ENHANCED HONEY BEES COLONY ALGORITHMS FOR PROTEIN TERTIARY STRUCTURE PREDICTION

HESHAM AWADH ABDALLAH BAHAMISH

UNIVERSITI SAINS MALAYSIA
2011

ENHANCED HONEY BEES COLONY ALGORITHMS FOR PROTEIN TERTIARY STRUCTURE PREDICTION

by

HESHAM AWADH ABDALLAH BAHAMISH

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

April 2011

ACKNOWLEDGEMENTS

All praise be unto Allah and may His peace and blessings be upon His final prophet and messenger Muhammad; peace be upon him.

First and foremost, I would like to extend my most profound gratitude and deepest thanks to my principal supervisor, Professor Dr. Rosni Abdullah, Universiti Sains Malaysia, for her guidance, commitment and encouragement throughout the entire period of this study. Her dedication, time spent on coordinating and supervising the whole thesis, and continuous guidance enabled me to remain focused on the study, right from the beginning of the study till the very end. Without Allah and Professor Dr. Rosni's constant support, none of this would have been possible.

My deepest thanks also go to my parents who have been an inspiration and I ask Allah to grant them His paradise. Special thanks go to my family, that is, my wife, Salwa, and sons, Abdullah and Mohammed, my brother, Gihad, and sisters, Nahed and Ghada, for their patience in my quest; their great support is highly appreciated.

My sincere thanks are also extended to my laboratory mates in the Parallel and Distributed Computing Centre (PDCC) laboratory, the academic and technical staff of the School of Computer Sciences for their assistance, and to my University in Yemen, Hadhramout University-Mukalla, for the financial support.

TABLE OF CONTENTS

CHAPTER 3 – LITERATURE REVIEW

CHAPTER 6 - PROTEIN CONFORMATIONAL SEARCH USING ARTIFICIAL BEES COLONY ALGORITHM

CHAPTER 7 - PARALLEL HONEY BEES COLONY ALGORITHMS FOR PROTEIN CONFORMATIONAL SEARCH

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABC	Artificial Bee Colony
ABCA	Artificial Bee Colony Algorithm
ACO	Ant Colony Optimization
AI	Artificial Intelligence
AMBER	Assisted Model Building and Energy Refinement
ANN	Artificial Neural Network
BCO	Bee Colony Optimisation
CHARMM	Chemistry at HARvard Molecular Mechanics
CM	Comparative Modelling
CN	Contact Numbers
CSA	Conformational Space Annealing
DDGA	Dual Distributed Genetic Algorithm
DE	Differential evolution
DFIRE	Distance-Scaled Finite Ideal-gas REference
DGA	Distribute Genetic Algorithm
DGQ	Drone Generation Based on the Queen Genotype
DHAD	Dihedral Angle Difference
DNA	DeoxyriboNucleic Acid
DPSO	Discrete Particle Swarm Optimisation
ECEPP	Empirical Conformational Energy Program for Peptides
EDMC	Electrostatically driven Monte Carlo
FCM	Fuzzy C-Means
FMBO	Fast Marriage in honey bee optimisation
fmGA	fast messy Genetic Algorithm
FR	Fold Recognition
GA	Genetic Algorithm
GA-FCM	Genetic algorithm- Fuzzy C-Means
GROMOS	GROningen MOlecular Simulation

GSAT	Greedy SAT
HA	Hybrid Algorithm
HBMK-means	Honey Bee Mating K-means
HBO	Honey Bees mating algorithm
HBPI	Honey-Bees Policy Iteration
HM	Homology Modelling
HP	Hydrophobic Polar lattice model
HPGA/GBX	Hybrid Parallel GA/Guide Blend Crossover
HSABCA	Hybrid Simplex Artificial Bee Colony Algorithm
HSE	Half-Sphere Exposure
IABC	Interactive Artificial Bee Colony
I-PAES	Immune Based Pareto Archived Evolution Strategy
LCMST	Leaf Constrained Minimum Spanning Tree
LVQ	Learning Vector Quantisation
MBO	Marriage in honey Bee Optimisation
MC	Monte Carlo
MCM	Monte Carlo with Minimisation
MD	Molecular Dynamics
METRO	Metropolis Criterion for the Acceptance of the New Queen
MLP	Multi-layered Perceptron
MM	Molecular Mechanics(
MMFF	Merck Molecular Force Field
MPI	Message Passing Interface
MPNS-GRASP	Multiple Phase Neighbourhood Searches – Greedy Randomized Adaptive Search Procedure
mRNA	messenger Ribonucleic acid
NE SIMPSA	Non-Equilibrium Simplex Simulated Annealing
NM-MBO	Nelder-Mead Marriage in honey bee optimisation
NMR	Nuclear Magnetic Resonance
PAES	Pareto Archived Evolution Strategy
PCB	Printed Circuit Board

PDB	Protein Data Bank
PSIPRED	Position Specific Iterated secondary structure PREDiction server
PSO	Particle Swarm Optimisation
PSP	Protein Structure Prediction
QEA	Quantum Evolutionary Algorithm
QRD	Queen Replacement by the Drone
REMC	Replica Exchange Monte Carlo
RMSD	Root Mean Square Deviation
RNA	Ribonucleic acid
SA	Simulated Annealing
SAT	Satisfiability problem
SI	Swarm Intelligence
SMMP	Simple Molecular Mechanics for Proteins
SOM	Self-Organizing Feature Maps
SOMHBMK	Self-Organizing Feature Maps Honey Bee Mating K-means
SUMSL	Secant-type Unconstrained Minimisation Solver
SVM	Support Vector Machine
TS	Tabu Search
TSP	Travelling Salesman Problem
VRP	Vehicle Routing Problem
WPS	Wolf Pack Search
WPS-MBO	Wolf Pack Search-Marriage in honey Bee Optimisation
α**BB**	Branch and Bound

