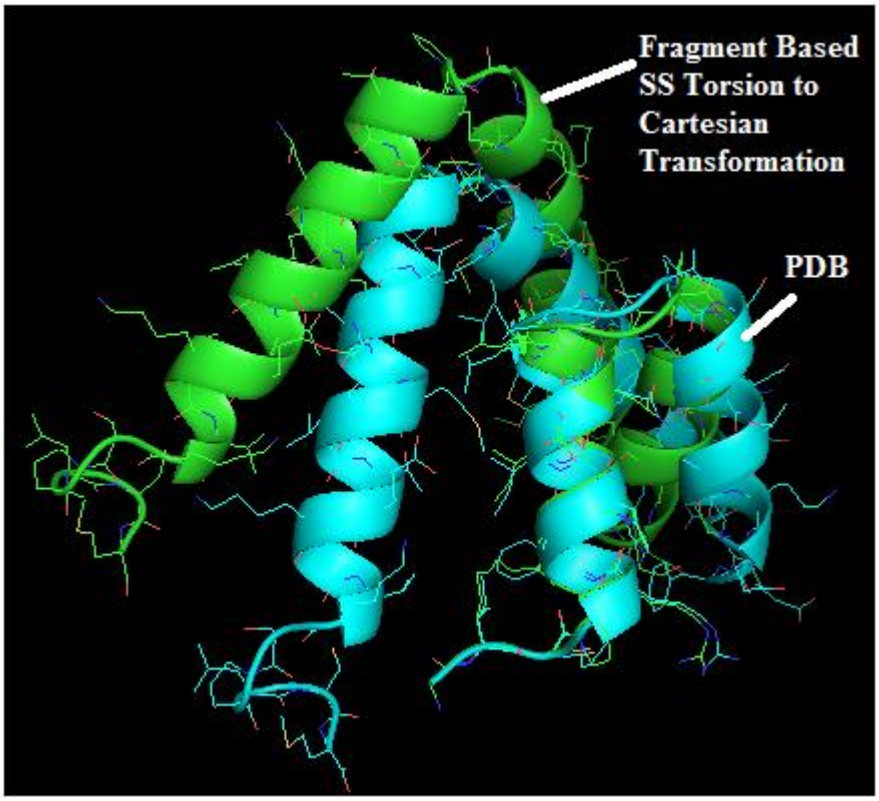


Figure 4.9. 1UTG Generated by Dihedral to Cartesian Transformation Algorithm.

CHAPTER FIVE

Experimental Results

5.1 – Fragment based SS and Mansour et al. SS Generated Results

In this section we compare our results to the results generated by Mansour et al. (2011). Table 5.1 tabulates the minimum Cα-RMSD values generated by both algorithms for 1CRN, 1ROP and 1UTG proteins. Figures 5.1, 5.2, and 5.3 display the tertiary structures of the three proteins in their native state (PDB) and generated by the two algorithms.

Table 5.1 shows that the Cα-RMSD for 1CRN dropped from 9.01 Å to 8.05 Å, for 1ROP from 12.14 Å to 5.43 Å and for 1UTG from 14.78 Å 12.34 Å. This shows that the approach utilized in this study significantly improves the three protein Cα-RMSD values. Furthermore, the structures generated, unlike the Mansour et al. work, have no discontinuities in them.

Figures 5.4, 5.5, and 5.6 show the PDB and the best structures generated by fragment based Scatter Search.

Table 5.1. Mansour et al. and Fragment Based SS Cα-RMSD Values.

Methodology / Proteins	1CRN	1ROP	1UTG
Mansour et al. (2011)	9.01 Å	12.14 Å	14.78 Å
Fragment based SS	8.05 Å	5.43 Å	12.34 Å

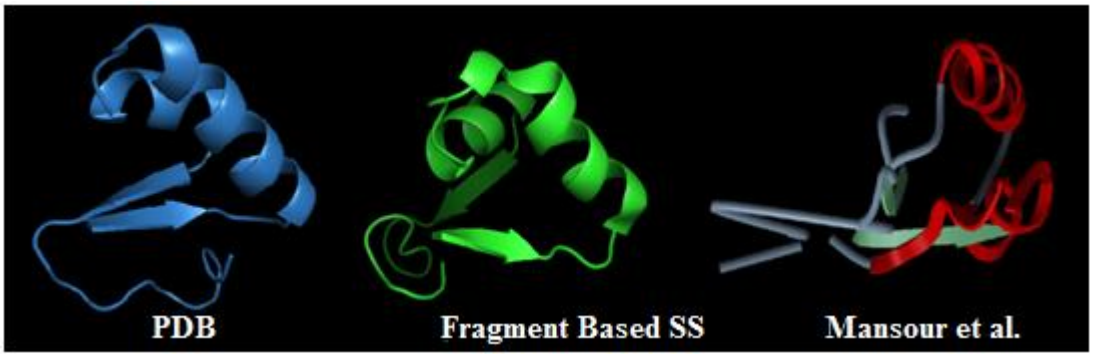


Figure 5.1. Structures generated by the two methods and the PDB structure for 1CRN.

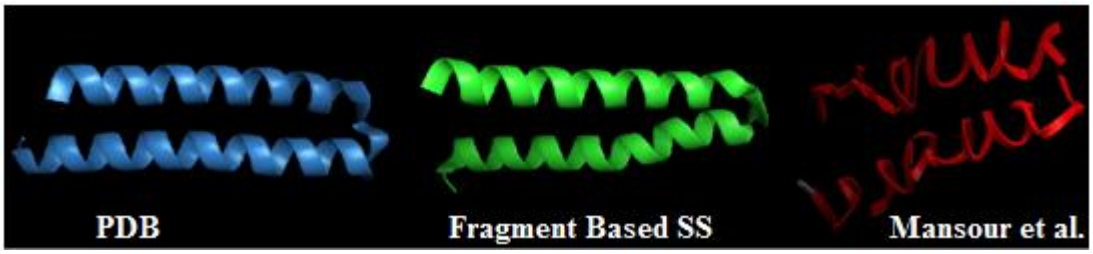


Figure 5.2. Structures generated by the two methods and the PDB structure for 1ROP.

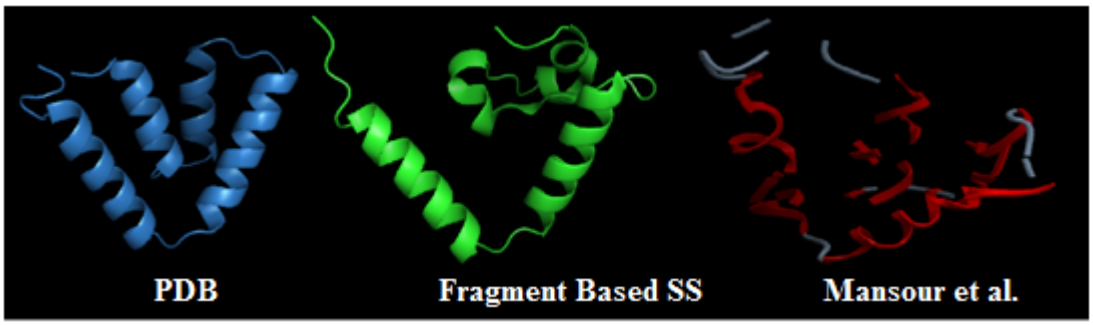


Figure 5.3. Structures generated by the two methods and the PDB structure for 1UTG.

