HARMONY SEARCH ALGORITHMS FOR AB INITIO PROTEIN TERTIARY STRUCTURE PREDICTION

MOHAMMED SAID SALEH ABUAL-RUB

UNIVERSITI SAINS MALAYSIA

2011

HARMONY SEARCH ALGORITHMS FOR AB INITIO PROTEIN TERTIARY STRUCTURE PREDICTION

by

MOHAMMED SAID SALEH ABUAL-RUB

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

February 2011

ACKNOWLEDGEMENTS

All the praises and thanks be to Allah, the Lord of the world for giving me the energy and the talent to finish my research; He guides me and grants me success in my life. I can not count His bounties on me.

I would, also, like to express my deepest gratitude and appreciation to my main supervisor, Professor Rosni Abdullah, for her invaluable encouragement and guidance. Her support and comments have provided me with the adequate strength that enabled me to undertake this challenge. I am also grateful to my second supervisor, Assoc. Prof. Dr. Ahamad Tajudin Khader for his comments and guidance throughout the period of conducting this research.

In addition, I would like to thank the academic and technical support staff of the School of Computer Sciences, USM, who provided me with the facilities needed to conduct my research. I also wish to extend my gratitude to Universiti Sains Malaysia for granting me a scholarship to pursue my Ph.D.

Last but not least, my sincere thanks to my family; my mother, my wives, my children, and my brothers and sisters who have always shown their faithful support during my study. I appreciate their everlasting patience during the long period of my study. Special thanks to my brother Professor Marwan for his continuous support in my study since bachelor degree until PhD and to my sister Majd for her valuable comments, suggestions, and language editing. I sincerely thank my friends Mohammed Al-Betar, Hesham Bahamish, Ali Kattan, and Khalid Jaber for their help and support. Finally, I am very thankful to all my friends and family who always make Dua'a (Supplications to Allah) for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

AHSA Adaptive Harmony Search Algorithm

bw bandwidth

CASP Critical Assessment of Techniques for Protein Structure Prediction

CAFASP Critical Assessment of Fully Automated Structure Prediction

CSA Clonal Selection Algorithm

DNA Deoxyribonucleic Acid

ECEPP Empirical Conformational Energy Program for Peptides

GA Genetic Algorithm

GHS Global-best Harmony Search

HM Harmony Memory

HMCR Harmony Memory Consideration Rate

HMS Harmony Memory Size

HPSACO heuristic particle swarm ant colony optimization

HSA Harmony Search Algorithm

HHSA Hybrid Harmony Search Algorithm

ILS Iterated Local Search

kcal/mol Kilocalorie per mole

LDA Linear Discriminate Analysis

MC Monte Carlo

MCM Monte Carlo with minimization

MD Molecular Dynamics

mRNA Messenger RNA

MQAPs Model Quality Assessment Programs

NGHS novel global harmony search algorithm

NMR Nuclear Magnetic Resonance

NP Non polynomial

PAR Pitch Adjustment Rate

PDB Protein Data Bank

PSP Protein Structure Prediction

PSPP Protein Structure Prediction Problem

PSR Particle Swarm Rate

PSO Particle Swarm Optimization

PTSP protein tertiary structure prediction

REM replica exchange Monte Carlo Method

RMSD Root-Mean-Square Deviation

RNA Ribonucleic Acid

rRNA Ribosomal RNA

SA Simulated Annealing

SHSA Standard Harmony Search Algorithm

SMMP Simple Molecular Mechanics for Proteins

SQP Sequential Quadratic Programming

tRNA Transfer RNA

LIST OF SYMBOLS

α	Alpha
Cα	Alpha Carbon
Å	Angstrom
β	Beta
Cβ	Beta Carbon
ω	Omega
ϕ	Phi
ψ	Psi
Cφ	Varphi

ALGORITMA GELINTAR HARMONI UNTUK RAMALAN STRUKTUR TERTIER PROTEIN AB INITIO

ABSTRAK

Meramal struktur tertier protein daripada jujukan linear struktur-struktur tersebut adalah suatu cabaran besar dalam bidang biologi. Tesis ini berkisar tentang ramalan struktur tertier protein ab initio. Algoritma Gelintar Harmoni (HSA) disesuaikan untuk ramalan struktur tertier protein di mana keseluruhan proses dimodelkan sebagai pengoptimunan permasalahan. HSA telah pun memperolehi penyelesaian-penyelesaian yang layak tetapi tidak sehebat yang dilaporkan di dalam penulisan. Beberapa kekurangan telah dikenalpasti dan diselesaikan dengan mengusulkan Algoritma Gelintar Harmoni Adaptasi (AHSA) dan Algoritma Gelintar Harmoni Hibrid (HHSA). AHSA memperkenalkan satu skema baru bagi mengawal dua parameter utama HSA, iaitu Kadar Pelarasan Pic (PAR) dan Kadar Pertimbangan Ingatan Harmoni (HMCR), yang sesuai untuk Masalah Peramalan Struktur Protein (PSPP). Eksperimen-eksperimen melibatkan dua penanda-aras terkenal iaitu 'Met-enkephalin' dan '1CRN' telah dijalankan. Hasil eksperimen telah menunjukkan bahawa AHSA dan HHSA berjaya mepertingkatkan prestasi ramalan struktur tertier protein. Kedua-dua algoritma mampu menentukan tenaga terendah bagi protein yang diberikan, dan didapati lebih baik dari keputusan yang dicatatkan oleh sesetengah algoritma terkini. Di samping itu, dua nilai tenaga optimum global baharu bagi protein Met-enkephalin dicatat oleh kedua-dua AHSA dan HHSA berdasarkan medan kuasa ECEPP/3 dan ECEPP/2; dengan $\omega = 180^{\circ}$.

HARMONY SEARCH ALGORITHMS FOR AB INITIO PROTEIN TERTIARY STRUCTURE PREDICTION

ABSTRACT

Predicting the tertiary structure of proteins from their linear sequence is really a big challenge in biology. This thesis considers the *ab initio* protein tertiary structure prediction. The Harmony Search Algorithm (HSA) has been adapted for the protein structure prediction by modeling the problem as an optimization problem. HSA has obtained feasible solutions but not as magnificent as those reported in the literature. However, some shortcomings were identified and addressed by proposing an Adaptive Harmony Search Algorithm (AHSA) and a Hybrid Harmony Search Algorithm (HHSA). The AHSA introduces a new scheme for controlling the two main parameters of HSA, i.e. Pitch Adjustment Rate (PAR) and Harmony Memory Consideration Rate (HMCR), suitable for the Protein Structure Prediction Problem (PSPP). Experiments on two popular benchmarks namely 'Met-enkephalin' and '1CRN' has been performed. The experimental results have proved that both AHSA and HHSA have improved the overall performance of *ab initio* protein tertiary structure prediction. Both AHSA and HHSA have converged the lowest energy of the given proteins, and their results have outperformed some of the lowest energies recorded by some state-of-the-art algorithms. Moreover, two new global optimal energy values of the the 'Met-enkephalin' protein has been recorded by both AHSA and HHSA based on ECEPP/3 and ECEPP/2 force fields with $\omega = 180^{\circ}$.

