UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MÁRCIO DORN

MOIRAE: A Computational Strategy to Predict 3-D Structures of Polypeptides

Thesis presented in partial fulfillment
of the requirements for the degree of
Doctor of Computer Science

Prof. Dr. Luis C. Lamb
Advisor

Prof^{a}. Dr^{a}. Luciana S. Buriol
Coadvisor

Porto Alegre, August 2012

CIP – CATALOGING-IN-PUBLICATION

Dorn, Márcio

MOIRAE: A Computational Strategy to Predict 3-D Structures of Polypeptides / Márcio Dorn. – Porto Alegre: PPGC da UFRGS, 2012.

217 f.: il.

Thesis (Ph.D.) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2012. Advisor: Luis C. Lamb; Coadvisor: Luciana S. Buriol.

1. 3-D protein structure prediction. 2. Artificial neural networks. 3. Genetic algorithms. 4. GA `local-search` operator. 5. Structural bioinformatics. I. Lamb, Luis C.. II. Buriol, Luciana S.. III. Título.

“Das interessanteste an unseren Universum ist, dass man es verstehen kann.”
“O Interessante em nosso universo que podemos entende-lo."
— Albert Einstein

“In Greek mythology, the Moirae (often known in English as The Fates) were three sisters who determined the fate of both gods and humans. The three women were dismal, responsible for spinning (Clotho), weaving (Lachesis) and cut (Atropos) what would be the thread of life of all individuals:
Clotho in Greek means "spinning", held the spindle and weaving the thread of life.
Lachesis in Greek means "sort", pulled and wrapped the cord tissue.
Atropos in Greek means "away", she cut the thread of life.
In this thesis Spinning represents the fact of acquire and combine structural patterns from experimental protein structures, Sort represents the genetic algorithm developed to search the conformation space in order to find the protein native-like structure and Away represents the developed strategy to keep out bad solutions.
”

AGRADECIMENTOS

Primeiramente gostaria de agradecer aos meus pais Marlise e Valdir Dorn por sempre estarem presentes em todas as etapas da minha vida, me apoiando, me incentivando e me inspirando. Sem a ajuda de vocês mais este sonho não teria se tornado realidade. Agradeço também as pessoas que quando do início do meu doutoramento se faziam presentes e que hoje infelizmente já não estão mais entre nós: Vó Helga Kupske e Vô Roberto Kupske (*in memoriam*). Saudades eternas de vocês. Aos avós paternos Reinert e Berta Dorn, anos já se passaram mas a saudade continua. Vocês sempre foram um exemplo de pessoas. Aos demais familiares Marlene, Carlos, Hildegard, Egon, Claudio, Mariza, Morgane e Mônica. Obrigado pelo apoio, carinho e amizade. Aos meus amigos Leonardo Klein, Everton Becker, Eleonor Mayer, Paula Pavan, Melina, Guaraci, Elaine Costa, Neri Costa e Dione Taschetto, com os quais, sempre pude contar. Aos amigos e colegas de laboratório e da UFRGS, Fabiane Dillenburg, Daniel Farenzena, Diego Noble, Rodrigo Machado, Cristiano Reis dos Santos, Rodrigo Kassick, Michelle Leonhardt, Fernando Stefanello, Arton Dorneles, Tiago Primo, Kelen Bernardi e Rafael Parizi pelas conversas, cafés e momentos de descontração.

Aos pais acadêmicos Luis Lamb e Luciana Buriol que me acolheram na UFRGS. Obrigado pela dedicação de vocês, pela paciência, pela confiança, pelo apoio, pelas conversas, pela compreensão e pelo incentivo. Vocês são exemplos de profissionais e de seres humanos. Obrigado por tudo! A todos os técnicos administrativos do INF e do PPGC, em especial Beatriz Haro, Claudia de Quadros, Silvania Vidal, Margareth Schäffer, Leivo Ortiz, Luis Otávio e Elisiane de Almeida. Obrigado também a todos os professores do INF em especial Álvaro Moreira e Lisandro Granville. Ao professor Hugo Verli (CBiot/UFRGS) pelo apoio, incentivo e disponibilidade para conversas e trocas de idéias. Aos professores Mario Inostroza (Universidade do Chile) e Leandro Wives (INF/UFRGS) pelos construtivos comentários sobre o conteúdo da tese. Aos colegas e amigos de outras instituições nacionais e internacionais pelo apoio e incentivo ao longo desta caminhada: André Braga (Universidade de Brasília), Carlos Humberto Llanos (Universidade de Brasília), Martin Baumman (KIT, Alemanha), Florian Wilhelm (KIT, Alemanha), Eva Ketelaer (KIT, Alemanha), Mathias Krause (KIT, Alemanha), Andrea Nestler (KIT, Alemanha), Mario Torres (Red Hat, Alemanha), Federica Ciacca (Itália), Maristela Brizzi (Unijuí), Priscila Grynberg (UFMG), Tálita Sono (UFMG) e Diogo Volpati (UNESP).

CONTENTS

LIST OF ABBREVIATIONS AND ACRONYMS

AMBER	Assisted Model Building with Energy Refinement
A3N	Artificial Neural Network-N-gram based method
ANN	Artificial Neural Network
ASA	Accessible Surface Area
BB	Branch and Bound
BOSS	Biochemical and Organic Simulation System
CATH	Class, Architecture, Topology, Homologous super-family
CASP	Critical Assessment of protein Structure Prediction
CCP	Contact Capacity Potentials
CE	Combinatorial Extension
CHARMM	Chemistry at HARvard Macromolecular Mechanics
CM	Comparative Modeling
CREF	Central-REsidue-Fragment-based method
CSA	Conformation Space Annealing
DDD	Dali Domain Dictionary
DSSP	Define Secondary Structure of Proteins
HMM	Hidden Markov Models
HOOMD	Highly Optimized Object Oriented System
HP	Hydrophobic Polar
IDP	Intrinsically Disordered Proteins
ILP	Integer Linear Programming
FSSP	Fold classification based on Structure-Structure alignment of Proteins
FR	Fold Recognition
GA	Genetic Algorithm
GB	Generalized Born
GB/SA	Generalized Born/Solvent Accessibility

GROMCAS	GROningen MAchine for Chemical Simulations
HP	Hydrophobic Polar
LINUS	Local Independent Nucleated Units of Structure
LS	Local Search
MC	Monte Carlo
MM	Molecular Mechanics
NAB	Nucleic Acid Builder
NACCESS	Atomic Solvent Accessible Area Calculations
NMR	Nuclear Magnetic Resonance
PD	Protein Design
PPA	Profile-Profile Alignment
PDB	Protein Data Bank
PDBj	Protein Data Bank Japan
PDBe	Protein Data Bank Europe
PDF	Probability Density Function
PPA	Profile-Profile Alignment
PLOP	Protein Local Optimization Program
PME	Particle Mesh E-Wald
PREDICT	Profile Enumeration Dictionary
PROFESY	PROFile Enumerating Dictionary
PSP	Protein Structure Prediction
PSS	Potential Smoothing and Search
QM	Quantum Mechanics
QMMM	Quantum and Molecular Mechanics
RAPTOR	Rapid Protein Threading Predictor
REMC	Replica Exchange Monte Carlo
SA	Solvent Accessibility
SANDER	Simulated Annealing with NMR-Derived Energy Restraints
SCOP	Structural Classification of Proteins
SIESTA	Spanish Initiative for Electronic Simulation with Thousand of Atoms
SNMP	Simple Molecular Mechanics for Proteins
3-D	Three-Dimensional

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Currently, one of the main research problems in `Structural Bioinformatics` is associated to the study and prediction of the `3-D` structure of proteins. The 1990's `GENOME` projects resulted in a large increase in the number of protein sequences. However, the number of identified `3-D` protein structures have not followed the same growth trend. The number of protein sequences is much higher than the number of known `3-D` structures. Many computational methodologies, systems and algorithms have been proposed to address the protein structure prediction problem. However, the problem still remains challenging because of the complexity and high dimensionality of a protein conformational search space. This work presents a new computational strategy for the `3-D` protein structure prediction problem. A first principle strategy which uses database information for the prediction of the `3-D` structure of polypeptides was developed. The proposed technique manipulates structural information from the `PDB` in order to generate torsion angles intervals. Torsion angles intervals are used as input to a genetic algorithm with a `local-search` operator in order to search the protein conformational space and predict its `3-D` structure. Results show that the `3-D` structures obtained by the proposed method were topologically comparable to their correspondent experimental structure.

Keywords: 3-D protein structure prediction, artificial neural networks, genetic algorithms, GA `local-search` operator, structural bioinformatics.

MOIRAE: Uma Estratégia Computacional para Predizer a Estrutura 3D de Polypeptídeos

RESUMO

Atualmente, um dos principais problemas de pesquisa na Bioinformática Estrutural está associado com o estudo e a predição de estruturas `3D` de proteínas. Os projetos `GENOMA` resultaram em um grande aumento no número de sequência de proteínas. Entretanto, o número de estruturas `3D` de proteínas não cresceram nas mesmas proporções. O número de sequências de proteínas é muito maior do que o número de estruturas `3D` que são conhecidas. Diversas metodologias computacionais, sistemas e algoritmos foram proposto como uma solução para o problema da predição de estruturas de proteínas. Entretanto, este problema continua sendo desafiador por causa de sua complexidade e pela grande dimensão do espaço de busca conformacional de uma proteína. Este trabalho apresenta uma nova estratégia computacional para o problema da predição de estruturas `3D` de proteínas. Uma estratégia computacional baseados em primeiros principios e que utiliza informações da experimental foi desenvolvida. A técnica proposta manipula informações estruturais do `PDB` como forma a gerar intervalos de ângulos de torção. Intervalos de ângulos de torção são utilizados como entrada em um algoritmo genético com um operador de `busca-local`. Este algoritmo é então utilizado para percorrer o espaço de busca conformacional e predizer a estrutura `3D` de proteínas. Os resultados encontrados mostam que as estruturas `3D` obtidas pelo método proposto são comparáveis topologicamente com as suas respectivas estruturas experimentais.

Palavras-chave: predição da estrutura `3D` de proteínas, redes neurais artificiais, algorítmos genéticos, AG operador de busca local, bioinformática estrutural.

1 INTRODUCTION

Currently, one of the main research problems in Structural Bioinformatics is the prediction of three-dimensional (3-D) protein structures. Knowledge of the protein structure allows the investigation of biological processes more directly, with higher resolution and finer detail. Proteins or Polypeptides are polymers made of 20 different amino acid residues. Each protein is defined by its unique sequence of amino acid residues that under physiological conditions folds into a specific shape known as its native state (ANFINSEN, 1973). Each amino acid residue includes an α carbon (C_α) with bonds to amino (NH) and carboxyl (COOH) groups, and a variable side-chain (R) that gives the specific physicochemical properties of each amino acid residue. A peptide is a molecule composed of two or more amino acid residues chained by a chemical bond called the peptide bond. This bond is formed when the carboxyl group of one residue reacts with the amino group of the other residue, thereby releasing a water molecule (H_2O) (LEHNINGER; NELSON; COX, 2005; TRAMONTANO, 2006; LESK, 2010; LILJAS et al., 2001). The linked carbon, oxygen and nitrogen atoms form the protein backbone. Two or more linked amino acid residues are referred to as a peptide, and larger peptides are generally referred to as polypeptides or proteins (CREIGHTON, 1990; LESK, 2002).

The 1990's GENOME projects resulted in a large increase in the number of protein sequences. Unfortunately, the number of identified 3-D protein structures did not follow the same trend. Currently, the number of protein sequences is far higher than the number of known 3-D structures. If we compare the number of non-redundant sequences[1] of protein sequences stored in GenBank with the number of 3-D protein structures with distinct folds[2] stored in the Protein Data Bank (BERMAN et al., 2000) (PDB[3]) we observe there is a large gap between the number of protein sequences we can generate and the number of new protein folds we can determine by experimental methods such as X-ray diffraction and Nuclear Magnetic Resonance (NMR).

Determining the 3-D structure of a protein is both experimentally expensive (due to the costs associated to crystallography, electron microscopy or NMR), and time consuming. A way to determine accurate protein structures quickly and at low-costs will benefit many life science fields such as medicine, biotechnology and the pharmaceutical industry. Tertiary protein structure prediction is currently one of the challenging problems in Structural Bioinformatics (TRAMONTANO, 2006; ZHANG; VERETNIK; BOURNE, 2005). Predicting the folded structure of a protein

[1] GenBank non-redudant sequences: 16,393,342 on August, 2, 2012.

[2] PDB distinct folds: 1,195 on August, 2, 2012.