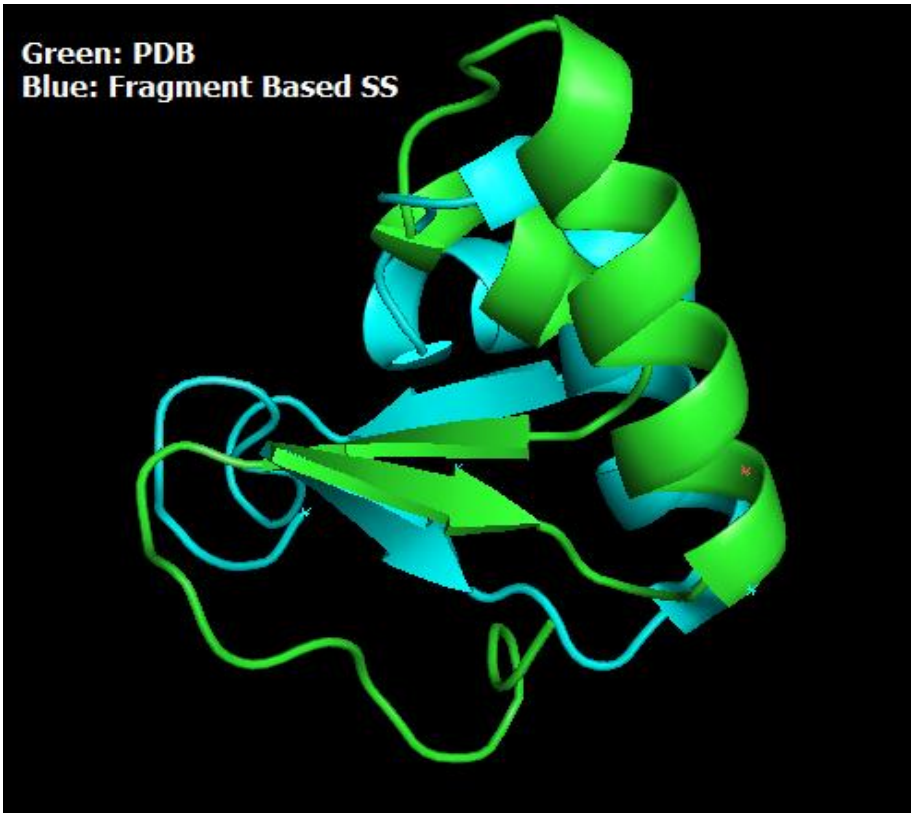


Figure 5.4. 1CRN Structures Generated by Fragment Based Scatter Search and PDB

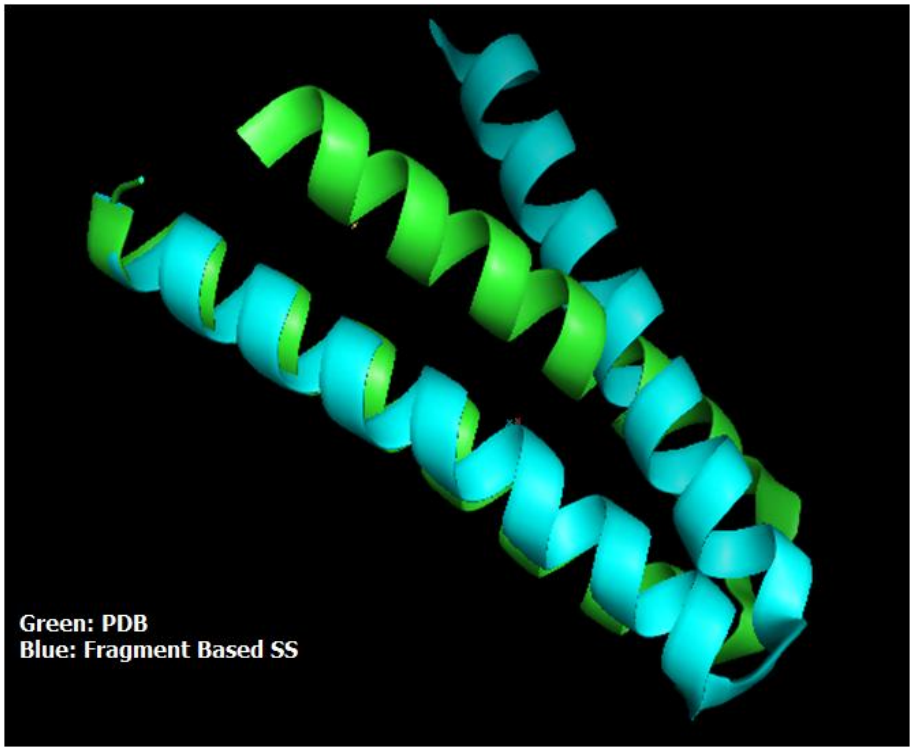


Figure 5.5. 1ROP Structures Generated by Fragment Based Scatter Search and PDB

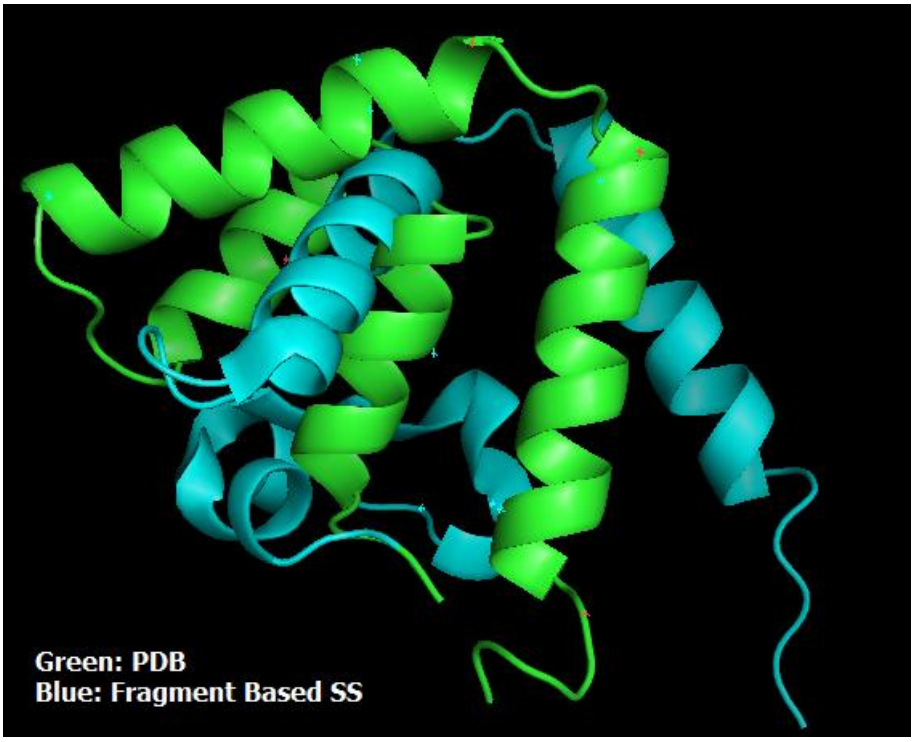


Figure 5.6. 1UTG Structures Generated by Fragment Based Scatter Search and PDB

5.2 – Fragment based Scatter Search using CHARMM36 and CHARMM22

In this section, to show the significance of using CHARMM36 function over the CHARMM22 utilized by Mansour et al. (2011), we run our code using CHARMM22 energy function. The results tabulated in table 5.2, show that for the three proteins the minimum Cα-RMSD values are higher when CHARMM22 is utilized. For 1CRN the minimum Cα-RMSD increased from 8.05 Å to 13.68 Å, for 1ROP from 5.43 Å to 9.36 Å, and for 1UTG from 12.34 Å to 14.40 Å.

Figures 5.7, 5.8 and 5.9 display a comparison of the structures generated for the three proteins.

Table 5.2. Cα-RMSD Values Generated from CHARMM36 and CHARMM22.

Method / Proteins	1CRN	1ROP	1UTG
Fragment based SS with CHARMM22	13.68 Å	9.36 Å	14.40 Å
Fragment based SS with CHARMM36	8.05 **Å**	5.43 Å	12.34 Å