PENINGKATAN ALGORITMA KOLONI LEBAH MADU UNTUK RAMALAN STRUKTUR TERTIER PROTEIN

ABSTRAK

Kepentingan peranan protein dalam proses biologi tubuh manusia telah diakui dan tidak diragui lagi. Protein mampu melaksanakan fungsi biologi apabila ia terlipat ke dalam struktur tertier. Kaedah ramalan struktur tertier protein secara eksperimen memerlukan masa yang lama dan sangat mahal. Tambahan pula, struktur protein juga sering kali tidak dapat ditentukan secara eksperimen. Saintis dari pelbagai bidang berusaha menghasilkan teori serta kaedah komputasi yang mampu menyelesaikan masalah ramalan struktur protein secara kos-efektif. Secara komputasi, masalah ramalan struktur protein dirumus sebagai suatu masalah optimum dan matlamat utama ialah untuk menggelintar ruangan carian untuk mencari protein yang sama bentuk dengan tenaga bebas terendah (struktur protein). Kajian ini bertujuan menyelidik serta meneroka buat kali pertama, kemampuan algoritma berdasarkan koloni lebah madu menggelintar ruangan carian untuk mencari protein yang sama bentuk dengan tenaga bebas terendah, serta menggabungjalinkan teknik selari ke dalam algoritma tersebut untuk meningkatkan keupayaan mencari protein yang sama bentuk. Prinsip-prinsip penggabungjalinan dalam koloni lebah madu (algoritma MBO) dan perlakuan pencarian makanan koloni lebah madu (algoritma ABC) telah diadaptasikan untuk menyelesaikan masalah mencari protein yang sama bentuk. Algoritma-algoritma selari telah dibangunkan untuk meningkatkan prestasi algoritma gelintaran. Daripada algoritma-algoritma yang diadaptasi, tenaga bebas terendah telah diperolehi bagi protein yang diuji. Tenaga bebas terendah Met-enkephalin telah diperolehi (-12.42 dan-12.9101 kcal/mol). Di samping itu, tenaga bebas terendah bagi C-peptida dan struktur terbaik bagi 12 bioaktif peptida juga telah ditemui. Prestasi algoritma MBO selari menunjukkan penambahan kelajuan hampir linear manakala algoritma ABC menunjukkan penambahan kelajuan linear.

ENHANCED HONEY BEES COLONY ALGORITHMS FOR PROTEIN TERTIARY STRUCTURE PREDICTION

ABSTRACT

There is no doubt about the role that protein plays in the biological processes inside the human body. Proteins are able to perform their biological functions when they fold into their tertiary structure. Experimental protein tertiary structure prediction methods are time consuming and expensive and it is not always possible to determine the protein structure experimentally. Scientists from many fields work to develop theoretical and computational methods which provide cost effective solutions to the protein structure prediction problem. Computationally, the protein structure prediction problem is formulated as an optimisation problem and the goal is to search the protein conformational search space to find the lowest free energy conformation (protein structure). The aim of this study is to investigate and explore for the first time, the capability of the honey bees colony-based algorithms in searching the protein conformational search space to find the lowest free energy conformation, and to incorporate parallel techniques into the protein conformational search algorithms to enhance the protein conformational search. The principles of marriage in the honey bees' colony (MBO algorithm) and the honey bee colony's foraging behaviour (ABC algorithm) were adapted to solve protein conformational search problem. Parallel algorithms were developed to enhance the performance of the search algorithms. The adapted algorithms were able to find the reported lowest free energy conformation for the test proteins. The lowest free energy conformation of Met-enkephalin was found (-12.42 and -12.9101 kcal/mol). Lower free energy conformations for C-peptide and good structures for the 12 bioactive peptides were found. The parallel MBO algorithm gained near linear speed-up and the parallel ABC algorithm gained linear speed-up.

CHAPTER ONE

INTRODUCTION

Protein is constructed from a list of 20 amino acids types. Proteins play a vital role in the biological processes of the human body. Significantly, a protein will only be able to perform its biological function when it folds into its tertiary structure. This tertiary structure is known as the biological active state or the native state. Moreover, many of the drugs become effective when their structures are closely associated with the structure of the proteins (Ogura et al., 2003).

In Bioinformatics, especially in protein data, three kinds of information are closely related (Satou et al., 1997). They are sequence, structure, and function. The sequence determines the structure, and the structure determines the function.

The determination of the protein sequence from the genes encoded in the DNA is known as the first genetic code while the determination of the protein structure from the amino acids sequence is considered as the determination of the second genetic code (Chan and Dill, 1993; Hardin et al., 2002). While the first genetic code has been solved, the second code is still not fully understood and needs more research in order to break it. Protein Structure Prediction (PSP) is currently perhaps the biggest problem in Bioinformatics (Keedwell and Narayanan, 2005).

The PSP problem is simply stated as “Given a protein sequence, what is its tertiary structure?”. Solving this problem is not as simple as its statement. The prediction of a protein’s structure from its amino-acid sequence is regarded as a great challenge in many scientific disciplines. It is a fundamental scientific problem and a great challenge in structural

Biology (Brock and Brunette, 2005), computational Biology, Chemistry (Floudas, 2007), and Bioinformatics (Helles, 2008; Kanehisa, 1998; Meidanis, 2003), and it is one of the unresolved problems in Biophysics. One of the most important objectives of Bioinformatics is the prediction of protein structure (Hu et al., 2008; Mount, 2004; Rizk, 2006). Protein structure can be determined by using experimental methods and computational methods. Figure 1.1 gives an overview of the PSP. As shown in this figure, the genome projects over the world produce new protein sequences which are stored in protein sequence databases. The structures of these sequences can be determined using structure prediction methods. These structures are stored in structure databases. The structures are used to solve different bioinformatics and medical problems such as protein function prediction and drug design. PSP methods are divided into experimental and computational methods.

