بسم الله الرحمن الرحيم

Al al-Bayt University
Prince Hussein Bin Abdullah College for Information Technology
Computer Science Department

Prediction of Protein 3D Structure from Contact Map

By

Nawal T. Alqllab

2009

Prediction of Protein 3D Structure from Contact Map

By

Nawal T. Alqllab

Supervisor: Dr. Jehad Al Nihoud
Co-supervisor: Dr. Hamed Al fawareh

A Thesis Submitted to the
Scientific Research and Graduate Faculty in Partial fulfillment of the
Requirements for the Degree of Master of Science
in Computer Science

<u>Members of the Committee</u>	<u>Approved</u>
Dr. Jehad Al Nihoud	…………………
Dr. Hamed Al fawareh	…………………
Dr. Mariam Nusair	…………………
Dr. mamoun Rbabaa	…………………
Dr. Akram Hamarsha	…………………

Al al-Bayt University
Mafraq, Jordan
2009

Dedications

To my husband Khaled
For your guidance, support, and love

To my brother Hussein
With my deep gratitude

To my mother and my aunt
For their encouragement and understanding during the period of my study

To my family and friends
For all what have done for me, their love, care, patience, and
encouragement to continue my way.

Acknowledgements

First of all, I would like to thank ALLAH for his grace to me, without his will, non of this would be possible.

I specially thank the following:

My supervisors Dr. Jehad Nihoud, and Dr. Hamed Fawareh for their guidance and help associate me during my study.

Dr. Isam Aldauood for his assistance and scientific hint to complete this work.

Also I would like to thank the members committee.

Thank you to all who have contributed to accomplish this work.

Table of Contents

List of Tables

List of Figure

Abstract

Prediction of Protein 3D Structure from Contact Map

Bioinformatics is an interesting topic for both the biologists and computer scientists. The ability to predict protein 3D structure from the amino acid sequence is not less than revolutionize in this area. A fundamental principle in all protein sciences is that protein structure leads to protein function. The 3D structure of protein can be represented using $N\times N$ symmetrical binary matrix C called contact map whose element $C\ (i, j)$ =1 if and only if the physical distance between amino acid *i* and *j* is less than or equal to a pre assigned threshold *T* otherwise $C\ (i, j)$ =0 Predicting 3D structure directly from primary sequence of protein is very complex problem, so various computational approach participate in analyzing and extracting rules to predict tertiary structure from contact map.

This thesis focuses on developing a method for predicting 3D structure of protein from contact map using MATLAB, which is contract with mathematical properties that can be derived from distance values between pairs of the amino acid, typically measured in Angstroms (A°). The proposed method focuses on choosing the threshold value for computing the contact map, which is affecting connectivity between the contact map and its 3D structure, not any threshold give accurate contact map which provides exact 3D structure. We found that the contact maps computed using threshold values (12-18) Å allow better 3D structure recovery than those computed at thresholds (7-9) Å. This approach aims to detect the dense areas that form the basic functional areas in the contact map by scanning module. Looking for the dense area is an important step that will improve the performance of the predicting 3D structure of protein from it CM based on prediction quality more than quantity of contacts. The experimental results in this thesis are obtained on data set of proteins related to different classes (mainly alpha, mainly beta, mixed alpha and beta) which are extracted from PDB. The average execution time of the proposed algorithm is varying depending on protein size.

The results show that the predicting structures can be determined using MATLAB in reliable and efficient manner.

CHAPTER ONE

INTRODUCTION

Bioinformatics, a rapidly evolving discipline, is the development and application of algorithms and methods to solve formal and practical problems arising from the management, bioinformatics analysis biological data and turn it into knowledge of biological systems.

The National Center for Biotechnology Information (NCBI) defines Bioinformatics as: “Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domain, and protein structure; and the development and implementation of tools that enable efficient access and management of different types of information” [1].

In bioinformatics, one of the major challenges facing the structural biology research is to determine the biological functions of genes identified through large-scale sequencing efforts. Predicting of the three dimension structure of protein provides valuable insight into function.

Unfortunately, the gap between the number of solved protein structures and the number of protein sequences continues to widen rapidly through the long and expensive processes required for solving structures experimentally. Prediction of structures from amino acid sequence is an emerging and promising method that may help to narrow this gap [3].

Traditional experimental techniques for deriving macromolecular structure data are X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and electron microscopy, these methods give data as a set of Cartesian coordinates representing the position of the atoms in these structure [4], But these methods areslow and don’t scale up to current sequencing speeds.

Furthermore, using experiments to determine how protein functions is a daunting task, so that predicting the 3D structure of protein from liner sequence of amino acids from contact map is an interesting topic for computer scientists.

In present, there is no model that gives 100% accurate prediction to this problem, but a number of intermediate stages that add a hope beam for solving this problem are available.

1.1 Historical Background

From 1866 when George Mendel discovered an effect element, called gene, which is responsible for passing and control of a single characteristic, until February 2001 when the first draft of human genome project was published, during that period, several events occurred, these events discovered a lot of theory and changed many ideas about molecular biology.

In the mid of eighteenth century, the common idea was said that chromosomal protein carry genetic information and the DNA plays a secondary role, but Avery and McCarty in 1944 provided actual experimental evidence that the DNA was the main constituent of genes which is responsible for inheritance [5]. In the mid of 1950s the first protein structure was determined through X-ray crystallography, after that NMR was used to determine nucleic acid structure which became a reality [7].

Protein structure is complex, but it should be noted that by the 1970s protein scientists had determined the basic principles. These principles of protein structure now form the stable foundation needed for researching many of the remaining questions in protein science [4].

Since the early 1980s, the number of structures of nucleic acids has grown exponentially. Over the last decade of the twentieth century there have been many important advances toward automated structure determination.

Continued efforts to uncover the underling principles of protein structure will result in much greater insights into the complex functions of these molecules.

1.2 Problem Definition

The most popular methods for deriving macromolecular structure data are X-ray crystallography, and Nuclear Magnetic Resonance (NMR) spectroscopy, but these are laborious and slow behind the rapid progress observed in structural genomics [4], so that the ability to predict 3D structure of protein using its amino acid sequence is a fundamental open problem in computational molecular biology.

Recently, there are many research efforts that provide guidelines for protein contact map prediction, these efforts used machine learning approaches such as neural network [11,12,13] and distance geometric [10,14,15,16].

Proteins structures are described by the coordinates of the atoms which are extracted from PDB, where residues are considered as unique entities to compute the contact map, a contact map is a binary matrix whose elements *C(i,j)*=1 if the distance between residues *i* and *j* is less than or equal to pre assigned threshold *T* and *C (i,j)*=0 otherwise. Threshold is a grand member effects on accuracy of the result in any method present in this approach.

There are three main representations for a conformation of a molecule: Cartesian coordinates, a distance geometry descriptor; and Internal coordinates. In Cartesian coordinates, each atom center is specified by x, y, and z coordinates [17]. Many molecular file formats, such as PDB use Cartesian coordinate's representation.

1.3 Objectives

X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy are the most widely used experimental methods to derive macromolecular structure data but they suffer from some limitations such as time consuming (minimum 6-12 month) and very expensive, also X-ray can not crystallize some types of proteins and difficult to determine their structures [24].

As a result, the current growth rate of the Cambridge Structural Data base (CCSD) is more than 15,000 new structures per year [25]. This growth rate is approximately 10 times the growth rate of the protein Data Bank (PDB one of the earliest scientific data base). So that the main reason that leads to develop non_ experimental models is that the current size of the protein sequences available precedes the available size of protein 3D structure, because the experimental methods are not efficient to predict all the available sequences rapidly [4].

Protein structure leads to protein function, the structures of proteins allow for the placement of particular chemical groups in specific places in 3D space, precise placement of chemical groups allows proteins to play important structural, transport, and regulatory functions in organisms, in order that Predicting of the three dimension structure of protein provides valuable insight into function [9], Recently, there are many research hard work to present many strategies for protein contact map prediction to get ride of the gap between the number of solved protein structures and the number of protein sequences in PDB.

This thesis aims to develop a method for predicting 3D structure of protein from contact map. A new approach takes short time to perform this mission and it is not expensive, the proposed method show that the protein structures can be determined by computing a set of consistent coordinates using MATLAB which is one of the most popular software programs used in computer science and it makes problem-solving easier and faster than other approaches [31].

1.4 Methodology

Predicting 3D protein structure from it's primary structure is the greatest open problem in the bioinformatics[1], solving this problem going from the contact map to the protein structure an efficient and fast algorithm is needed, many of methods introduced to reconstruct contacts prediction in several way[13, 34, 35,36, 37].

Traditionally, the contact map is a Boolean matrix that is created from distance map using a pre assigned threshold value t. Distance map D is a $N{\times}N$ matrix where N is the number of residues in a protein and $D[i,j]$ is the distance between coordinate of the α carbon in two residues i and j which measured in Angstroms *A*◌́ .Two residues i and j in a protein are come in contact with each other if the 3D distance $D[i,j]$ is less than or equal to some threshold value.

This thesis proposed method to predict a protein structure from contact map using MATLAB tools. The approach produces several procedures to finds a set of three dimensional coordinates consistent with contact map of threshold t.

Proposed method contains three modules. The SCANNER MODULE reads the protein ID from the list (extract from PDB). Then accepts the contact map CM of protein as an input, and produces a new contact map NCM by Scanning method. This process based on prediction quality more than quantity of contacts. The PRODUCER MODULE produces distance matrix procedure, which find a possible set of distance between nodes. Then compute a 3D point used nonlinear function from MATLAB tools. New contact map is extract based on new coordinates and compare with native contact map to find number of differences, the final module CORRECTOR MODULE generate a set of coordinates consistent with the given contact map. Then the module used MATLAB plot 3D function to map a protein 3D structure.

1.5 Related Work

For the past few years, several a approaches have been developed in order to help predicting a protein 3D structure to understand protein functionality ,these approaches used machine learning approaches such as neural network and support vector machine [11,12,13] and distance geometric [14,15,16].

This section gives a brief discussion about the advantages and disadvantage of these tools.

In 2008, Vassura et al. [14] produce a software tool for reconstructing a protein 3D structure form contact map. The tool based on distance geometry which, finds a set of three dimensional coordinates consistent with some given contact map of threshold T. The contact map of a given protein is a binary matrix CM such that *CM(i,j)* = 1 iff the Euclidean distance between residues *i* and *j* is less than or equal to a pre-assigned threshold *t*. The tools divide the system into two phases, the first phase; to generate a random initial set of 3D coordinates. This phase used metric matrix embedding algorithm to obtain good starting coordinates, before that they split the initial contact map in sub matrices. The sub matrices are then separately used to create sets of coordinates then merge it to give an initial solution. The merging procedure use rotation and translation to decrease the number of errors. While the second phase refines the set of coordinates by applying correction and perturbation procedure. The refinement applies until the set of coordinates is consistent with the given contact map. This phase applies iteratively two local techniques to obtain a new set of coordinates more consistent with the given CM in this step correction procedure doesn't add new errors to the coordinates set but eventually reduces the possibility to move some coordinate not yet well placed residue.

The tool shows that contact maps computed using threshold values greater than those commonly used for distances allow better 3D structure recovery than those computed at lower thresholds (7-9 Å). Repeated application of their method show that the contact map thresholds rang from 10 to 18 Angstrom allow to reconstruct 3D models that are very similar to the protein native structure. The disadvantage of this method apply used distance geometry which deals with the characterization of mathematical properties which is very complex and need daunting task to understand it.

In 2004, Zhang and Jing [12] used a Neural Network to predict protein contact map and find its 3D structure. The tool focus in grained contact map prediction. The approach concentrates on find a 3D structure from liner sequences of protein. The major task in this approach is to propose and verify precise and robust adaptation rule to predict contact map.

The approach taken was to extract data from PDB. Then choose the proteins have a single chain with number of amino acid less than 50, because of the difficulty in Neural Network to training with long chain of protein.

This tool used distance formula to compute distance matrix and normalize the distance matrix by convert all the distance into (0, 1). In addition, it used a set of threshold value to extract a pair node is in contact (i.e. $CM(i,j)$=1) . We can summarize the approach described in this tool by four different neural networks to get contact map as follows:

- Back propagation neural network
- Learning vector quantization neural network

- Radial basis function neural network
- Reinforcement network

The tool used 20 amino acids as inputs and output scheme. It proposed an easy input encoding scheme which used 5 bit to encode each amino acid and used fixed length of protein. The approaches keep global information to get better prediction. The disadvantages of this approach are time expensive and limitation on the length of protein sequences. The advantages of this approach it has higher resolution than just one contact map.

In 2002, Jing hu et al. [16] present techniques describe how data mining can be used to extract valuable information from contact map and focus on discover an extensive set of non local dense patterns and compile a library of such non local interaction, and cluster patterns based on their similarities and evaluate the quality.

This tool used contact map to discover 3D structure by test each two amino acid to determine 3D distance by coordinate of α carbon atom.

A pairs of amino acid in contact if distance less than threshold value =7 Å. The method used in this tool is divided into four stages:

- mining dense patterns
- pruning mined patterns
- clustering the dense patterns
- Integration of these patterns with biological data.

In the first stage they scan the DB of CM with 2D slide window. The tool used different window size to capture denser contact close to diagonal. The second stage extracted and isolated the pattern less dense and less distance from the diagonal by weighted the minimum density and verifying window size. Also this stage pruned redundant pattern by using slide window to capture all possible area in a matrix.

In clustering stage, the pattern generated into groups of similar interaction by used agglomerative clustering method. To find non local interaction it calculated a distance between each pair of pattern and between each pair of cluster, before they start clustering. This stage determined threshold for cluster. Then compare all pair of cluster and mark the closest. If the distance between two clusters is less than threshold t merged them into a single cluster. Finally, return to the first stage to continue the clustering. If the distance between the closest pair is greater than certain threshold, the clustering stops.

Their experiments used non redundant set of 2702 proteins from PDB, binary contact maps were generated using several contact thresholds. They discovered 9929 dense patterns in sliding window. The tool results showed that they can give 35% accuracy and 37% coverage for protein structure. The results are encouraging, but it's still far from providing sufficient accuracy for reliable 3D structure prediction.

In 1999, Jorge and zhijun, [30], developed a tool based on Gaussian smoothing to develop an efficient and reliable code to solve the distance geometry problem in protein structure. The algorithm in this tool work with the sparse set of distance constraints while other algorithm work for distance geometry which tend to work with dense set of constraints.

The problem in this approach is the distance geometry for determination of protein structures. The distance geometry is specified by a subset of all atom pairs. The distance between *i* and *j* atoms in a subset determine the lower and upper bounds to find a set of positions of the specified atoms. This problem is formulated in terms of finding the global minimum of the function.

The approach in this tool used Gaussian smoothing to transform function F into smoother function with fewer minimizes. The optimization algorithm applied to the transformed function and continuation techniques. The optimizations are used to trace the minimizers of the smooth function back to the original function. The advantage of this approach is work per iteration and proportional to a subset for sparse distance. The computational experiments show that the tool provides an efficient approach to the solution of the distance geometry problem and show an interesting issue is the dependence of the structures on the distance data.

1.6 Thesis Structure

This thesis consists of five chapters, during these chapters we tried to simplify the concepts and methods that are used for describing the problem and the proposed technique for the prediction process. The detailed description consists of:

- **Chapter one:** presents a biological introduction, which reviews bioinformatics concept and its biological problem, and presents historical background. This chapter presents in general problem definition, thesis methodology, then shows some related work and measures its accuracy and drawback, and describes thesis structure.
- **Chapter two:** presents experimental techniques for deriving macromolecular structure, and then focuses on proteins and proteins structures, also its show contact map concept and its role in expressing the 3D structure of protein.
- **Chapter three:** discusses the protein 3D structure problem, then presents the proposed technique to improve the prediction accuracy, and explains its modules and how it work with MATLAB tools.
- **Chapter four:** shows and discusses the experimental results. Four proteins are used to as examples and are tested using the proposed technique.
- **Chapter five:** draw the conclusion and the future possible work for the technique.

CHAPTER TWO

Experimental Approaches & Protein Structure

Reviewing of the literature on prediction a protein 3D structure will be as two parts experimental and non experimental methods.

Threading, homology and Ab_initio are non experimental methods come as a solution to predict 3D structure in efficient way which, developing an algorithms; these methods appear to overcome the limitation of the experimental methods. In the following sub sections, experimental and non experimental models are highlighted.

2.1 X-ray crystallography

X-ray crystallography crystallized the protein by electrons to create a diffraction pattern which determines the atomic structure of the protein, these process calculate the coordinate of atoms based on the measured electron density.

X-ray present accurate coordinate of atoms, however it is being laborious, low resolution (2.9 A◌̊ derived structure) and there are some type of proteins are difficult to crystallize by X-ray [28].

In order that significant time and effort are required to solve and complete a macromolecular crystal structure, these problems in X-ray created demand for computerized methods to improve the rate and resolution at which new structures are determine. Automation in macromolecular X-ray crystallography has been a goal for many researchers. Along standing area of computation within structural biology are the algorithms for de convoluting the X-ray diffraction pattern.

2.2 NMR Spectroscopy

The scientific goals of researchers are the prediction of structure and function of from sequences and simulations of the functions of a living cell. NMR spectroscopy is located to play an important role in this field, because of its ability to provide atomic resolution structural and information about proteins.