[3] www.rcsb.org/pdb

only from its amino acid sequence remains a challenging problem also in mathematical optimization (LANDER; WATERMAN, 1999) and is classified in computational complexity theory as NP-complete problem (CRESCENZI et al., 1998). The challenge arises due to the combinatorial explosion of plausible shapes, where a long amino acid chain ends up in one out of a huge number of 3-D conformations.

1.1 Motivation

Over the last years several computational strategies have been proposed as a solution to the Protein Structure Prediction (PSP) problem (WOOLEY; YE, 2010). These methods can be divided into four classes (FLOUDAS et al., 2006): (I) First principle methods without database information (OSGUTHORPE, 2000); (II) First principle methods with database information (ROHL et al., 2004; SRINIVASAN; ROSE, 1995); (III) Fold Recognition (FR) methods (BOWIE; LUTHY; EISENBERG, 1991; JONES; TAYLOR; THORNTON, 1992; BRYANT; ALTSCHUL, 1995); and (IV) Comparative Modeling (CM) methods (MARTÍ-RENOM et al., 2000). The first group of methods, which cannot rely on sequence similarity to known structures, aims at predicting new folds only through computational simulation of physicochemical properties of the folding process of the proteins in nature. This class of methods uses a concept of a free energy (Anfinsen's Hypothesis) to find the native state of a protein (ANFINSEN et al., 1961; ANFINSEN, 1973).

Groups II, III and IV represent the methods that are capable of making fast and effective prediction of protein 3-D structures when known template structures and fold libraries are available (KOLINSKI, 2004). In first principle methods with database information, general rules of protein structures are extracted from protein databases and used to build starting point 3-D protein structures. ROBETTA (ROHL et al., 2004; SIMONS et al., 1999B), I-TASSER (ZHANG, 2007) and LINUS (SRINIVASAN; ROSE, 1995) are examples of methods belonging at this group. Comparative modeling by homology can be applied whenever it is possible to detect a sequence evolutionary relationship between the target protein and the template protein of which the 3-D structure is known (SÁNCHEZ; SALI, 1997). The structure of these proteins are similar in the sense that amino acid residues with identical physicochemical properties and structure occupy the same position in homologous proteins. Fold Recognition methods are motivated by the notion that structure is more stable than sequence, i.e., proteins with no similar sequences could have similar folds. Fold Recognition methods are focused on predicting the 3-D folded structure of protein amino acid sequences for which comparative methods provide no reliable predictions. Fold-recognition *via* threading is limited to the fold library (KOLINSKI, 2004) derived from the PDB.

The most significant progress in last CASP[4] (9th edition) was identified by template-based modeling methods (methods that use database information) (KOOP et al., 2007; COZZETTO et al., 2009; ZHANG, 2008B; XU et al., 2011). Nevertheless, as observed in the experiments, the major challenge remains in the development of better methods for template production and identification (SODING, 2005); accurate structure for those regions are not easily derived from an obvious templates. In CASP9 not much progress in first principle (*ab initio*) methods without database information was observed (JAUCH et al., 2007; BEN-DAVID et al., 2009; FLOUDAS

[4] Critical Assessment of Structure Prediction. predictioncenter.org

et al., 2006; XU et al., 2011).

1.2 Contributions

Despite the significant progress in last CASP, it is still necessary the development of new strategies for extracting, representing and manipulating data from experimentally determined 3-D protein structures, as well the development of computational strategies to use this information in order to predict, from the amino acid sequence of a protein, its corresponding 3-D structure. The development of computer prediction methods which reduce the computational effort and allow the prediction of the three-dimensional structure of proteins is presented as one of the main challenges in Structural Bioinformatics and molecular biology of the XXI century.

This thesis presents a new strategy for the 3-D protein structure prediction problem. A first principle strategy which uses database information for the prediction of the 3-D structure of polypeptides was developed. The proposed technique manipulates structural information from the PDB in order to generate torsion angles intervals. Torsion angles intervals are used as input to a genetic algorithm with a local-search operator in order to search the protein conformational space and predict its native-like 3-D structure. The main contributions of this work are:

- The development of a new computational strategy to collect and represent structural information from experimentally determined protein structures;
- The development of a genetic algorithm with a local-search operator to search the protein three-dimensional conformational space in order to find the native-like 3-D structure of protein sequences;
- The development of a fragment-based strategy combined with ab-initio concepts to predict 3-D structures of proteins.

1.3 Thesis organization

This thesis is structured as follows. Chapter 2 provides the fundamental concepts of proteins, amino acids, peptide bond, structural levels and structural databases (readers familiar with these fundamental concepts can clearly skip this chapter). Chapter 3 shows fundamental concepts of protein kinematics and describes the structural representation of the 3-D structure of proteins employed in this work. Chapter 4 describes the four classes in which the 3-D protein structure prediction methods and algorithms are classified. In addition, we present details of the main prediction methods and outline the computational strategies that they use. Chapters 5 and 6 describes the developed computational strategy for the PSP problem. Chapter 7 presents the experiments, results and the discussion of the obtained results. Finally, Chapter 8 concludes the thesis and point out directions for further research.

2 ON PROTEINS

2.1 Introduction

From a structural perspective, a protein is an ordered linear chain of building blocks known as amino acids. Each protein is defined by its unique sequence of amino acid residues that causes the protein to fold into a particular three-dimensional (3-D) shape. This shape or fold gives the protein its specific biochemical properties, i.e., its function (LILJAS et al., 2001; LESK, 2010).

Knowledge of the protein structure allows the investigation of biological processes more directly, with higher resolution and finer detail. The sequence-protein-structure paradigm (also knows as the "lock-and-key" hypothesis) says that the protein can achieve its biological function only by folding into a unique, structured state determined by its amino acid sequence (ANFINSEN, 1973). Nevertheless, currently it has been recognized that not all protein functions are associated to a folded state (DUNKER et al., 2008; UVERSKY, 2001; TOMPA; CSERMELY, 2004; TOMPA, 2002; WRIGHT; DYSON, 1999; DUNKER et al., 2001). For some cases proteins must be unfolded or disordered to perform their functions (GUNASEKARAN et al., 2003). These proteins are called intrinsically disordered proteins (IDP) and represent around 30% of the protein sequences. Despite the presence of IDP proteins an important aspect of understanding and interpreting the function of a given protein involves characterizing molecular interactions. These interactions can be intra-molecular (ionic bonds, covalent bonds, metallic bonds) or intermolecular (hydrogen bonds and other non-covalent bonds such as van der Waals force). The knowledge of the 3-D structure of the polypeptides gives researchers very important information to infer the function of the protein (BRANDEN; TOOZE, 1998; LASKOWISKI; WATSON; THORNTON, 2005,B). There are a variety of proteins that plays functions on the cell (LESK, 2010): structural proteins; enzymes that catalyze chemical reactions; antibodies that recognize and repel invading pathogens; regulatory proteins; sensors; transporters and transducers that convert chemical to mechanical energy.

The determination of protein structure is both experimentally expensive (due to the costs associated to crystallography or NMR), and time consuming. The difficulty in determining and finding out the 3-D structure of proteins has generated a large discrepancy between the volume of data (sequences of amino acid residues) generated by the GENOME projects[1] and the number of 3-D structures of proteins which are known nowadays. This not only clearly illustrate the need for, but also motivate

[1] DOE Genomic Science. genomics.energy.gov .

further research in computational protein structure prediction methods.

2.2 Amino acid residues and their structures

An amino acid residue is a small molecule containing an amino group (H_3N^+), a carboxyl group ($COOH^-$), and a hydrogen atom attached to a central alpha carbon (C_α). In addition, each amino acid also has an R organic group (also called side-chain) attached to the C_α (Fig. 2.2). In chemistry an amino acid residue is represented as H3NCHRCOOH. The group R distinguishes one amino acid from another and confers the chemical properties of each amino acid residue (Fig. 2.1).

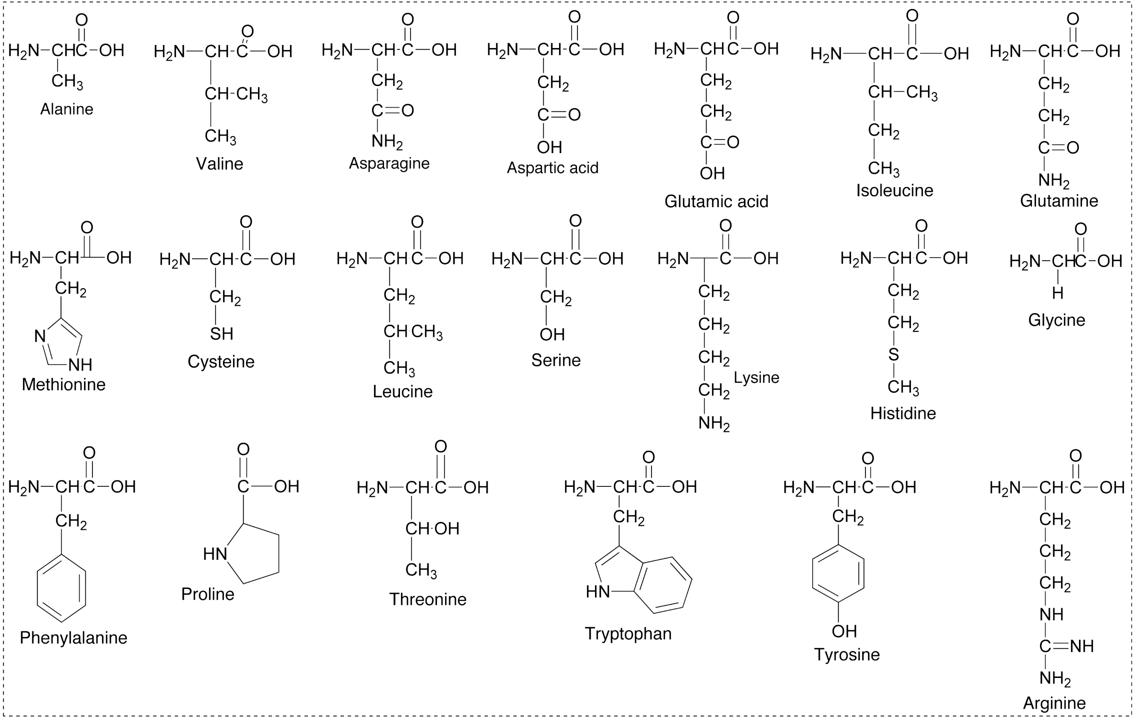


Figure 2.1: Chemical representation of the 20 amino acid residues. The side-chains of the amino acids vary in terms of size; electric charge and polarity. The physical properties of the amino acid side-chains influence interactions in the 3-D polypeptide structure.

In nature there are 20 distinct amino acid residues (Fig. 2.1, Tab. 2.1), each one with its own chemical properties (LODISH et al., 1990). Depending on the polarity of the side-chain, amino acids vary in their hydrophilic or hydrophobic character. The side-chains of the amino acids vary in Lesk (LESK, 2002): size (number of atoms); electric charge (some side-chains bear a net positive or negative charge at normal pH); polarity (some side-chains are polar; they can form hydrogen bonds to another polar side-chains, or to the main-chain, or to water). The importance of the physical properties of the side-chains comes from the influence they have on the amino acid residues interactions in the structure. The distribution of the hydrophilic and hydrophobic amino acids are important to determine the tertiary structure of the polypeptide.

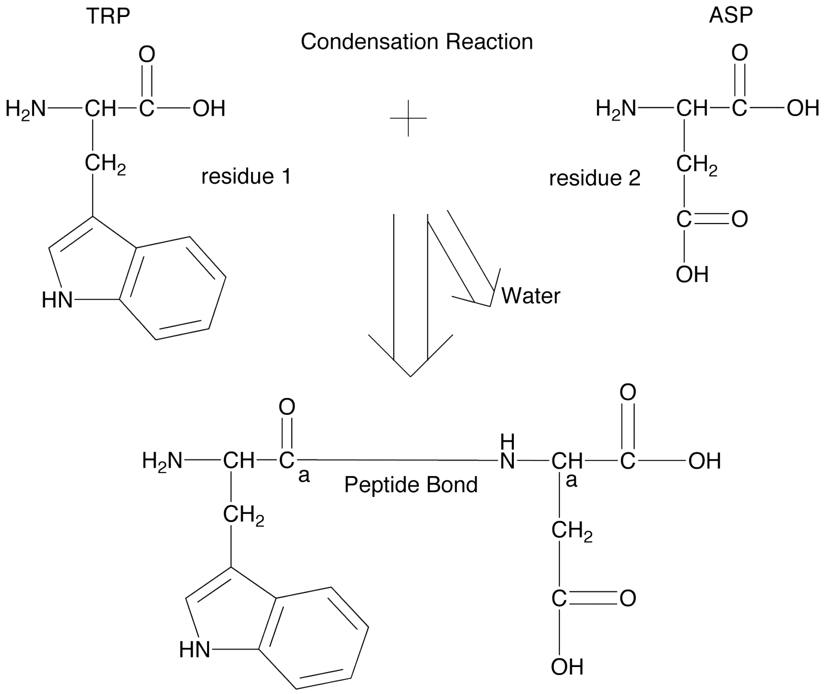


Figure 2.2: Chemical representation of two amino acid residues and the condensation reaction. The carboxyl group of one amino acid (amino acid 1) reacts with the amino group of the amino acid 2. A molecule of water is removed from two amino acids to form a peptide bond. N is nitrogen, C and C_α are carbons.