CHAPTER 1

INTRODUCTION

1.1 Background

The field of bioinformatics has experienced an explosive growth in the last few years. It is a rapidly developing branch of biology and is highly interdisciplinary. Bioinformatics has many practical applications in different areas of biology and medicine. The main goal of bioinformatics is to increase the understanding of biological processes by using informatics. To achieve this goal, bioinformatics focuses on developing and applying computationally intensive techniques including pattern recognition, data mining, machine learning and visualization algorithms (Jasinski, 2006).

Major research efforts has been introduced to the field of bioinformatics in many disciplines such as: sequence alignment, gene finding, gene therapy, gene expression prediction, drug design and discovery, protein structure alignment, protein structure prediction, and protein-protein interactions. The techniques of the previous disciplines result in huge biological data which need fast data analysis and data management techniques for processing them. One of the main disciplines of analyzing biological data is structure prediction of the existing protein sequence data. It is difficult and slow to predict the 3D structure of a protein (Laskowski and Thornton, 2008).

Currently there are only thousands of known 3D structures compared to millions of known protein sequences. Therefore, it is important to contribute in the research of protein structure prediction of the huge existing primary sequences. This field has become a very active field in

research nowadays. This thesis has mainly considered the protein tertiary structure prediction problem. Many algorithms have been introduced to solve this problem these last few years. However, this research has focused on the optimization algorithms which deal with the protein structure prediction as an optimization problem such as Genetic Algorithm (GA), Molecular Dynamics (MD), and Simulated Annealing (SA). Optimization can be defined as the process of finding the best value from many possible values under certain constraints (if any). One of the new and successful optimization and search algorithms is harmony search algorithm (HSA). It is a meta-heuristic algorithm, mimicking the improvisation process of music players (Geem et al., 2001).

HSA has been implemented successfully in a wide variety of optimization problems such as medical imaging, timetabling, Sudoku puzzles, web document clustering (Ingram and Zhang, 2009). With comparison to traditional optimization techniques, HSA has provided many advantages such as: it requires fewer mathematical requirements without initial value settings for decision variables, it considers all the existing vectors to generate a new vector -whereas the methods like genetic algorithm (GA) only considers the two parent vectors, and HSA does not necessarily require to encode and decode the decision variables into binary strings (Mahdavi and Abolhassani, 2009). This thesis has adapted HSA for Protein Structure Prediction Problem (PSPP) to show the efficiency of this algorithm to this research area, and then try to enhance its performance by using adaptive parameters and hybridizing it with some ideas from existing successful methods.

1.2 Motivation

The very fast and exponential growth of protein sequence data has opened the gate for new research in computer science to produce effective methods to process this data. It is essential to know the 3D structure of proteins in order to understand their biological functions (Wright

and Dyson, 1999). The difficulty of determining the three dimensional structure of proteins has led to an increasing gap between the huge number of protein sequences and the limited number of protein structures. Figure 1.1 shows the exponential growth of protein sequences compared to protein structures. The number of available protein structures in the Protein Data Bank (PDB) database is 2 to 3 orders of magnitude smaller than that of the available protein sequences. In August 17, 2010 there are 67,322 structures while in August 10, 2010 there are 11,636,205 sequences according to rcsb PDB and UniProt databases respectively. Therefore, an affordable approach and a high throughput method are urgently needed in order to understand the biological systems and to reduce the gap between protein sequences and protein structures.

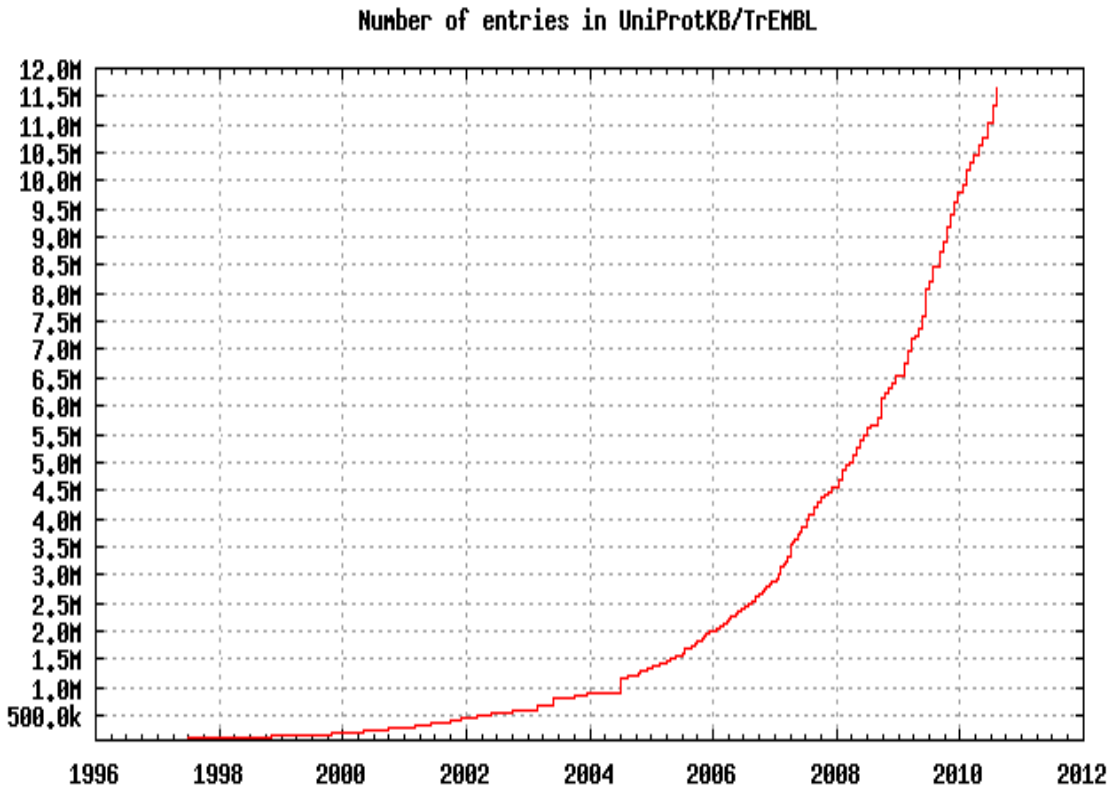

Figure 1.1: The growth of the protein sequences (Swiss-Prot database)

Several algorithms have been developed to solve *ab initio* protein structure prediction problem such as: genetic algorithm, Monte Carlo, basin paving (Lee et al., 2009). However, harmony search algorithm has been implemented successfully in a wide variety of optimization

problems (Ingram and Zhang, 2009) but it has not been investigated for protein structure prediction yet.