NMR contributes about 5% of protein data bank, NMR also a key tool in mechanistic enzomology and in studies of protein folding and stability [4].

In NMR the molecules are exposed to static magnetic field causing the nuclei of atoms to vibrate, and then the molecules are subjected to a second oscillating magnetic field, generating a characterizing spectrum for all the atoms for each molecule which becomes a spatial atomic map (3D structure).

NMR like X-ray remains slow and do not scale up to current sequencing speeds, but on the other hand NMR experiments provide complementary data to the crystallographic analysis. The challenge of interpreting NMR derived distance constraints into 3D structures, further introduced computational technologies to biological structures.

The raw data from X-ray and NMR are most often a set of Cartesian coordinates representing the position of the atoms in these structures, these experimental methods give a set of atomics proximately which need methods to embed these distance measures into 3D structure that satisfy these constrains, distance geometry and other nonlinear optimization methods have been developed for this purpose.

2.3 Major three approaches to predict a protein 3D structure

The predicting of protein structure from its sequences with more accuracy is the base goal of protein modeling. Protein modeling is the only way to obtain structural information if experimental techniques fail, some types of protein (membrane proteins) are difficult to crystallize by X-ray and simply too large for NMR analysis.

2.3.1 Homology modeling

Homology modeling is considered to be a reliable mode, homology modeling called comparative modeling since it compare between the sequence *(A)* which unknown structure with all sequence *(B)* of known structure stored in the PDB, if sequence *(B)* contain region that match sequence *(A)* with 50% identical residues then the structure of target sequence *(A)* will be similar to the fragment of the structure to the aligned region in the homologous sequence *(B)* [28]. This mean if there no homologous sequence to the target sequence the model will fail, in addition to if there are homologous sequence with less than 50% identical residues the structure will be denied.

Homology modeling is easy and reliable approach but it is restricted to the known structure proteins and it is neglects the identity of the protein because homology modeling predict 3D structure not a unique since it depends on the homologous sequences.

2.3.2 Threading

Homology modeling assumes strong similarity between the target structure sequences and knows structure sequence, so that threading model or fold recognition appear to overcome this quandary and predict a sequences with less than 50% identical residues to the known structure in PDB.

The basic idea in threading is a particular fold is assumed for the target sequence then evaluating the feasibility and favorably using some energetic and physical consideration assesses the quality and the acceptance of this folded [8].

Threading suffer from low quality compared with the homology modeling, however the two methods are restricted to known structure homology protein, in order that Ab_initio appears as a new approach to predict protein structure with out any dependency in the known structures.

2.3.3 Ab_initio

Ab_initio is a term used to define methods to predict the native conformation of protein from the amino acid sequence using only a computational model without extrinsic comparison to existing data. Ab_initio it may be sometimes interchangeable with the Latin term de novo [32].

De novo play a good role to extract rules that govern the transformation process. An important practical challenge in this is of large scale genome sequence project which are producing large numbers of protein sequence for which no 3D structural information is available.

Ab_initio results are an unreliable prediction but it may play fundamental role when the overall folding problem are solved.

2.4 Protein

Most of essential structure and functions of cells is refereed to Proteins. Proteins play a vital role in keeping the body working properly. For example, they are used to support the skeleton, control sense, move muscles, digest food, and defend against infections and process emotions.

There are more than 100,000 proteins that come in all shapes and sizes; however, they are all made up of the same set of 20 amino acids order in different way, its primary sequence. The structure of a protein is determined by the folding of this primary sequence [18].

Any consideration of protein function must be grounded in an understanding of protein structure. A fundamental principle in all of all protein science is that protein structure leads to protein function, and protein functions are divers, so it's no surprise that protein structures are divers' also [16].

For this who wishes to study protein structure this diversity represents a challenge. In 1958, the first three dimensional protein structure (the oxygen storage protein myoglobin) determined by John Kendrow and his co-workers[7], subsequent studies of the myoglobin structure revealed that the protein did have some regularities these regularities were also observed in other protein structures.

2.5 Protein structure

Protein structure has been organized into four levels which facilitates description and understanding of proteins: primary, secondary, tertiary, and quaternary structure [19]. This hierarchy makes protein structure studies more tractable.

2.5.1 Primary structure

Proteins are liner polymers composed of 20 simpler building blocks, called amino acid, which function as the molecular machines of living organism; proteins can contain any combination and number of the 20 amino acids in any order.

The concept of protein as liner amino acid polymers was inutility proposed by Fischer and Hofmeiter in 1902, [20]. Amino acids are small molecules that contain an amino group (NH2), a carboxyl group (COOH), and hydrogen atom attached to central alpha (α) carbon, see figure 2.1. Also amino acid have a side chain R group attached to the (α) carbon, R group distinguishes one amino acid from anther.

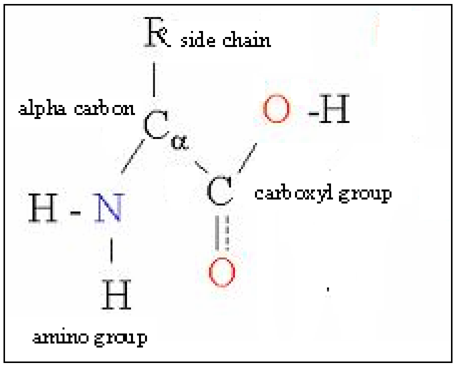

Figure 2.1 Basic amino acid structures[4]

The side chain gives the specific chemical properties of amino acid. In 1940 the exact set of amino acids used in protein was determined [20], this set can be grouped into three classes depend on the chemical properties by their side chain: hydrophobic, polar, and charged. Table 2.1 show lists of the amino acids, three letters and one letter abbreviation code, and their class. Amino acid form bonds with each other through reaction of their carboxyl and amino acid groups, called the peptide bond, see figure 2.2.

The specific characteristics of the peptide bond have important implications for three dimensional structures which formed by polypeptide bond, so any protein sequence folds into a particular 3D structure, and no more than one protein sequence folds into the same 3D structure [4].

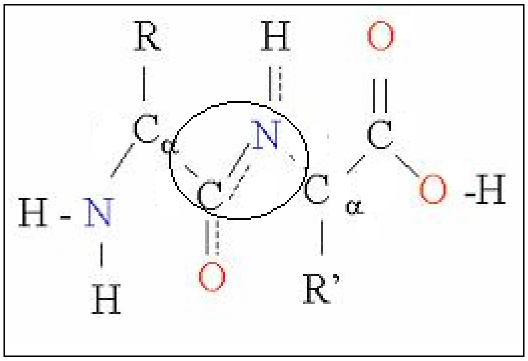

Figure 2.2 peptide bond between amino acid[4]

Table 2.1 the 20 amino acids and their abbreviation codes and their classes [4]

No	Name	3-letter	1-letter	Class
1.	Alanine	Ala	A	Hydrophobic
2.	Cysteine	Cys	C	Polar
3.	Aspartate	Asp	D	Charged
4.	Glutamate	Glu	E	Polar
5.	Phenylalanine	Phe	F	Hydrophobic
6.	Glycine	Gly	G	Hydrophobic
7.	Histidine	His	H	Charged + Polar
8.	Isoleucine	Ile	I	Hydrophobic
9.	Lysine	Lys	K	Charged
10.	Leucine	Leu	L	Hydrophobic
11.	Methionine	Met	M	Hydrophobic
12.	Asparaine	Asn	N	Polar
13.	Proline	Pro	P	Hydrophobic
14	Glutamine	Gln	Q	Polar
15	Arganine	Arg	R	Charged
16	Serine	Ser	S	Polar
17	Theronine	The	T	Polar
18	Valine	Val	V	Hydrophobic
19	Trypotphan	Trp	W	Hydrophobic+ Polar
20	Tryrosine	Tyr	Y	Polar

2.5.2 Secondary structure

The secondary structure of a protein consists of regular conformation of the polypeptide chain, this structure occurs when the sequence of amino acids are linked by hydrogen bond, [21] see figure2.3.

There are two types of secondary structure: **Alpha helix and Beta sheets,** these two types are the basis for structure and function prediction.

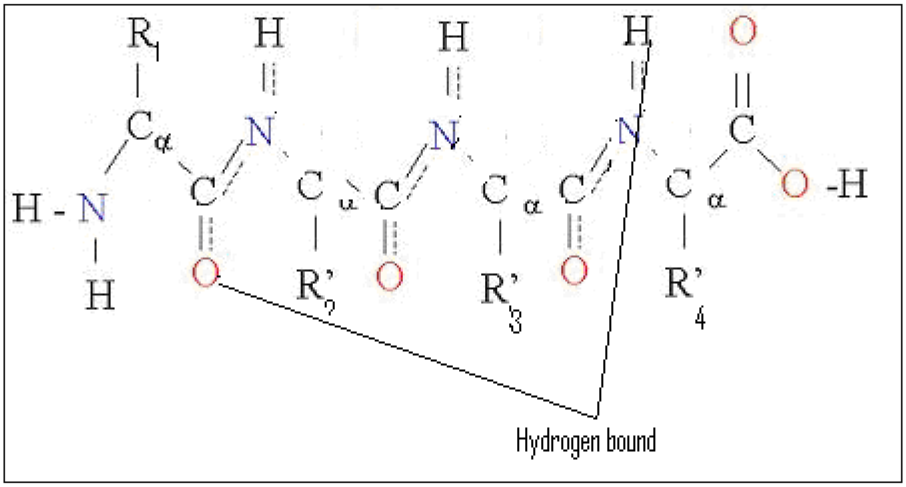

Figure 2.3 hydrogen bond in α helix[4]

1) α helix :

Helix is created by carving of the polypeptide backbone such that a regular coil shape is produced. The structure of this helix resulted from hydrogen bonding interaction between the carbonyl oxygen (CO) of each amino acid and the amino group (NH) of amino acid that is four position carboxyl terminuses to it along the helix [22].

2) β sheets :

β sheets are formed by hydrogen bond between a adjacent polypeptide chains, polypeptide chain in the sheet called β strands.

There are two types of β sheets:

- **Parallel:** β sheet is parallel if the sheets arrange in the same direction with respect to their amino terminal *N* and carboxyl- terminal *C* ends.
- **Anti Parallel:** in this type the sheets alternate their amino and carboxyl terminal end, such that a given sheets interacts with sheets in the opposite orientation.

Also there is a section of polypeptide chain that connects the secondary structures and has irregular structures called loop or coil, [22]. Most of these structures were predicted and observed in the 1960s and 1970s [4].

2.5.3 Tertiary structure

The tertiary structure of a protein is defined as the global three-dimensional structure of its polypeptide chain, which describes the spatial relationship of different secondary structures within polypeptide chain and how these structures fold into the 3D form of a protein.

In 1936, Alfred Mirsky and Linus Pauling described numerous important features of protein tertiary structure" our conception of a native protein molecule …is the following: the molecule consists of one polypeptide chain which continues without interruption throughout the molecule……..; this chain is folded into a uniquely defined configuration, in which it is held by hydrogen bonds between the peptide nitrogen and oxygen atoms and also between the free amino and carboxyl groups of the diamino and dicarboxyl amino acid residues" [23].

There are various of helix, sheet, and loop elements can combine in variety ways to produce a complete 3D structure, see figure 2.4, these combination interaction are fetched through interaction between the side chains of the amino acid residues of the protein, the side chain play active role in creating the final tertiary structure.

Years of experimentation made it possible to understand how secondary structure element combines in 3D space to yield the tertiary structure of protein [4].

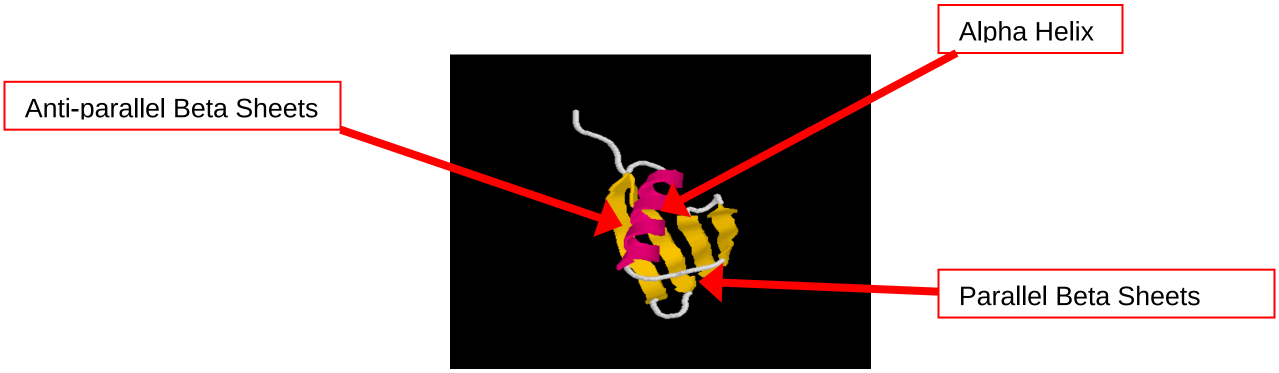

Figure 2.4: 3D structure of 2IGD protein (from PDB)

As more and more protein structures have been determined, Cyrus Chothia and in 1976, Michael Levitt derived classification grouped proteins based on their secondary structure element into four groups: **all α, all β, α/ β (mixture), and α + β** (parallel β connected by α helix) [6].

2.5.4 Quaternary structure

The tertiary structure of a protein describes the structure of a single polypeptide chain, but many proteins contain more than one polypeptide chains. These proteins have a quaternary structure.

A quaternary structure is formed by assembling these polypeptide chains into one super structure.

In 1926 Svedberg observed the first quaternary structure, but the quaternary structure concept was not importance until 1960s when the experiments on enzyme regulation showed that protein subunits were essential to understanding higher levels of cellular function [4]

2.6 Contact Map

Contact map is a great interest for its application in fold recognition and 3D structure determination. A contact map is representation tool of the protein 3D structure.

Traditionally, the contact map is created from the distance map where a distance matrix computed to produce the Boolean values by used a pre assigned threshold value *t*. Distance map *D* is a $N{\times}N$ matrix where *N* is the number of residues in a protein and *D[i,j]* is the distance between coordinate of the α carbon in two residues *i* and *j* which measured in Angstroms $A\dot{o}$.Two residues i and j in a protein are come in contact with each other if the 3D distance *D[i,j]* is less than or equal to some threshold value.

Contact map *C* for a protein sequence with *N* residues is $N{\times}N$ asymmetric Boolean matrix whose element *C(i,j)*=1 if residues *i* and *j* are contact and *C(i,j)*=0 otherwise.

The contact map provides useful information, contacts represent certain secondary structure and it captures non local interaction giving clues to its tertiary structure [16]. When the contact map cluster to contacts area you can see in figure 2.5 α helices appear in the contact map as a band along the main diagonal and β sheets are thick bands parallel or anti_parallel to the main diagonal, figure 2.6 shows the 3D structure of the same protein [16].

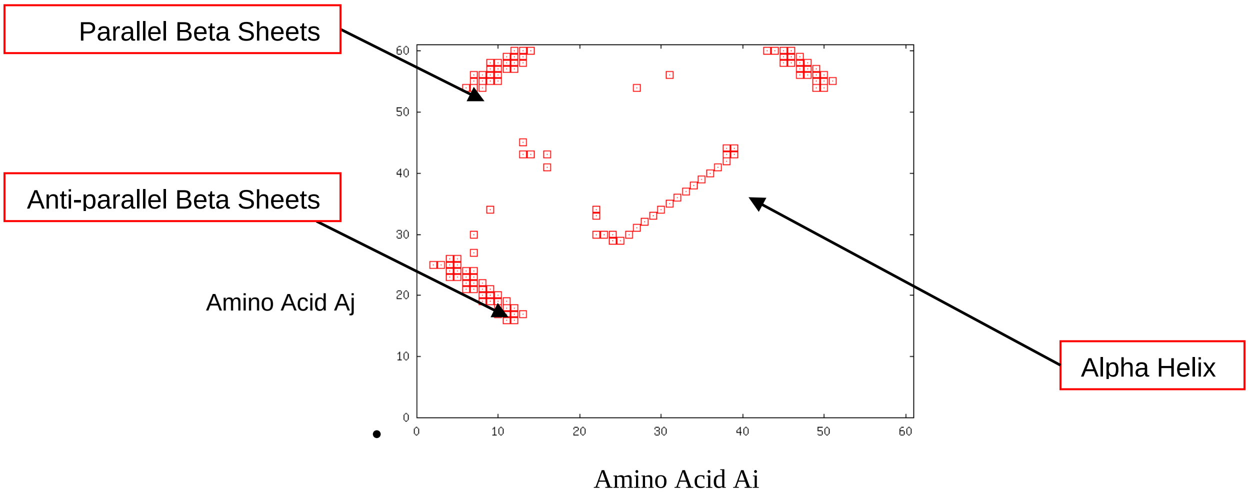

Figure 2.5 : contact map of 2IGD protein [16]

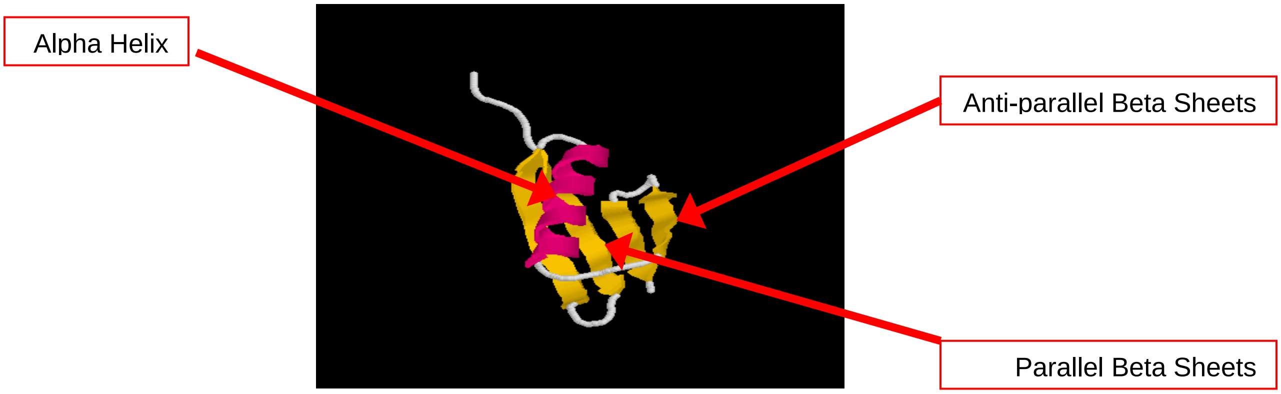

Figure 2.6: 3D structure of 2IGD protein(from PDB)

Determining an ideal value to use it as a threshold for certain contact map consider a challenge, various researcher are using different threshold such as 7 , 8 , 9 and even 5 Å and they show that this threshold are suitable for offering contact between contact map and it 3D structure [16,11]. In other research [14, 18] they observed that the suggested value of threshold which less than 10 A◌̊ decrease the number of contact until converts the whole map into thick lines, and in the otherwise increasing value converts map into contact state and they proved by their recant extensive experiment results that the contact map threshold ranging from 10 to 18 A◌̊ allow to reconstruct 3D models that are similar to the protein native structure.