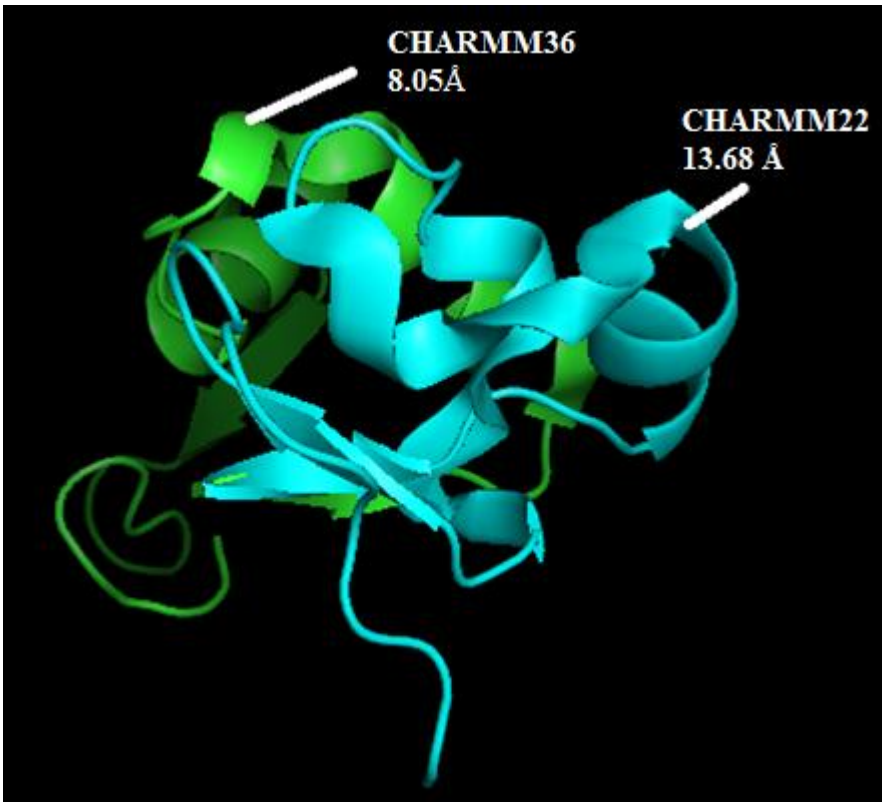


Figure 5.7. CHARMM22 and CHARMM36 Structures for 1CRN

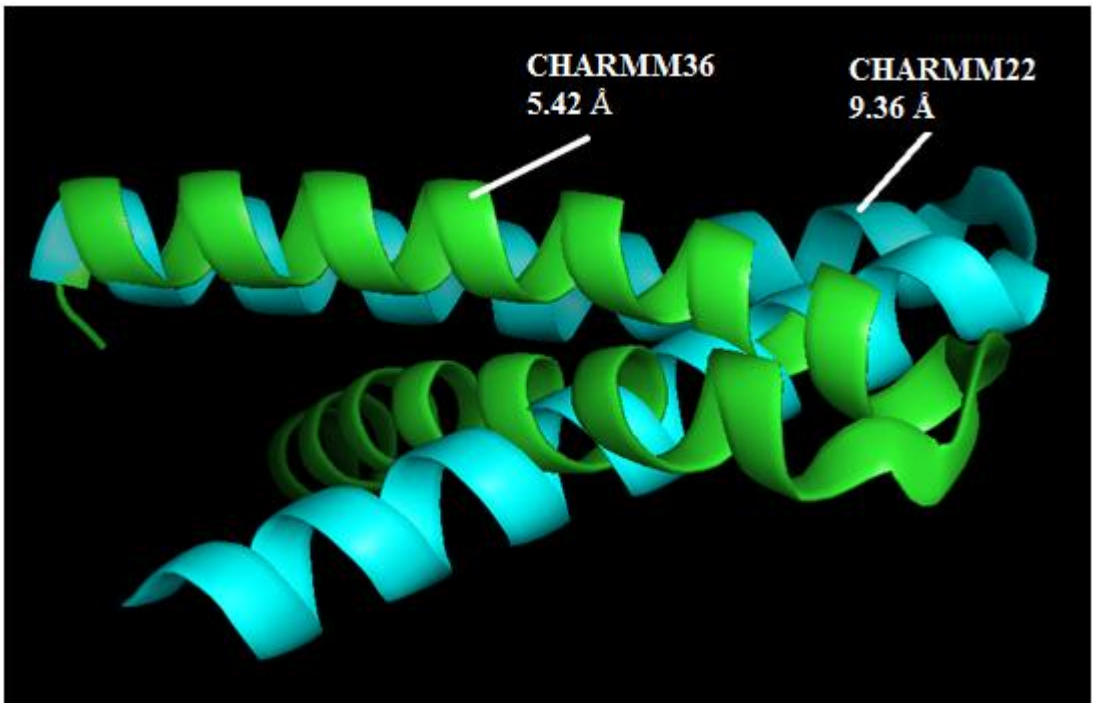


Figure 5.8. CHARMM22 and CHARMM36 Structures for 1ROP

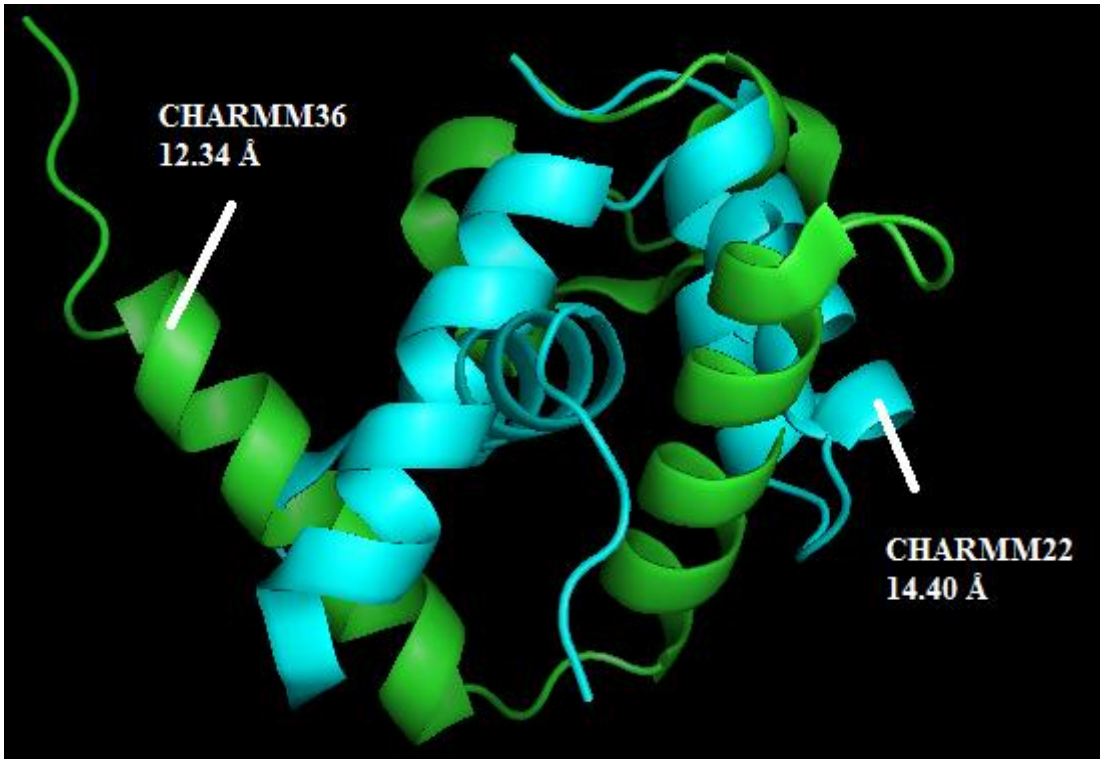


Figure 5.9. CHARMM22 and CHARMM36 Structures for 1UTG

5.3 – Fragment based Scatter Search, ROSETTA and I-TASSER Generated Structures

In these set of experiments we compare the generated structures from our code with those generated by I-TASSER and ROSETTA. Figures 5.10, 5.11 and 5.12 display the visualizations of the generated proteins.

Since I-TASSER and ROSETTA do not set the first amino acid coordinates of the structures to the coordinates of the corresponding PDB protein, after generating the structures from their servers we translated the coordinates to calculate the Cα-RMSDs and to visualize them.