1.3 Problem Statement

This research intends to help biologists predict the tertiary structure of a protein from its sequence. Determining the tertiary structure of a protein is essential to understand the protein function (Wright and Dyson, 1999). However, predicting the tertiary structure of a protein from its sequence is still a challenging problem even for small proteins (Verma and Wenzel, 2007). A successful *ab initio* method for protein structure prediction depends on a powerful conformational search method to find the minimum energy using an energy function. However, finding the lowest free energy conformation of a protein is a NP-hard problem (Unger and Moult, 1993a), and the protein tertiary structure prediction problem has proved to be NP-complete (Berger and Leighton, 1998). This implies that no exhaustive search methodology is feasible to solve this problem. This fact has opened the gate for the non-deterministic search techniques like simulated annealing, genetic algorithm, tabu search, and ant colony to be the most successful techniques to solve this problem (Lee et al., 2009).

Molecular Dynamics (MD) and Monte Carlo (MC) are two common methods to explore protein conformational search space. For protein prediction, these two methods require an enormous amount of computational resources to explore the conformational space. A main technical difficulty of Monte Carlo simulations is that the energy landscape of protein conformational space is quite rough and rugged due to the fact that it contains many energy barriers, that may trap the MC simulation procedures (Lee et al., 2009). Different conformational search methods, however, have been developed to overcome these problems.

Until now, there is no single powerful search method that outperforms other methods for

all cases; nevertheless, there can be such a method that can outperforms other methods in some cases (Lee et al., 2009). The main focus of this research is to enhance the accuracy of protein tertiary structure prediction. The current studies, however, do not investigate harmony search algorithm in the context of protein structure prediction. Therefore, this research intends to adapt HSA for the PSPP to demonstrate its advantages and disadvantages, and based on that enhance its performance by incorporating some other metaheuristic components within the adapted HSA. The advantages of HSA make it worthy to be investigated to solve the PSPP.

The main questions of this research are:

1. Can Harmony Search Algorithm be adapted for the protein tertiary structure prediction problem?
2. Can the performance of the adapted HSA be enhanced to yield more accurate results by using adaptive parameters control?
3. Can the performance of the adaptive HSA be further enhanced by using some hybrid approaches?

1.4 Research Objectives

The main aim of this thesis is to investigate HSA to solve the PSPP. So, new alternatives to solve the PSPP are provided. Therefore, the objectives of this thesis are:

- To adapt HSA for the protein tertiary structure prediction problem.
- To improve the adapted HSA by controlling its two main parameters, HMCR and PAR.
- To hybridize the (AHSA) with a local search method to enhance the accuracy of the final results.

1.5 Research Scope

This research covers the protein structure prediction problem. Protein structure has many levels; primary structure, secondary structure, tertiary structure, and quaternary structure. This thesis focuses on the tertiary structure level of the protein structure. However, there are several categories of computational methods to predict the protein tertiary structure, this research considers *ab initio* method which predicts the tertiary structure of the protein from its sequence alone -without any previous knowledge. *Ab initio* protein structure prediction method should have three main components: problem representation, searching tool and an energy function. This research focuses on the searching part of the method. In short, this research is limited to investigating HSA to *ab initio* PSP.

1.6 Methodology

As mentioned in the previous section, the main objective of this research is to investigate harmony search algorithm to predict the protein tertiary structure from its sequence. This section provides a brief overview of the methodology of this research to achieve the research objectives. The methodology is described in more detail in chapter 4.

To answer the first question of this research, *ab initio* protein structure problem has been modeled as an optimization problem. HSA has been adapted for the PSPP; this method is called in this research as Standard Harmony Search Algorithm (SHSA). Although the performance of the (SHSA) is not magnificent, the algorithm has deserved more research in order to answer the second research question. For the second research question, an Adaptive Harmony Search Algorithm (AHSA) has been introduced. The AHSA includes modification in some parameters and operators of the SHSA. Some weaknesses have been revealed after implementing the AHSA related to the exploitation property; this has led the research to answer

its third question. For the third research question, a local search algorithm, called Iterated Local Search (ILS), has been hybridized with the AHSA to improve its local exploitation. Moreover, the global-best memory consideration, an idea from Particle Swarm Optimization (PSO), is applied as a selection mechanism.

Finally, a comparative analysis has been conducted to compare the three proposed algorithms, SHSA, AHSA, and HHSA with some other methods; as well as the performance of three algorithms among each other. Figure 1.2 describes the methodology of this research.