2.6 Database Sourcing

The protein data base includes different types of information a associated with a protein such as atomic coordinates , primary and secondary structure there are three popular data base : National Center for Bio Technology (NCBT), European Bioinformatics Institute (EBI) and Genome Net, Japan.

The protein Data Bank (PDB) was established at Brookhaven National Laboratory (BNL) in 1971 as an archive for biological molecular crystal structure [3].

In the beginning the archive held seven structures. In 1980s the number of structures increased due to the improvement in technology such as X-ray and NMR methods, in the beginning of 2002 there were more than 17,000 entries in the PDB files. Recently there are more than 30,000 PDB files exit in this ftp server [12].

Table 2.2 PDB Format ftp://ftp.rcsb.org/pub/pdb/data/biounit/coordinates/all,

```
•    ATOM    1  N    MET  B   1    18.343  41.550  -5.088  1.00  66.05       N
•    ATOM    2  CA   MET  B   1    17.015  41.170  -4.537  1.00  66.19       C
•    ATOM    3  C    MET  B   1    16.762  39.683  -4.438  1.00  53.39       C
 •    ATOM    4  O    MET  B   1    16.422  39.005  -5.404  1.00  45.93       O
•    ATOM    5  CB   MET  B   1    15.782  41.966  -5.065  1.00  72.30       C
•    ATOM    6  CG   MET  B   1    14.956  42.662  -3.969  1.00  74.44       C
•    ATOM    7  SD   MET  B   1    15.892  43.609  -2.752  1.00  78.85       S
•    ATOM    8  CE   MET  B   1    15.758  45.288  -3.400  1.00  10.00       C
•    ATOM    9  N    ASN  B   2    16.941  39.216  -3.219  1.00  50.29       N
       |       |  |    |    |   |      |       |         |      |     |          |
       1       2  3    4    5   6      7       8         9     10    11         12
```

1. Record name "ATOM”
2. Integer serial “Atom serial number”.
3. Atom name “name of amino acid”.
4. Residue name “name of residue”.
5. Character chain “ID Chain identifier”.
6. Integer Residue Sequence “Residue sequence number in protein”.
7. Real x Orthogonal “coordinates for X in Angstroms”.
8. Real y Orthogonal “coordinates for Y in Angstroms”.
9. Real z Orthogonal “coordinates for Z in Angstroms”.
10. Real occupancy “Occupancy”.
11. Real temperature Factor “Temperature factor”.
12. String element “Element symbol, right-justified”.

The most useful data in PDB file is the orthogonal coordinates x, y, z, from which we will build contact map by compute the distance between every pair node in a protein and compare with threshold value to predict 3D structure of this protein.

CHAPTER THREE

Proposed Method

Predicting the 3D structure of protein from linear sequence of amino acids is an interesting topic for computer scientists. Each protein may contain thousands of atoms in different shapes, a fact which makes it helpful to automatically predict a protein through software tools. These tools for replacing a tradition experiment technique. This problem becomes even more complicated when the developer uses a complicated protein. Contact map help developers by giving them information about the protein system.

Reconstructing a protein from its contact map using MATLAB is a proposed method to assist in enhancing constituents and predicts a protein 3D structure.

This section shows the pseudocode of reconstruction algorithm which takes a contact map (CM) of a chosen protein with a pre assigned threshold value.

```
Reconstruction_Algorithm (CM, T)
 NCM = scan_CM (CM) % scan of CM based on number of neighbors
 D=distance_matrix(NCM,T)%compute the distance between two atoms
 D=shortdist(D)%compute the shortest distance between two atoms

 C=nonlin_coordinat(Dist)% compute the coordinates by MATLAB tool
 NCM = new_contact_map(CM,C,T)% extract new contact map
 e =compar_contact_maps(NCM,CM);

      Q=positive number
      While differences between NCM and native CM is not
              Acceptable and Q not equal zero
        NC=coorect_coordinat(CM,C,T)%correct coordinates of some
                                     not well placed residues
        NCM = new_contact_map(CM,NC,T)
        e =compar_contact_maps(NCM,CM)
        Q=Q-1;
      End
End
  map_3D_structuer (NC)%map 3D structure of this protein by MATLAB
tool
END
```

Figure 3.1: Reconstruction Algorithm

The proposed reconstructing method is divided into three modules namely: **SCANNER MODULE, PRODUCER MODULE, and CORRECTER MODULE** as shown in figure 3.2.

The figure show methodology of the reconstruction method which start with scanning module, scanning module takes CM of proteins from the list which extract from PDB, then by scan function generate new contact map with new contact point after that producer module takes NCM as input and generate arbitrary distance fit to the NCM depending on some literature survey, Shortest path function is used to obtain the best set of distance must be satisfy the triangle inequality, then FSOLVE function from MATLAB tool used to give the best set of three dimensional coordinates fit for D, take random set as starting point then applies until set of coordinate is acceptable. Then compare extract CM with the native.

Correct module takes the set of coordinate as input to find the possible radius mobility of some not yet well placed residues and move these residues to new position with new coordinates used mobility and correct direction function, this process iteratively applies until control parameter *Q* becomes zero (*Q* is number of tray to correct coordinates) or until **ε** percentage of error becomes acceptable.

In the final, plot 3D function from MATLAB take the new set of coordinates and map 3D structure of protein.

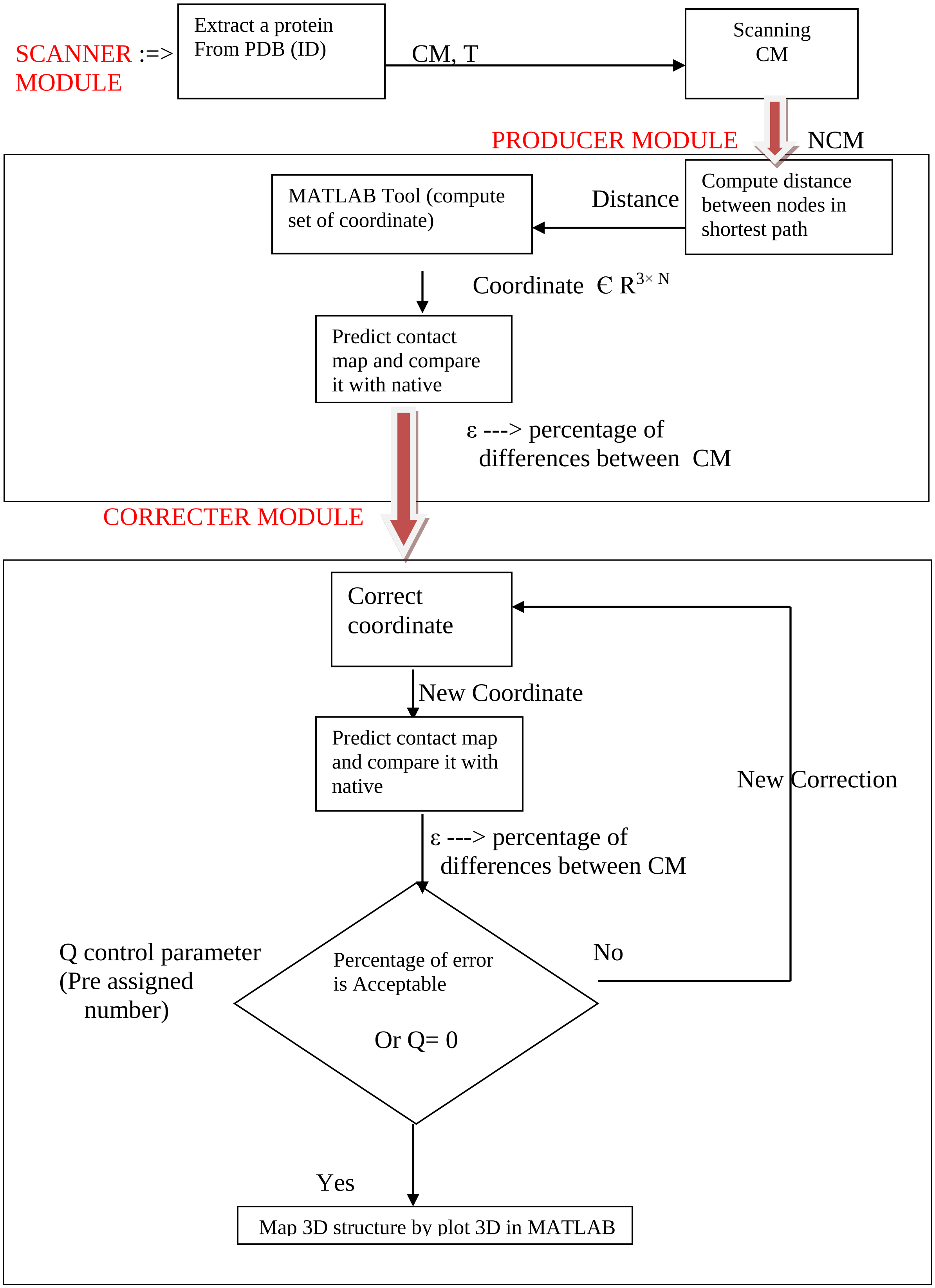
SCANNER :=>
MODULE
Extract a protein From PDB (ID)
CM, T
Scanning CM
PRODUCER MODULE
NCM
MATLAB Tool (compute set of coordinate)
Distance
Compute distance between nodes in shortest path
Coordinate Є R$^{3\times N}$
Predict contact map and compare it with native
ε ---> percentage of differences between CM
CORRECTER MODULE
Correct coordinate
New Coordinate
Predict contact map and compare it with native
New Correction
ε ---> percentage of differences between CM
Q control parameter (Pre assigned number)
Percentage of error is Acceptable
Or Q= 0
No
Yes
Map 3D structure by plot 3D in MATLAB

Figure 3.2: Data flow of the approach

3.1.1 SCANNER MODULE

The Scanner module reads a protein from PDB, and constructs a protein contact map table as input to this method, then produces a New Contact Map NCM.

Scanning the contact map for a protein is much more reliable to predict the more important areas of the contact map which we call it dense area in NCM. This process based on prediction quality more than quantity of contacts. In all previous studies shows that predict 50% of the contact map with 5% errors much reliable than predicting 100% of contact map with 25% errors [2].

The main objective of this thesis is to detect the dense areas that form the basic functional areas in the contact map. Looking for the dense area is an important step that will improve the performance of the predicting 3D structure of protein from it CM.

Scanning module used **SCAN_CM Procedure** to reprocess all contact residues. This module assumes that two atoms *i* and *j* are in contact if and only if they share a high number of neighbors, i.e. *C (i, j)* =1 are in contact and share with less than 10 neighbors that are closest to a specific point or *C (i, j)* =0 are consider in contact if they share with greater than 20 neighbors that are closest to a specific point. Find a number of neighbors will increase the probability of selected contact residues. In other cases; this approach will decrease the probability of wrongly predicted contact pairs outside the dense area. The details of scanning CM pesedocode is shown in figure 3.3**.**

```
Scan_CM(CM)
%take native CM of n node as input
for i=1 to n

   for j=i+1 to n
       count=0
           for k=1 to n
             if i and k contact && k and j contact
              count=count+1
               if i and j is contact && count< 10
                    or i and j is not contact && count> 20

                then node i and j is contact in NCM
               Break and take anther node
Return NCM
```

Figure 3.3 Scanning procedure

3.1.2 PRODUCER MODULE

The Producer Module takes a new contact map (NCM) to produces a possible set of distance between nodes $D Є R^{N\times N}$ depending on threshold value range from 7 to 18 A˚ consistent. In addition by using some literature survey about the physical conformation of the proteins this module can know the average distance between adjacent alpha carbons *D [i,j]* which is 3.84 A˚(i.e. | $i - j$ |=1). Also, the other distance of contact node can be obtained from classified protein by **count_distance procedure** in the otherwise the distance of nodes which are not contact set as random number depending on threshold value. **Shortest_path_dist procedure** is used to obtain the best set of distance which satisfy the triangle inequality, **(i.e. for all *i* , *j* , *k* node from 1 to n, distance [*i* , *j*] <= distance [*i* , *k*] + distance [*j* , *k*]).**

```
distance_matrix(NCM, Threshold)

  for i=1 to n
    for j=i to n
        if i and j contact in NCM
           Distance between i and j= count_distance(T,i,j)
          else
     Distance between i and j = T* random(1,1)+T
Return D matrix
End
```

Figure 3.4: Distance matrix procedure

```
count_distance(Thresholde,i,j)
%the set of distances in this procedure are taken from literature
survey
if i equal j
   x=0
if |i-j| equal 1
   x=3.8
if |i-j| equal 2
   x=6+ random(1,1)
if |i-j| equal 3
   x=7 +random(1,1)
if  |i-j| greater than 3
   x=(0.91-(T/100))*T
return X
End
```

Figure 3.5: Count Distance procedure

To compute a 3D point the Producer Module used a consistent distance matrix D with supported by **nonlinear_coordinat procedure**. This procedure used **FSOLVE** function from MATLAB tools, FSOLVE finds a root (zero) of a system of nonlinear equations, FSOLVE calls compute distance function which accepts random set of coordinates (vector x) as starting points also a distance matrix D as parameter to solve nonlinear system using Euclidian distance equation between every pairs of amino acid protein sequence are selected to solve the nonlinear system. This process applies iteratively until the best set of three dimensional coordinates fit for distance matrix D.

The procedure accepts the results if the root (zero) of the system is found, otherwise a new random initial set of coordinate is generated and the procedure restart from beginning by using MATLAB tool.

```
nonlin_coordinat(D)
C is coordinate matrix
For a=1:10
x0=random set of coordinate take as starting point
C = FSOLVE(@(x) compute_coordinat(C,D),x0)
If set of coordinates accept
Breake and  Return C

End
```

Figure 3.6: Nonlinear coordinate procedure

```
compute_coordinat(x,D)

for i=1 to n
  for j=i to n


F=(x((i-1)*3+1)-x((j-1)*3+1))^2+(x((i-1)*3+2)-x((j-
1)*3+2))^2+(x((i-1)*3+3)-x((j-1)*3+3))^2-Dist(i,j)^2;

  End
```

Figure 3.7: Compute coordinate procedure

The Producer Module used **new_contact_map** and **compare_contact maps** procedures to the current set of coordinate to extract new contact map and compare two contact map (native CM with predict CM) to find error percentage.

```
New_contact_map(CM,Coordinate,Threshold)

for i=1 to n
   for j=i to n
         if i equal j
               D(i,j)=zero and  NCM(i,j)contact

          Else
       D(i,j)= sqrt ((C(1,i)-C(1,j))^2+(C(2,i)-C(2,j))^2+(C(3,i)-
C(3,j))^2);


               if D(i,j) less than or equal Threshold
                    NCM(i,j)is contact
               else
                    NCM(i,j)is not contact
               End
  End
```

Figure 3.8: New contact map procedure

```
compar_contact_maps(NCM,CM)
  e=0
  for i=1 to n
       for j=1 to n
           if NCM(i,j)not equal CM(i,j)
               e=e+1
Return error=e/n*n
End
```

Figure 3.9: Compare contact maps procedure

3.1.3 CORRECTOR MODULE

Corrector Module takes the producer module output and applies **Correct_coordinate** procedure to find the possible radius mobility of some not yet well placed residues and move these residues to new position with new coordinates, we used the same correction in [26], but without rotation and translation.