2.3 The peptide bond

A peptide is a molecule composed of two or more amino acid residues chained by a chemical bond called the peptide bond (Fig. 2.2). This peptide bond is formed when the carboxyl group of one residue reacts with the amino group of the other residue, thereby releasing a water molecule (H_2O). Two or more linked amino acid residues are referred to as a peptide, and larger peptides are generally referred to as polypeptides or proteins (CREIGHTON, 1990; LESK, 2002) (Fig. 2.3). Peptides generally contain fewer than 20-30 amino acid residues, whereas polypeptides contain as many as 4000 residues.

In a peptide or polypeptide all atoms from the group R are referred to as side-chain and the remaining atoms are referred to as the peptide backbone. The specific characteristics of the peptide bond have important implications for the 3-D fold that can be adopted by polypeptides. The peptide bond (C-N) has a double bond and is not allowed rotation of the molecule around this bond. The rotation is only permitted around the bonds N-Cα and C_α-C. These bonds are known as PHI (ϕ) and PSI (ψ) dihedral angles and are free to rotate (LESK, 2002; LODISH et al., 1990). This freedom is mostly responsible for the conformation adopted by the polypeptide backbone. However, the rotational freedom around the ϕ (N-C_α) and ψ (C_α-C) angles is limited by steric hindrance between the side-chain of the amino acid residue and the peptide backbone (BRANDEN; TOOZE, 1998; LESK, 2002; SCHEEF; FINK, 2003). As a consequence, the possible conformation of a given polypeptide is quite limited and depends on the amino acid chemical properties. The peptide bond itself tends to be planar, with two allowed states: trans, $\omega \simeq 180°$ (usually) and cis,

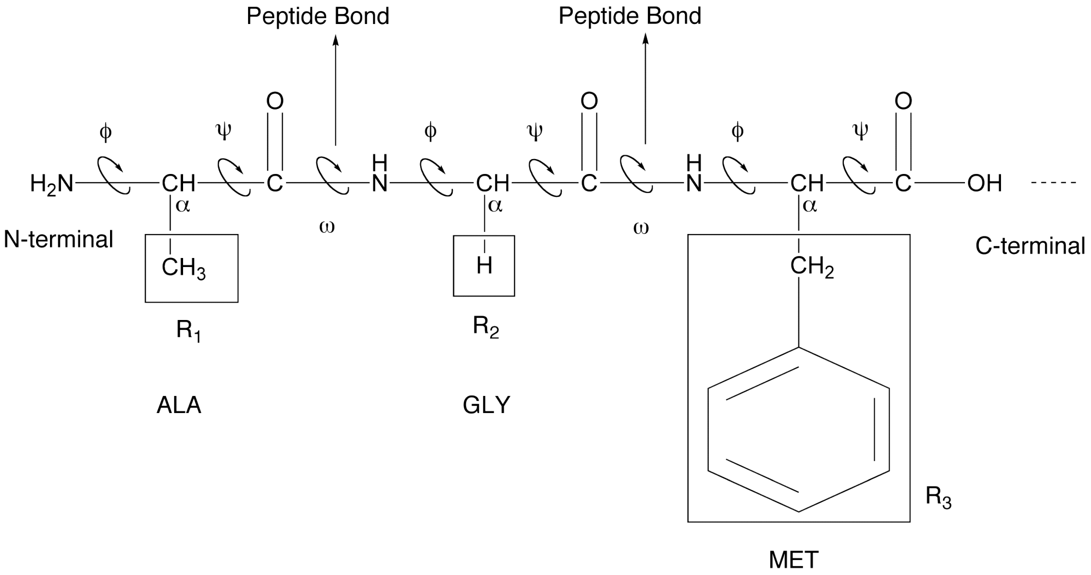


Figure 2.3: Schematic representation of a model peptide. N is nitrogen, C and C_α are carbons.

$\omega \simeq 0°$ (rarely) (BRANDEN; TOOZE, 1998; LESK, 2002). The sequence of ϕ, ψ and ω angles of all residues in a protein defines the backbone conformation or fold (HOVMOLLER; OHLSON, 2002).

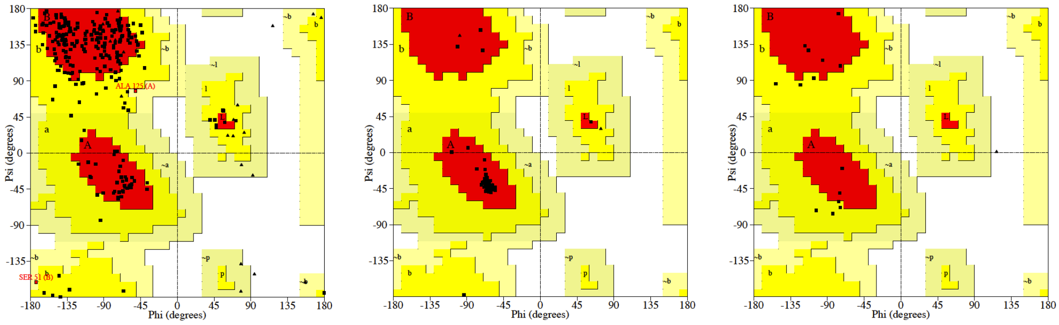
The angles ϕ and ψ can have any value between -180° and +180°. However, some combinations are prohibited by steric interferences between atoms from the main-chain and atoms from the side-chain (two atoms cannot occupy the same space) (HOVMOLLER; OHLSON, 2002). The allowed and prohibited values for the torsion angles ϕ and ψ are graphically demonstrated by the map of Sasisekharan-Ramakrishnan-Ramachandran[2], or simply Ramachandran plot (RAMACHANDRAN; SASISEKHARAN, 1968) (Fig. 2.4). The red, yellow, and light yellow regions represent the favored, allowed, and "generously allowed" regions as defined by the software PROCHECK (LASKOWSKI et al., 1996). Black dots represent each amino acid residue of the protein. There are two main allowed regions of residue conformation: α and β. These regions correspond to the mayor types of secondary structures: α-helix and β-sheet. Secondary structures will be detailed in Section 2.4.

2.3.1 Side-chain conformations

Similar to the polypeptide backbone, side-chain have also dihedral angles. The number of angles χ of the side-chain depends on the amino acid type. Table 2.1 present the number of χ dihedral angles of each of the 20 amino acid residues. The group R of each amino acid residue gives its unique proprieties. The amino acid residues can be divided into groups according to their proprieties (LILJAS et al., 2001):

- Non-polar: amino acid residues with hydrophobic side-chain. Side-chains which have pure hydrocarbon alkyl groups (alkane branches) or aromatic

[2] The Ramachandran plot is a plot of the torsional angles phi and psi of the residues (amino acids) contained in a peptide.



(a) PDB ID: 1AIL (b) PDB ID: 3CRE (c) PDB ID: 1NIL

Figure 2.4: Sasisekharan-Ramakrishnan-Ramachandran plot of protein with PDB ID: 1AIL (q), PDB ID: 3CRE (b) and PDB ID: 1NIL (c). (a) Protein 1AIL composed in major part of residues in a α-helix state. (b) Protein 3CRE composed in major part of residues in a β-sheet state. (c) Protein 1NIL composed in mayor part by residues in coil state. Illustrations were prepared with PROCHECK (LASKOWSKI et al., 1996).

(benzene rings) are non-polar. The side-chain of this amino acid residues contributes to the formation of hydrophobic interactions within the protein;

- Charged Polar: amino acid residues with side-chains which have various functional groups such as acids, amides, alcohols, and amine are polar. Glutamate, Aspartate are amino acids that are usually negative at physiological pH. Arginine, Lysin are amino acids that are usually positive at physiological pH. Histidine is sometimes positive at physiological pH;

- Uncharged Polar: acid residues that are more water soluble than non-polar amino acids, they contain functional groups that form hydrogen bonds with water.

These proprieties contribute to determine the fold of the protein and also to determine the surface proprieties of the molecule. This is important for selective interactions with other molecules and catalysis of chemical reactions (LESK, 2010). The side-chains of one amino acid residue reacts with the side-chain of other amino acid residue in many different ways. The hydrophobic side-chains interact with other hydrophobic side-chains (LILJAS et al., 2001). Polar side-chains can form hydrogen bonds to each other residue or to the main chain atoms. Charged groups frequently interact with side-chains with opposite charge on the surface of the protein.

Different conformations of any side-chain are called Rotamers (LESK, 2002). Rotamers libraries are collections of side-chain dihedral angles used specifically for optimize the position of side-chain dihedral angles. It consists of information of side-chain orientations of protein structures determined experimentally. Due to the steric constraints of the dihedral angles the side-chain of one amino acid residue can assume only certain conformations (LESK, 2002). Rotamers libraries are useful for modelling protein structures because it reduce the number of possible conformations that an amino acid side-chain can assume. The most common library is the Dunbrack rotamers library[3] (DUNBRACK JR.; COHEN, 1997; DUNBRACK JR.; KARPLUS,

[3] Dunbrack Rotamer library. dunbrack.fccc.edu/bbdep.

2003).

Table 2.1: The 20 amino acid residues. Number of χ angles in each of the 20 amino acids residues (3rd Column). Amino acid residues can be divided in groups according to their properties (4-8 columns).

Amino	3-letter	1-letter	N°.	Non-polar	Charged	Uncharged
Acid	code	code	χ angles		polar	polar
Alanine	ALA	A	-	●		
Arginine	ARG	R	4		●	
Asparagine	ASN	N	2			●
Aspartic Acid	ASP	D	2		●	
Cysteine	CYS	C	1			●
Glutamic Acid	GLU	E	3		●	
Glutamine	GLN	Q	3			●
Glycine	GLY	G	-			
Histidine	HIS	H	2		●	
Isoleucine	ILE	I	2	●		
Leucine	LEU	L	2	●		
Lysine	LYS	K	4		●	
Methionine	MET	M	3	●		
Phenylalanine	PHE	F	2	●		
Proline	PRO	P	-	●		
Serine	SER	S	1			●
Threonine	THR	T	1			●
Tryptophan	TRP	W	2			●
Tyrosine	TYR	Y	2			●
Valine	VAL	V	1	●		

2.4 Description of protein structures

Proteins can be studied in four levels (LEHNINGER; NELSON; COX, 2005; LODISH et al., 1990): (i) primary structure, (ii) secondary structure, (iii) tertiary structure and (iv) quaternary structure. This hierarchy facilitates the description and the understanding of proteins. However, it does not aim at describing precisely the physical laws that produce protein structures; it is an abstraction that aims at making protein structure studies more tractable (SCHEEF; FINK, 2003).

2.4.1 Primary structure

The primary structure simply describes the sequence of amino acid residues in a linear order (BRANDEN; TOOZE, 1998; LEHNINGER; NELSON; COX, 2005; LESK, 2002; LODISH et al., 1990). Each amino acid residue binds to other amino acid residue through a peptide bond. The beginning of the primary structure corresponds to its N-terminal region and the end of its primary structure is the C-terminal region (Fig. 2.3).

2.4.2 Secondary structure

Proteins are linear polymers that can assume several conformations. The stable arrangement of amino acid residues of the polypeptide forms structural patterns (LEHNINGER; NELSON; COX, 2005). These structural patterns represent the secondary structure of a polypeptide.

The secondary structure is defined by the presence of hydrogen bond patterns between the hydrogen atoms of the amino groups and the oxygen atoms of the carboxyl groups in the polypeptide chain. A regularity in the spatial conformation is maintained through these intermolecular interactions. There are two most commonly secondary structures: α-helices (PAULING; COREY; BRANSON, 1951) and β-sheets (PAULING; COREY, 1951). There are other periodic conformations (coils and turns), but the α-helix and β-sheets are the most stable and can be considered as the main elements present in 3-D structures.