As Tabulated in table 5.3, the Cα-RMSD results generated by our code are the lowest for the three proteins. These experiments confirm that since RMSD is a global measure, a small disorientation in one part of a protein results in a large RMSD value. For all the three tested proteins, the generated structures by I-TASEER and

ROSETTA have high Cα-RMSD values even though the overall conformation is accurate.

Table 5.3. Cα-RMSD Values Generated by Fragment based SS, I-TASSER and ROSETTA

Method / Proteins	1CRN	1ROP	1UTG
I-TASSER	12.14 Å	26.14 Å	19.94 Å
ROSETTA	11.35 Å	23.28 Å	18.20 Å
Fragment based SS	8.05 Å	5.43 Å	10.34 Å

Figure 5.10. 1CRN Structures Generated.

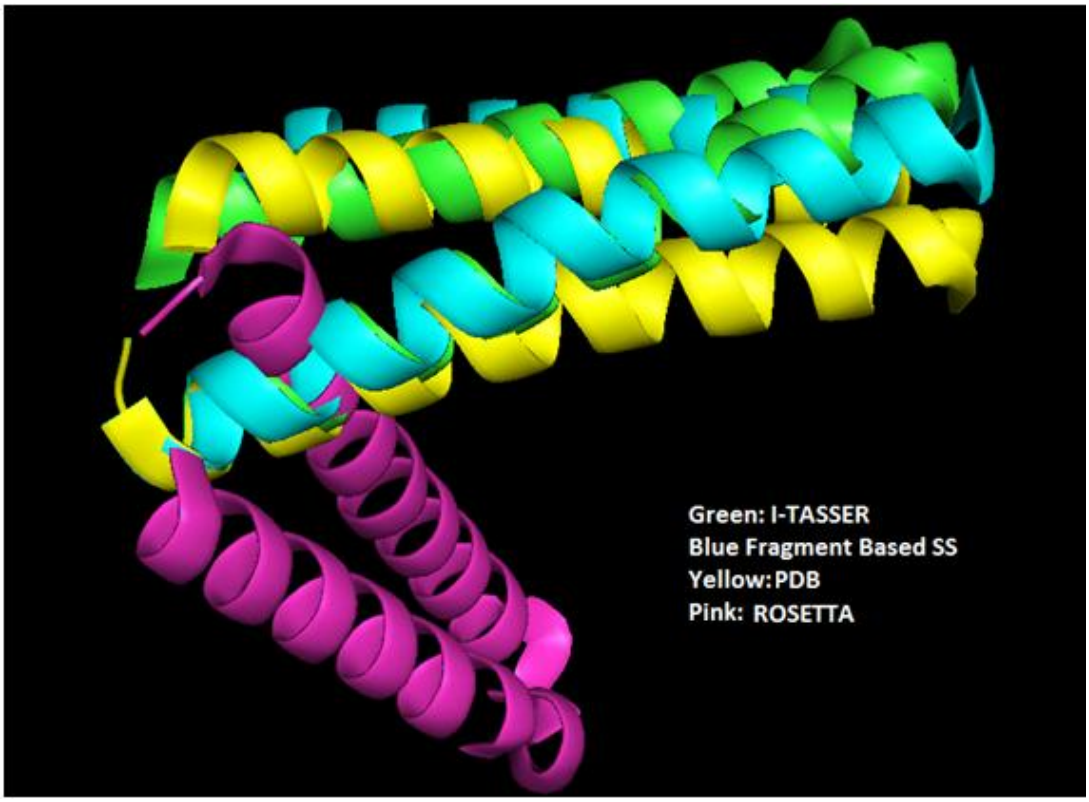


Figure 5.11. 1ROP Structures Generated.