Mobility procedure takes maximum distance between residue *i* and *j* if the two nodes are contact and minimum distance otherwise, to calculate radius of mobility of residue *i*.

$D0 = \min\{d(i,j) \mid d(i,j) > t \text{ and } CM[i, j] = 0\}$
$D1 = \max\{ d(i,j) \mid d(i,j) \leq t \text{ and } CM[i, j] = 1\}$.

Then the **Mobility procedure** takes minimum distance between *D0* and *D1* (i.e. $M(i) = \min \{D0 - t, t - D1\}$).

To determine the direction of move of residue *i* without do any effect of correct residues **Correct_direction** procedure is used.

Then the module extracts new contact map depending on new correction and compare two Contact maps. In order that, this process iteratively applies until control parameter *Q* becomes zero (*Q* is number of tray to correct coordinates) or until **ε** percentage of error becomes acceptable. If the consistent set of coordinates is found the module used **plot 3D** function from MATLAB .The plot3 function displays a three-dimensional plot of a set of data points.plot3(X_1, Y_1, Z_1), where X_1, Y_1 , and Z_1 are vectors or matrices, plots one or more lines in three-dimensional space through the points whose coordinates are the elements of X_1, Y_1 , and Z_1.

```
coorect_coordinat (CM,C,T)
  For i=1 to n
   For j=1 to n
      If CM (i, j)contact and D(i, j) greater than Threshold
         OR
      CM(i, j)not contact and D(i, j)less than or equal Threshold
         R = mobility (i)
         New Coordinate (i) =correct_direction (i)

END
```

Figure 3.10: Correct coordinate procedure

```
Mobility (i)

 For j=1 to n
  If CM (i, j) contact and D (i, j) less than or equal Threshold
                      D1=max (D1, D (i, j))
      Else
  If (CM (i, j) not contact and D (i, j)greater than Threshold
           D0=min(D0, D (i, j))

   M (i) =min (D0-T, T-D1)

END
```

Figure 3.11: Mobility procedure

```
correct_direction(i)
   for j=1 to n
         if CM(i,j)contact and D(i,j)greater than Threshold
          OR
        CM(i,j)not contact and D(i,j)less than or equal Threshold

     if CM(i,j)contact
           V= V – C(i)-C(j)/D(i,j))
                else
           V=V + C(i)-C(j)/D(i,j))


      k =C(i)+ ( V *(r/norm(V)))
```

Figure 3.12: Correct direction procedure

```
map_3D_structuer(New Coordinate)

PLOT3(New Coordinate) %PLOT3 function from MATLAB TOOL
End
```

Figure 3.13: Map 3D structure procedure

This approach focuses on choosing the threshold value for computing the contact map, which is affect on connects between the contact map and its 3D structure, not any threshold give accurate contact map which provides exact 3D structure. The experimental results show that the contact maps computed using threshold values(12-18) Å allow better 3D structure recovery than those computed at thresholds (7-9) Å.

CHAPTER FOUR

Experimental Results

This chapter presents experimental results that show the efficiency of our proposed method for predicting a protein structure. We took the list of proteins of different lengths related to the most popular classes from the PDB.

For each protein in the selected list we generate different contact maps by changing the threshold value and analyze the result when we scan the contact map with a pre assigned threshold to show the accuracy of extracting contact map from dense area instead of whole area, and show the effect of threshold on the 3D structure of a protein. This chapter shows some experimental results for different proteins. In addition, the results have been analyzed and compared with the original proteins

4.1 Experimental Result 1: 2IGD protein

2IGD a protein has a single chain (A) and contains one α_helix and two parallel β_sheet. Figure 4.1 shows the ۳D structure of the protein and its plot 3D .

The following table 4.1 shows that the PDB ID of a protein is 2IGD. 2IGD has 61 residues and classify as Alpha Beta protein in CATH Classification.

Table 4.1: 2IGD Protein properties(from PDB)

PDB ID	**2IGD**
Length	**61**
Type	**Polypeptide (L)**
Chain	**A**
CATH_Classification	**Alpha Beta Protein**
Amino Acid Sequences	**MTPAVTTYKLVINGKTLKGETTTKAVDATAEKAFKQYANDNGVWTYDDATFTVTE**
Experimental Method	**X-ray diffraction, 1.10 resolution**
Polymer	**1**
Molecule	**Protein G**

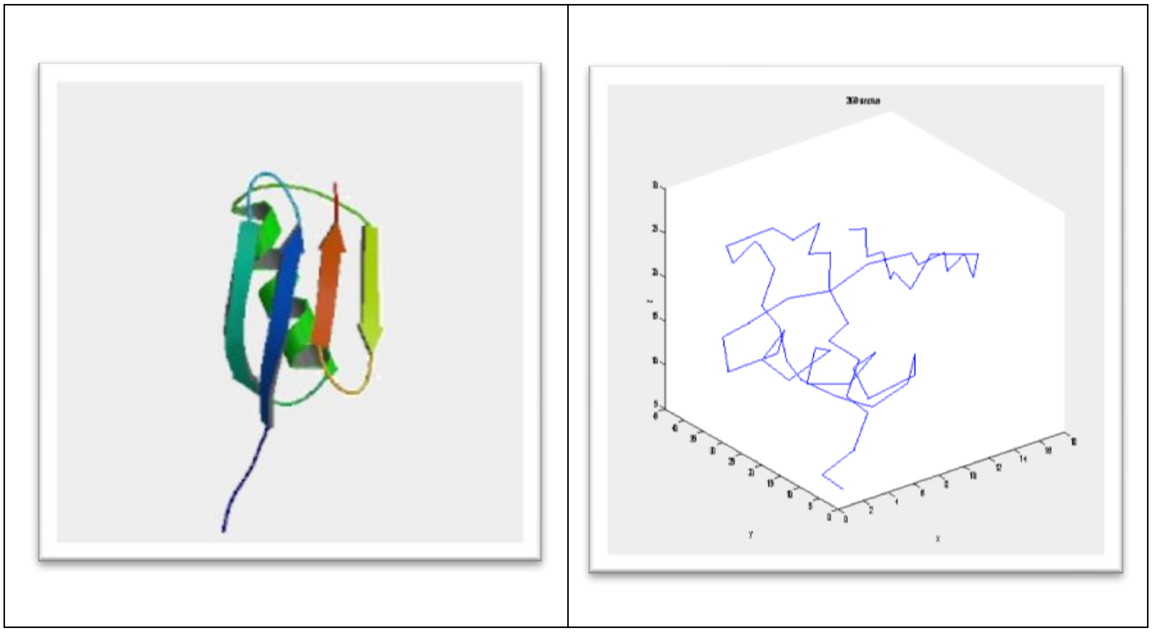

Figure 4.1: (left) 2IGD 3D structure (from PDB)**, (right) plot of 3D structure**

We ran the experiment for each protein and generate 12 different contact maps by changing the contact threshold from 7 to 18 Angstrom as shown in Table 4.2. Furthermore, the table shows the percentage of error before and after correction with average time for 2IGD protein. The analysis of the result shows that the correction procedure reduces the percentage of error when it applies iteratively. The correction procedure continues until the best set of coordinates is found comparing with the native contact map. The method help to predict the 3D structure more accurately.

Table 4.2 Recovery of 3D structure from contact map

Threshold Value	Percentage of error	Percentage of error after correct ion	Average Time In seconds
7	0.13	0.11	85
8	0.12	0.11	110
9	0.12	0.10	128
10	0.12	0.10	100
11	0.13	0.11	80
12	0.14	0.12	91
13	0.13	0.12	116
14	0.14	0.11	86
15	0.12	0.11	120
16	0.12	0.10	98
17	0.12	0.09	192
18	0.13	0.10	80

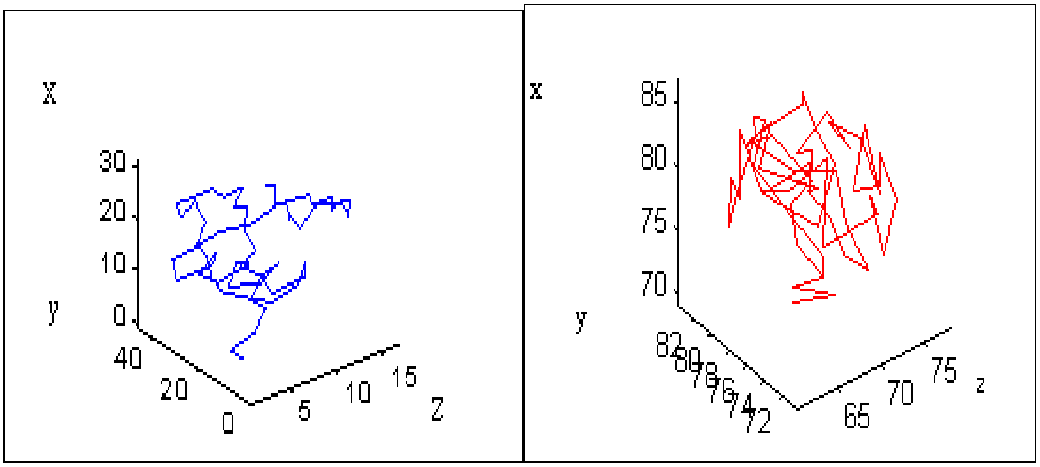

Figure 4.2: (left) 2IGD 3D structure (native), (right) 2IGD prediction of 3D structure at threshold 7

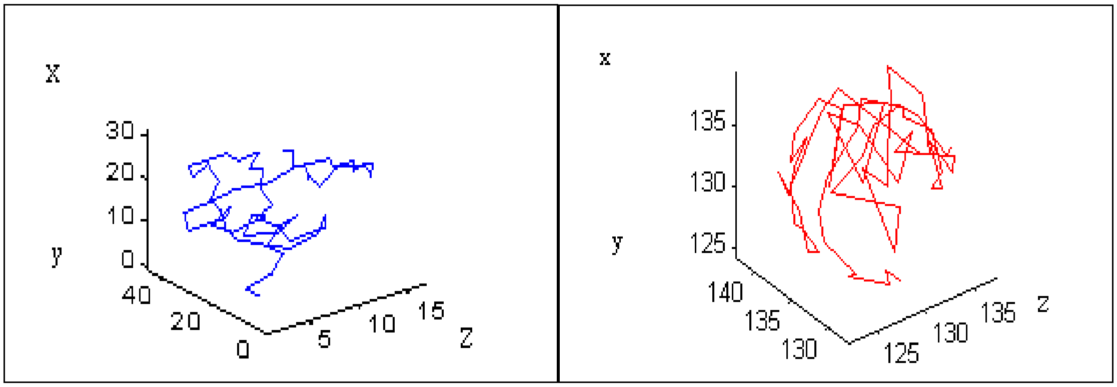

Figure 4.3: (left) 2IGD 3D structure (native), (right) 2IGD prediction of 3D structure at threshold 9

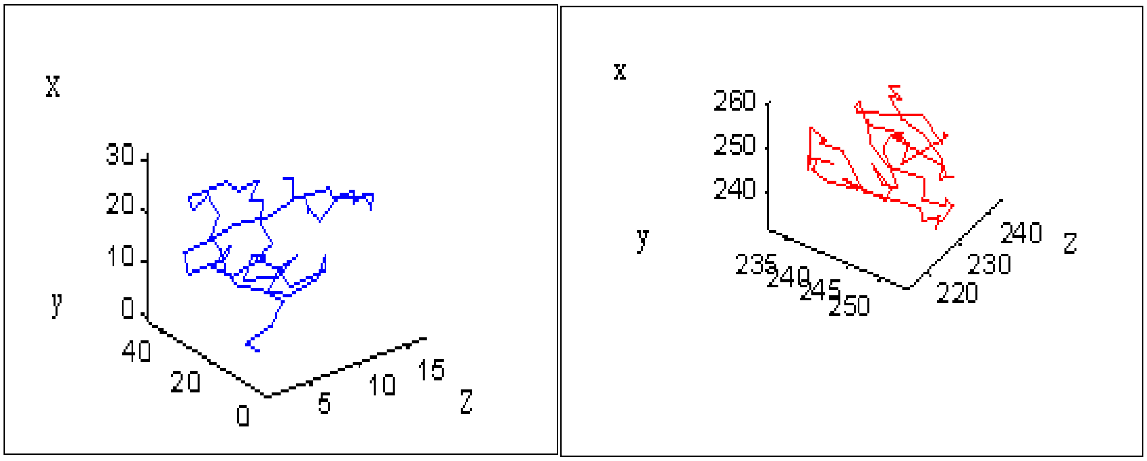

Figure 4.4: (left) 2IGD 3D structure (native), (right) 2IGD prediction of 3D structure at threshold 16

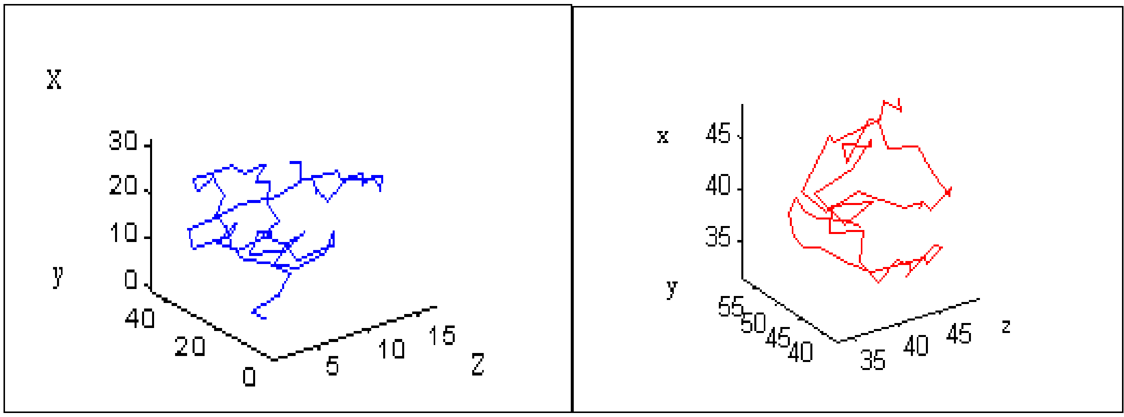

Figure 4.5: (left) 2IGD 3D structure (native), (right) 2IGD prediction of 3D structure at threshold 17

The contact map computed with a threshold equal to 7-9 Angstrom does not contain enough global information of the protein structure to differentiate the protein from others. So that the prediction structure is not clear compared with the native structure as shown in figure 4.2 and 4.3. When the contact map is computed at a threshold of 16 and 17 Angstrom as shown figure 4.4 and 4.5, more features appear and the recovered 3D structure is more similar to the native one and more accurate. This finding encouraged us to do a search in the different threshold values to improve the percentage of error. We found that a better 3D reconstruction is obtained when a high threshold value is adopted (12 -18) when contact maps are computed.

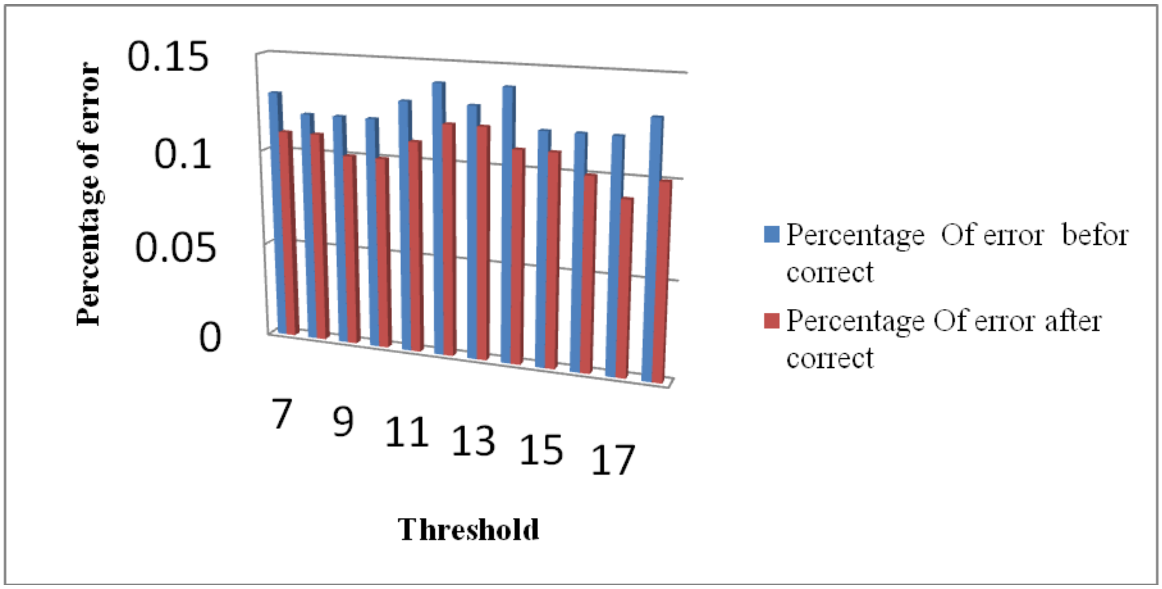

Figure 4.6: Threshold & percentage of errors for 2IGD protein

Figure 4.6 shows chart to illustrate and compare the percentage of error after and before the correction at different thresholds. Correction procedure improve the coordinates to obtain the best set consists with native contact map in this approach. The percentage of error is the difference between the predicting contact map and the native contact map.