α-**Helices**: this structure is stabilized by one hydrogen bond between the Nitrogen (N) atom of a peptide bond and the Oxygen (O) atom of the carboxyl group in the fourth amino acid residue of the N-terminal region (LEHNINGER; NELSON; COX, 2005; PAULING; COREY; BRANSON, 1951) (Fig. 2.5 a - red). Each successive turn of helix is held with the adjacent turns by three or four hydrogen bonds. These hydrogen bonds when combined, ensure the stability of the helical structure. Some residues form α-helices better than others. The α-helices have on average 3.6 amino acid residues per turn. The amino acids Alanine, Glutamine, Leucine, and Methionine are commonly found in α-helices. The amino acids Proline, Glycine, Tyrosine, and Serine usually do not occur in α-helix structures (BAXEVANIS; QUELLETTE, 1990; BRANDEN; TOOZE, 1998). Proline is commonly thought of as a α-helix breaker because its bulky ring structure disrupts the formation of n + 4 hydrogen bonds (BAXEVANIS; QUELLETTE, 1990). The number of amino acid residues in an α-helix is highly variable and may be in the range of 5 to 40 amino acid residues - commonly, α-helices present 10 amino acid residues (PAULING; COREY; BRANSON, 1951). The amino acid residues present in a α-helix have their dihedral angles (ϕ and ψ) ranging from around -30° to -120° for ϕ and from -60° to -20° for ψ in the Ramachandran plot (HOVMOLLER; OHLSON, 2002) (Fig. 2.4 - a).

β-**sheets**: when the polypeptide structures are arranged side by side they form a regular structure similar to a series of sheets (PAULING; COREY, 1951) (Fig. 2.5 a - green). The β-sheets consist of extended polypeptide chains with neighboring chains extending parallel/anti-parallel to each other. The amine and carboxyl groups of peptide bonds point towards each other in the same plane, so hydrogen bonding can occur between adjacent polypeptide chains. The amino acid residues present in a β-sheet have their dihedral angles (ϕ and ψ) ranging from around -180° to -45° for ϕ and from 45° to 225° for ψ in the Ramachandran plot (HOVMOLLER; OHLSON, 2002). Adjacent β-sheets can form hydrogen bonds in anti-parallel or parallel arrangements (BRANDEN; TOOZE, 1998; LEHNINGER; NELSON; COX, 2005; LESK, 2002; LODISH et al., 1990). In an anti-parallel model the successive β-strands alternate directions so that the N-terminal of one sheet is adjacent to the C-terminal of the next. In a parallel arrangement, all successive N-terminal regions are oriented in the same direction. The hydrogen bonding patterns are different in the anti-parallel and parallel β-sheets (PAULING; COREY, 1951) (Fig. 2.4 - b).

`Coil` and `turns`: the third type of secondary structure is an irregular secondary structure, called `coil` or `turn` (Fig. 2.5 a - gray). These structures are formed in regions where the polypeptide changes its directions, i.e., after a regular secondary structure in the form of α-`helix` and β-`sheets`. `Turns` and `coils` differ mainly by the number of amino acid residues in the irregular structure. `Coils` present a smaller number of amino acid residues than `turns`. `Turns` and `coils` are the structural elements that bind successive regular secondary structures. For irregular structures there is no specific region in the `Ramachandran plot`. The combination of angles ϕ and ψ can occur in any area of the `Ramachandran plot` and that includes regions of β-`sheets` and α-`helices` (HOVMOLLER; OHLSON, 2002) (Fig. 2.4 - c). Because of this particularity, `turns` and `coils` are difficult to predict by computational methods.

2.4.3 Tertiary structure

The tertiary structure of a protein is represented by the distribution of secondary structures in a `3-D` space (Fig. 2.5 a). The three-dimensional shape assumed by a protein is also called `native` or `functional structure`. The native structure of a protein is formed by the variation of thermodynamic factors, i.e., covalent interactions, hydrogen bonds, hydrophobic interactions, electrostatic interactions, `van der Waals`, and repulsive forces (GIBAS; JAMBECK, 2001; LEHNINGER; NELSON; COX, 2005; LESK, 2002; LODISH et al., 1990). In addition, the side-chains play an important role in creating the final structure of the polypeptide (SCHEEF; FINK, 2003). Through the tertiary structure of a protein it is possible to analyze or infer the function of the protein in the cell. It is possible to identify the active site, binding sites on a receptor, or a recombination site for the action of another protein (LEHNINGER; NELSON; COX, 2005). The tertiary structure of a protein is related to its topology (or fold). The topology of a protein is given by the type of succession of secondary structures that are connected to and from the shape in which these structures are organized in a `3-D` space.

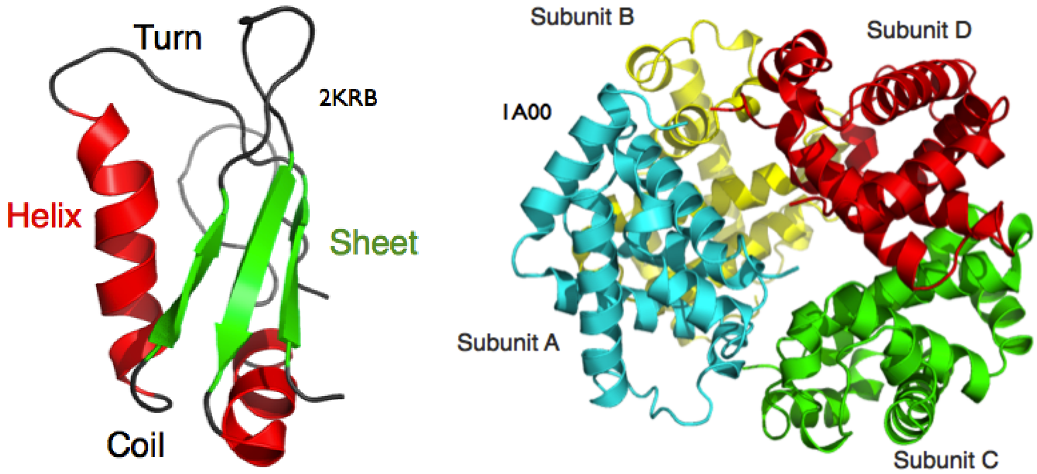
2.4.3.1 Stabilization of the native state

There are many factors that stabilizes the native states of proteins (LESK, 2002):

- `Covalent interactions`: occurs when atoms share one pair of electrons. Covalent bonds alter the nature of the atoms involved. In proteins the covalent bonds are responsible for keeping together an amino acid to other by peptide bonds in the main-chain. Another example are the disulphide bridges, between cysteine residues (LESK, 2002).

- `Hydrogen bonds`: stabilizes and orient chemical groups with regard to one another. Secondary structures α-`helices` and β-`sheets` achieve hydrogen bonds formation by the backbone atoms. A hydrogen bond is formed by a proton interacting with two adjacent electronegative atoms with electron lone pairs called the donor and acceptor.

- `Van der Waals`: refers to intermolecular forces arising from polarization of the molecules. These interactions are very weak and act only when the molecules are very close to each other. The large number of `van der Waals` interactions



(a) Secondary and tertiary structure

(b) Quaternary structure

Figure 2.5: The three-dimensional structure of a protein. (a) The elements of secondary structure are usually folded into a compact shape using a variety of `loops` and `turns`. α-`helices`, β-`sheets` and irregular structures (`coil` and `turns`) are highlighted. (b) Quaternary structure of `Hemoglobin`. Ribbon structure illustrations were designed with `PYMOL` (The `PYMOL` molecular graphics system. Delano Scientific, San Carlos, CA, USA - `www.pymol.org`).

in the macromolecules contributes significantly to structure stability (LILJAS et al., 2001).

- `Hydrophobic effect`: the hydrophobicity of an amino acid residue represent the measure of the interaction between the amino acid side-chain and water. The hydrophobicity scale of different amino acid residue inside proteins contributes to the protein stability. The accessible surface are of a protein measure the thermodynamic interaction between the protein and the water (LESK, 2002) and its contribution in protein folding.

2.4.4 Quaternary structure

A protein may have different polypeptide chains (or subunits) forming a quaternary structure (Fig. 2.5 B). The quaternary structure of a protein is the arrangement of various tertiary structures. This structure is maintained by the same forces that determine the secondary and tertiary structures (hydrogen bonding, hydrophobic interactions, hydrophilic interactions) (LEHNINGER; NELSON; COX, 2005; LESK, 2002; LODISH et al., 1990).

2.5 Protein taxonomy

Proteins can be classified into groups based on their structural and evolutionary relationships. The evolutionary relationship between proteins is a fundamental factor in prediction methods. The structure of a protein is similar to another one in the sense that amino acid residues with identical physiochemical properties occupy the same position in homologous proteins and consequently present in some cases a similar `3-D` structure. Evolutionary related proteins have similar sequences and naturally occurring homologous proteins have similar protein structure. Three-dimensional protein structures are evolutionary more preserved than sequences (KACZANOWSKI; ZIELENKIEWICZ, 2010).

A widely used classification scheme consists of three groups: `family, super family, and fold` (MURZIN et al., 1995). The most general classification of protein families is based on the secondary and tertiary polypeptide structures (LESK, 2002; LEVITT; CHOTHIA, 1976; MOUNT, 2001). This classification allows a quantitative measure of the structural differences of proteins and distinguishes them according to their folding patterns:

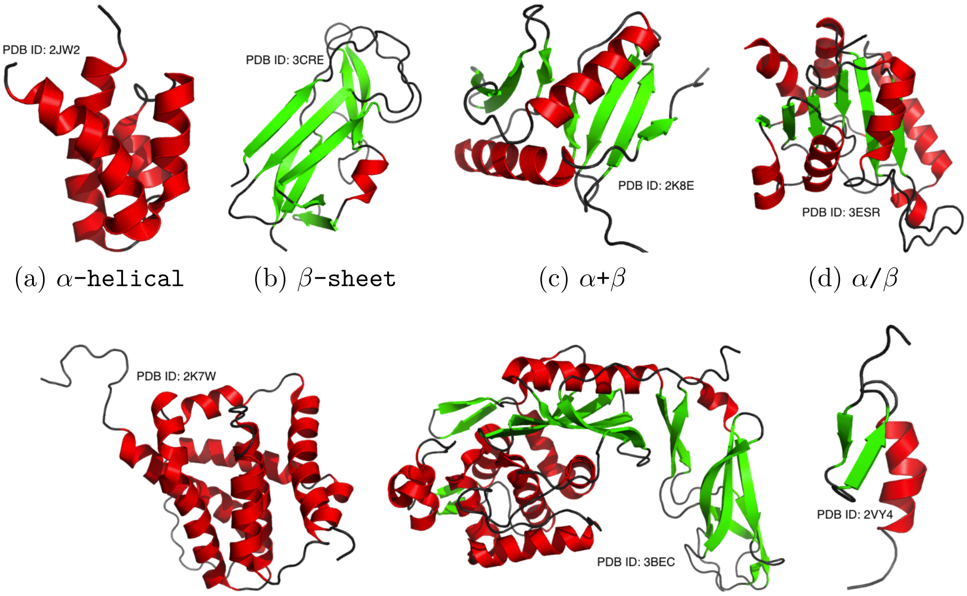
- `Family`: protein structures that display a clear evolutionary relationship;
- `Super family`: protein structures that exhibit probable and common evolutionary origin;
- `Fold`: protein structures that present strong structural similarity.

Various databases are constructed based on the classification of protein structures. The most common databases are: `SCOP`[4] - Structural Classification of Proteins (LO CONTE et al., 1999), `CATH`[5] - Class, Architecture, Topology, Homologous super-family (ORENGO et al., 1997, 1999), `FSSP/DDD` - Fold classification based on Structure-Structure alignment of Proteins/Dali Domain Dictionary (HOLM et al., 1992) , and `CE` - the Combinatorial Extension Method (SHINDYALOV; BOURNE, 1998). `SCOP` is one of the most important and widely used classification databases. It organizes the structure of proteins based on their evolutionary origin and structural similarity. `SCOP` has seven classes in which proteins or polypeptides are grouped (Fig. 2.6):

1. α-`helical`: secondary structure exclusively or almost exclusively α-`helical`;
2. β-`sheet`: secondary structure exclusively or almost exclusively β-`sheet`;
3. α+β: α-`helices` and β-`sheets` separated in different parts of the molecule; absence of β-α-β super-secondary structure;
4. α/β: helices and sheets assembled from β-α-β units;
5. `small proteins`: proteins with only some secondary structures maintained by disulphide bond or ligand;
6. `multi-domain proteins` (α and β): folds consisting of two or more domains belonging to different classes;
7. membrane and cell surface proteins and peptides, that do not include proteins in the immune system.