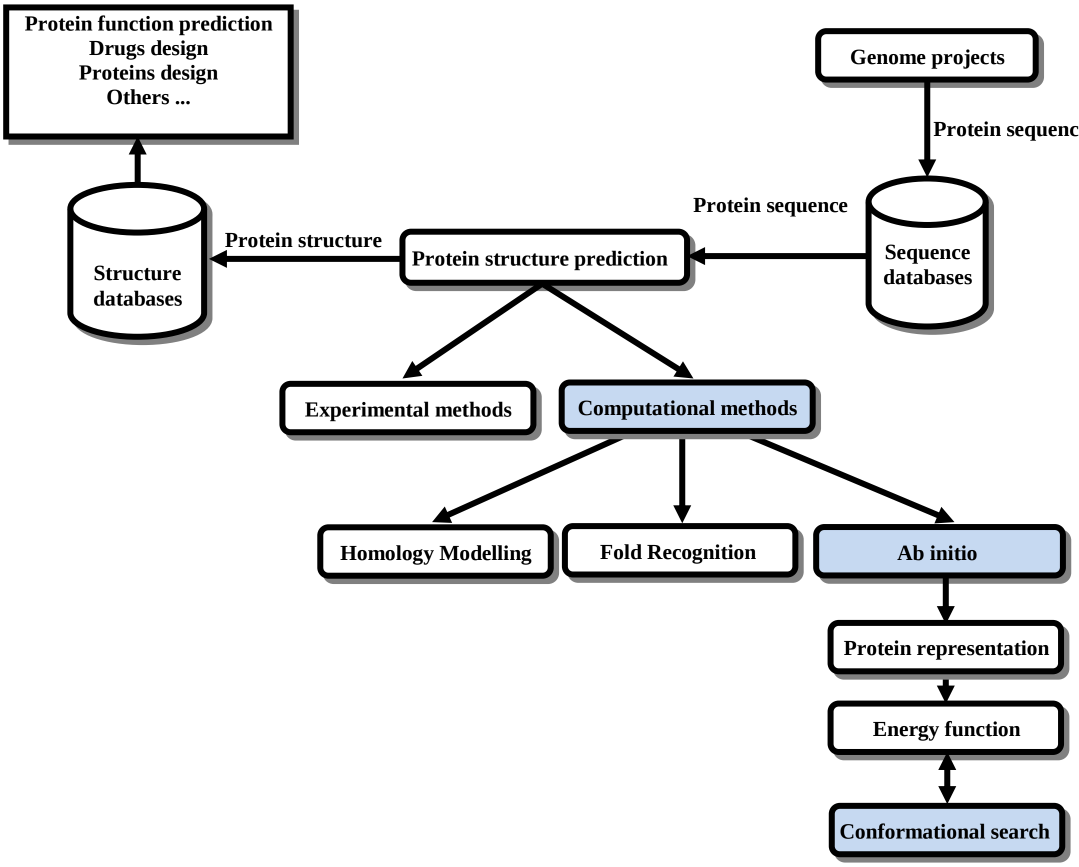

Figure 1.1: Overview of protein structure prediction

Experimental protein structure determination methods such as Nuclear Magnetic Resonance (NMR) and X-ray crystallography are the main trusted and mostly used methods. They are the main sources of information about protein structure (Chen et al., 1994). However, they have some drawbacks. They are difficult to use, time-consuming, laborious and expensive. As they need special equipment and human efforts, the determination of the structure of a single protein may take from several months up to years of laboratory work. In addition, not all protein structures can be determined using experimental methods (Evans et al., 1995; Zhang, 2002a). As an example, the NMR method can determine the structure of proteins which are not longer than 100 amino acids (Jones, 2000) while protein crystallisability is a prerequisite for the X-ray method. So it cannot be applied to all proteins because not all proteins can be crystallised (Chen et al., 1994; Chen et al., 2007).

Because of these limitations, the experimental protein structure determination is still slower than sequence determination. Various genome projects are identifying more new genes than the number of protein structures being determined by experimental methods (Karl-Heinz, 2003). This has led to an obvious big gap between the number of known protein sequences deposited in sequence databases and the number of determined protein structures deposited in structure databases. For example, the protein sequences in Swissprot database **(**UniProtKB/TrEMBL Release 2010_10 of 5 October 2010) have 12098541 entries**,** while the protein structures in Protein DataBase (PDB) (up to 12 October 2010) register 68562 structures. Therefore, other fast methods of protein structure prediction are needed to resolve this gap.

The prediction of the protein structure from the protein sequence demands a continuous development of new methods to solve the problems especially when there is less experimental information. Because of the challenges in the determination of protein structures experimentally, scientists from many fields such as Biology, Computer Science, Mathematics, Biochemistry, and Physics work to develop theoretical and computational methods to predict the protein tertiary structures. Computational PSP is the only alternative to deduce the protein structure and understand its function (Sundararajan and Kolaskar, 1996b).

Computational methods have become one of the most important research topics in modern molecular biology (Liu et al., 2005). Theoretical methods are important and necessary to help biologists in obtaining protein structure information (Liu et al., 2005; Suzuki and Okuda, 2008; Zhang, 2002b). They provide a cost-effective solution to the PSP problem (Beiersdorfer et al., 1997; Das et al., 2008; Greenwood and Shin, 2002; Ogura et al., 2003). Furthermore, using computational methods enables the structure prediction of a large number of protein sequences which cannot be determined experimentally (Baker and Sali, 2001). Computational methods are traditionally classified into three approaches. These are Homology Modelling (HM) or Comparative Modelling (CM), Fold Recognition (FR) or Threading and Ab initio.

Ab initio methods are based on the Anfinsen thermodynamic hypothesis (Anfinsen, 1973) which states that the tertiary structure of the protein is the conformation with the lowest free energy. Predicting the protein structure using the ab initio methods is one of the top ten challenges in Bioinformatics (Meidanis, 2003).

Based on the Anfinsen thermodynamic hypothesis, the PSP problem is formulated as an optimisation problem (Morales et al., 2000; Garduno-Juarez et al., 2003; Ogura et al., 2003; Crivelli and Head-Gordon, 2004; Bortolussi et al., 2005; Vengadesan and Gautham, 2005; Yun-Ling and Lan, 2006). The goal is to search the protein conformational search space to find the lowest free energy conformation. This conformation is the structure associated with a stable state. Protein conformation refers to the protein tertiary structure and can be traditionally defined as the organisation of its atoms in the three dimensional space. These atoms can be inter-converted purely by rotation about a single bond. Most molecules can adopt more than one conformation. Protein has an infinite number of non superimposable conformations (Karl-Heinz, 2003).

In order to predict the protein structure using ab initio methods, a proper representation of protein conformation is required. An energy function is used to calculate the conformation energy while a conformational search algorithm is utilised to search the conformation search space to find the lowest free energy conformation.

1.1 Problem Statement

The PSP problem is a hard combinatorial optimisation problem (Greenwood and Shin, 2002; Lee et al., 1997) as it involves searching the conformational search space for the lowest free energy. Conformational search algorithms explore the protein conformational search space with a major goal to find the lowest free energy conformation (Zhang, 2002a). Searching the protein conformational space is a grand challenge in protein tertiary structure prediction due to the large number of possible conformations and the local minimum problem. In general, if a protein has n atoms, the degree of freedom is $3n$-6. Accordingly, a protein with 100 amino acids where each amino acid has 20 atoms, the number of degrees of freedom is equal to ([(100*20)*3]-6=5994) (Schulze-Kremer, 2000). In other words, by considering five torsional angles for each of the 100 amino acids and taking five values for each angle, the number of possible conformations will be 25^{100}.

Thus, the PSP problem is considered as an NP-hard problem (Khimasia and Coveney, 1997, Morales et al., 2000, Garduno-Juarez et al., 2003) or even NP-Complete problem (Seung-Yeon et al., 2003, Bortolussi et al., 2005). Protein conformational search algorithms need an exponential time to search the protein conformational search space which is similar as searching for "*a needle in a haystack*" (Dill et al., 1993). It is impractical to test all the feasible conformations to find the lowest free energy conformation. Therefore, the success in predicting protein tertiary structure is dependent on the efficiency of the searching method to pass over different conformations without testing all conformational possibilities (Zhou and Abagyan, 2002) and without regard to the folding processes (Day et al., 2003; Morales et al., 2000).

There is a need for a robust and efficient searching algorithm. Since the problem is a combinatorial optimisation problem, many optimisation algorithms have been developed to search the protein conformational space. Some of the most common algorithms are Monte Carlo (MC) (Evans et al., 1995; Ripoll and Thomas, 1990), Simulated Annealing (SA) (Fadrná and Koca, 1997; Ogura et al., 2003; Tanimura et al., 2004; Yun-Ling and Lan, 2006), and Genetic Algorithms (GA) (Beiersdorfer et al., 1997; Garduno-Juarez et al., 2003; Gates et al., 1995; Khimasia and Coveney ,1997a; Madhusmita et al., 2008; Schulze-Kremer, 1994; Schulze-Kremer, 1996; Unger and Moult, 1993; Xiang, 2000).