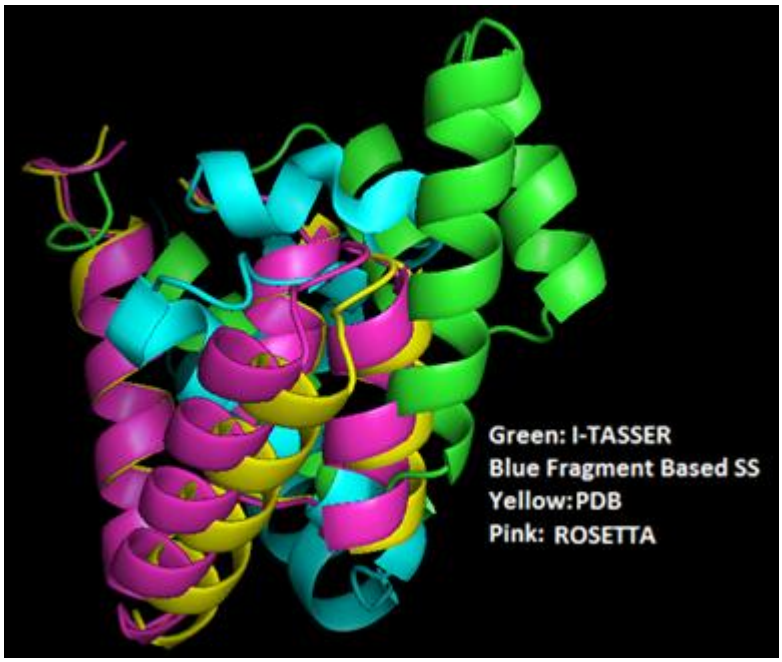


Figure 5.12. 1UTG Structures Generated..

5.4 – Side Chain Assembly Generated Structures

In this section we show visualizations of the side chains assembled in phase two of the algorithm for the three proteins. Figures 5.13, 5.14, 5.15 and 5.16 show that the

side chains are correctly assembled. The produced structures have all-atom RMSD values 9.62 Å, 9.33 Å and 11.56 Å for 1CRN, 1ROP and 1UTG respectively.

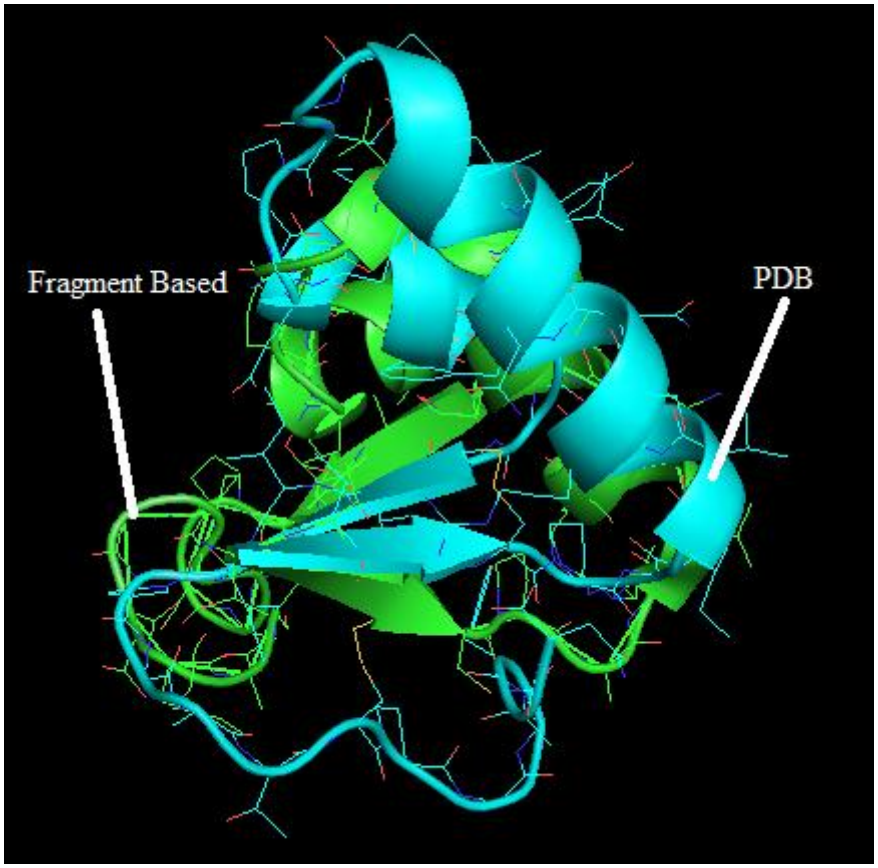


Figure 5.13. Fragment Based and PDB Structures for 1CRN.

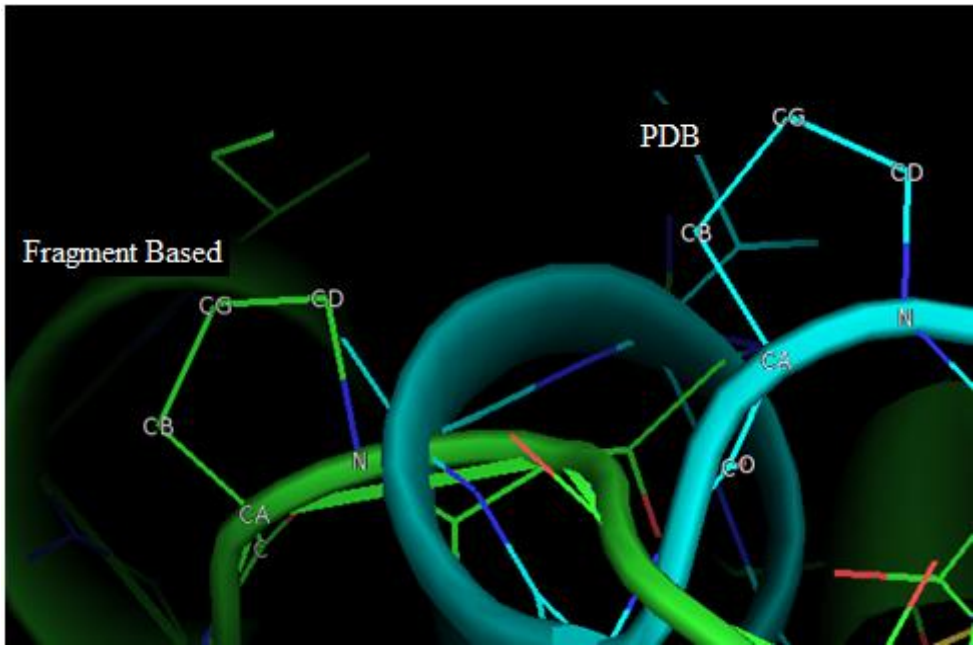


Figure 5.14. Fragment Based and PDB Side Chain Atoms for 1CRN PRO 36 Amino Acid.