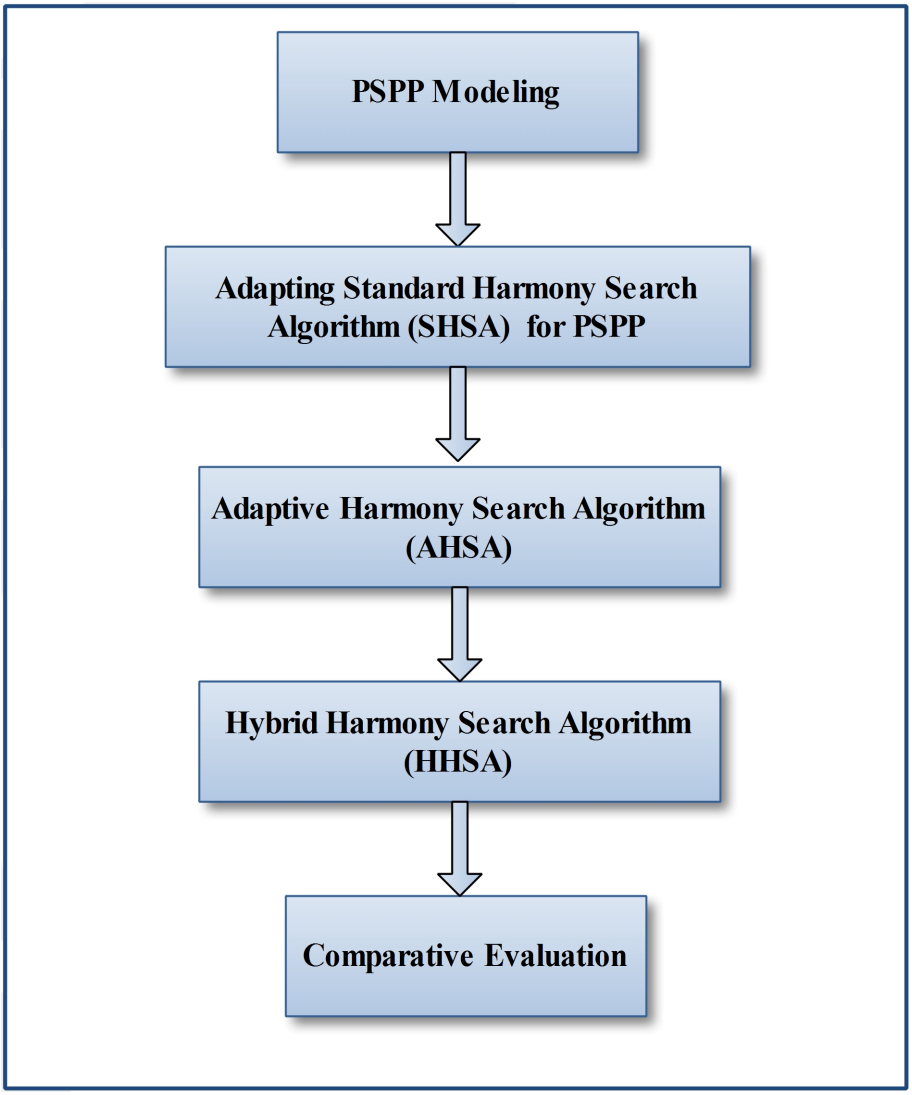

Figure 1.2: A flowchart of the research methodology

1.7 Main Contributions

The main contributions to the thesis are:

1. Adapting Harmony Search Algorithm for protein tertiary structure prediction problem. This will be the first attempt to apply Harmony Search Algorithm for this problem. Henceforward, called Standard Harmony Search Algorithm (SHSA).

2. Introducing two modified harmony search-based algorithms to enhance the performance of the SHSA as follows:

   (a) An Adaptive Harmony Search Algorithm (AHSA) that introduces a new scheme for tuning the two main parameters of SHSA; HMCR and PAR.

   (b) A Hybrid Harmony Search Algorithm (HHSA) that enhances the performance of the AHSA by hybridizing it with two metaheuristic components:

   (i) An iterated local search algorithm to increase the ability of the AHSA to find the local optimal solution in the search space of the new harmony.

   (ii) Global best concept of the Particle Swarm Optimization (PSO) to improve the speed of convergence of the proposed algorithm.

1.8 Thesis Outline

This thesis contains eight chapters organized as follows: Chapter 2 covers a background of proteins and protein structure prediction methods. It also presents an overview of the harmony search algorithm. Chapter 3 is divided into two main sections; the first one reviews the current and related work done in protein structure prediction problem. It also discusses different methods of *ab initio* protein structure prediction modeling. The second section discusses the applications of harmony search algorithm. Chapter 4 defines the modeling and representation

of the PSPP, and it explains an overall methodology of this research. Chapters 5, 6, and 7 discuss the experimental design and results of the SHSA, AHSA and HHSA respectively. Finally, the last chapter provides an overall conclusion and possible future work.

CHAPTER 2

BACKGROUND

2.1 Introduction

Bioinformatics refers to the field concerned with the analysis of biological information using computers and statistical techniques. Research in bioinformatics includes method development for retrieval and analysis of the biological data. It is a rapidly developing branch of biology and is highly interdisciplinary. Bioinformatics has many practical applications in different areas of biology and medicine. In order to apply computing techniques into biological research, computer scientists need to understand the basic terms of the biological research. Thus, this chapter introduces the basic biological terms that are essential for the computer scientists to understand the data that they are dealing with. The focus will be on protein, protein structure, and protein structure prediction.

This chapter consists of three main parts, the first part discusses the basic types of biological data; namely, deoxyribonucleic acids (DNA), ribonucleic acid (RNA), and protein primary sequences; providing more details about protein and its different structure levels. The second part defines the protein structure prediction and explores the two main categories of methods for protein structure prediction; the experimental methods and the computational methods (including the method used in this research which is *ab initio*). The last part of this chapter provides an overview of the harmony search algorithm which is adapted to solve the protein structure prediction problem.

2.2 Types of Biological Data

The basic types of data produced from biological experiments are primary sequence data which can be categorized into three main types; namely, deoxyribonucleic acids (DNA) which is a double-stranded nucleic acid that contains the genetic information, ribonucleic acid (RNA), which is a nucleic acid molecule similar to DNA but containing ribose rather than deoxyribose, and protein primary sequences which is a polypeptide chain made up of different amino acids linked together in a definite sequence. This section gives a detailed description of protein and protein structure prediction methods.

2.2.1 Protein

Proteins are the major components of living organisms; they perform a wide range of essential functions in cells. For example, the haemoglobin in our red blood cells is a protein which is responsible for transporting oxygen around our body. It is made up of four polypeptide chains; two α chains and two β chains as shown in Figure 2.1.

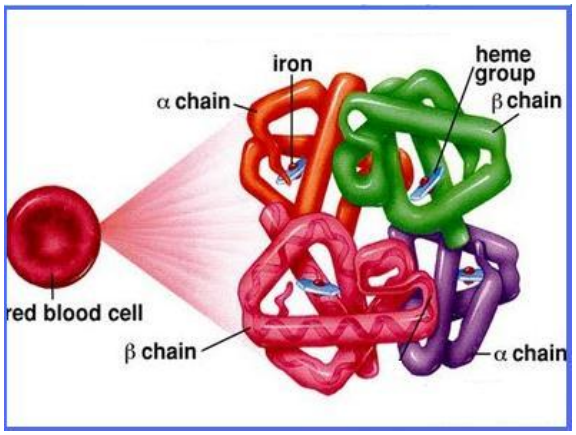

Figure 2.1: Hemoglobin structure (Mader and Wiemerslage, 2000)

Moreover, proteins catalyze the biochemical reactions, regulate and control the metabolic activities, and maintain structural integrity of organisms. Proteins can be classified in different

ways based on their biological functions -as can be seen in Table 2.1. A protein is a polypep-

Table 2.1: Classification of proteins according to biological function (Rosenberg, 2005)

Type	**Example**
Enzymes- Catalyze biological reactions	β-galactosidase
Transport and Storage	Hemoglobin
Movement	Actin and Myosin in muscles
Immune Protection	Immunoglobulins (antibodies)
Regulatory Function within cells	Transcription Factors
Hormones	Insulin, Estrogen
Structural	Collagen

tide chain made up of different amino acids linked together in a definite sequence. Proteins, commonly, contain twenty amino acids; each amino acid has a similar -yet- unique structure. Different proteins have different amino acids; the amino acids sequence, however, is known as the primary structure of the protein. The sequence of those 20 common amino acids found in proteins can be referred to in two ways: the three letters code and the one letter code -as shown in Table 2.2.