Recent years have showed the application of Swarm Intelligence (SI) based algorithms in solving Bioinformatics problems (Das et al., 2008). In the PSP problem, the idea of using the cooperative and collective behaviour of social insects to search the protein conformational search space was addressed by Huber and van Gunsteren (1998). Ant Colony Optimisation (ACO) (Daeyaert et al., 2007; Fidanova and Lirkov, 2008; Hu et al., 2008; Shmygelska, 2006; Shmygelska et al., 2002a; Shmygelska and Hoos, 2003; Shmygelska and Hoos, 2005) and Particle Swarm Optimisation (PSO) (Datta et al., 2008) were used to predict the structure of the protein.

SI based algorithms that are inspired by the behaviour of the honey bees colony can be classified into different classes. Marriage in honey Bees Optimisation (MBO) algorithm is inspired by the process of reproduction (marriage) in the honey bees colony, and Artificial Bee Colony (ABC) algorithm is inspired by the foraging behaviour of the honey bees colony. These algorithms have been applied to many applications and optimisation problems.

Previously unsolved old problems can be insightfully investigated using algorithms inspired from honey bee behaviour (Olague and Puente, 2006). Using the principles of honey bees colony, the difficult combinatorial optimisation problems such as protein tertiary structure prediction can be solved. In this study, we ask whether honey bees colony inspired conformational search algorithms, which is based on the foraging behaviour of the honey bees colony and process of reproduction behaviour, can be used to find the lowest free energy conformation of proteins.

1.2 Justification for using SI Algorithms to Solve the PSP Problem

The justifications for using the SI algorithms to solve the PSP problem are:-

1) SI algorithms are adapted and being successfully applied to optimisation problems in a variety of fields that involve combinatorial complexity (Denby and Le Hégarat-Mascle, 2003).
2) Collective behaviour can speed up the search in combinatorial optimisation problems (Dorigo et al., 1996; Haynes, 1997).
3) SI algorithms have attracted researchers working on bioinformatics problems over the world (Das et al., 2008). They play a role in the bioinformatics task, i.e. the PSP (Das et al., 2008).

1.3 Motivation

Through the knowledge of the protein tertiary structure, much valuable information can be revealed. This information is essential in helping scientists get a better understanding of the protein functionality and the understanding of many diseases that occur as a result of protein mis-folding (Greenwood and Shin, 2002; Schlick, 2002). With this in mind scientists can design new drugs that interact with targeted proteins and modify their functions (Chen et al., 1994), and design new drugs that can cure diseases (Greenwood and Shin, 2002; Schlick, 2002; Yun-Ling and Lan, 2006). From a practical point of view, the sequence structure gap is the main factor motivating the need for predictions of protein structure (Mala, 2008).

As the conformational search problem is computationally expensive, parallelisation of the sequential algorithms is needed to enhance their performance.

1.4 Objectives

The main aim of this research is to enhance the protein tertiary structure prediction problem using a spectrum of SI algorithms. In particular, the present study focuses on the adaptation of algorithms inspired by the honey bees colony to search the protein conformational search space for the lowest free energy conformation. Since searching the protein conformational search space is computationally expensive, there is potential that the adapted algorithms be parallelised. As such, the study focuses on the following specific objectives.

1) To enhance the protein conformational search by adapting the concepts of marriage in honey bees colony (MBO algorithm).
2) To enhance the protein conformational search by adapting the concepts of the foraging behaviour of honey bees colony (ABC algorithm).
3) To incorporate parallel techniques into the protein conformational search algorithms to speed up the protein conformational search.

1.5 Scope of the Study

This study focuses on using computational PSP methods in solving the protein tertiary structure prediction problem, that is, using the ab initio method, in particular, in the protein conformational search problem. The representation of the protein conformation is the torsion angles of the main chain and the side chain of the amino acids. The size of the proteins, which used in this study, is ranged from 5 to 20 amino acids.

1.6 Main Contributions

This study adapts honey bees colony based algorithms to solve the PSP problem. The novel contribution of this research is the use of honey bees colony based algorithms and torsion angles representation with secondary structure information to determine protein tertiary structure. This study makes the following contributions:

1. Refines the generic MBO algorithm and introduces three new modifications to its structure.
2. Adapts the refined MBO algorithm to solve the protein conformational search problem as the first applications of the MBO algorithm for this problem.
3. Parallelises the MBO algorithm as the first attempt to parallelise the MBO algorithm and applies it to solve the protein conformational search problem.
4. Introduces two new modifications to the ABC algorithm and adapts it to solve the protein conformational search problem.
5. Parallelises the ABC algorithm and applies it to solve the protein conformational search problem.

1.7 Thesis Organisation

The body of this thesis consists of eight chapters. The organisation of the rest chapters is as follows:

Chapter 2:

This chapter gives a background about the protein and an overview of computational PSP methods. It also gives an overview on SI and honey bees colony. The MBO and ABC algorithms are also presented.

Chapter 3:

This chapter reviews the protein conformational search algorithms. It also reviews the parallel protein conformational search algorithms and the applications and modifications of MBO and ABC algorithms.

Chapter 4:

This chapter provides information on the research framework. It contains information on the data pre-processing and the datasets used in the research and the methodology employed for the different parts of the work.

Chapter 5:

This chapter introduces the refined MBO algorithm and the proposed three modifications. It describes the adaptation of the refined MBO and modified MBO algorithms to solve the protein conformational search problem. The major components of the algorithms are described and the experimental results and evaluations are presented.

Chapter 6:

This chapter presents the adaptation of ABC algorithm to solve the protein conformational search problem. It explains the two new proposed modifications to the ABC algorithm and the adaptation of the modified ABC algorithm to solve the protein conformational search problem.

Chapter 7:

This chapter presents the parallel design and implementations of the MBO and ABC algorithms and discusses their results.

Chapter 8:

In this chapter, the study closes with a summary of the results and some concluding remarks. Suggestions for future work are also presented.

CHAPTER TWO

BACKGROUND

2.1 Introduction

PSP problem is one of the most difficult problems faced by researchers today. PSP is one of the most compelling challenges for scientists in Bioinformatics. It is still one of the fundamental unsolved problems in Bioinformatics and computational structural biology and in many other research areas (Das et al., 2008). So far, there is no radical solution available to this problem. The main difficulty of this problem is centred in finding a correct way to calculate the protein energy as well as exploring the large conformational search space for the lowest free energy protein conformation. A wide variety of computational methods has been developed to predict the protein structure.

This chapter starts by giving a background about the protein in section 2.2 and an overview of computational PSP methods in section 2.3. An overview on Swarm Intelligence is given in section 2.4, and honey bees colony are described in section 2.5. Sections 2.6 and 2.7 provide an overview on the MBO and ABC algorithms. A summary of the chapter is given in section 2.8.