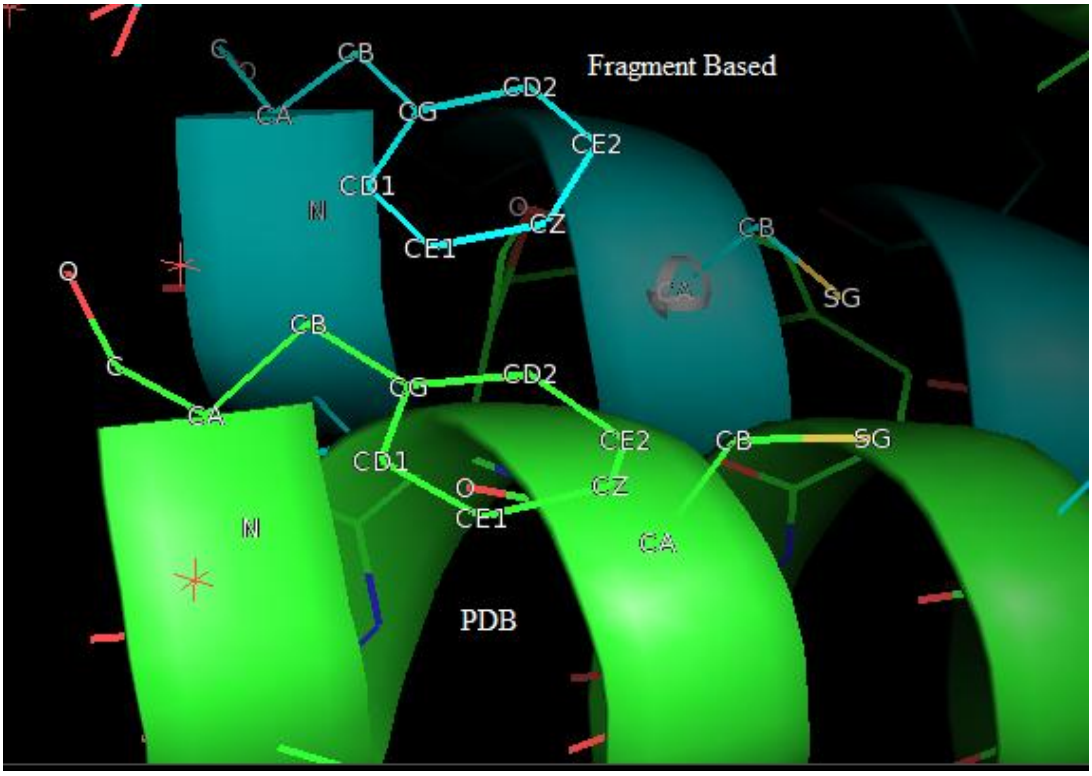


Figure 5.15 Fragment Based and PDB Side Chain Atoms for 1ROP PHE 56 and CYS 52 Amino Acids.

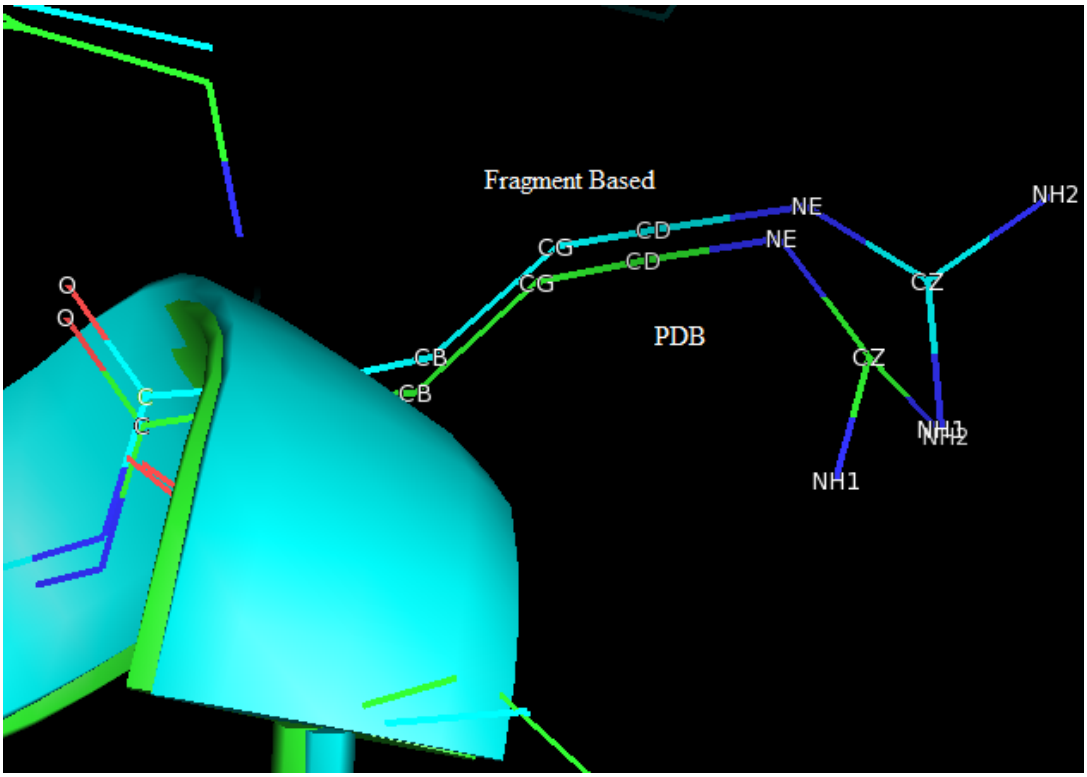


Figure 5.16. Fragment Based and PDB Side Chain Atoms for 1UTG ARG 5 Amino Acid.

5.6 – Generated Structures Energy Values

Table 5.4 displays the energy values of the final High Quality RefSet corresponding to the three proteins.

Table 5.4. Energy Values of the Final High Quality RefSet

1CRN		1ROP		1UTG	
Energy	**Cα-RMSD**	**Energy**	**Cα-RMSD**	**Energy**	**Cα-RMSD**
19854.40	15.69	20407.98	12.19	29841	12.34
19868.25	20.88	20407.99	5.43	29841.06	16.07
19871.78	20.47	20408.02	9.79	29841.22	17.50
19879.02	12.62	20408.03	14.22	29841	17.94
19882.92	19.68	20408.07	7.71	29841.06	19.11
19883.00	10.81	20408.07	13.62	29841.01	20.08
19884.91	20.20	20408.08	9.67	29841.03	22.05
19885.69	20.32	20408.08	11.26	29840.98	22.50
19891.29	8.05	20408.09	9.25	29841.14	22.76
19918.31	17.97	20408.4	10.82	29841.07	22.79

From the generated values we conclude that since our energy function is not 100% accurate, in the final high quality RefSet the solution with the lowest energy is not always the lowest Cα-RMSD value solution. For 1UTG the lowest energy value

solution (29841 kcal/mol) had the lowest Cα-RMSD value (12.34 Å). For 1ROP, the lowest Cα-RMSD (5.43 Å) solution is the second lowest energy valued (20407.99 kcal/mol) solution. As for 1CRN, the lowest RMSD (10.81 Å) solution is the ninth lowest energy solution (19883.00 kcal/mol).