Table 2.2: The twenty amino acids in both 3 letters code and 1 letter code (Waterman, 1995)

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

To illustrate, we can refer to a small peptide which contains 8 residues using the three-letter code as: AspIleGluPheArgValLeuHis or as: DIEFRVLH using the one-letter code. Proteins are not linear molecules of amino acid sequence like DIEFRVLH -for example. Rather, this sequence folds into a complex three-dimensional structure which is unique to each protein. This three-dimensional structure allows proteins to function. Thus, in order to understand the protein function, we must understand protein structure (Hill et al., 2000). Most of the amino acids have a carboxyl group and an amino group, the general structure of amino acid is shown in Figure 2.2 [1]; where "R" represents a side chain specific to each particular amino acid, and each amino acid has a different side chain.

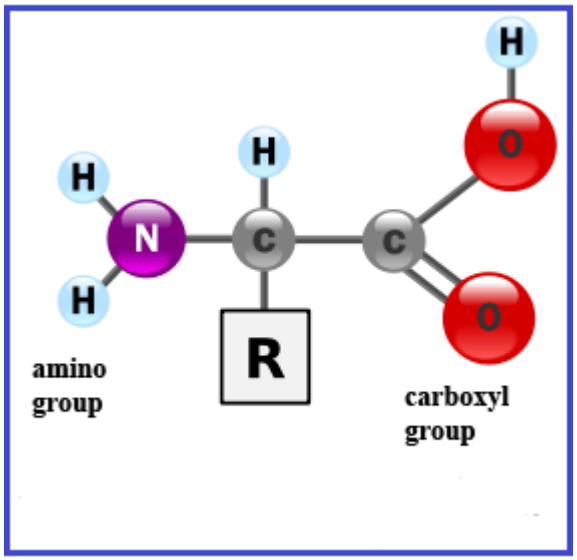

Figure 2.2: Amino acid structure

[1] adapted from *http://homepages.ius.edu/dspurloc/c122/casein.htm*

Amino acids are usually classified by properties of the side chain into four groups: acidic, basic, hydrophilic and hydrophobic. Table 2.3 shows the chemical properties of the side chains for the different 20 amino acids.

Table 2.3: The chemical properties of the side chains for the 20 common amino acids

Amino Acid	**Side chain type**
Alanine	hydrophobic
Arginine	basic
Asparagine	hydrophilic
Aspartate	acidic
Cysteine	hydrophilic
Histidine	basic
Isoleucine	hydrophobic
Glutamine	hydrophilic
Glutamate	acidic
Glycine	hydrophilic
Leucine	hydrophobic
Lysine	basic
Methionine	hydrophobic
Phenylalanine	hydrophobic
Proline	hydrophobic
Serine	hydrophilic
Threonine	hydrophobic
Tryptophan	hydrophobic
Tyrosine	hydrophilic
Valine	hydrophobic

The side chains vary extremely in their complexity and properties; (Akwete Adjei, 1997) for example, the side chain of glycine is simply hydrogen. Figure 2.3 shows the chemical structure of the common amino acids. The protein sequences available in the databases have differ-

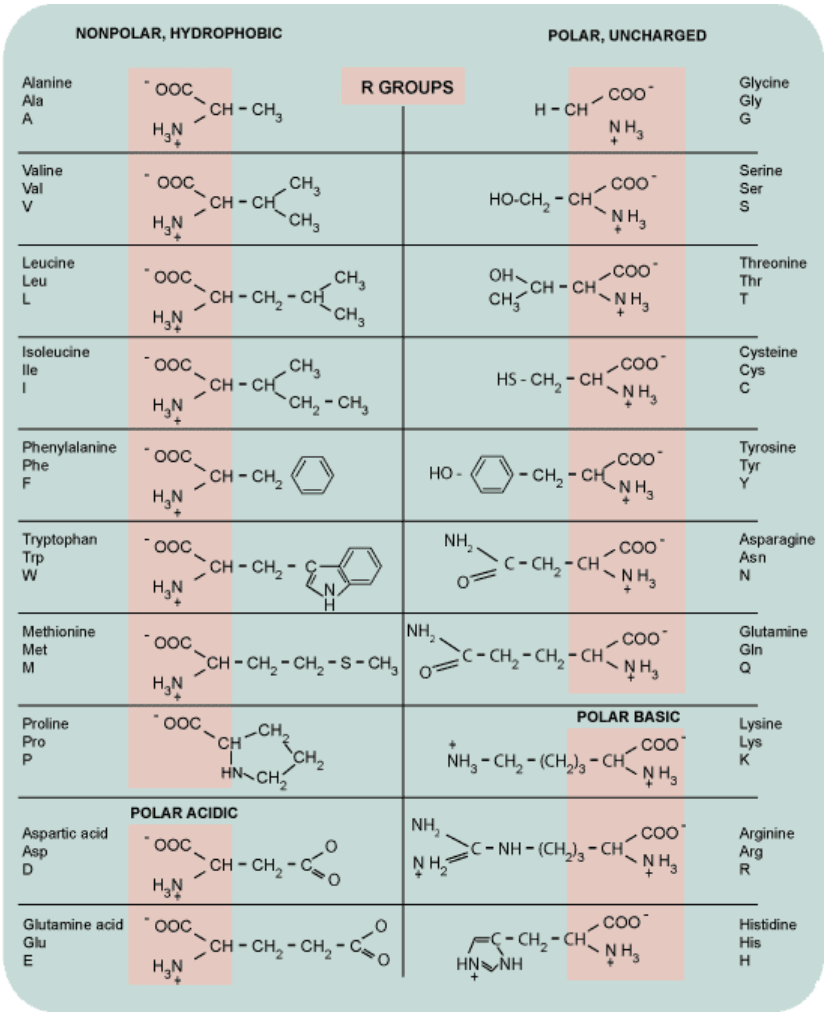

Figure 2.3: The chemical structure of the common amino acids adapted from (Rosenberg, 2005)

ent sizes (size or length of a protein means the number of amino acids). The shortest sequence is Q16047_HUMAN; it has 4 amino acids while the longest sequence is Q3ASY8_CHLCH; it has 36805 amino acids. The average sequence length in UniProtKB/TrEMBL databases is 321 amino acids. Table 2.4 shows the repartition of the sequences by size, Figure 2.4 shows the length distribution of the protein sequences available in UniProt Database.