2.2 Protein Background

Proteins are the main building blocks and machineries for all living organisms. They play important roles in the activities inside the cells of the living organisms. Inside the human body, there are thousands of protein types. Proteins are the key components of the human body. They build up the cellular components and mediate biological and metabolic processes. Each cell of the human body contains a number of proteins that play various essential biological functions such as the enzymatic activity of the cell, attacking diseases, transporting and sending biological signal transmissions. These functions are fundamental to the life through which the human body performs its functions properly.

The protein is formed inside the cell when the Deoxyribonucleic acid (DNA) transcribes the encoded genes into messenger Ribonucleic acid (mRNA) which is translated by the ribosome into a sequence of amino acids that compose the protein. This is known as the central dogma of molecular biology which is shown in Figure 2.1.

Proteins are polymers of connected amino acids whose composition is encoded in genes. These amino acids are the basic building blocks of the protein. There are twenty amino acid types in nature. Each of them is denoted by a different letter (or three letters) as shown in Table 2.1. Proteins differ only by the sequential order and the number of amino acids. The length of the protein molecule can vary from a few to many thousands of amino acids.

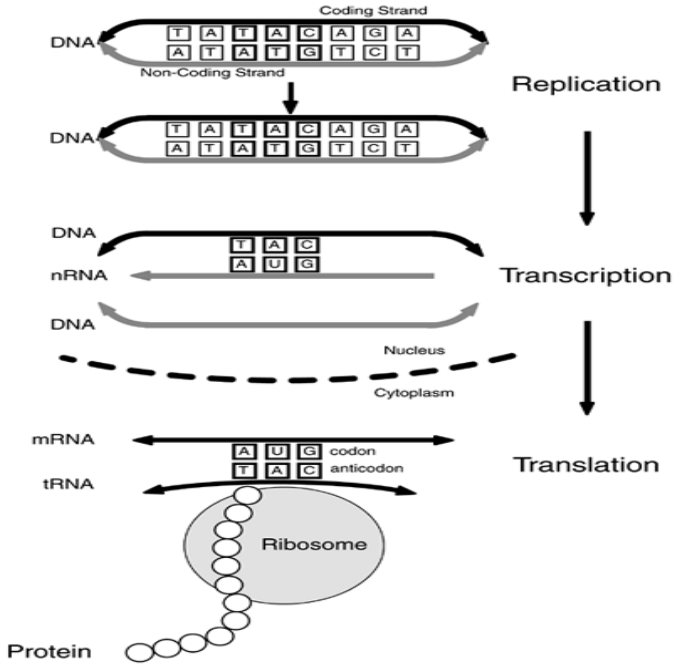

Figure 2.1: The central dogma of molecular biology (Bergeron 2003)

Table 2.1: The twenty amino acids

Name	1-Letter	3-Letter
Alanine	A	Ala
Arginine	R	Arg
Asparagine	N	Asn
Aspartic	D	Asp
Cysteine	C	Cys
Glutamic	E	Glu
Glutamine	Q	Gln
Glycine	G	Gly
Histidine	H	His
Isoleucine	I	Ile
Leucine	L	Leu
Lysine	K	Lys
Methionine	M	Met
Phenylalanine	F	Phe
Proline	P	Pro
Serine	S	Ser
Threonine	T	Thr
Tryptophan	W	Trp
Tyrosine	Y	Tyr
Valine	V	Val

Each amino acid consists of two parts: a main chain or backbone and a side chain or R chain. The main chain is the same in all the amino acid types. The differences are in the side chain which determines the chemical properties of the amino acid. The main chain contains a central carbon (Cα) which is bonded with an amino group (--NH2), a hydrogen atom (H) and a carboxylic acid group (-COOH). The side chain is attached to the central carbon which is denoted by (R) as shown in Figure 2.2. There are 20 different types of side chains in nature. Some are simple, made of only one atom and some are complicated containing many atoms.

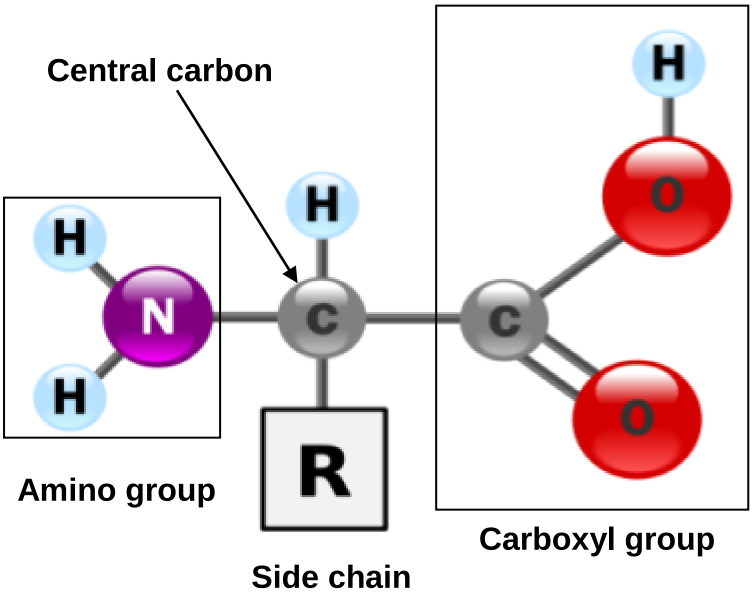

Figure 2.2 : Amino acid
Source: http://www.mcat45.com/content/protein

Amino acids are connected to each other by a peptide bond. The peptide bond is formed between two amino acids when the carboxyl group of the first amino acid interacts with the amino group of the second amino acid. A water molecule is released due to this interaction as shown in Figure 2.3.