[4] Structural Classification of Proteins. `scop.mrc-lmb.cam.ac.uk/scop`.

[5] Protein Structure Classification. `www.cathdb.info`.



(a) α-helical (b) β-sheet (c) α+β (d) α/β

(e) Membrane and cell (f) Multi-domain (g) Small

Figure 2.6: SCOP protein classes. Protein structures are based on similarities of their amino acid sequences and three-dimensional structures. (a) α-helical protein, (b) β-sheet protein, (c) α+β protein, (d) α/β, (e) Membrane and cell surface proteins and peptides, (f) Multi-domain proteins (α/β), (g) Small proteins. Ribbon structure illustrations were prepared with PYMOL.

2.6 Structural databases and structural parameters

One of the most important database in the field of 3-D protein structure prediction is the Protein Data Bank[6] (PDB) (BERMAN et al., 2000). The application of empirical approaches to protein structure prediction is entirely dependent on experimental databases. The PDB contains publicly available 3-D structures of proteins, nucleic acids, and a variety of other complex biomolecules experimentally determined by X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy. There are other protein databases: PDBj[7], PDBe[8]. Currently, the PDB stores 80,402 protein structures[9].

In addition to the information available from the PDB several programs are also available. These programs calculate additional structural parameters from the entries of the PDB: main chain torsion angles (ϕ, ψ) the surface area accessible to a water molecule, distance between all residue pairs in the form of a matrix, secondary structure, super-secondary structure folding patterns. Probably the most widely used software in this field are DSSP (KABSCH; SANDER, 1983, 1984), PROMOTIF (HUTCHINSON; THORNTON, 1996), PROCHECK (MORRIS et al., 1992; LASKOWSKI et al., 1993, 1996), NACCESS [10] and TORSIONS [11] (by A.C.R. Martin,

[6] PDB. Protein Data Bank. www.pdb.org.

[7] PDBj. Protein Data Bank Japan. www.pdbj.org.

[8] PDBe. Protein Data Bank in Europe. www.ebi.ac.uk/pdbe.

[9] Jun, 2012.

[10] Hubbard, S.J. and Thornton, J.M. 'NACCESS', computer program, 1993, Department of Biochemistry and Molecular Biology, University College London

[11] Bloomsbury Center for Bioinformatics. www.bioinf.org.uk/software/swreg.html.

UCL-London).

2.7 Chapter conclusions

In this chapter the basic concepts of `Structural Bioinformatics` were presented: amino acid residues, peptide bonds, torsion angles, protein structural description and structural databases. This chapter serves as basis for understanding other topics discussed in the next chapter of this thesis. Next chapter (Chapter 3) addresses the protein kinematics problem and presents the most common representations of polypeptides structures that are in literature: `Cartesian` position of the atoms and dihedral angles. Chapter 3 also discusses the polypeptide structure representation adopted in this work and the developed mechanisms to manipulate its structures.

3 ON THE PROTEIN KINEMATICS

3.1 Introduction

Protein structures can adopt a variety of shapes. The structure of one protein is defined by its amino acid sequence that folds spontaneously during or after biosynthesis. The relation of the amino acid sequence of an protein and the conformation was first proven by Anfinsen's experiments (ANFINSEN et al., 1961; ANFINSEN, 1973) and depends on the solvent, the concentration of salts and the temperature. As described in Chapter 2 the native and functional state of a protein depends of many factors such as covalent interactions, hydrogen bonds, van der Waals interactions, solvent and hydrophobic contacts.

The computational representation of a 3-D protein structure is a challenging task due to the difficulty in representing the protein structure and simulating the factors that contribute for the native structure stability. This representation is related to the level of detail used to represent the 3-D protein structure. The higher the number of details, higher is the capacity of representing the protein in its native state. The most detailed representation includes all atoms of the proteins and solvent molecules. The geometric representation is one of the most important elements of 3-D protein structure prediction methods and is directly related to the reduction or increase of the protein conformational search space. Using all atoms to represent the protein is computationally expensive and thus, simplified representations are often used (CHIVIAN et al., 2003). This Chapter describes the most common strategies to computationally represent and manipulate the 3-D structure of proteins.

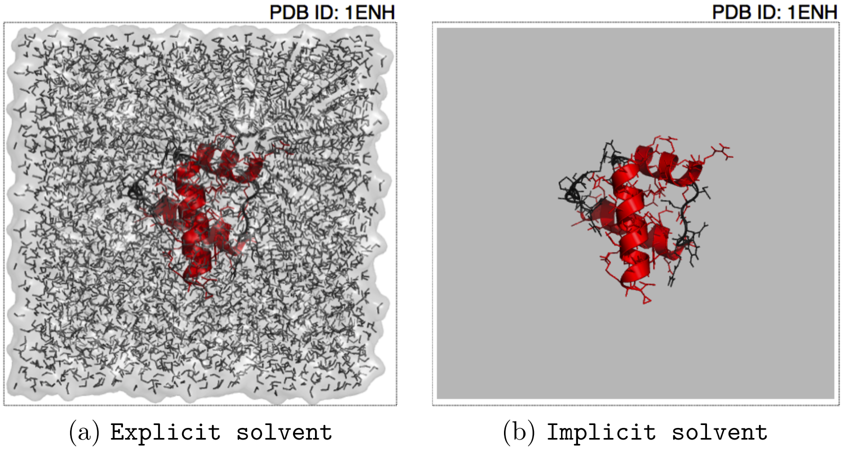
3.2 Three-dimensional protein structure representation

There are two most common representations of polypeptides structures found in the literature. The first model represents the 3-D protein structure through the Cartesian position of the atoms. In this case, a polypeptide chain can be represented as a set P of atoms in the three dimensional space (R^3) (Eq. 3.1).

$$P = [\overrightarrow{a_1}, \overrightarrow{a_2}, \ldots, \overrightarrow{a_n}], \tag{3.1}$$

where n is the total number of atoms in the molecule. The geometry of a polypeptide structure is described by assigning to each i-th atom a 3-D coordinate vector $\overrightarrow{a_i}$ (Eq. 3.2).

$$\overrightarrow{a_i} = \left(\; a_{i.x}, a_{i.y}, a_{i.z} \; \right) \tag{3.2}$$



(a) `Explicit solvent` (b) `Implicit solvent`

Figure 3.1: Schematic representation of polypeptide solvation models. (a) Explicit water solvent (Gray). In the explicit solvation model the water molecules are placed around the simulated solute. (b) Implicit continuous solvation model. Implicit solvation represents the solvent as a continuum medium instead of individual solvent molecules. Ribbon structure illustrations were designed with `PYMOL` - `www.pymol.org`.

This in turn gives rise to $3n$ degrees of freedom, where n is the number of atoms of the polypeptide structure. This number increases when solvent is considered (for example, $\texttt{H}_2\texttt{O}$) (Fig. 3.1). Using all atoms to represent the protein and solvent is computationally expensive. There are some simplifications that can be employed (CHIVIAN et al., 2003): (i) the replacement of all-atom model of the protein and the solvent environment (`explicit solvent`) (Fig. 3.1-a) by a model where the solvent is modelled by potential fields (`implicit solvent`) (Fig. 3.1-b) (STILL; YEO H.C. AMD KOLATKAR; CLARKE, 1990; TSUI; CASE, 2001; MACKERREL, 2010); (ii) the use of the united-atom model which eliminates hydrogen atoms that don't have the capability to participate in hydrogen bonds (KHALILI et al., 2005); (iii) the use of virtual atoms (OSGUTHORPE, 2000). Figure 3.1-b illustrates the implicit solvent model. The major advantage of implicit models when compared with explicit solvation is the large speed up by removing the explicit representation of some several thousand water atoms.

The second model represents the polypeptide structure by means of the set of dihedral torsion angles and is based on the fact that bond lengths are nearly constant in a polypeptide chain (Fig. 3.2-A) (NEUMAIER, 1997). In this representation, two atoms $\overrightarrow{a_i}$ and $\overrightarrow{a_j}$ that are joined by a chemical bond can be represented as a bond vector $\overrightarrow{s}$ (Eq. 3.3) where the length of the bond vector can be computed with the Euclidean norm (Eq. 3.4).

$$\overrightarrow{s} = \overrightarrow{a_i} - \overrightarrow{a_j} \tag{3.3}$$

$$\|\overrightarrow{s}\| = \sqrt{s_x^2 + s_y^2 + s_z^2} \tag{3.4}$$

For two adjacent bonds $\overrightarrow{a_b} - \overrightarrow{a_c}$ and $\overrightarrow{a_d} - \overrightarrow{a_e}$, there are three bond vectors (Eq. 3.5).

$$\begin{aligned}\overrightarrow{u} &= \overrightarrow{a_c} - \overrightarrow{a_b} \\ \overrightarrow{v} &= \overrightarrow{a_e} - \overrightarrow{a_d} \\ \overrightarrow{r} &= \overrightarrow{a_d} - \overrightarrow{a_c}\end{aligned} \tag{3.5}$$

The bond angle formed by the bonds between the atoms $\overrightarrow{a_b} - \overrightarrow{a_c} - \overrightarrow{a_d}$ (Fig. 3.2-b) can be computed by Equation 3.6.

$$arcsin\ \alpha \frac{\|\overrightarrow{u} \times \overrightarrow{r}\|}{\|\overrightarrow{u}\|\ \|\overrightarrow{r}\|}, \tag{3.6}$$

where $\overrightarrow{u} \times \overrightarrow{r}$ is determined by Equation 3.7 (cross product in $\mathtt{R}^3$).

$$\overrightarrow{u} \times \overrightarrow{r} = \begin{pmatrix} u_y r_z - u_z r_y \\ u_z r_x - u_x r_z \\ u_x r_y - u_y r_x \end{pmatrix} \tag{3.7}$$

Similarly the bond angle formed by the bonds between atoms $\overrightarrow{a_c} - \overrightarrow{a_d} - \overrightarrow{a_e}$ (Fig. 3.2-b) can be computed by Equation 3.8.

$$arcsin\ \beta \frac{\|\overrightarrow{v} \times \overrightarrow{r}\|}{\|v\|\ \|r\|}, \tag{3.8}$$

where $\overrightarrow{v} \times \overrightarrow{r}$ is determined from Equation 3.9.

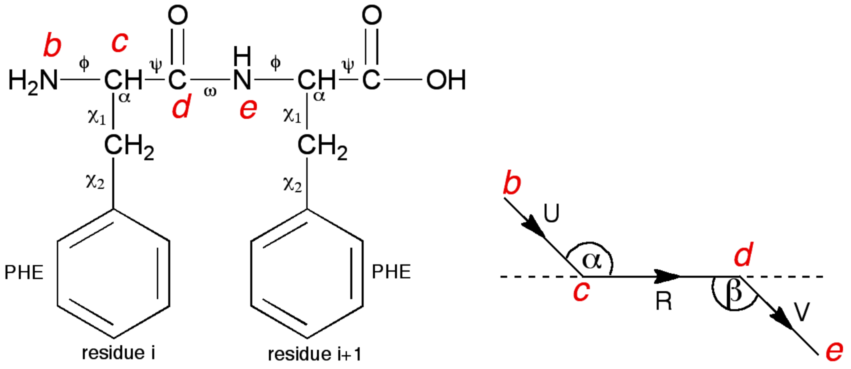
$$\overrightarrow{v} \times \overrightarrow{r} = \begin{pmatrix} v_y r_z - v_z r_y \\ v_z r_x - v_x r_z \\ v_x r_y - v_y r_x \end{pmatrix} \tag{3.9}$$

As describe in Section 2.3, the full set of bond lengths, bond angles, and dihedral angles fix the geometry of the polypeptide molecule. The use of dihedral angles has the advantage over the `Cartesian` model for having the degree of freedom reduced. For the backbone representation of a polypeptide this gives rise to $3m$ degrees of freedom, where m is the number of amino acid residues. The main disadvantage of the used of dihedral angles is that a small change in one dihedral angle causes drastic changes in the polypeptide structure.

Some other strategies can be employed in order to further reduce the number of degrees of freedom in a dihedral representation. The peptide bond tends to be planar, with two allowed states for the ω (`OMEGA`) torsion angle: `trans 180.0`° (usually) and `cis 0`° (rarely). This means that only the ϕ (`PHI`) and ψ (`PSI`) torsion angles can be used to represent the protein backbone. Side-chains of the polypeptide structure can also be represented by torsion angles. As described in section 2.3.1, the number of χ (`CHI`) dihedral angles of each side-chain depends on the amino acid residue type, and side-chain rotamers can be used to reduce the conformation search space.