Table 2.4: Repartition of protein sequences by size (UniProt Database)

Protein length	Number of proteins	Protein length	Number of proteins
1-50	250228	951-1000	52321
51-100	924138	1001-1100	69015
101-150	1064115	1101-1200	48676
151-200	1028833	1201-1300	33186
201-250	1030667	1301-1400	21951
251-300	998232	1401-1500	17645
301-350	907370	1501-1600	12695
351-400	705807	1601-1700	9294
401-450	593429	1701-1800	7431
451-500	496037	1801-1900	5968
501-550	339913	1901-2000	5025
551-600	260966	2001-2100	4052
601-650	189541	2101-2200	4207
651-700	147627	2201-2300	3321
701-750	126824	2301-2400	2615
751-800	113570	2401-2500	2275
801-850	84302	> 2500	19696
851-900	76461		

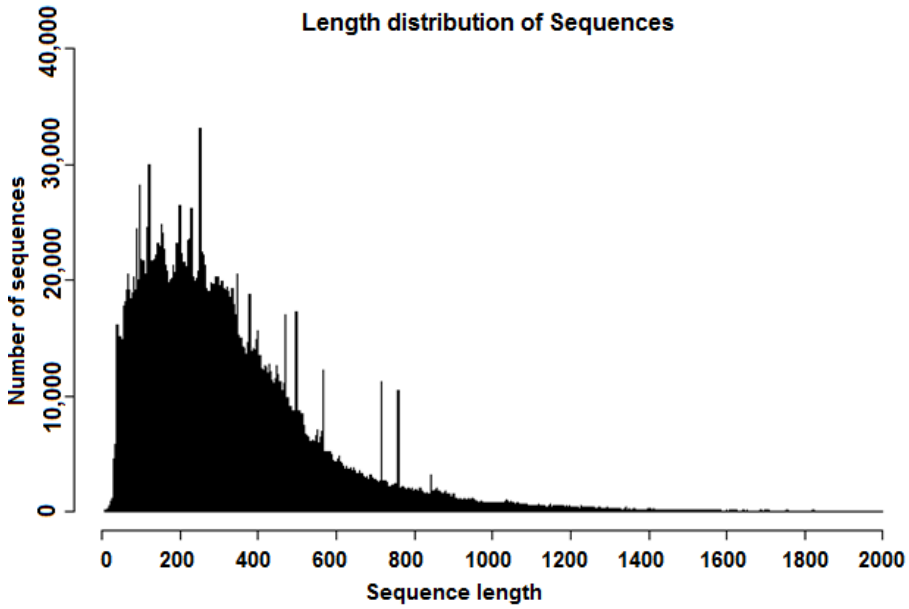

Figure 2.4: Length distribution of protein sequences in UniProtKB/TrEMBL Release 2010_09

2.2.2 Levels of Protein Structures

Protein structure can be described in four hierarchical levels of complexity (Golan, 2008) - Figure 2.5 [2] illustrates this:

[2] adapted from $http://en.wikipedia.org/wiki/File:Main_protein_structure_levels_en.svg$

1. Primary structure: this level refers to the linear sequence of amino acids. The sequence of amino acids in each protein is determined by the gene that encodes it. The gene is transcribed into a messenger RNA (mRNA), and the mRNA is translated into a protein by the ribosome.

2. Secondary structure: this structure refers to the formation of a regular pattern of twists of the polypeptide chain . It is a "local" ordered structure brought about via hydrogen bonding mainly within the peptide backbone. The two most common secondary structure elements in proteins are the alpha (α) helix and the beta (β) sheet.

3. Tertiary structure: this structure refers to the three dimensional structure of the protein sequence it can be described as the global folding of a single polypeptide chain. The folding of the polypeptide chain is stabilized by multiple weak, and non-covalent interactions including: hydrogen bonds, electrostatic interactions among charged amino acid side chains between positive and negative sites on macromolecules, and hydrophobic interactions. When the polypeptide chain folds, the side chains of the polar residues get exposed to the outer surface while the side chains of the non-polar amino acids will hide within the structure.

4. Quaternary structure: this structure involves uniting more than one polypeptide chain to form a multi-subunit structure. This subunits can be formed from the same polypeptide chain or from different ones. For example, Hemoglobin, which transfers oxygen in the blood, is a tetramer which is composed of two polypeptide chains of one type (141 amino acids) and two of a different type (146 amino acids). Not all proteins exhibit quaternary structure; usually, each polypeptide within a multi-subunit protein folds more-or-less independently into a stable tertiary structure. The folded subunits, then, unite together to form the final structure. For some proteins, quaternary structure is required for full activity of the protein.