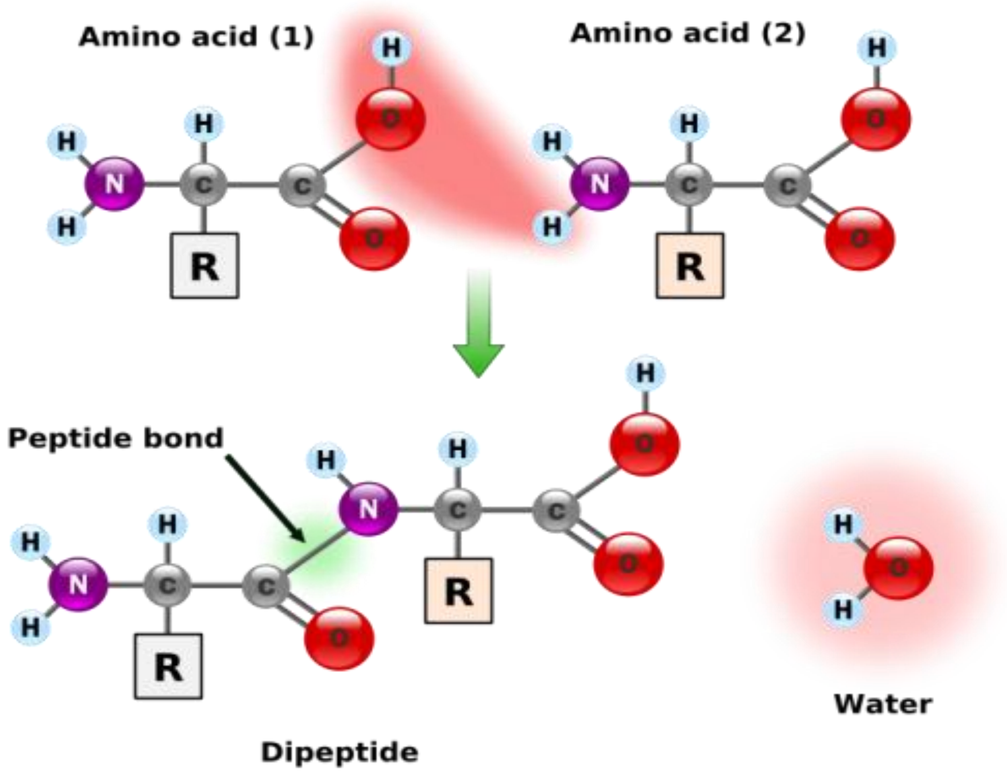

Figure 2.3: Peptide bond
Source: http://www.mcat45.com/content/protein.

Proteins can be delineated through four different hierarchical levels as shown in Figure 2.4. These are as follows:-

Primary structure: is the chain of amino acids sequence.

Secondary structure: is formed due to the interactions between the atoms of the main chain which results in local structures such as α-helix and β-sheet.

Tertiary structure: is the three dimensional arrangement of the atoms of the amino acids as the secondary structural elements are packed together due to polarity and the interactions between the side chains.

Quaternary structure: a protein which consists of several protein subunits (domains) held together.

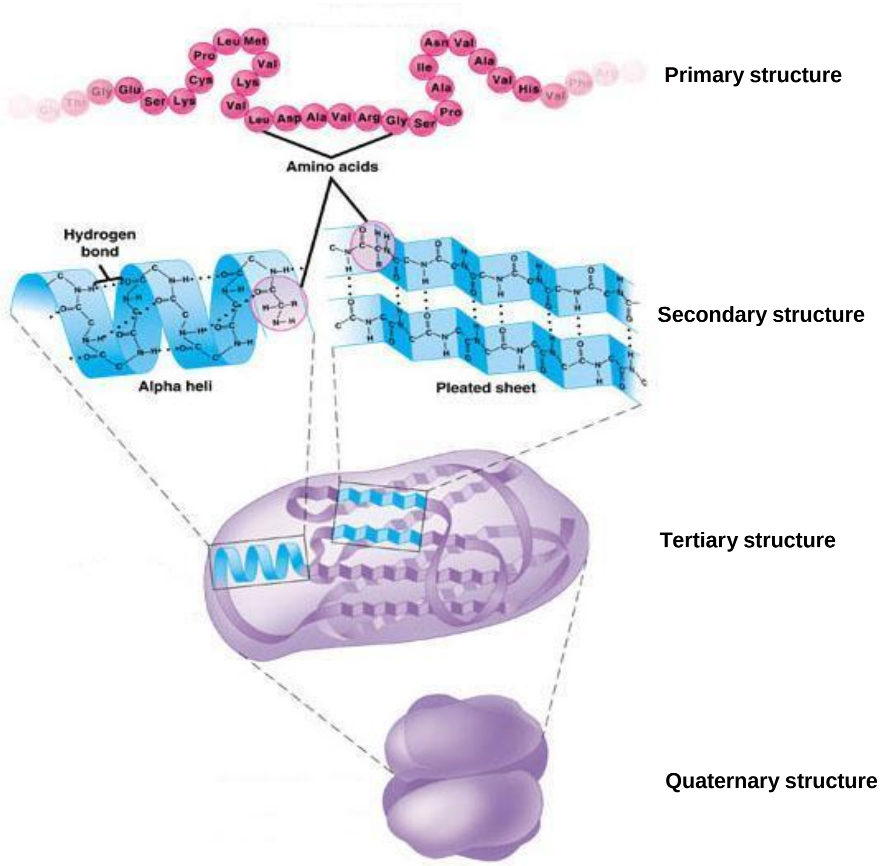

Figure 2.4: The four protein structure levels
Adapted from https://peggleston-bioreview.wikispaces.com/Life+Molecules?f=print

2.3 Computational Protein Structure Prediction Methods

Computational PSP methods are classified into three classes based on the sequence similarity to the target sequence and the utilisation of protein information available in structure databases (Bonneau and Baker, 2001; Yi-Yuan et al., 2005; Zhang 2002b). These classes are:

a) Homology Modelling,

b) Fold Recognition and

c) Ab initio.

Homology Modelling and Fold Recognition use sequence similarity in the prediction processes but the ab initio method does not. Computational methods also can be grouped into two groups: “non-optimisation” or knowledge-based methods (Homology Modelling and Fold Recognition) and “optimisation methods” (ab initio).

2.3.1 Homology Modelling

Homology Modelling (HM) also known as Comparative Modelling (CM) is the easiest, most reliable, and the most successful computational protein tertiary structure prediction method (Augen, 2004; Jones, 2004; Pedersen, 1999; Zhang, 2002a). HM is based on the observations of the structure experimental data which indicate that the protein sequence determines the protein structure and that the similarity in the protein sequence imposes the similarity in the protein structure (Zhang, 2002b). This similarity could be interpreted as the new proteins which evolve progressively by adding, deleting or changing the location of the amino acids while retaining the structure and function of the protein during this process (Zhang, 2002a).

Figure 2.5 depicts the HM processes. HM methods do not have to care of the folding mechanics of a protein. They build a model of tertiary structure based on the identifiable sequence association between the target protein and another protein or proteins of known structure. The prediction starts by searching for suitable structure templates for the target protein sequence. This is performed by comparing the sequence of the target protein with the sequences of proteins of known structures in the structure databases. The sequence of the target protein is then aligned to the structural templates. The protein backbone is built from the alignment, the loops are added and the side-chains are placed. Finally, the model is further refined.