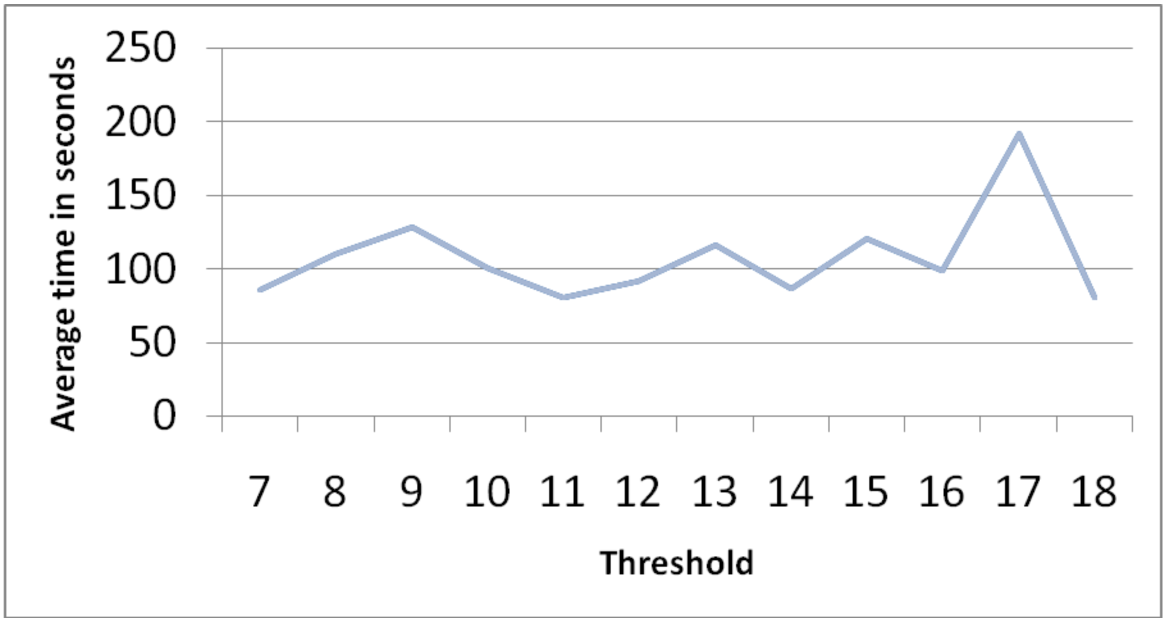

Figure 4.7: Threshold & ِِِAverage Time for 2IGD

Figure 4.7 shows that the running time of predicting protein structure is not affected by threshold value, The average time decreases and increases in arbitrary way because the algorithm take random number as starting points to **FSOLVE** function to solve the nonlinear system.

4.2 Experimental Result 2: 6PTI protein

6PTI is a protein has a single chain and contains two small α_helices and one anti_ parallel β_sheet. 6PTI belong to few secondary structures in CATH Classification and has 58 amino acids in chain A, as shown in the table 4.3.

Figure 4.8 shows the helices and sheet in 6PTI 3D structure and plot of 3D structure.

Table 4.3: 6PTI Protein properties(from PDB)

PDB ID	**6PTI**
Length	**58**
Type	**Polypeptide (L)**
Chain	**A**
CATH_Classification	**Few secondary structure**
Amino Acid Sequences	**APCLGPPTTGPCLAAIIATPTAALA GLCGTPVTGGCAALAAAPLSAGACMATCGGA**
Experimental Method	**X-ray diffraction, 1.70 resolution**
Polymer	**1**
Molecule	**PANCREATIC TRYPSIN INHIBITOR PRECURSOR**

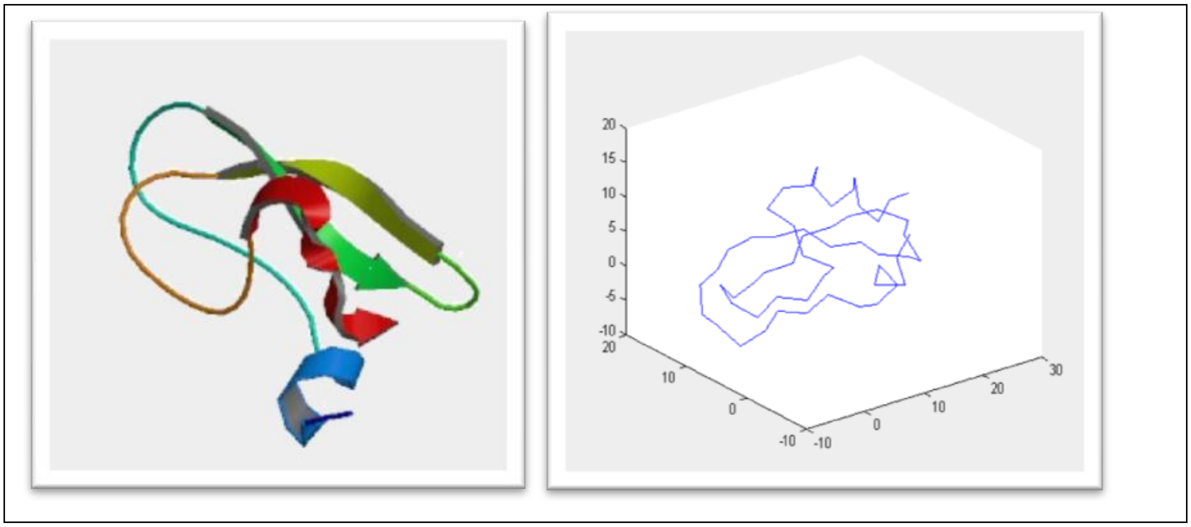

Figure 4.8: (left) 6PTI 3D structure (from PDB)**, (right) plot of 3D structure**

Table 4.4 shows the 12 threshold value from 7 to 18 Angstrom used to generate different contact maps and show the execution time. The analysis of the result shows that the correction procedure applies iteratively to decrease the percentage of error as shown in the table. The method help to predict the 3D structure more accurately.

Table 4.4 Recovery of 3D structure from contact map

Threshold Value	Percentage of error	Percentage of error after correction	Average Time In seconds
7	0.14	0.12	66
8	0.13	0.11	85
9	0.12	0.11	144
10	0.13	0.11	154
11	0.12	0.10	155
12	0.12	0.10	146
13	0.10	0.09	140
14	0.12	0.10	150
15	0.12	0.10	160
16	0.13	0.11	150
17	0.12	0.10	120
18	0.12	0.11	100

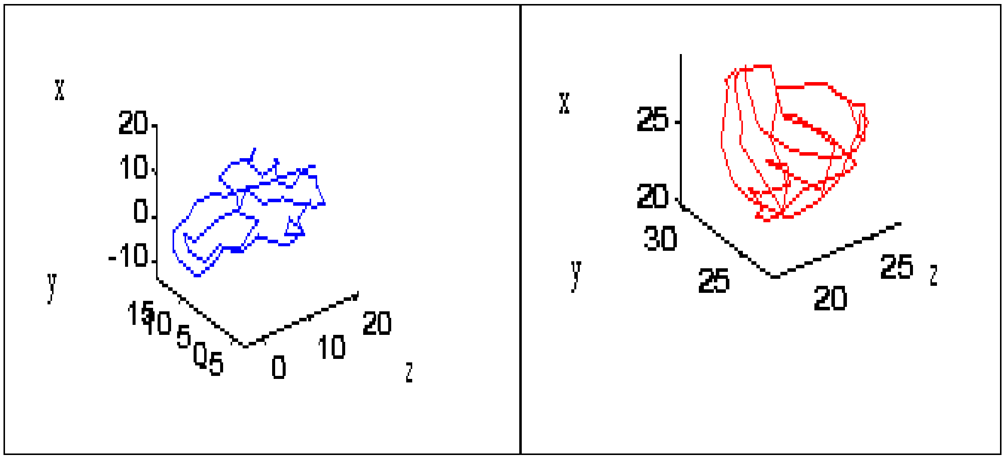

Figure 4.9: (left) 6PTI 3D structure (native), (right) 6PTI prediction of 3D structure at threshold 7

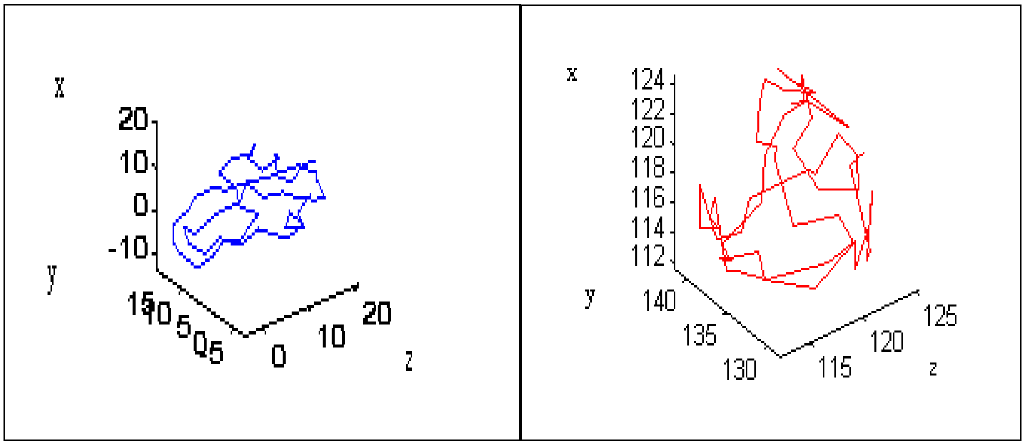

Figure 4.10: (left) 6PTI 3D structure (native), (right) 6PTI prediction of 3D structure at threshold 8

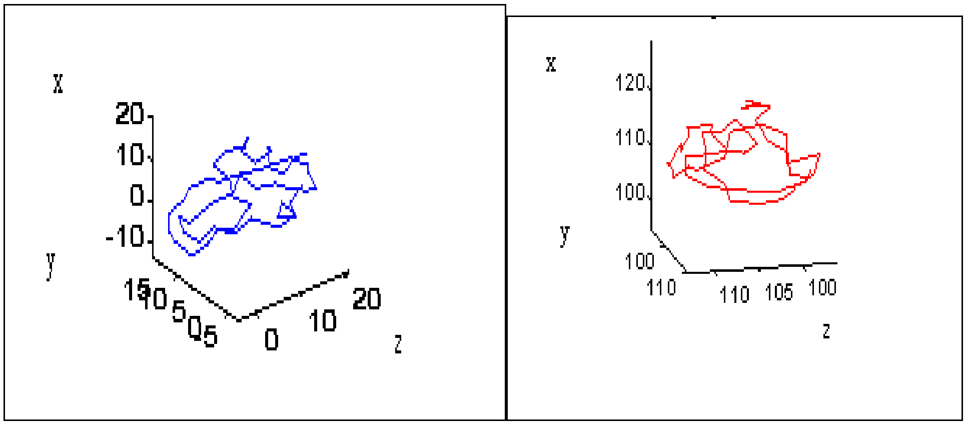

Figure 4.11: (left) 6PTI 3D structure (native), (right) 6PTI prediction of 3D structure at threshold 12

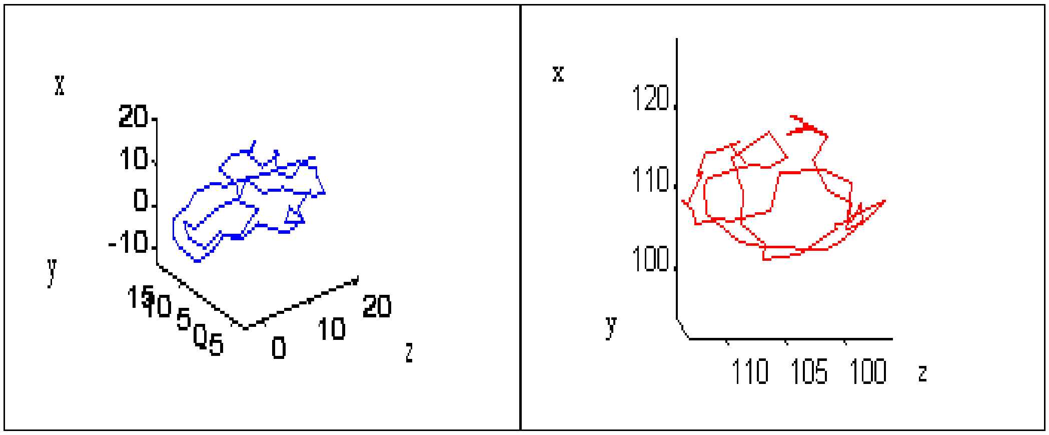

Figure 4.12: (left) 6PTI 3D structure (native), (right) 6PTI prediction of 3D structure at threshold 16

As shown in the previous experiment the 3D structure which predict from thresholds equal to 7-9 Angstrom is not clear compared with the native structure as shown in figure 4.9 and 4.10. While the figure 4.11 and 4.12 shows the predicting 3D structure of 6PTI at a threshold 12 and 16, which are more similar to the native one.

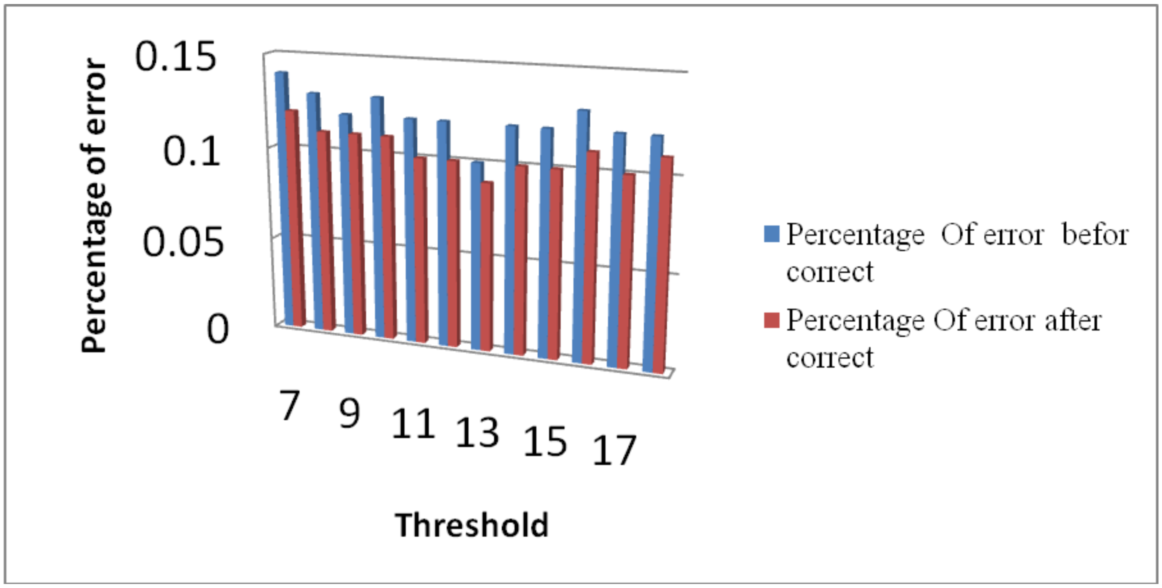

Figure 4.13: Threshold & percentage of errors for 6PTI protein

In this approach the correction procedure try to find the possible radius mobility of some not yet well placed residues and move these residues to new position with new coordinates, The experimental result shows that the contact maps computed after correction are better than those computed before correction and the percentage of error is decreased in different threshold value.

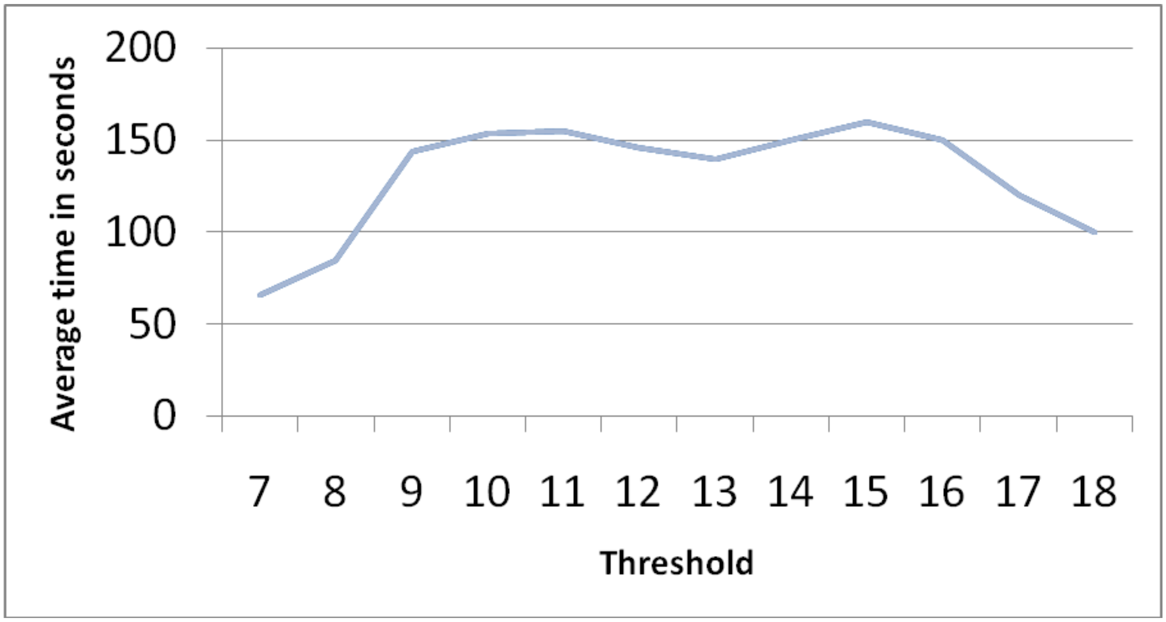

Figure 4.14: Threshold & average Time for 6PTI

When we ran the proposed algorithm we observed that the running time of predicting protein structure is not affected by threshold value. The analysis result shows that the algorithm take random number as starting points to FSOLVE function, FSOLVE try to find a root (zero) of a system of nonlinear equations. This process applies iteratively until the best set of three dimensional coordinates fit for distance matrix D. So that the execution time vary in arbitrary way as shown in figure 4.14.