3.3 Protein structure kinematics

In order to utilize torsion angles we need to transform dihedral angles in Cartesian coordinates of the polypeptide atom (main-chain and side-chain). It means that given a set of dihedral angles we need to calculate the `X`, `Y` and `Z` coordinates of each



(a) Peptide (b) Bond vector

Figure 3.2: Schematic representation of a model peptide and its torsion angles. N is nitrogen, C and C_α are carbons. (a) A peptide composed by two Phenylalanine chained by a peptide bond (ω). (b) Bond vectors and dihedral angles; b, c, d and e represent atoms; U, R and V represent bond vectors.

atom in the macromolecule and vice-versa. This process can be done by rotation matrices. Transformations are applied to residues and molecules to move them into new orientations and/or positions. Let be i a bond and θ the degree of rotation around this bond. After a rotation $R(i, \theta)$ the Cartesian coordinates of atoms that are subsequently the bond need to be update. In Equation 3.10, $[x, y, z, 1]$ represents the position of a generic atom in homogeneous form, $[x^{'}, y^{'}, z^{'}, 1]$ describes the atom position after the rotation, and $R(i, \theta)$ is a 4x4 matrix that encodes a rotation of θ degrees around an axis. In order to perform successive rotations about different bonds, this procedure is repeated, updating the Cartesian coordinates for each rotation.

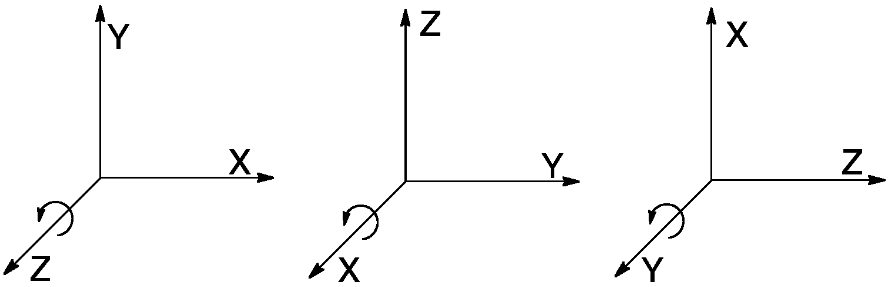
$$[x^{'}, y^{'}, z^{'}, 1]^T = R(i, \theta).[x, y, z, 1]^T \tag{3.10}$$

In a 3-D coordinate representation, an atom is translated from position P=(x,y,z) to position P' = (x',y',z') with a matrix operation (Eq. 3.11).

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3.11}$$

This operation can be represented in a compact way as $P' = T \cdot P$, where P' represents the point position after the translation, T represents the translation of the point P. Parameters t_x, t_y and t_z specifies translation distances for coordinate directions x, y and z. An object is translated in three dimensions by transforming each of the defining points of the object. In order to obtain a translation in the opposite direction t_x, t_y, t_z should be negated ($-t_x$, $-t_y$, $-t_z$).

To generate a rotation transformation for an object, an axis of rotation and the angular rotation must be designed. A rotation is described in the homogeneous coordinate form by Equation 3.12 that performs a rotation in the z-axis (Fig. 3.3 - a). θ represents the rotation angles.



(a) `Z-axis rotation` (b) `X-axis rotation` (c) `Y-axis rotation`

Figure 3.3: Cyclic perturbation of the `Cartesian` coordinates axes to produce the three sets of coordinates-axis rotation equations.

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} cos\theta & -sin\theta & 0 & 0 \\ sin\theta & cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3.12}$$

Transformation matrices for rotations about the other two coordinates axis can be obtained with a cyclic permutation of the coordinate parameters `x,y` and `z`. A rotation in `x-axis` is described in the homogeneous coordinate form by Eq. 3.13 (Fig. 3.3 - b). A rotation in `y-axis` is performed by Eq. 3.14 (Fig. 3.3 - c).

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos\theta & -sin\theta & 0 \\ 0 & sin\theta & cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3.13}$$

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} cos\theta & 0 & sin\theta & 0 \\ 0 & 1 & 0 & 0 \\ -sin\theta & 0 & cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3.14}$$

Table A.1 (Appendix) shows the set of atoms necessary to perform a rotation in the χ_1 of an amino acid residue. Tables A.2, A.3, A.4 (Appendix) show, respectively, the set of atoms necessary to perform a rotation in the dihedral angles χ_2, χ_3 and χ_4 of an amino acid residue.

In the same way the main-chain dihedral angles can be calculated as follows:

- `PHI`: is obtained by the right-handed rotation around the `N-CA` bond. A value equal to zero is obtained when the `CA-C` bond is `cis` to `C-N` bond.
- `PSI`: is obtained by the right-handed rotation around the `CA-C` bond. A value equal to zero is obtained when the `C-N` bond is `cis` to `N-CA` bond.
- `OMEGA`: is obtained by the right-handed rotation around the `C-N` bond. A value equal to zero is obtained when the `CA-C` bond of the preceding residue is `cis` to `N-CA` bond.

In this procedure the coordinates of the main chain atoms are used to calculate the dihedral angles `phi` and `psi`. The `psi` angle is missed for the last residue in each chain. The `phi` angle is missed for the first residue in each chain (Fig. 3.2).

3.4 Chapter conclusions

In this chapter the basic concepts related to protein structure representation and the protein kinematics problem were presented: `Cartesian` position of the atoms and dihedral angles. The first model gives rise to $3n$ degrees of freedom, where n represent the number of atoms of the polypeptide structure, against $3m$ degrees of freedom for the dihedral angles representation, were m represents the number of amino acid residues. Next chapter (Chapter 4) describes the four classes in which the currently `3-D` protein structure prediction methods and algorithms are classified. These methods use different of kind models to represent the polypeptide structure. However all of them uses concepts of `Cartesian` position or dihedral angles to represent the polypeptide structure.

4 TERTIARY PROTEIN STRUCTURE PREDICTION METHODS

4.1 Introduction

Predicting the folded structure of a protein only from its amino acid sequence remains a challenging problem in mathematical optimization (LANDER; WATERMAN, 1999). The challenge arises due to the combinatorial explosion of plausible shapes each of which represent a local minimum of an intricate non-convex function of which the global minimum is sought. In nature, proteins typically present 50 to 500 amino acid residues (LESK, 2002; TRAMONTANO, 2006).

The prediction of the `3-D` structure of polypeptides based only on the amino acid sequence (primary structure) is a problem that has, over the last 40 years, challenged computer scientists, biochemists, mathematicians and biologists (BAXEVANIS; QUELLETTE, 1990). The `Protein Structure Prediction Problem` is one of the main research problems in `Structural Bioinformatics` (CREIGHTON, 1990). The main challenge is to understand how the information encoded in the linear sequence of amino acid residues is translated into the `3-D` structure, and from this acquired knowledge, to develop computational methodologies that can correctly predict the native structure of a protein molecule.

Many methods and algorithms have been proposed, tested and analyzed over the years as a solution to this complex problem, see e.g (ZHANG, 2008; WU; SKOLNICK; ZHANG, 2007; XU; PENG; ZHAO, 2009; HILDEBRAND et al., 2009; KRIEGER et al., 2009; ZHANG, 2007; MOULT, 2005; ZHOU; SKOLNICK, 2009; OSGUTHORPE, 2000; CUTELLO; NARZISI; NICOSIA, 2006; TRAMONTANO, 2006; BUJNICKI, 2006; ROHL et al., 2004; SIMONS et al., 1999; JONES; TAYLOR; THORNTON, 1992; JONES, 2001; SRINIVASAN; ROSE, 2002, 1995). In the literature, one can find several classifications of the `3-D` protein structure prediction methods. Floudas (FLOUDAS et al., 2006) classifies the computational methods for protein structure prediction into four groups:

1. first principle (`ab initio`) methods without database information;
2. first principle methods with database information;
3. comparative homology; and
4. fold recognition.

Regardless of the group, all developed `3-D` protein structure prediction methods have to be tested for the ability to predict new protein structures. Every two years since 1994 a worldwide experiment called CASP[1] (critical assessment of structure prediction) is performed to test protein structure prediction methods. Structural biologists who are about to publish a structure are asked to submit the corresponding sequence for structure prediction. The predictions are then compared with the newly experimentally determined structures (by NMR or X-ray crystallography methods). CASP provides research groups with an opportunity to objectively test their structure prediction methods and delivers an independent assessment of the state-of-the art in protein structure modelling to the research community and software users.

CASP competition involves a large number of research groups using a variety of methods from the four groups described in this chapter. A total of 117 protein experimental structures were available to evaluation and assessment in the last CASP (ninth edition conducted in 2010). These structures were divided into domains, each of which was treated as a separated evaluation unit (MOULT et al., 2011). A total of 248 groups participated in CASP9 experiments. These groups have generated a total of 86,891 models, of which 62,665 were 3-D coordinate sets, 1,220 were sequence alignments converted into coordinates to assessment (KRYSHTAFOVYCH; FIDELIS K. ANDMOULT, 2011B). The remaining submissions are for residue-residue contacts (4,162), structural disorder (5,210), binding site identification (5666), estimation of three-dimensional model quality (7,116) and refinement of initial models (1,709) (MOULT et al., 2011). In the last CASP in addition to the evaluation of the overall accuracy of the 3-D models, many aspects of the structured models were analyzed: prediction accuracy of a model (KRYSHTAFOVYCH; FIDELIS; TRAMONTANO, 2011), prediction of a structural disorder (MONASTYRSKYY et al., 2011), intra-molecular contact identification (KRYSHTAFOVYCH et al., 2011B), identification of binding sites (SCHMIDT et al., 2011), and the analysis of accuracy of quaternary structure (MARIANI et al., 2011). The most significant progress in CASP9 was identified by template-based modelling methods (methods that use database information) (KOOP et al., 2007; COZZETTO et al., 2009; ZHANG, 2008B; XU et al., 2011). There was evidence of improved accuracy for targets of mid range difficulty, likely attributable to improved methods that combine information from multiple templates (WU; ZHANG, 2007; CHENG, 2008). The major remaining challenge in this class of methods is the development of better methods for template production and identification (SODING, 2005); accurate structure for those regions are not easily derived from an obvious template.

In CASP9 was not shown much progress in Free Modeling methods (first principle (ab initio) methods without database information) (JAUCH et al., 2007; BEN-DAVID et al., 2009; FLOUDAS et al., 2006; XU et al., 2011). Among the methods that have been tested in CASP9, I-TASSER presented a significant improvement in its predictions. This improvement is shown mainly because I-TASSER incorporates two components (XU et al., 2011): REMO (LI; ZHANG, 2009) and FG-MD (LI; ZHANG, 2011). REMO is a method for atomic structure construction and improvement of hydrogen-bonding network and FG-MD is fragment-guided molecular dynamics based method that uses constrained molecular dynamics simulation to adjust the position of each atom in the protein.

Each of the four classes of protein structure prediction methods that will be detailed below have some limitations. The analysis of CASP9 experiments reveals that the best results are achieved by methods which combine principles of homology

[1] CASP. predictioncenter.org.

modeling, first principles without database information, first principle with database information and threading methods. First principle methods without database information (ab initio methods) have limitations with respect to the size of the conformational search space (KARPLUS, 1997; LEVINTHAL, 1968). It is not possible to simulate all folding process of long sequences of amino acid residues. Methods that use fragments still have two major limitations. The first one is related to the challenge of dealing with large conformational search spaces caused by different combination of such fragments. The second refers to the challenge of reducing the potential energy in regions where combination of fragments occur. Despite the high quality predictions, comparative modelling by homology and fold recognition have also some limitations: inability to perform prediction of new folds. This is explained by the fact that this methodology can only predict structures of protein sequences which are similar or nearly identical to other protein sequences of known structures in the PDB. The second limitation is that it is not possible to study the folding process of the protein, i.e., the path that an unfolded protein traverses to the functional state (native state).

4.2 Ab initio methods: first principle methods without database information

Ab initio methods, the first principle methods without database information, are founded on thermodynamics and based on the fact that the native structure of a protein corresponds to the global minimum of its free energy (ANFINSEN et al., 1961; ANFINSEN, 1973; TRAMONTANO, 2006). Ab initio structure prediction methods aim at predicting the native conformation of a protein considering only the amino acid sequence (BONNEAU; BAKER, 2001). Osguthorpe (OSGUTHORPE, 2000) defines “ab initio folding” as the class of methods that are based on energy functions that describe the physics of a current conformational state and where only this function is used to search the native structure of the polypeptide. In pure ab initio methods the use of structural templates from a database such as the PDB is not allowed. The structural information from determined structures is only used in the parameterization of empirical all-atoms potentials used in force-fields (potential energy functions) such as AMBER (CORNELL et al., 1995), CHARMM (BROOKS et al., 1983; FIELD et al., 1998), GROMOS (CHRISTEN et al., 2005), GROMACS (SPOEL et al., 2005), OPLS (JORGENSEN; MAXWELL; TIRADO-RIVES, 1996) and ECEPP/2 (MOMANY et al., 1975), among others. Ab initio protein folding is considered a global optimization problem where the goal is to identify the values of a variable set (torsion angles, position of all atoms or a specific set of atoms in the protein structure) that describe the minimum energy of the polypeptide conformation.