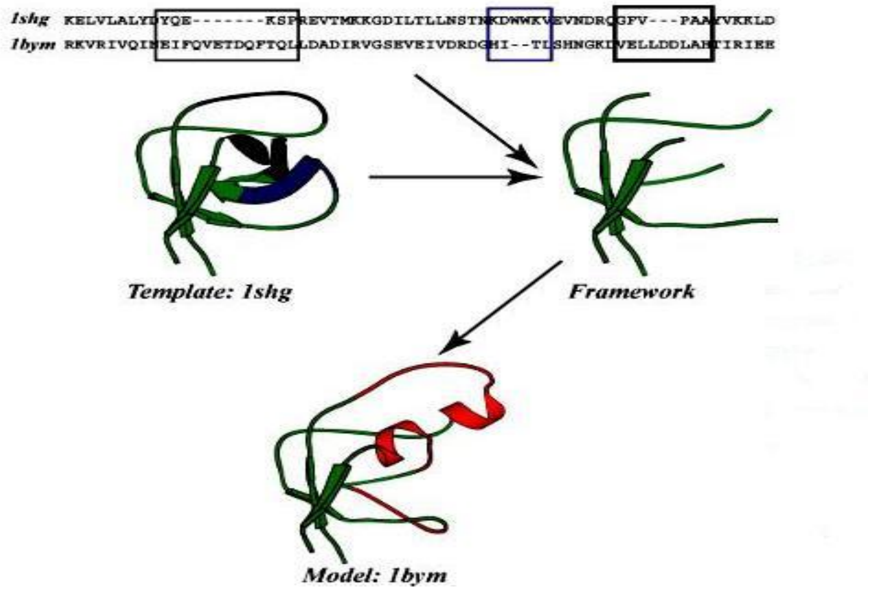
1shg KELVLALYDYQE-------KSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFV---PAAYVKKLD
1bym RKVRIVQINEIFQVETDQFTQLLDADIRVGSEVEIVDRDGHI--TLSHNGKDVELLDDLAHTIRIEE
Template: 1shg
Framework
Model: 1bym

Figure 2.5: Homology Modelling
http://koehllab.genomecenter.ucdavis.edu/teaching/ecs129/09

In order to have successful and accurate structure prediction using HM, the target protein should have a clear evolutionary relationship to at least another protein with known structure which is already stored in the structure databases (Bergeron, 2002; Skolnick and Kolinski, 2001; Skolnick et al., 2006; Zhang, 2008). In other words, HM is limited to predict the structure of protein families with at least one known structure. HM cannot help in understanding how and why a protein folds into a specific structure (Lee et al., 2009). This is because understanding the effects of different forces that play important roles in the formation of secondary and tertiary structure cannot be obtained by using HM (Pillardy et al., 2001; Volker et al., 1999).

The quality of the prediction using HM depends on the degree of similarity between the target protein and the proteins in the structure databases (Floudas, 2007; Pillardy et al., 2001). The higher the similarity is, the higher the prediction quality (Shortle, 1999). The sequence alignment is the bottleneck of the HM (Schonbrun et al., 2002). Achieving a good quality alignment plays an important role in the success of the HM (Schonbrun et al., 2002) and in the accuracy of the predicted structure (Shortle, 1999; Zhang, 2002b).

2.3.2 Fold Recognition or Threading

In cases where the HM methods fail to find similar protein sequences to the target protein sequence, then the Fold Recognition (FR) or threading methods can take its place to predict the protein structure based on the similarity between the sequence of the target protein and the structure of known protein folds.

FR methods are based on the fact that the number of protein folds in nature is limited, and that the structure of the target protein should be similar to one or some of these folds (Lotan, 2004). When the target protein is structurally similar to some known protein folds, these proteins are said to be remote homologous. FR tries to identify the remote homologue from the known protein folds. FR chooses the fittest fold to the target sequence by aligning the target sequence with the known protein structure folds (sequence-structure alignment) from a set of alternatives according to some energy function (Pedersen, 1999). Figure 2.6 gives an overview of FR.

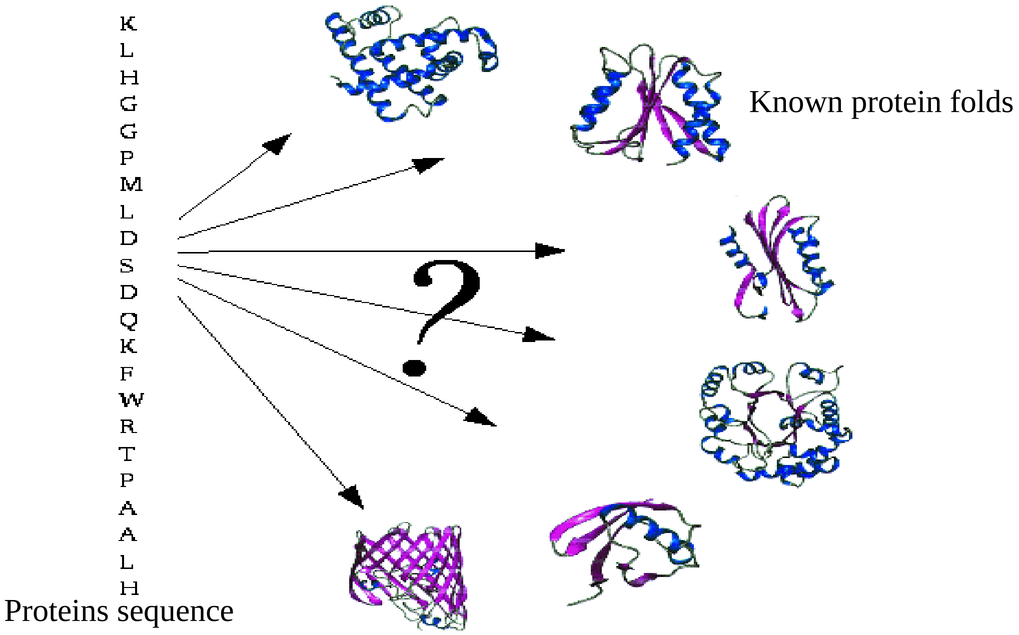

Fig. 2.6: Fold Recognition
http://biology.polytechnique.fr/proteinsathome/documentation2.php.

Similar to HM, the sequence similarity plays an important role in the quality of the prediction of the FR methods. FR methods fail to predict the precise fold when the similarity of the sequence is low. For that, new folds cannot be predicted because the prediction is based on already known folds (Ginalski et al., 2005). FR is limited by the high computational cost of the energy functions that are used to determine the correct fold (Zhang, 2002a). Moreover, FR does not provide a general understanding of the role of particular interactions in the formation of protein structure and the mechanisms of protein folding (Pillardy et al., 2001). Finally, according to Zhang (2008) the progress and development in the FR methods have reached a steady state.

2.3.3 Ab Initio

Generally, both HM and FR methods fail to predict the protein structure when the similarity between the sequence of the protein and sequences/known folds of known structures is low or cannot be detected. In this case, ab initio provides a valuable complement to these methods because it can be applied more generally to predict the structure of any protein sequence (Floudas et al., 2006; Ye, 2007).

The word “ab initio” or “de novo” means “from the first principles” or “from the beginning”. Ab initio PSP methods try to predict the protein tertiary structure from the amino acids’ sequence using physical principles. They try to fold the protein from a random conformation to the native conformation i.e. the tertiary structure (Skolnick and Kolinski, 2001). Ab initio methods are based on the Anfinsen thermodynamic hypothesis (Anfinsen, 1973). Anfinsen hypothesis is the most widely accepted and used hypothesis in PSP (Ngan et al., 2008). It explains the process of protein folding and it was formulated in a Nobel Prize winning experiment. This experiment revealed that the protein amino acids have all the necessary information of the forces that fold the protein into its native conformation, which is the conformation with the lowest free energy (Chan and Dill, 1993). Therefore, the natural conformation of the protein in the real world corresponds to the free energy minimal conformation.