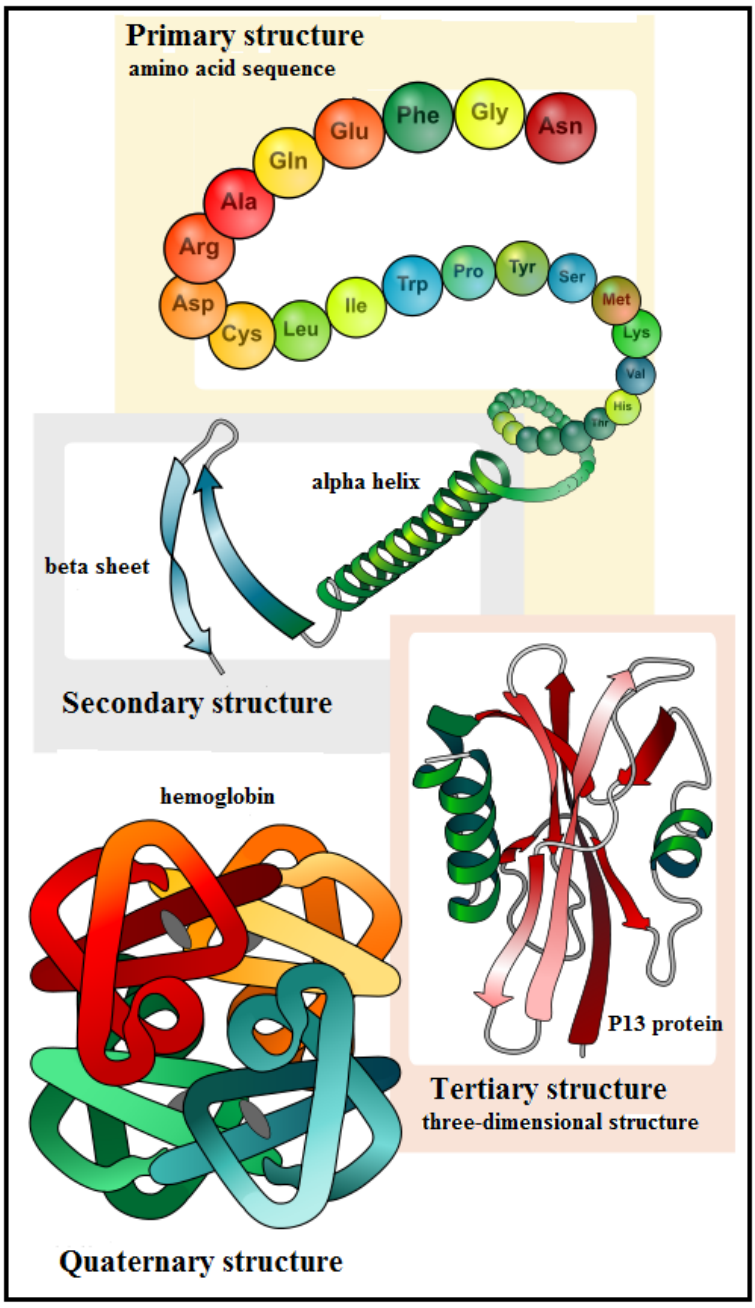

Figure 2.5: The four different levels of protein structure

2.3 Protein Structure Prediction

Predicting the three-dimensional structure of a protein from its linear sequence is a great challenge in the current computational biology. The problem can be described as the prediction of the three-dimensional structure of a protein from its amino acid sequence, or the prediction of a protein's tertiary structure from its primary structure. The protein tertiary structure problem has

been proven to be NP-complete (Berger and Leighton, 1998) (Hoque et al., 2009). There are two methods for protein structure prediction: the experimental methods and the computational methods.

2.3.1 Experimental Methods

In the meantime, there are two main experimental methods available for protein structure prediction: X-ray crystallography and Nuclear Magnetic Resonance (NMR). Unfortunately, these methods are not efficient enough because they are expensive and time-consuming although most of the protein structures available in protein data bank (PDB) are determined by the experimental method.

2.3.1(a) X-ray Crystallography

X-ray crystallography is a method that acts as an atomic microscope, using X-rays instead of visible light to determine the three-dimensional structure of proteins. It is used to determine the arrangement of atoms within a crystal, in which a beam of X-rays strikes a crystal and diffracts into many specific directions.

A crystallographer can produce a three-dimensional picture of the density of electrons within the crystal from the angles and intensities of the diffracted beams. However, the result of a crystallographic experiment is not really a picture of the atoms, it is, rather, a map of the distribution of electrons in the molecule, i.e. an electron density map. This electron density map provides a pretty good picture of the molecule. From this electron density, the mean positions of the atoms in the crystal can be determined -as well as their chemical bonds, their disorder and other various information. Figure 2.6 [3] illustrates the workflow of solving the molecule structure by X-ray crystallography.

[3] adapted from $http://www.answers.com/topic/x-ray-crystallography$

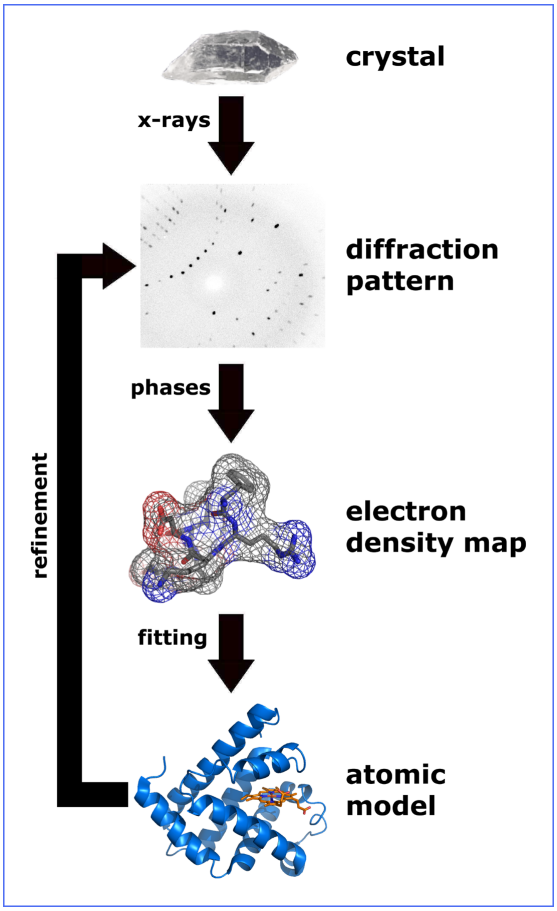

Figure 2.6: Workflow of X-ray crystallography

Crystal structures of proteins began to be solved in the late 1950s, beginning with the structure of sperm whale myoglobin by Max Perutz and Sir John Cowdery Kendrew, their studies in this field enabled them to win the Nobel Prize in Chemistry in 1962. Since that success, more than 55,333 protein structures have been determined by X-ray crystallography which implies that it is the most successful experimental method in protein structure prediction; more than 85% of the protein structures available in protein data bank have been determined using X-ray crystallography. Figure 2.7 shows the number of protein structures predicted by X-ray crystallography yearly in addition to the total number of the protein structures until August 2010.