Ab initio methods simulate the protein conformational space using an energy function, which describes the internal energy of the protein and its interactions with the environment in which it is inserted. The goal is to find a global minimum of free energy that corresponds to the native or functional state of the protein (OSGUTHORPE, 2000; TRAMONTANO, 2006). Ab initio methods can predict new folds because they are not limited to templates from the PDB. However, these methods have some limitations with respect to the size of the conformational search space (KARPLUS, 1997; LEVINTHAL, 1968). This problem is frequently referred to by many authors as the Levinthal's "paradox" (ZWANZIG; SZABO; BAGCHI, 1991) following studies carried out by Levinthal in 1969 (LEVINTHAL, 1968). The "paradox" is that most small proteins fold spontaneously on a millisecond or even microsecond time scale. Levinthal has demonstrated that a random conformation energetically favorable among all possible conformations is not possible. Levinthal has also suggested that the native structure might have a higher energy, if the lowest energy was not kinetically accessible. In his experiments, Levinthal noted that due to the very large number of degrees of freedom in an unfolded polypeptide chain, a protein molecule has an enormous number of possible conformations (thus rendering an NP-Complete problem) (CRESCENZI et al., 1998; FRAENKEL, 1993; HART; ISTRAIL, 1997; NGO; MARKS; KARPLUS, 1997; LEVINTHAL, 1968).

In general an ab initio method requires three elements (CHIVIAN et al., 2003; OSGUTHORPE, 2000): (i) a geometric representation of the protein chain, (ii) a potential function and (iii) an energy surface searching technique. In the sequel, each of these elements are described in further detail.

Geometric Representation: this representation corresponds to the number of atoms that are used to represent the protein. The most detailed representations include all atoms of the protein and the surrounding solvent molecules (for example, H_2O). Using all atoms to represent the protein is computationally expensive. Such representations can be simplified in a number of ways (CHIVIAN et al., 2003): the all-atom model of both the protein and the solvent environment (explicit solvent) is usually replaced by employing an all-atom model, with the solvent modelled by potential fields of various descriptions (implicit solvent). In general, the united-atom model is frequently used to reduce the computational cost (KHALILI et al., 2005). In this model, explicit hydrogen atoms - with the exception of those that have the capability to participate in hydrogen bonds - are eliminated. Virtual-atoms can also be used to represent one residue and reduce the computational cost (OSGUTHORPE, 2000). In turn, Rotamers (DUNBRACK JR.; COHEN, 1997; DUNBRACK JR.; KARPLUS, 2003) can also be used to represent a limited set of conformations that side-chains can adopt in the polypeptide structure.

Almost all ab initio folding methods use some form of simplified geometry model, in which single virtual atoms of the model represent a number of atoms in the all-atom model (OSGUTHORPE, 2000). The geometric representation is one of the most important elements of an ab initio method and is directly related to the reduction or increase of the associated computational complexity. An all-atom model can demand enormous computational effort during a simulation. On the other hand, simplified representation models can preserve the main structure characteristics and reduce the computational time demanded by a protein folding simulation.

Potential Functions: The second element of an ab initio method is a energy function (Eq. 4.1). Energy functions are used in Molecular Mechanics (MM) simulations (JORGENSEN; TIRADO-RIVES, 2005; MACKERELL JR., 2004), Protein Design (GORDON; MARSHALL; MAYO, 1999; POKALA; HANDEL, 2000) and Protein Structure Prediction (LAZARIDIS; KARPLUS, 2000).

There are two categories: MM potentials and protein structure-derived potential functions (scoring functions) (ZHANG; SKOLNICK, 2004). The first category aims at modelling the forces that determine protein conformations using physically based parameterized functional forms from small molecule data or in vacuo quantum mechanics (QM) calculations (CHIVIAN et al., 2003). The second category is empirically derived from experimental structures from the PDB (CHIVIAN et al., 2003; HAO; SCHERAGA, 1999; KOPPENSTEINER; SIPPL, 1995; LAZARIDIS; KARPLUS, 2000; MOHANTY et al., 1999; SIPPL; HENDLICH; LACKNER, 1992; SIPPL, 1995; LU; SKOLNICK, 2001; GOHLKEA; HENDLICHA; KLEBE, 2000). These two classes of potentials represent the forces that determine the macromolecular conformation: solvation [2], electrostatic [3], van der Waals interactions[4], covalent bonds[5], angles, torsions (4.1) (BOAS; HARBURY, 2007; CHIVIAN et al., 2003; PARK; HUANG; LEVITT, 1997; POKALA; HANDEL, 2000).

The main advantage of using a knowledge-based energy function is that it can model any behavior observed in known protein crystal structures, even when there is no good physical understanding of their behavior (BOAS; HARBURY, 2007). The disadvantage is that these functions cannot predict new behaviors absent in the training set obtained from the PDB.

A potential energy function incorporates two types of terms: bonded and non - bonded (MACKERREL, 2010). The bonded terms (bonds, angles and torsions) are covalently linked. The bonded terms constrain bond lengths and angles near their equilibrium values. The bonded terms also include a torsional potential (torsion) that models the periodic energy barriers encountered during bond rotation. The non-bonded potential includes: ionic bonds, hydrophobic interactions, hydrogen bonds, van der Waals forces, and dipole-dipole bonds.

In a simple and general way the most common Potential Energy Function has the form of Equation 4.1.

$$E_{\texttt{total}} = \sum_{\texttt{bonds}} \texttt{B(C)} + \sum_{\texttt{angles}} \texttt{A(C)} + \sum_{\texttt{torsions}} \texttt{T(C)} + \sum_{\texttt{non-bond}} \texttt{NB(C)} \tag{4.1}$$

where,

- bonds denote an harmonic potential representing the interaction between atomic pairs where atoms are separated by one covalent bond, i.e., 1,2-pairs;

- angles denote the angular vibrational motion occurring between an 1,2,3-triple of covalently bonded atoms;

- torsions denote the torsion angle potential (also known as dihedral angle); it describes the angular spring between the planes formed by the first three and last three atoms of a consecutively bonded 1,2,3,4-quadruple of atoms;

- non-bond involves interactions between all 1,2-pairs of atoms, usually excluding pairs of atoms already involved in a bonded term;

[2] Solvation is the process of attraction and association of molecules of a solvent with molecules or ions of a solution.

[3] Composed by hydrogen bonds, salt bridges and van der Waals interactions. It provides attractive forces between molecules.

[4] van der Waals are the attractive or repulsive forces between molecules or between parts of the same molecule.

[5] A covalent bond is a form of chemical bonding that is characterized by the sharing of pairs of electrons between atoms, and other covalent bonds.

- C denotes a protein conformation;
- $\mathtt{E}_{\text{total}}$ is the potential energy obtained by the sum of the bonded terms (bonds, angles and torsion) and non-bonded terms (ionic bonds, hydrophobic interactions, hydrogen bonds, van der Waals forces, and dipole-dipole bonds).

There is a number of potential energy functions used in computational molecular biology. AMBER (CORNELL et al., 1995), CHARMM (BROOKS et al., 1983; FIELD et al., 1998) and ECEPP (MOMANY et al., 1975) are the most widely used potential energy functions in PSP and Protein Folding problems. A review of potential energy functions is found in Halgren (HALGREN, 1995). 4.2 presents the CHARMM force field (FIELD et al., 1998).

$$\begin{aligned}
E_{\text{total}} = {} & \sum_{\text{bonds}} K_b(b-b_0)^2 + \sum_{\text{UB}} K_{UB}(S-S_0)^2 + \sum_{\text{angle}} K_\theta(\theta-\theta_0)^2 \\
& + \sum_{\text{dihedrals}} K_\chi(1+\cos(\eta-\delta)) \\
& + \sum_{\text{impropers}} K_{imp}(\varphi-\varphi_0)^2) + \sum_{\text{nonbond}} \epsilon\left[\left(\frac{R_{minij}}{r_{ij}}\right)^{12} - \left(\frac{R_{minij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\epsilon_1 r_{ij}}
\end{aligned} \tag{4.2}$$

where $\mathtt{K}_b$, $\mathtt{K}_{UB}$, $\mathtt{K}_\theta$, $\mathtt{K}_\chi$ and $\mathtt{K}_{imp}$ are the bond, Urey-Bradley angle (HAGLER et al., 1979; LIFSON; WARSHEL, 1968), dihedral angle and improper dihedral angle force constants, respectively; b, S, θ, χ and φ are the bond length, Urey-Bradley 1.3 distance, bond angle, dihedral angle, and improper torsion angle, respectively. The subscript zero represents the equilibrium value for the individual terms. Coulomb and Lennard-Jones 6-12 terms contribute to the external or non-bonded interactions; ϵ is the Lennard-Jones (the depth of the potential well) and $\mathtt{R}_{min}$ is the distance at the Lennard-Jones minimum, $\mathtt{q}_i$ is the partial atomic charge, ϵ_1 is the effective dielectric constant, and $\mathtt{r}_{ij}$ is the distance between atoms i and j.

Energy surface search techniques: The most widely used search techniques applied in ab initio methods are: genetic algorithms (PEDERSEN; MOULT, 1997; TUFFERY et al., 1991), Monte Carlo (MC) simulations (SIMONS et al., 1997), evolutionary algorithms (BOWIE; EISENBERG, 1994), and Molecular Dynamics (MD) simulations (GUNSTEREN; BERENDSEN, 1990; RAPAPORT, 2004; KOZA, 1992). These methods use the parameters of the potential energy function to search for the conformation with the minimal energy (DESJARLAIS; CLARKEB, 1998). The conformational search techniques are grouped into two categories: optimization algorithms (PAPADIMITRIOU; STEIGLITZ, 1998; VAZIRANI, 2001) or stochastic algorithms (Monte Carlo, genetic algorithms, evolutionary algorithms) and deterministic methods (FLOUDAS, 2004; HORST; TUY, 2010) (Molecular Dynamics simulations).

In MC-based (SIMONS et al., 1997) methods a starting structure is perturbed by a random change in the position of the atom structure, torsion angles or rotamers. If one change decreases the potential energy of the structure then it is accepted. Otherwise, the Metropolis criterion is used to accept or reject the changes. Genetic algorithms are based on populations of solutions by iterative cycles of operations (HOLLAND, 1975; POKALA; HANDEL, 2000). Deterministic methods (FLOUDAS, 2004) always converge to an optimal solution. The main advantage of stochastic algorithms is that they can deal with problems of great complexity because they do not require an exhaustive search. However, there is no guarantee that these methods converge to the global minimum (VOIGT; GORDON; MAYO, 2000).

4.2.1 Overview of ab initio approaches

There are many computational packages that are used in ab initio calculations. These simulation packages are frequently used in the protein folding problem and in other molecular modelling problems such as molecular docking (LENGAUER; RAREY, 1996; KITCHEN et al., 2004), which predicts the preferred orientation of a molecule with respect to another molecule when bound to each other to form a stable complex (LENGAUER; RAREY, 1996). There are also ab initio algorithms developed specifically for the tertiary PSP problem. The most common simulation packages are: AMBER (Assisted Model Building with Energy Refinement) (CASE et al., 2005; PEARLMAN et al., 1995), CHARMM (Chemistry at HARvard Molecular Mechanics) (BROOKS et al., 1983; MACKERELL JR. et al., 1998), UNRES (LIWO et al., 1998, 1999, 1997), GROMACS (Groningen MAchine for Chemical Simulation) (SPOEL et al., 2005; HESS et al., 2008) and TINKER (Software Tools for Molecular Design) (PONDER; RICHARDS, 1987; KUNDROT; PONDER; RICHARDS, 1991). Ab initio protein structure prediction methods include: LINUS (Local Independent Nucleated Units of Structure) (SRINIVASAN; ROSE, 2002, 1995), ASTROFOLD (KLEPEIS; FLOUDAS, 2003) and BHAGEERATH (JAYARAM et al., 2006; NARANG et al., 2005, 2006). Ab initio packages and prediction methods are detailed in Table B.2 (Appendix).