Based on Anfinsen thermodynamic hypothesis the PSP problem is formulated as a combinatorial minimisation optimisation problem (Bortolussi et al., 2005; Crivelli and Head-Gordon, 2004; Garduno-Juarez et al., 2003; Morales et al., 2000; Ogura et al., 2003; Yun-Ling and Lan, 2006). Basically, ab initio protein tertiary structure prediction methods perform a conformational search guided by an energy function (Floudas et al., 2006; Lee et al., 2009). The aim is to search the protein conformational search space to find the lowest free energy conformation. In order to achieve that, three main components of the ab initio method must be considered (Bonneau and Baker, 2001; Hardin et al., 2002; Huang et al., 2000; Jones, 2000; Lee et al., 2009; Osguthorpe, 2000; Pedersen, 1999; Zhang, 2002b). These components are:

(1) A proper protein *representation*.

(2) An *energy function* compatible with the protein conformation representation is used to calculate the conformation energy.

(3) A *conformational search algorithm* which is utilised to search the conformation search space to find the lowest free energy conformation.

Since the PSP problem is formulated as an optimisation problem, optimisation is one of the promising approaches to solve this problem (Hoek, 1994). A wide range of optimisation methods have been developed to tackle this problem. Optimisation methods represent the conformation of a protein as a set of parameters. These parameters form the protein conformational search space. The protein conformational search space consists of all possible conformations of the protein. The prediction of the protein tertiary structure using ab initio methods is performed by searching the protein conformational search space to locate the global minimum energy conformation. This is accomplished by generating many conformations by making changes to the parameters. The generated conformations are evaluated by employing the energy function. The search is performed iteratively and the conformation corresponding to the global minimum is finally chosen to be the structure of the protein (Jones, 2000; Pillardy et al., 2001).

Protein tertiary structure prediction using ab initio methods is the "holy grail" of the PSP field (Helles, 2008; Jones, 2000). Ab initio PSP remains a difficult challenge today (Ngan et al., 2008). Developing an accurate ab initio PSP method is one of the top ten challenges in bioinformatics (Meidanis, 2003) and a major goal of theoretical molecular biology (Friesner and Gunn, 1996). It is a true computational challenge to predict the protein tertiary structure using only the protein sequence information. It is the most complicated prediction approach (Feldman, 2003). According to Yang (2008) predicting the structure of protein with larger than 150 amino acids using ab initio methods is a non-trivial task and considered a challenge due to a lack of accuracy yield by energy functions and large conformational search space (Chivian et al., 2003) which includes multiple local minimum solutions. Because of these complexities, it is generally believed that the prediction of the protein tertiary structure from first principles is impossible (Okamoto, 2000). On the contrary, other researchers are of the opinion that the problem can be optimally solved (Pillardy et al., 2001).

Ab initio methods are not limited to predicting the structures of proteins which belong to protein families that have known structures. However, ab initio methods are computationally expensive and provide low to moderate accuracy. Regardless of the accuracy of the ab initio methods, these methods are useful since the predicted structure with errors could be used to predict some aspects of the protein function (Sanchez et al., 2000).

Ab initio methods can be classified into (i) knowledge–based ab initio and (ii) classical ab initio (Forman, 2001) or Simulation methods (Zhang, 2002a). Knowledge-based methods use constraints and rules which are inferred from the data of known structures. Simulation methods, however, do not use databases and predict the structure based on physical principles. Their accuracy is low and the success is limited to small proteins (less than 100 amino acids) (Lee et al., 2009). The following subsections describe the three ab initio components.

Table 2.2 summaries the advantages and disadvantages of the computational PSP approaches. As the focus of this study is on ab initio PSP methods, in the following subsections, the three main components of the ab initio PSP are described in details.

Table 2.2: The Advantages and disadvantages of the computational PSP approaches

Approach	Advantages	Disadvantages
Homology Modelling	• Most accurate.	• Cannot predict structure of new proteins.
Fold Recognition	• Prediction is done based on a limited number of protein folds.	• Cannot predict structure of new proteins. • Computationally expensive.
Ab initio	• Able to predict the structure of any protein.	• Low to moderate accuracy. • Computationally expensive.

2.3.3(a) Protein Representation

Many of the real world problems are considered as optimisation problems. Usually, when attempting to solve these problems using computational methods, an obvious representation of the problem and their control variables are required. It is important that this representation should cover possible solutions, and at the same time, it should not cover more details since this increases the search space and in consequence the run time of the optimisation algorithm (Matthias, 1998).

In order to predict the protein tertiary structure starting from its amino acid sequence and to be able to understand the nature and process of the formation of the protein structure using computational methods (Kolinski and Skolnick, 2004), it is essential and very important to determine an appropriate protein model or representation. Theoretical protein models or representations describe and summarise the information of the structure in the real world to the required level of details (Pedersen, 1999).

Protein must be clearly represented as much as possible. So, protein representation should have enough information to make the explanation of computational PSP experiments feasible and as unambiguous as possible (Kolinski and Skolnick, 2004; Skolnick and

Kolinski, 2001). The protein representation is important because it determines the size of the search space. The protein representation plays an important role in determining the computation time required to calculate the energy of the protein. In addition, they should enable the generation of a sufficient number of conformations to be searched (Osguthorpe, 2000) and should also cover every possible conformation (Matthias, 1998).

There is a wide variety of protein representations in different levels of details. They can be classified based on two points of view: the number of particles which represent the protein structure or the level of detail (all-atoms, united atoms, virtual atoms with one, two or at least two atoms per residue) and the type of the phase space to be searched continues (off-lattice) or discrete (lattice) (Kolinski and Skolnick, 2004; Osguthorpe, 2000).

There is an essential trade-off between the completeness of a protein representation and its intricacy. More complete protein structure representations introduce more conformational degrees of freedom, making them more complex thereby increasing the size of the protein conformational search space (Depristo, 2004; Kolinski and Skolnick, 2004). On the other hand, reduced or simplified protein representations try to simplify the PSP problem by reducing the complexity of the protein representation. This can be achieved by reducing the number of degrees of freedom available to the amino acid (Volker et al., 1999).

Reduced protein representations are very important tools in PSP (Kolinski and Skolnick, 2004). They represent the geometry of the peptide bond and the various secondary structure elements but treat side chain and intermolecular force in an approximate manner (Levitt and Warshel, 1975). However, according to Bonneau and Baker (2001), the differentiation of the accurate native conformation from the similar conformations is one the most difficult tasks that researchers face. This is due to the insensitivity of the energy function of the reduced model (Bonneau and Baker, 2001).