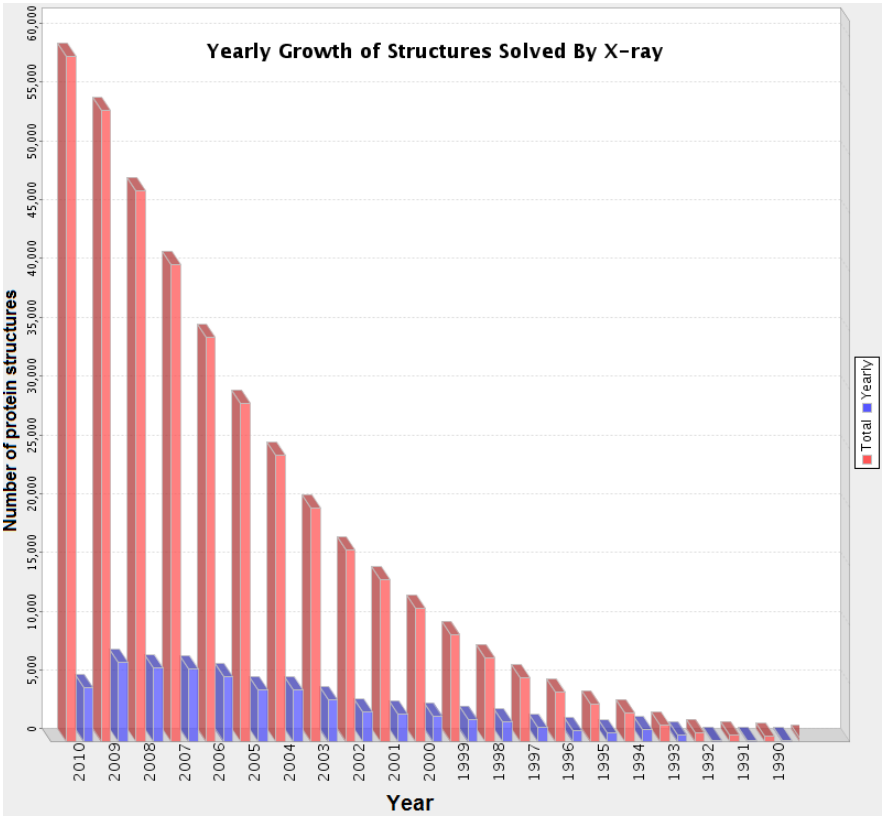

Figure 2.7: Yearly Growth of Structures solved by X-ray crystallography (PDB , August 2010 release)

2.3.1(b) Nuclear Magnetic Resonance (NMR)

Nuclear Magnetic Resonance is a technique relies on the fact that some atomic nuclei are magnetic by nature, like hydrogen -for example. The magnetic nuclei can adopt states of different energies in the magnetic field. Applying radio-frequency radiation can induce the nuclei to flip between these energy states, which can be measured and depicted in the form of a spectrum. Figure 2.8 [4] shows the process of solving protein structure by NMR. The NMR properties of a nucleus depend on its chemical environment; the energy difference between the orientations and the frequency of absorbing energy by the nucleus are some of the NMR properties. Magnetic nuclei are affected by each other through chemical bonds and over short distances through space. This can be manipulated to assign resonance signals to particular nuclei in a complex structure and derive constraints for the distances that separate them (Pietzsch, 2006).

[4] adapted from *http://www.science.org.au/events/sats/sats2004/mackay.html*

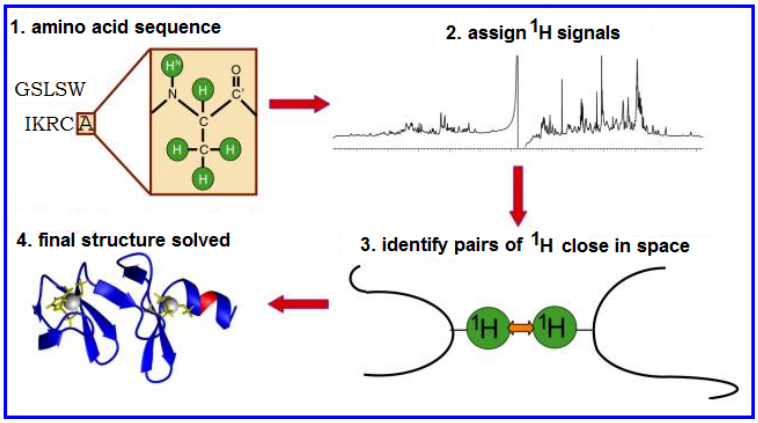

Figure 2.8: Protein structure by NMR

Since it has been first used to determine the three-dimensional structure of a protein in 1984, more than 8546 protein structures have been determined by NMR which is more than 12% of the protein structures available in protein data bank. Figure 2.9 shows the number of protein structures predicted by NMR yearly in addition to the total number of the protein structures until August 2010.

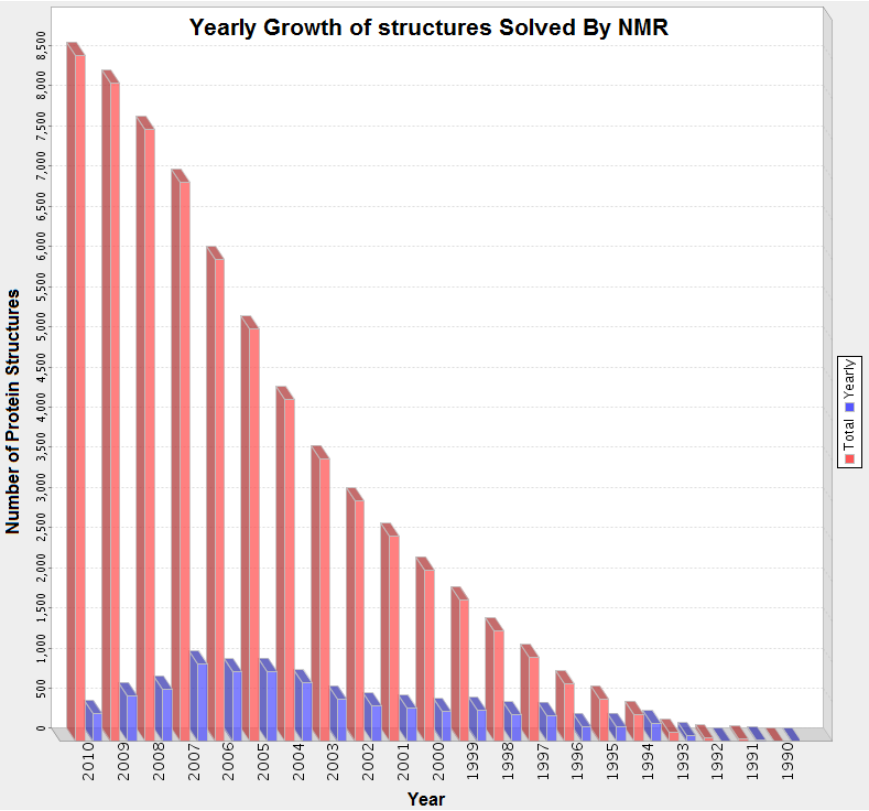

Figure 2.9: Yearly Growth of Structures solved by NMR (PDB , August 2010 release)