Regarding the execution times, 1CRN halted after 425minutes, 1ROP after 561 minutes and 1UTG after 697 minutes. It should be also noted that 90% of the execution time is utilized by phase one of the algorithm.

CHAPTER SIX

Conclusions and Recommendations

In this thesis an *ab initio* fragment based protein structure prediction algorithm is presented. Given a protein sequence and its corresponding fragments, the algorithm first assembles the backbone of the candidate solutions then the side chains of the best generated solution.

Several improvements are made to the recently developed Scatter Search algorithm by Mansour et al. (2011). The most recent version of the CHARMM force field, CHARMM36, is employed. VSWITCH and FSHIFT methods are used to calculate interactions between non-bonded atom pairs that shift and switch interatomic forces at long distances making large simulations computationally feasible without destabilizing the macromolecule. Results are evaluated on three proteins and compared with those generated by I-TASSER, ROSETTA, and Mansour et al. (2011).

The proposed algorithm generates structures with Cα-RMSD values comparable to that of existing approaches. For 1CRN, the lowest Cα-RMSD value reached is 8.05Å, for 1ROP 5.43 Å, and for 1UTG 12.34 Å.

The major limitation of this work is the inaccuracy of the energy function. Assuming that bond lengths and bond angles are constant, does not reflect the real behavior of proteins where these values fluctuate with a small angle change. Another limitation is the dihedral to Cartesian transformation method utilized that is not 100% accurate.

Further future work can focus on using more precise energy functions. The CMAP term ignored for simplicity, should be added to the energy function. In addition, hydrogen atoms, ignored, can be added to the solution representation, thus adding the hydrogen bonding term to the energy function. Also adding bonds, angles, improper torsions ignored as a result of the constant binding geometry assumption.

Moreover, the GDT metric can to use to assess the structures and a parallel version of the algorithm can be developed to decrease the execution time.

References

Abual-Rub, M.S., & Abdullah, R. (2008). A survey of protein fold recognition algorithms. *Journal of Computer Science, 4*(9), 768-776. Doi:10.1.1.153.1613

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science*, *181*(4096), 223-230.

Banner, D.W., Kokkinidis, M., & Tsernoglou, D. (1987). Structure of the ColE1 rop protein at 1.7 A resolution. *Journal of Molecular Biology*, *196*(3), 657-675.

Bartesaghi, A., & Subramaniam, S. (2009). Membrane protein structure determination using cryo-electron tomography and 3D image averaging. *Current Opinion in Structural Biology*, *19*(4), 402-407. Doi:10.1016/j.sbi.2009.06.005.

Bonetti, D.R.F., Delbem, A.C.B., Travieso, G., & de Souza, P.S.L. (2010, July 18-23). *Optimizing van der Waals calculi using Cell-lists and MPI.* Paper presented at the Proceedings of the IEEE Congress on Evolutionary Computation, Barcelona (pp. 1-7). USA:IEEE.

Bradley, P., Misura, K.M.S., & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science*, *309*(5742), 1868-1871.

Brooks, B.R., Bruccoleri, B.E., Olafson, B.D., States, D.J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, *4*(2), 187-217.

Buchan, D.W., Ward, S.M., Lobley, A.E., Nugent, T.C., Bryson, K., & Jones, D.T. (2010). Protein annotation and modelling servers at University College London. *Nucleic Acids Research*, *38*, 563-568. Doi:10.1093/nar/gkq427.

Carugo, O., & Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Science*, *10*(7), 1470-1473.

Chen, L., Liu, G., Wang, Q., & Hou, W. (2009, October 16-18). *Homology modeling of the three-dimensional structure of bovine serum albumin.* Paper presented at the 3rd International Conference on Biomedical Engineering and Informatics, Yantai, (pp. 2377-2381). USA: IEEE.

Cramer, F. (2007). Emil Fischer's lock-and-key hypothesis after 100 years - towards a supracellular chemistry. In J.-P. Behr (Ed.), *Perspectives in supramolecular chemistry: The lock-and-key principle* (pp. 1-23). UK: John Wiley & Sons.

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

Doong, S.-H. (2007, January 3-6). *Protein homology modeling with heuristic search for sequence alignment.* Paper presented at Proceedings of the 40th Hawaii International Conference on System Sciences, Waikoloa, *(*pp. 128-137). USA: IEEE.

Floudas, C.A. (2007). Computational methods in protein structure prediction. *Biotechnology and Bioengineering*, *97*(2), 207-213.

Glover, F. (1998). A template for scatter search and path relinking. In J.K. Hao, E. Lutton, E. Ronald, M. Schoenauer, & D. Snyers (Eds.), *Artificial evolution, lecture notes in computer science* (pp. 1-51). Heidelberg:Springer.

Gront, D., Kulp, D.W., Vernon, R.M., Strauss, C.E.M., & Baker, D. (2011). Generalized fragment picking in rosetta: Design, protocols and applications. *PLoS ONE, 6*(8). Doi:10.1371/journal.pone.0023294

Hasel, W., Hendrickson, T.F., & Still, W.C. (1988). A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahedron Computer Methodoloqy, 1*(2), 103-116.

Hu, X.-M., Zhang, J., Xiao, J., & Li, Y. (2008). Protein folding in hydrophobic-polar lattice model: A flexible ant-colony optimization approach. *Protein and Peptide Letters*, *15*(5), 469-477.

Ilari, A., & Savino, C. (2008). Protein structure determination by x-ray crystallography. *Methods in Molecular Biology*, *452*, 63-87.

Jaroszewski, L., Li, Z., Cai, X.H., Weber, C., & Godzik, A. (2011). FFAS server: Novel features and applications. *Nucleic Acids Research*, *39*, 38-44.

Johnson, C., & Katikireddy, A. (2006, July 8-12). *A Genetic algorithm with backtracking for protein structure prediction*. Paper Presented at Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, Seattle, (pp. 299-300). New York, NY: ACM.

Jones, D.T., & McGuffin, L.J. (2003). Assembling novel protein folds from super-secondary structural fragments. *Proteins, 53*(S6), 480-485.

Kaufmann, K.W., Lemmon, G.H., DeLuca, S.L., Sheehan, J.H., & Meiler, J. (2009). Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry, 49*(1), 2987-2998.