4.3 Experimental Result 3: 451C Protein

451C is a protein has a single chain and contains several α_helices distributed in different region. Figure 4.15 shows the helices and plot 3D structure of 451C.

451C is the ID of a protein in the PDB, the following table 4.5 shows some properties for 451C.

Table 4.5: 451C Protein properties(from PDB)

PDB ID	**451C**
Length	**82**
Type	**Polypeptide (L)**
Chain	**A**
CATH_Classification	**Mainly Alpha**
Amino Acid Sequences	**GAPGVLPLALGCVACHAIATLMVGPATLAVAALPAGGAGA** **GAGLAGAILAGSGGVTGPIPMPPAAVSAAGAGTLALTVLSGL**
Experimental Method	**X-ray diffraction, 1.60 resolution**
Polymer	**1**
Molecule	**CYTOCHROME C551**

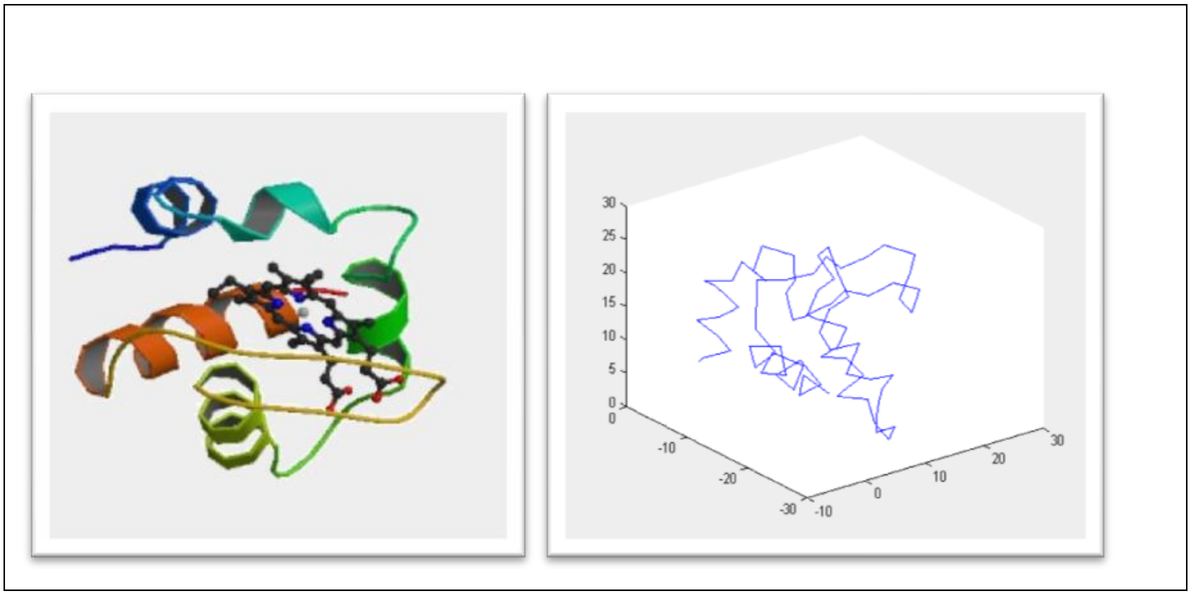

Figure 4.15: (left) 451C 3D structure (from PDB)**, (right) plot of 3D structure**

Table 4.6 Recovery of 3D structure from contact map

Threshold value	Percentage of error	Percentage of error after correction	Average Time In Mint
7	0.11	0.07	13.30
8	0.10	0.08	17.40
9	0.11	0.09	21.14
10	0.12	0.11	20.50
11	0.13	0.11	23.16
12	0.11	0.09	28.43
13	0.12	0.10	25.30
14	0.12	0.10	22.15
15	0.11	0.08	27
16	0.09	0.08	24.10
17	0.09	0.09	20.55
18	0.10	0.08	18.25

As shown in the previous experiment that the correction procedure improves the coordinates to predict 3D structure more accurately. Table 4.6 shows the percentage of error before and after the correction with average time of 451C protein.

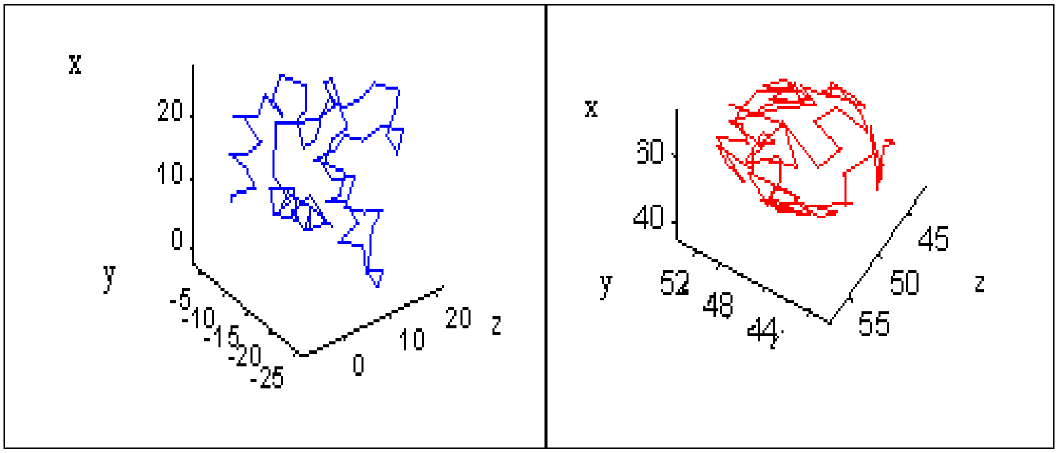

Figure 4.16: (left) 451C 3D structure (native), (right) 451C prediction of 3D structure at threshold 7

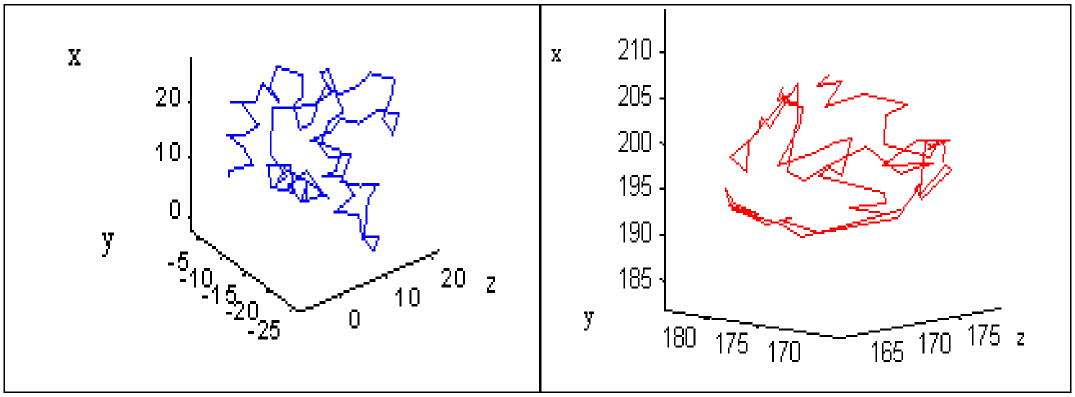

Figure 4.17: (left) 451C 3D structure (native), (right) 451C prediction of 3D structure at threshold 8

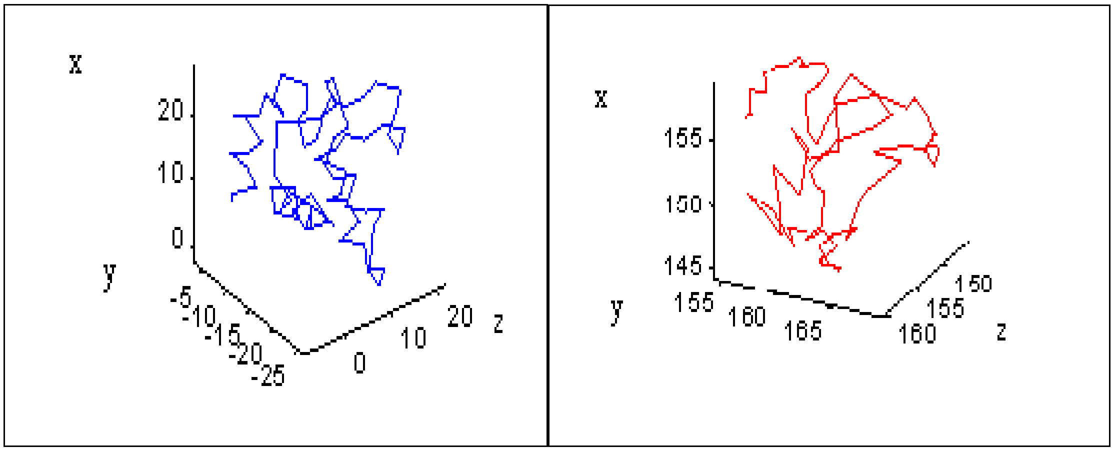

Figure 4.18: (left) 451C 3D structure (native), (right) 451C prediction of 3D structure at threshold 16

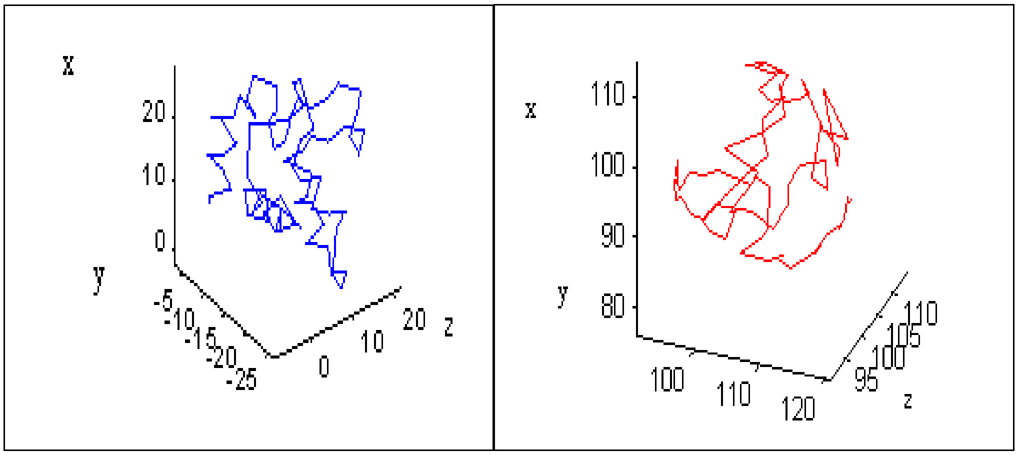

Figure 4.19: (left) 451C 3D structure (native), (right) 451C prediction of 3D structure at threshold 18

Determining an ideal value to use it as a threshold for certain contact map is considered a challenge. The experiment result show that the contact map computed with a threshold less than 10A◌̇ decrease the number of contact until it converts the whole map into thick lines so that the prediction structure is not clear enough as shown in figure 4.16 and 4.17, otherwise increasing the threshold value converts map into contact state. We find that the contact maps computed using threshold values (12-18) Å allow better 3D structure recovery than those computed at thresholds (7-9) Å, figure 4.18 and 4.19 show 3D structure of 451C at threshold 16 and 18 Å.

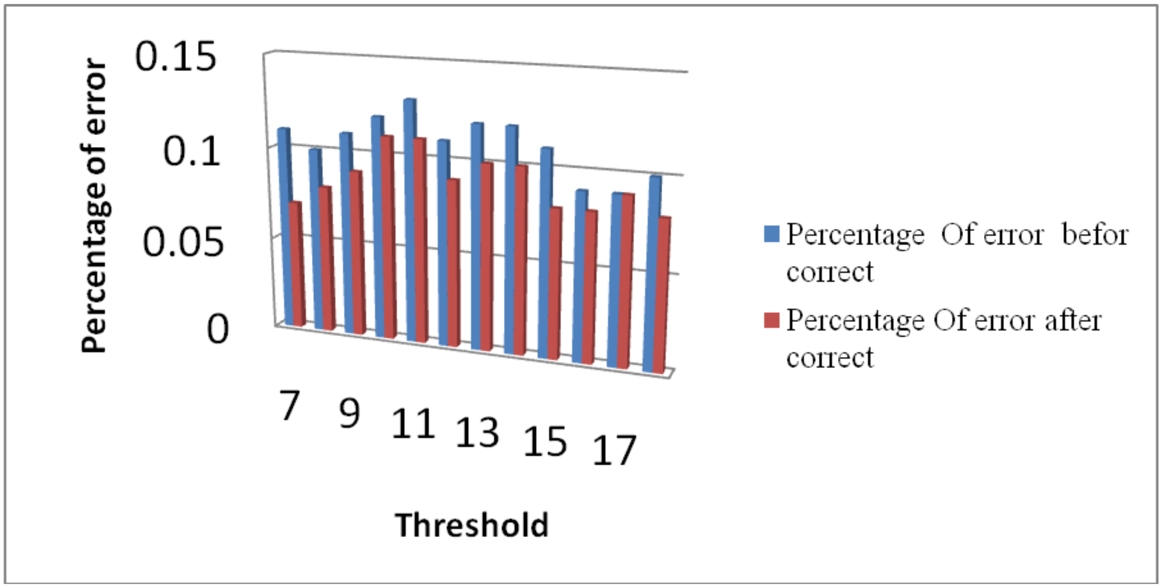

Figure 4.20: Threshold & percentage of errors for 451C protein

The chart in figure 4.20 shows the improvement of the correction procedure on the percentage of error in various thresholds for 451C protein.

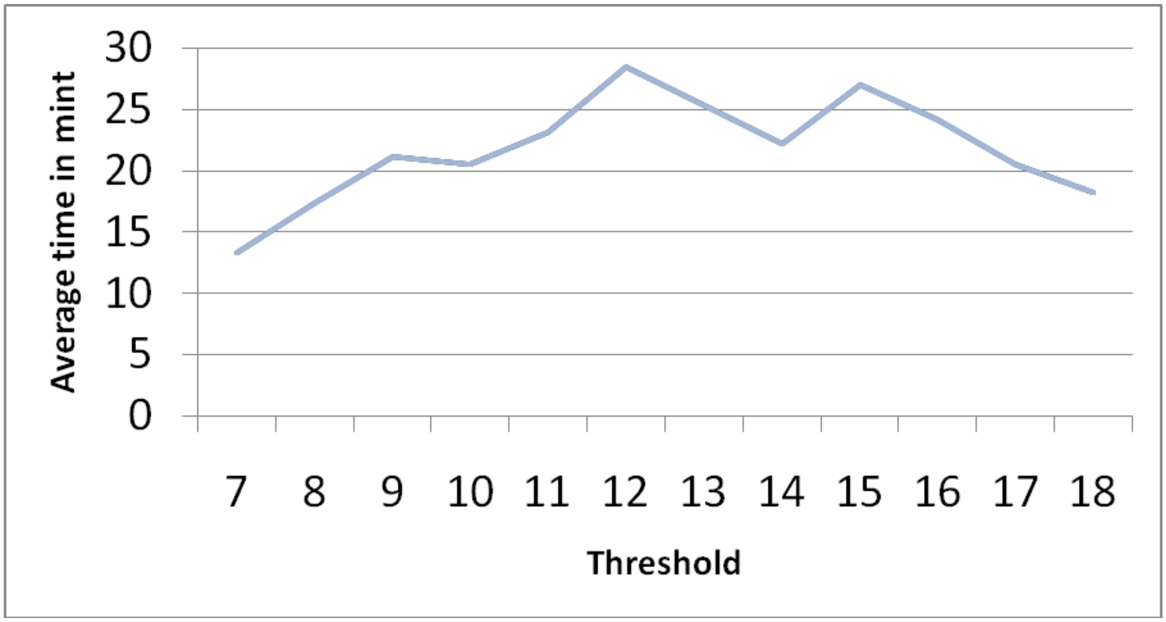

Figure 4.21: Threshold & Average Time for 451C

As shown in the previous experiment the running time of predicting protein structure is not affected by threshold value, figure 4.21 shows the average time at each threshold from 7 to 18 for 451C protein.