Kopp, J., & Schwede, T. (2004). Automated protein structure homology modeling: A progress report. *Pharmacogenomics*, *5*(4), 405-416.

Lee, J., Wu, S., & Zhang, Y. (2009). Ab initio protein structure prediction. In R. D. John (Ed.), *From protein structure to function with bioinformatics* (pp. 1-26). London: Springer.

Levinthal, C. (1968). Are the pathways for protein folding? *Journal of Chemical Physics*, *65*(1), 44-45.

Liang, F., & Wong, W.H. (2001). Evolutionary Monte Carlo for protein folding simulations. *Journal of Chemical Physics, 115*(7), 3374-3381.

Liwo, A., Pillardy, J., Czaplewski, C., Lee, J., Ripoll, D.R., Groth, M., … Scheraga, H.A. (2000, April 8-10). *UNRES: A united-residue force field for energy-based prediction of protein structure-origin and significance of multibody terms.* Paper presented at Proceedings of the fourth annual international conference on Computational molecular biology, Tokyo, Japan (pp. 193-200). USA: ACM.

Lobley, A., Sadowski, M.I. & Jones, D.T. (2009). pGenTHREADER and pDomTHREADER: New methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*, *25*(14), 1761-1767.

MacKerell, A.D., Bashford, D., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., … Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry*, *102*(18), 3586-3616.

Maisuradze, G.G., Senet, P., Czaplewski, C., Liwo, A., & Scheraga, H.A. (2010). Investigation of protein folding by coarse-grained molecular dynamics with the UNRES force field. *Journal of Physical Chemistry, 114*(13), 4471-4485.

Mansour, N., Ghalayini, I., El-Sibai, M., & Rizk, S. (2011, March 23-25). *Evolutionary algorithm for predicting all-atom protein structure.* Paper presented at Proceedings of the ISCA Third International Conference on Bioinformatics and Computational Biology, New Orleans, Louisiana, USA (pp. 7-12). USA: ACM.

Mansour,N., Kanj, F., & Khachfe, H. (2010, August 10-12). *Evolutionary algorithm for protein structure prediction.* Paper presented at Sixth International Conference on Natural Computation, Yantai, Shandong (pp. 3974-3977). USA: IEEE.

Mukherjee, A.B., Zhang , Z., & Chilton, B.S. (2007). Uteroglobin: A steroid-inducible immunomodulatory protein that founded the Secretoglobin superfamily. *Endocrine Reviews, 28*(7), 707-25.

Neylon, C. (2008). Small angle neutron and X-ray scattering in structural biology: Recent examples from the literature. *European Biophysics Journal*, *37*(5), 531-541.

Paluszewski, M., Hamelryck, T., & Winter, P. (2006). Reconstructing protein structure from solvent exposure using Tabu search. *Algorithms for Molecular Biology*, *1*(20). Doi:10.1186/1748-7188-1-20

Parsons, J., Holmes, J.B., Rojas, J.M., Tsai, J., & Strauss, C.E. (2005). Practical conversion from torsion space to Cartesian space for in silico protein synthesis. *Journal of Computational Chemistry*, *26*(10), 1063-1068.

Patton, A.L., Punch, III, W.F., & Goodman, E.D. (1995, July 15-19). *A standard GA approach to native protein conformation prediction.* Paper presented at Proceedings of the 6th International Conference on Genetic Algorithms, Pittsburgh (pp. 574-581). San Francisco: Morgan Kaufmann.

Ponder, J.W., & Case, D. (2003). Force fields for protein simulations. *Advances in Protein Chemistry, 66*, 27-85.

Ramachandran, G.N., & Sasisekharan, V. (1968). Conformations of polypeptides and proteins. *Advances in Protein Chemistry, 23*, 283-437.

Ramachandran, G.N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology, 7*, 95-99.

Rohl, C.A., Strauss, C.E., Misura, K.M., & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods in Enzymology*, *383*, 66-93.

Roy, A., Yang, J., & Zhang, Y. (2012). COFACTOR: An accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research, 40*, 471- 477. Doi:10.1093/nar/gks372

Roy,A., Kucukural, A., & Zhang,Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols*, *5*(4), 725-738.

Sanger, F., & Thompson, E.O.P. (1953). The amino-acid sequence in the glycyl chain of insulin. 1. The investigation of lower peptides from partial hydrolysates. *Biochemical Journal, 53*(3), 353-366.

Sanger, F., & Thompson, E.O.P. (1953). The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymatic hydrolysates. *Biochemical Journal, 53*(3), 366-374.

Setubal, J., & Meidanis, J. (1997). *Introduction to computational molecular biology*. Boston: PWS Publishing Company.

Shmygelska, A., & Hoos, H.H. (2005). An ant colony optimization algorithm for the 2D and 3D hydrophobic polar protein folding problem. *BMC Bioinformatics, 6*(30). Doi:10.1186/1471-2105-6-30

Sikder, A.R., & Zomaya, A.Y. (2005). An overview of protein-folding techniques: Issues and perspectives. *International Journal of Bioinformatics Research and Applications, 1*(1), 121-143.

Steinbach, P.J., & Brooks, B.R. (1994). New spherical-cutoff methods for long-range forces in macromolecular simulation**.** *Journal of Computational Chemistry, 15*(7), 667-683**.**

Takano, T. (1977). Structure of myoglobin refined at 2.0 Å resolution. II. Structure of deoxymyoglobin from sperm whale. *Journal of Molecular Biology, 110*(3), 569-584.

Teeter, M.M. (1984). Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin. *Proceedings of the National Academy of Sciences of the United States of America, 81*(19), 6014-6018.

Schrödinger. (2013). The PyMOL Molecular Graphics System. Retrieved from http://www.pymol.org/

Unger, R., & Moult, J. (1993). Genetic algorithms for protein folding simulations. *Journal of Molecular Biology, 231*(1), 75-81.

Wesson, L., & Eisenberg, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in Solution. *Protein Science, 1*(2), 227-235.

Wodak, S., & Janin, J. (1980). Analytical approximation to the accessible surface area of proteins. *Proceedings of the National Academy of Sciences of the United States of America, 77*(4), 1736-1740.

Wu, H. (1931). Studies on denaturation of proteins. XIII. A theory of denaturation. 1931. *Chinese Journal of Physiology, 5*, 321-344.

Wüthrich, K. (1990). Protein structure determination in solution by NMR spectroscopy. *The Journal of Biological Chemistry, 265*(36), 22059-22062.

Zhang, X., & Li, T. (2007, July 6-8). *Improved particle swarm optimization algorithm for 2D protein folding prediction.* Paper presented at the 1st International Conference on Bioinformatics and Biomedical Engineering, Wuhan (pp. 53-56). USA: IEEE.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics, 9*(40). Doi:10.1186/1471-2105-9-40