4.4 Experimental Result 4: 2CPG Protein

2CPG is a protein has three chains (A,B,C), figure 4.22 shows the 3D structure which contains several α_helices distributed in different region and two β_sheet.

The following table 4.7 shows some properties of 2CPG. A protein belong to Mainly Alpha protein in CATH Classification

Table 4.7: 2CPG Protein properties(from PDB)

PDB ID	**2CPG**
Length	**43**
Type	**Polypeptide (L)**
Chains	**A,B,C**
CATH_Classification	**Mainly Alpha**
Amino Acid Sequences	**MLLALTITLSGSVLGALGLMAAGMGLSLSAMISVALGATLLGG**
Experimental Method	**X-ray diffraction, 1.60 resolution**
Polymer	**1**
Molecule	**TRANSCRIPTIONAL REPRESSOR COPG**

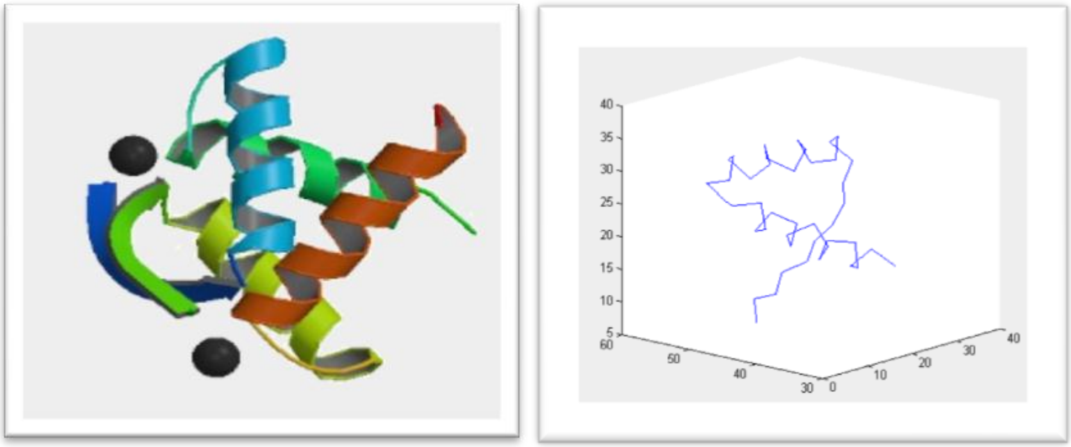

Figure 4.22: (left) 2CPG 3D structure (from PDB)**, (right) plot of 3D structure**

Table 4.8 Recovery of 3D structure from contact map

Threshold Value	**Percentage of error**	**Percentage of error after correct**	**Average Time Seconds**
7	**0.10**	**0.07**	**55**
8	**0.08**	**0.08**	**54**
9	**0.09**	**0.07**	**51**
10	**0.09**	**0.07**	**40**
11	**0.10**	**0.07**	**34**
12	**0.09**	**0.05**	**35**
13	**0.09**	**0.06**	**33**
14	**0.09**	**0.07**	**20**
15	**0.09**	**0.06**	**20**
16	**0.07**	**0.06**	**18**
17	**0.07**	**0.06**	**20**
18	**0.08**	**0.06**	**17**

The result of 12 different contacts map generated by changing the contact threshold from 7 to 18 Angstrom shows in Table 4.8, these contact maps computed to obtain the best 3D structure in this approach. Furthermore, the table shows the percentage of error which improved by correction procedure with average time.

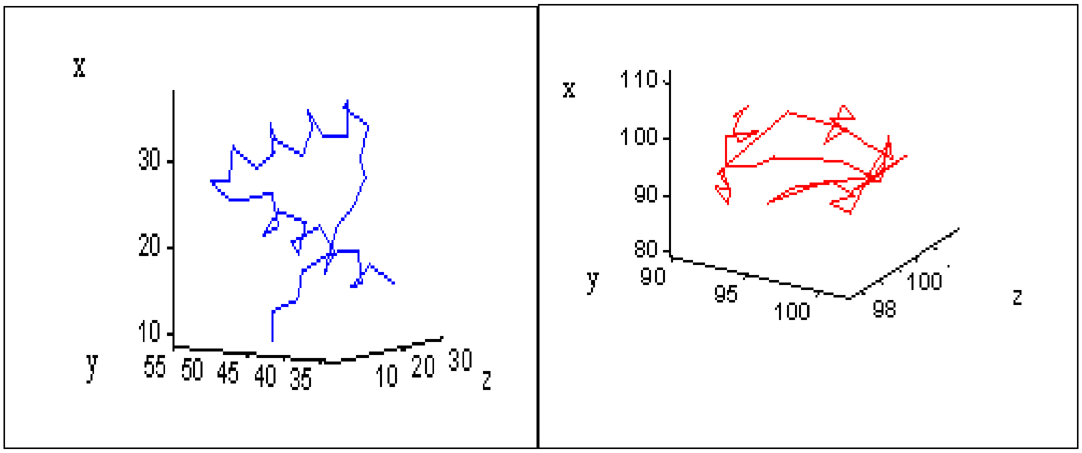

Figure 4.23: (left) 2CPG 3D structure (native), (right) 2CPG prediction of 3D structure at threshold 7

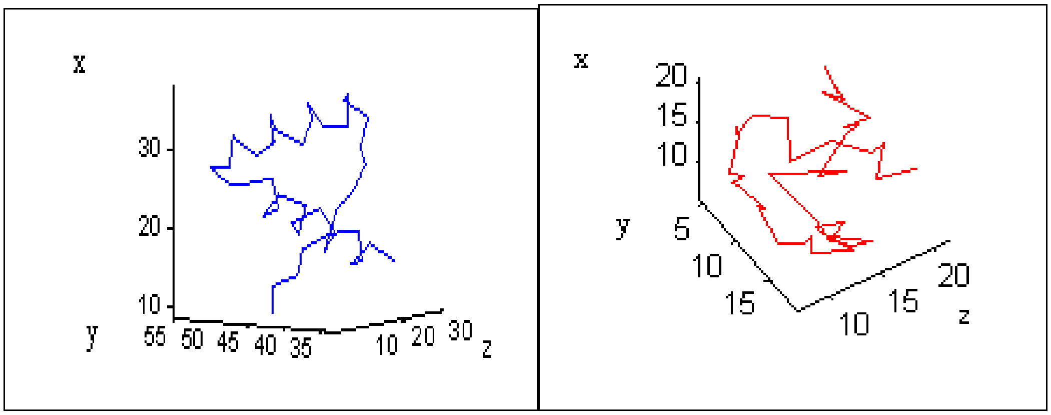

Figure 4.24: (left) 2CPG 3D structure (native), (right) 2CPG prediction of 3D structure at threshold 9

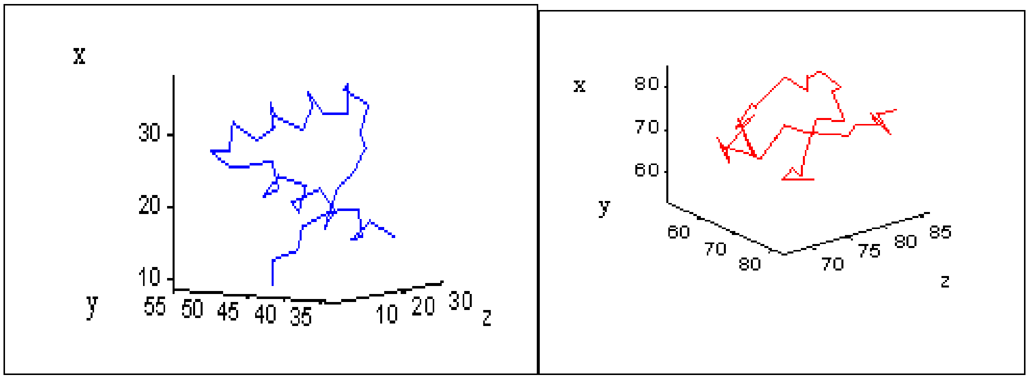

Figure 4.25: (left) 2CPG 3D structure (native), (right) 2CPG prediction of 3D structure at threshold 12

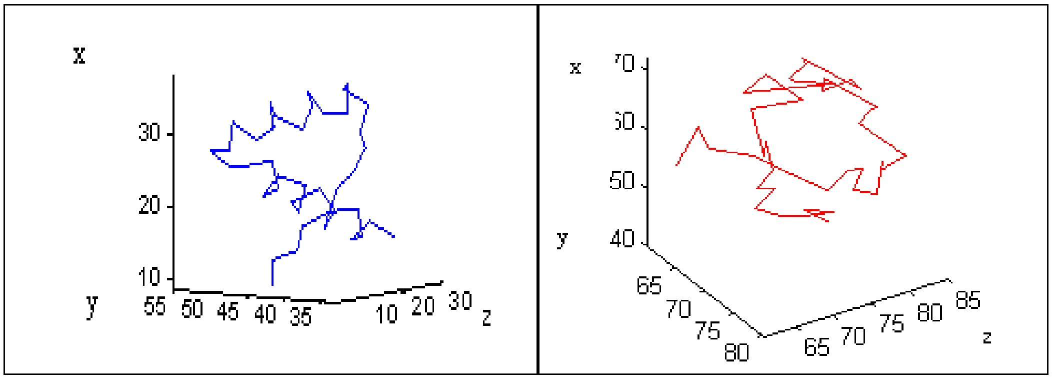

Figure 4.26: (left) 2CPG 3D structure (native), (right) 2CPG prediction of 3D structure at threshold 16

Figure 4.23 and 4.24 show the prediction structure of 2CPG protein at thresholds 7 and 9, through the experiment we found that the dense area of contact map computed with a threshold equal to 7-9 Angstrom does not contain enough contact node. So that the prediction structure is not clear compare with the structure predict at a threshold of 12-18 Angstrom, figure 4.25 and 4.26 show the 3D structure at 12 and 16 threshold.

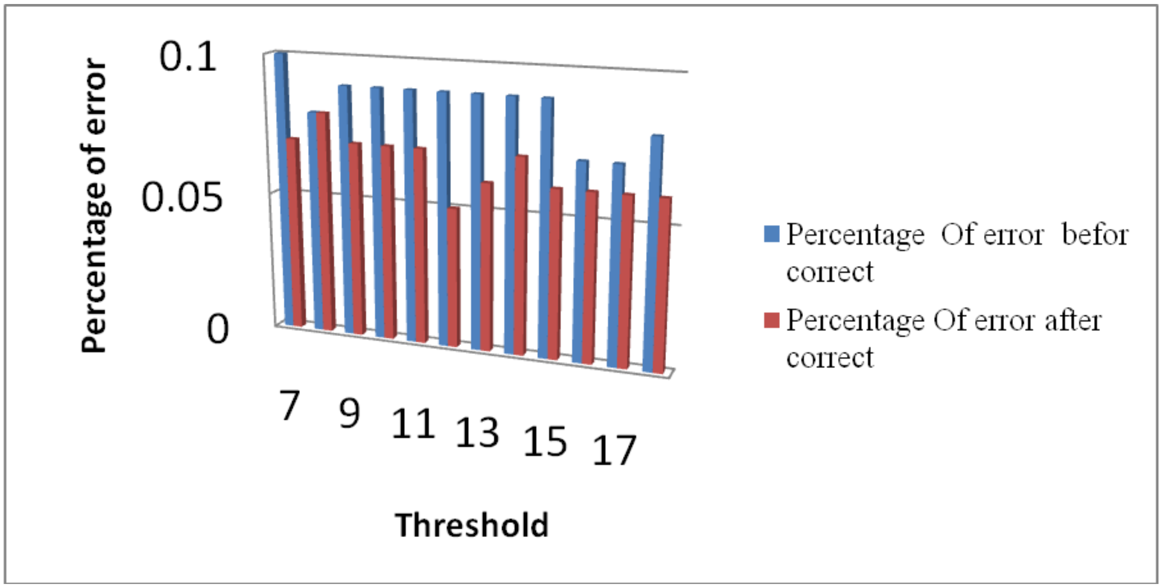

Figure 4.27 : Threshold & percentage of errors 2CPG protein

Figure 4.26 shows the percentage of error before and after correction when we ran the proposed algorithm on 2CPG protein at different thresholds.

The original result predict from the algorithm show that the set of coordinates extracted from FSOLVE function plot 3D structure is similar to the native structure, the correction procedure improve the result to became more accurately. As a future work if rotation and translation are applied to the coordinates we predict that the accuracy will be improved dramatically.

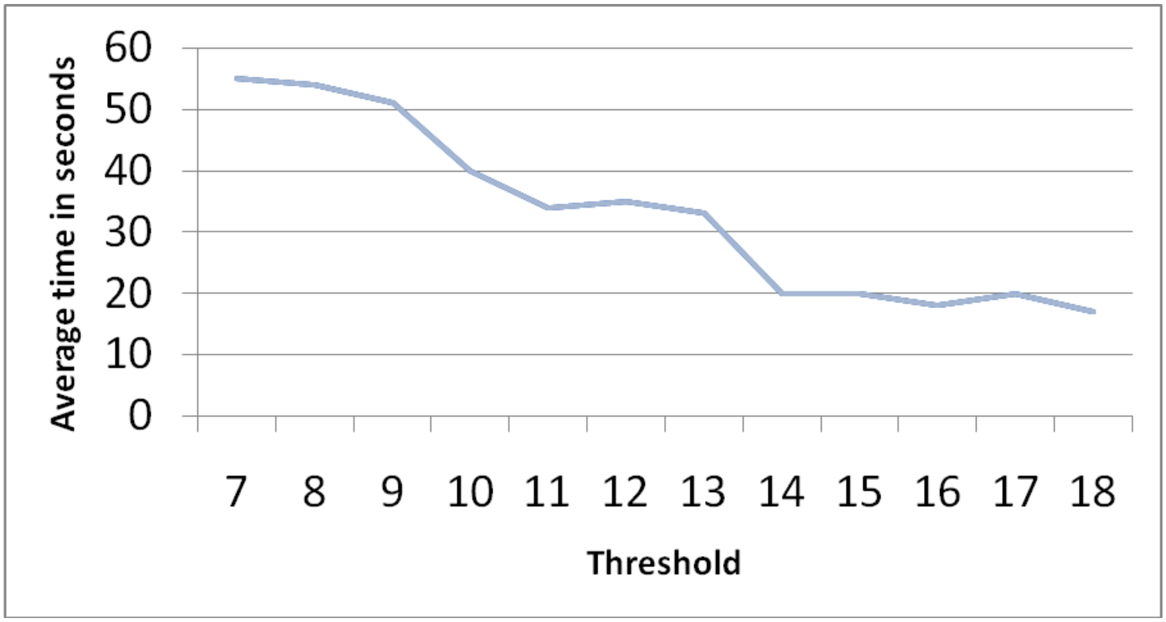

Figure 4.28 : Threshold & Average Time for 2CPG

Figure 4.27 show the running time of predicting protein structure with various threshold values. As shown in the figure the average time at threshold 7 is 55 second and it is decreased to 20 second at threshold 17 these result is not fixed and it is may be decrease or increase in arbitrary way because the algorithm take random number as starting points to FSOLVE function to solve the nonlinear system.

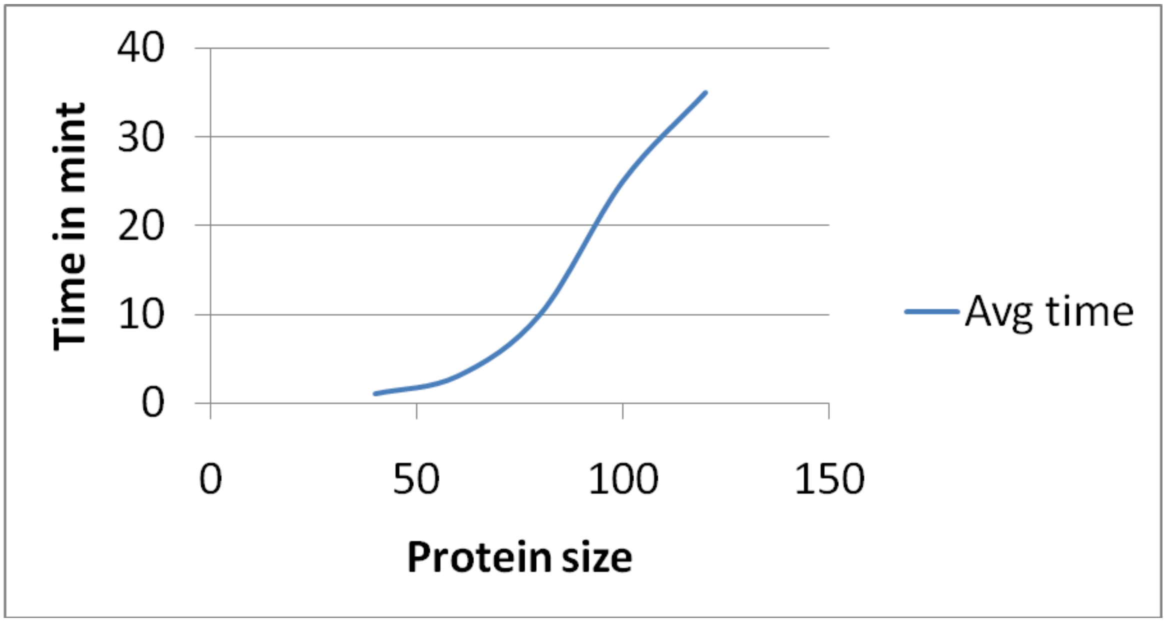

Figure 4.29: protein size & Average time

An analysis of our algorithm show that the protein length affected in running time of proteins. When the protein size is very long the average of running time is increase. For example the maximum running time in different thresholds of 2CPG protein with length 43 is one mint, while the maximum running time in different thresholds of 451C protein with length 82 is 28 mints.

CHAPTER FIVE

Conclusion & Future Work

5.1 Conclusion

Predicting a protein structure is one of the approaches that have been used in folding a protein 3D structure. For the past few years, several efforts have been developed in order to help predict a protein 3D structure to understand protein functionality. These efforts used machine learning approaches such as neural network and support vector machine and distance geometric.

This thesis used contact map matrix as a starting point to predict 3D structure of a protein, and show that the contact maps computed using threshold values (12-18) Å allow better 3D structure recovery than those computed at thresholds (7-9) Å.

The experimental results show that the scanning of contact map for a protein is much more reliable to predict the more important areas of the contact map. This process based on prediction quality more than quantity of contacts. Looking for the dense area is an important step that will improve the performance of the predicting 3D structure of protein from it CM.

The main contribution of this thesis is using the MATLAB which introduce an efficient and very fast way to solve the problem with an improvement in prediction accuracy by FSOLVE function to solve the nonlinear system and give the best set of three dimensional coordinates fit for Distance between nodes and map 3D structure of a protein by PLOT3 function.

5.2 Future Work

We would like to suggest some interesting issues and ideas that could not be reached because of limited time and recourses and other constraints, and they will aids an important and enhancement on the proposed approach as future work:

- It is possible to increase the detection accuracy of dense area of contact map through divide the contact map into clusters and separately use the sub matrices to create sets of coordinates then merge it in order to select best solution in important area.
- MATLAB is a program that is very useful to solve large problems, so that I believe it is important to pay the most attention to students to learn and understand MATLAB tools and used bioinformatics tool in MATLAB to solve any problem in molecular biology.
- Develop new approach merge between distance geometry and MATLAB tools to recover the three dimensional protein structures and give a set of lower and upper bounds to residues inter atomic distances.
- Take a group of proteins from NMR and the same group from X-ray then predicts the 3D structure of these proteins and compares the percentage of error to improve the prediction accuracy.
- Improve the approach by add new procedure to rotate and translate the coordinate to obtain the best set consists with the native structure with zero error.

REFEREANCES

1. Bioinformatics Web-Comprehensive Educational Resource on Bioinformatics Web Site: online available from URL http://www.geocities.com/bioinformaticsweb/index.html

2. Zaki .M, Jin .S, and Bystroff .C. “Mining Residue Contacts in Proteins Using Local structure Predictions”. IEEE International Symposium on Bioinformatics and Biomedical Engineering; 2000.p.p 168-175

3. Wikipedia, the free encyclopedia, bioinformatics web, 2008. http://en.wikipedia.org/wiki/Bioinformatics

4. Philip .E and Helge.W , "Structural Bioinformatics"(hand book), san Diego supercomputer center, Pharmacology department, Univ. California san Diego, wiley-liss publisher; 2003.

5. Richard and Lopez . “Maclyn McCarty (1911-2005)-Obituary”. Journal : Biographical-Item. Nature 2005; vo. 433: p.p 372-372.

6. Orengo .C, Michie .A, Jones .S, Swindells .M, and Thornton .J."CATH_ Hierarchic Classification of protein domain structure".1997; 5,p.p 108-193.

7. Kendrew J, Bodo G, Dintzis H, Parrish R, Wyckoff H, and Pillips D.”A three dimensional model of the myoglobin molecule obtained by X-ray analysis”. Nature 1958, p.p 662-666.

8. Leonid .A. Mirny and Eugene .I. “Protein Structure Prediction by Threading”. Harvard University, Department of Chemistry and Chemical Biology,1998.

9. Ying. Z and George .K. “Prediction of Contact Maps Using Support Vector Machines”, 2002,p.p1-5

10. Crippen, G. “Distance geometry vs. coordinates for protein folding calculations”. Proteins, 2005,p.p 82-89.

11. Pollastri G, and baldi P. “Prediction of contact maps by recurrent neural networks architectures and hidden context propagation from all cardinal corners”. Bioinformatics, 2002 ;vo.1,p.p1-19.

12. Jing. H, Ming. Z and Zhen. S,"CS 6890 Project Report ", 2004, p.p 2-23.

13. Piero. F, Osavaldo. O, Rtta.C and Alfonso. V,"Prediction of Contact Maps With Neural Network and Correlated mutation", biology department, Univ. Bologna, Italy, 2001, protein engineering vo.14, no.11, p.p.835-843.

14. Marco. V, Luciano. M, Filippo. M, Pietro. D and piero.F, "Reconstruction of 3D Structures From Protein Contact Maps", CS department, bioinformatics group, Univ. Bologna, Italy, 2008: p.p.1-12.

15. John. M," Predicting protein three-dimensional structure", Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 1999,vo.10, p.p583–588.

16. Jing. He, xiaolan. S, Mohammed. Z, ying. S and Chris. B, "Mining Protein Contact Map", BIOKDD02: workshop on data mining on bioinformatics, with SIGKDD02mConference 2002: p.p 3-10.

17. Vicky .C. "Updating Torsion Angles of Molecular Conformations", Department of Computer Science, Virginia Tech, 2005.

18. Mireille .G. “Distance geometry, hellix packing, and contact map congeruncy advisors”. Queen’s university; ,2006.p.p 3-19

19. Casbon. J. “Protein secondary structure prediction with support vector machine”. M.Sc thesis, Sussex, published, 2002.

20. Fruten .J, “Molecules and life: historical essays on the interplay of chemistry and biology”. 1972.

21. Thomasson .W.”Unraveling the mystery of protein folding”. FASEB,1997

22. Westhead .D, Parish .J, and Twyman .R. “Instant notes bioinformatics” . bios scientific publishers, 2002.

23. Linus . P and Alfred .M. “On the structure of native, denatured and coagulated proteins”, Proc natl acad sci USA 1936, p.p 47-439

24. Gutpa .N, Managal .N, and Biswas .S. “Evolution and similaraity Evaluation of protein structure in contact map space”. Proteins:structure , function , and bioinformatics 2004;p.p 196-204.

25. Allen .F, Kennard .O, Taylor .R, “Systematic analysis of structural data as a research technique in organic chemistry”;Acc chem. Research 16, 1983, p.p 53-146.

26. Marco. V, Luciano. M, Filippo. M, Pietro. D and piero.F, "Reconstruction of 3D Structures From Protein Contact Maps", CS department, bioinformatics group, Univ. Bologna, Italy, 2006: p.p 13-15.

27. Vassura. M, Margara.L and Pietro .D, " Fault Tolerance Reconstruction of 3D Structure from Protein Contact Maps", 2007,p.p 2-9.

28. Orengo .C, Jones .S, and Thornton .J. Bioinformatics genes,proteins and computer.1st edtion. BIOS Scientific publisher.2003.

29. Blumental .L. “Theory and application of distance geometry”, Chelsea, new york,1970.

30. Jorge.M, zhijun .W. “Distance geometry optimization for protein structure”.1996,p.p.1-14.

31. Etter, Delores, David Kuncicky and Holly Moore. “Introduction to MATLAB® 7”. New York: Prentice Hall. 2005.

32. Zhiyong Z. “An Overview of Protein Structure Prediction From Homology to Ab Initio”,Final Project For Bioc218, Computational Molecular Biology,2002.

33. Olmea.O, Burkhard .R, and Alfonso .V.” Effective use of sequence correlation and conservation in fold recognition”. J. Mol. Biol. 1999,p.p1221–1239.

34. Pollastri.G, Rost .Band Baldi.P.”Prediction of coordination number and relative solvent accessibility in proteins”, proteins: structure, function, and genetics,2002, p.p 142-153 .

35. Thomas.D, Casari.G, and Sander.C. “The prediction of protein contacts from multiple sequence alignments”. Protein Engineering.1996.

36. Vendruscolo .M, Kussell.E, and Domany.E. “Recovery of protein structure from contact maps”. Folding & Design, 1997, 2, p.p295-306.

الملخص

توقع الشكل الثلاثي للبروتين من خريطة الاتصال

يعتبر علم (Bioinformatics) من العلوم الحديثة التي أثارت إهتمام كل من علماء الحاسوب وعلماءالاحياء. حيث أن القدره على توقع شكل البروتين الثلاثي من سلسلة الأحماض ألأمينيه المكونه لهذا البروتين يعتبر ثورة علميه في هذا المجال. يعد تركيب اي بروتين مؤشرا قويا على الوظيفة التي يقوم بها هذه هي القاعده الأساسيه في كل العلوم التي تختص بدراسة البروتينات. حيث يمكن تمثيل الشكل الثلاثي للبروتين من خلال خارطة الاتصال والتي تعرف بمصفوفه ثنائيه تبين أماكن الاتصال بين الأحماض الامينيه في البروتين والتي تحدد من خلال مسافة معينة بين الاحماض داخل البروتين، يعتبر اي حمضين متصلين إذا كانت المسافه بينهما أقل او تساوي المسافه المحدده مسبقا (Threshold) وإلا فإن الحمضين غير متصلين. ونظرا لأهمية هذا الموضوع وصعوبته تم تقديم الكثير من الابحاث المختلفه في هذا المجال والتي تحلل وتستخرج قواعد تفيد في توقع الشكل الثلاثي للبروتين من خلال خارطة الاتصال.

من هذا المنطلق تم التركيز في هذه الاطروحة على البروتينات وتوقع الشكل الثلاثي لها باستخدام برنامج(MATLAB). إن الطرق التقليدية التي كانت تستخدم في تحديد شكل البروتين تستغرق وقتاً طويلا لإكتشاف شكل البروتين الواحد وتحتاج الى أموال باهظة كما أنها تفتقر الى القدره على اكتشاف شكل العديد من البروتينات ، لذلك أصبح البحث عن طرق جديدة امرا مهما.

تركز خطة عمل هذة الاطروحة على تطوير خوارزمية تستخدام خارطة الإتصال بمسافات مختلفة للتحكم بنقاط الإتصال للوصول الى مجموعة الاحداثيات الثلاثية الأفضل لرسم الشكل الثلاثي للبروتين بأسهل الطرق واسرعها ، حيث وُجد ان خرائط الاتصال التي تم حسابها اعتمادا على مسافات (Threshold) تتراوح بين ١٢ الى ١٨ اعطت رسما ادق من تلك المسافات (Threshold) التي تتروح بين ٧ الى ٩ .

لقد أظهرت النتائج التجريبيه التي تم تطبيقها على بروتينات بانواع مختلفة أن الوقت المستغرق في تنفيذ الخوارزميه المقترحه قليل جدا وأنه يتأثر بطول البروتين المستخدم. وكما اثبتت النتائج كفاءة هذه الخوارزميه المقترحه في توقع الاشكال الثلاثية الابعاد للبروتينات.

Appendix A

```
function  Reconstruct(CM,T)

CM1 = scan_CM (CM); % scan of CM based on number of
neighbors
D = distance_matrix(CM1,T); %compute the distance between
two atoms
Dist =shortdist(D);% short path distance

C=nonlin_coordinat(Dist);%compute the coordinat of all
atoms by nonlinear system
NCM = new_contact_map(CM,C,T);%extract new contact map
based on a new coordinat
e =compar_contact_maps(NCM,CM);

 Q=10;

          while E> 0.10 && Q > 0
              NC=coorect_coordinat(CM,C,T);
              NCM = new_contact_map(CM,NC,T);
              e =compar_contact_maps(NCM,CM);
              Q=Q-1;

              C=NC;
          end


 map_3D_structuer(NC);    % map 3D structre of this protein

end
%****************************************************************
function    CM1=scan_CM(CM)
n=length(CM);
CM1=zeros(n:n);
for i=1:n
    CM1(i,i)=1;
   for j=i+1:n
       count=0;

         for k=1:n

           if ((CM(i,k)==1) && (CM(k,j)==1))
              count=count+1;
               if (CM(i,j)==1)&& (count==10)) ||
                   CM(i,j)==1) && (count==20)
```

```
                  CM1(i,j)=1;
                  CM1(j,i)= 1 ;
               break;

            end
         end
       end
   end
end

end
%****************************************************************
function D = distance_matrix(CM1,T)
n=length(CM1);
D=zeros(n:n);
for i=1:n
    for j=i:n
        if CM1(i,j)==1
            D(i,j)= count_distance(T,i,j);
          else
            D(i,j)= rand(1,1)+T;
        end
             D(j,i)=  D(i,j);
    end
end
end
%****************************************************************
 function x= count_distance(T,i,j)
if i==j
   x=0;
elseif abs(i-j)==1
    x=3.8 ;
     elseif abs(i-j)==2
     x=6+rand(1,1);
     elseif abs(i-j)==3
     x=7+rand(1,1);
     elseif abs(i-j)>3
     x=(0.91-(T/100))*T ;
end
end
%****************************************************************


function Dist=shortdist(D)
n=length(D);
Dist=zeros(n:n);
for i=1:n
      for j=i+1:n
        for k=1:n
           if (D(i,j)<= (D(i,k)+ D(k,j)))
              Dist(i,j)=D(i,j);
           else
```

```
               Dist(i,j)=(D(i,k)+ D(k,j));
              end
        end

   end
end

end

%*****************************************************************
function C= nonlin_coordinat(Dist)
C=[];
 For a=1:10
n=length(Dist);
x0=40*rand(n*3,1).10*rand(1,1);
x = fsolve(@(x) compute_coordinat(x,Dist),x0);
      for l=1:n
          s=((l-1)*3)+1;
      C(1:3,l)=x(s:s+2,1);
      end

  end
end

%*****************************************************************
function F = compute_coordinat(x,Dist)
n=length(Dist);
F=[];
c=0;
 for i=1:n
 for j=i:n
     c=c+1;
F(c)=(x((i-1)*3+1)-x((j-1)*3+1))^2+(x((i-1)*3+2)-x((j-
1)*3+2))^2+(x((i-1)*3+3)-x((j-1)*3+3))^2-Dist(i,j)^2;
end
 end

end
%*****************************************************************
function  NCM= new_contact_map(CM,C,T)
n=length(CM);
NCM=zeros(n:n);

global d;
d=zeros(n:n);
for i=1:n
   for j=i:n
         if i==j
               d(i,j)=0;
    NCM(i,j)=1;
   Else
```

```
d(i,j)= sqrt ((C(1,i)-C(1,j))^2+(C(2,i)-C(2,j))^2+(C(3,i)-
C(3,j))^2);

               if d(i,j)<= T
                    NCM(i,j)=1;
               else
                    NCM(i,j)=0;
               end
         end
           d(j,i)=  d(i,j);
           NCM(j,i)=NCM(i,j);

       end
end
end
%****************************************************************

function e =compar_contact_maps(NCM,CM)
  e=0;
  n=length(CM);
for i=1:n
       for j=1:n
           if NCM(i,j)~=CM(i,j)
               e=e+1;
           end
       end
end

e=e/(n*n);
disp(e);
end
%****************************************************************

function NC=coorect_coordinat(CM,C,T)
n=length(CM);
global d;
for i=1:n
   for j=1:n
      if (CM(i,j)==1 && d(i,j)> T )||( CM(i,j)==0 && d(i,j)<=T)
         r=mobilty(CM,d,i,T);
         NC(1:3,i)=correct_direction(CM,C,T,i,r,d);
         if norm((C(1:3,i)-c))<= r
             NC(1:3,i)=c;
          else
            NC(1:3,i)=C(1:3,i);
         end
       break;
      end
   end
end
%****************************************************************
```

```
function  r =mobilty(CM,d,i,T)
n=length(CM);
D0=inf;
D1=0;
   for j=1:n
      if (CM(i,j)==1 && d(i,j)<= T )
          D1=max(D1,d(i,j));
      else
          if  (CM(i,j)==0 && d(i,j)>T)
           D0=min(D0,d(i,j));
          end
      end
   end
   m(i)=min(D0-T,T-D1);
   r=m(i);
end

%*****************************************************************

 function k =correct_direction(CM,C,T,i,r,d)
        V=[0;0;0];
        n=length(CM);
        for j=1:n
       if (CM(i,j)==1 && d(i,j)> T )||( CM(i,j)==0 && d(i,j)<=T)
         if CM(i,j)==1
       V= V - (((C(1,i)-C(1,j))+(C(2,i)-C(2,j))+(C(3,i)-
C(3,j)))/d(i,j))
          else
        V=V + (((C(1,i)-C(1,j))+(C(2,i)-C(2,j))+(C(3,i)-
C(3,j)))/d(i,j));
                end
            end
        end

      k =C(1:3,i)+ ( V *(r/norm(V)));

    end



%*****************************************************************


function  map_3D_structuer(NC)
plot3(NC(1,1:end),NC(2,1:end),NC(3,1:end));
end
%*****************************************************************
```