AMBER (CASE et al., 2005; PEARLMAN et al., 1995) is an example of ab initio package that allows users to carry out and analyze MD simulations for proteins, nucleic acids and carbohydrates. Basically, it is composed of two parts: (i) a set of molecular mechanical force fields for the simulation of biomolecules and (ii) a set of molecular simulation programs. The first part covers the set of empirical parameters used in the simulations. The second part is concentrated in the methods used for energy minimization and molecular dynamics. There are three main steps in an AMBER ab initio simulation task: (1) system preparation; (2) simulation and (3) trajectory analysis. The main preparation programs in AMBER are: antechAMBER, which assembles force fields for residues or organic molecules that are not part of the standard libraries; and LEaP, which constructs bio-polymers from the component residue, solvates the system, and prepares the list of force fields and their associated parameters. SANDER (Simulated Annealing with NMR-Derived Energy Restraints) is the main Molecular Dynamics program which carries out energy minimization, MD, and NMR refinements. SANDER provides direct support for several force fields for proteins and nucleic acids, and for several water models and other organic solvents. Further, AMBER presents some programs to analyze Molecular Dynamics trajectories; one of the most important is ptraj which can be used to calculate angles between atoms, to compute and average structure over all configurations read in, to calculate pair distances in selected atoms, to analyze hydrogen bonds, to compute correlation and other functions. AMBER provides support for explicit and implicit solvent models (RICHARDS, 1977). In case of explicit solvents it provides support for water models, methanol, chloroform, N-methylacetamide and urea/water mixtures. The implicit solvent model has several advantages over the explicit water representation; the main advantage is related to the fact that implicit models are often computationally less expensive. AMBER implements implicit solvent models with Poisson-Boltzmann (FOGOLARI; BRIGO; MOLINARI, 2002) and Generalized Born approach (STILL; YEO H.C. AMD KOLATKAR; CLARKE, 1990; ONUFRIEV; BASHFORD; CASE, 2002) and explicit solvent models with Particle-Mesh E-Wald summation (PME) (DARDEN; YORK; PEDERSEN, 2009; TOUKMAJI; BOARD, 1996). A good general overview of the AMBER codes can be found in (CASE et al., 2005).

CHARMM (BROOKS et al., 1983; MACKERELL JR. et al., 1998) provides a molecular dynamics simulation and analysis package as well as a widely used set of force fields for molecular dynamics. The package allows generation and analysis of a wide range of molecular simulations. The most basic kinds of simulation are: minimization of a given structure and production runs of a molecular dynamics. There are more advanced features: free energy perturbation, quasi-harmonic entropy estimation, correlation analysis, and combined quantum and molecular mechanics (QM/MM) methods. CHARMM is one of the most used programs for molecular dynamics. It is also considered the oldest program for MD simulations and has accumulated a huge number of features. The CHARMM force fields for proteins include: united-atom model, CHARMM19, all-atom CHARMM22 (FIELD et al., 1998) and its dihedral potential corrected variant CHARMM22/CMAP (MACKERELL; FEIG; BROOKS, 2004). CHARMM22 is parameterized for the TIP3P explicit water model (JORGENSEN et al., 1983) and frequently used with implicit solvents. For DNA, RNA, and lipids, CHARMM27 (MACKERELL; BANAVALI; FOLOPPE, 2001) is used.

GROMACS (HESS et al., 2008; SPOEL et al., 2005) provides an MD program with source code specially directed towards the simulation of biological macromolecules in aqueous and membrane environments. It does not have a force field of its own, but it is compatible with other force fields such as AMBER (CORNELL et al., 1995), OPLS (JORGENSEN; MAXWELL; TIRADO-RIVES, 1996), GROMOS (CHRISTEN et al., 2005) and ENCAD (LEVITT, 1983). The package provides micro-canonical Hamiltonian Mechanics (LAVALLE, 2006), stochastic dynamics and energy minimization algorithms. GROMACS package includes a set of analysis tools that permit trajectory and structural fluctuation analysis. A molecular system is defined by its size and shape, the number of types of molecules it contains, and the coordinates and velocities of each atom. The forces and energies are computed on the basis of three different types of interactions: bonded interactions (between two, three or four particles), non-bonded interactions (between pairs of particles) and special interactions (that can define or impose position, angle or distance constraints on the motion of the system). GROMACS implements Quantum Mechanics and Molecular Mechanics approaches that are frequently used to simulate chemical reactions in solution or in enzymes. Both approaches have interfaces to several Quantum Chemistry Packages: MOPAC (DEWAR, 1983), GAMESS-UK (GUEST et al., 2005), GAUSSIAN[6]. GROMACS uses a united-atom model in order to reduce the complexity of representing the molecular structure and for removing some degrees of freedom. The package has long been used in the protein folding problem (SPOEL et al., 1996; SPOEL; VOGEL; BERENDSEN, 1996; SPOEL; BERENDSEN, 1997; SPOEL, 1998).

TINKER (KUNDROT; PONDER; RICHARDS, 1991; PONDER; RICHARDS, 1987) is a software package used in empirical force field molecular mechanics and Molecular Dynamics calculations. It implements a variety of algorithms includ-

[6] Gaussian. www.gaussian.com

ing distance geometry with fast metrization and Gaussian trial distances (HUANG; SAMUDRALA; PONDER, 1998); Elber's (ELBER; KARPLUS, 1987) reaction path method, global optimization via Potential Smoothing (PAPPU; HART; PONDER, 1998) and search algorithms, Molecular Dynamics (GUNSTEREN; BERENDSEN, 1990) with simulated annealing and stochastic dynamics (GUARNIERI; STILL, 1994); Particle Mesh E-Wald (PME) summation (DARDEN; YORK; PEDERSEN, 2009; TOUKMAJI; BOARD, 1996); Monte Carlo minimization; atomic multipole treatment of electrostatics with explicit dipole (WILLIAMS, 1998); Eisenberg - McLachlan ASP (EISENBERG; MCLACHLAN, 1986; WESSON; EISENBERG, 1992) and GB/SA (QIU et al., 1997; STILL; YEO H.C. AMD KOLATKAR; CLARKE, 1990) continuum solvation models and truncated Newton TNCG local energy minimization (PONDER; RICHARDS, 1987; DEMBO; STEIHAUG, 1983; EISENSTAT; WALKER, 1996). The routines from TINKER [7] package provide many functions that can be used in the protein folding problem (PONDER, 2010): (1) energy minimization and structural optimization via conjugate gradient, variable metric or truncate Newton method over Cartesian Coordinates, torsion angles or rigid bodies; (2) molecular, stochastic and rigid body dynamics with periodic boundaries[8] and control of temperature and pressure; (3) analysis of energy distribution within a structure; (4) simulated annealing with various cooling protocols; (5) normal mode vibrational analysis; (6) conformational search and global optimization; (7) transition state location and conformational pathways; (8) fitting of energy parameters to crystal data; (9) distance geometry with pairwise metrization; (10) molecular volumes and surface areas; (11) free energy changes for structural mutations, and (12) global optimization via energy surface smoothing including Potential Smoothing and Search (PSS) method.

Bhageerath (JAYARAM et al., 2006; NARANG et al., 2005, 2006) is an ab initio protein structure prediction algorithm. It reduces the search space to generate probable candidates for the protein native structure using a set of eight modules. Module one (generate PDB from FASTA[9] sequence) involves the formation of a 3-D structure from the amino acid sequence with the secondary structure information. Module two (generate trial structures) involves the generation of a large number of trial structures with a systematic sampling of the conformational space of loop dihedrals. Module three (pad through biophysical filters) has the objective of reducing the number of improbable candidates through the application of a screening procedure based on persistence length [10] and Radios of Gyration[11] filters. The resultant structures are refined in module four by a Monte Carlo sampling procedure in dihedral space to remove steric clashes between atoms of the main chain and side-chains. In Module five the energy of the structures is minimized (step descent and conjugate gradient approaches) to further optimize the side-chains. Module six consists of ranking the structures using all atom energy based empirical

[7] TINKER dasher.wustl.edu/tinker .

[8] In Molecular Dynamics, periodic bond conditions are usually applied to simulate bulk gasses, liquids, crystals or mixtures.

[9] FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes.

[10] Persistence length: is the maximum length of the uninterrupted polypeptide chain persisting in a particular direction.

[11] Radius of Gyration: describes the overall spread of the molecule and is defined as the root mean square distance of the collection of atoms from their common gravity center.

scoring function (NARANG et al., 2006). Module seven reduces the probable candidates based on the protein regularity index (THUKRAL et al., 2007). The last module selects the 10 best structures using a topological equivalence criterion and the accessible surface area (RICHARDS, 1977).

UNRES (LIWO et al., 1998, 1999, 1997) is a united-residue force field for energy-based prediction of protein structure. In the UNRES model, a polypeptide chain is presented as a sequence of α-carbon (C_α) atoms linked by virtual bonds with attached united side-chains and united peptide groups. Each united peptide group is located in the middle between two consecutive C_α, with a peptide group being located between a $C_{\alpha i}$ and $C_{\alpha i+1}$. UNRES considers as interaction sites only the united peptide groups and the united side-chains, the C_α are used only to define the geometry of the polypeptide chain. UNRES force-field is widely used in protein folding simulations and protein structure prediction (OLDZIEJ et al., 2005; LIWO et al., 2010; MAISURADZE et al., 2010; CZAPLEWSKI et al., 2009; SHEN; LIWO; SCHERAGA, 2009; HE et al., 2009; NANIAS; CZAPLEWSKI; SCHERAGA, 2009; SHEN et al., 2008).

ASTROFOLD (KLEPEIS; FLOUDAS, 2003) is a combinatorial and global optimization framework for the ab initio prediction of 3-D structures of proteins. It is composed by four main steps: (1) α-helix prediction; (2) β-sheet prediction; (3) loop modelling and (4) tertiary structure prediction. In the first step, the principle of hierarchical folding is used to predict α-helices (KLEPEIS; FLOUDAS, 2002), where the polypeptide sequence is divided into sub-sequences and optimization techniques are employed in order to find the conformation of a given peptide with the lowest energy (KLEPEIS; IERAPETRITOU; FLOUDAS, 1998). Two algorithms are used to generate low energy assembles: (1) a deterministic branch and bound algorithm (αBB) (KLEPEIS; ANDROULAKIS; FLOUDAS, 1998B; KLEPEIS; FLOUDAS, 1999) and (2) Conformation Space Annealing (CSA) (LEE; SCHERAGA, 1999; LEE et al., 2000, 2001). After predicting the α-helices the remaining residues are analyzed in order to identify the formation of β-sheets. The β-sheet prediction is based mainly on the hydrophobic information and on the prediction of tertiary hydrophobic contacts to identify parallel and anti-parallel structures (KEPLEIS; FLOUDAS, 2002B). The formulation of hydrophobic interactions between β-sheets residues produces an Integer Linear Programming (ILP) problem that is solved through an iterative solution and integer cut constrains (KEPLEIS; FLOUDAS, 2002B). In the loop modeling and restraint step the structure prediction problem is formulated based on the development of atomic distance and dihedral-angle restraints derived from the α-helix and β-sheet prediction results. The dihedral angles bonds are assigned according to the predicted structure class: α-helix, β-sheet or loop. The loop region has a large structural variability and its prediction is a complex computational task. ASTROFOLD uses a physics-based ab initio protein structure approach (KLEPEIS; FLOUDAS, 2003B) in order to predict the loop segments. After determining the appropriate bounds on dihedral angles and inter-atomic distances, a combination of an αBB global optimization algorithm, stochastic global optimization and MD in torsion angle space (KLEPEIS; FLOUDAS, 2003B) is used in order to find the polypeptide structure with the lowest internal energy.

LINUS (SRINIVASAN; ROSE, 2002, 1995) (Local Independent Nucleated Units of Structure) is an implementation of a hierarchical fold model (ROSE, 1979; ROSE;

WOLFENDEN, 1993) to predict the fold of a protein. It is based on the idea that globular proteins are organized as a structural hierarchy (GRIPPEN, 1978; ROSE, 1979) and that a complex fold can be decomposed into secondary structure elements (α-`helix`, β-`sheet`, `coil`, `loops`) together with their superstructure (RICHARDS; KUNDROT, 1988). The `LINUS` algorithm accumulates favorable structures that are acceptable in a fixed interval of allowed interactions, and repeats this in stages as the size of the interval increases. In each stage the polypeptide chain is allowed to randomly move by under the influence of an energy function. A hierarchy is established in order to recognize favorable conformations at an early stage; they are then constrained in order to persist during the algorithm stages. It considers two types of interactions during the simulation stage: repulsive (two non-bonded atoms can not occupy the same space at the same time), and attractive interactions (hydrogen bonds and the tendency of apolar residues to cluster). The algorithm starts with a target amino acid sequence and sequentially (one residue per time from the `N-terminal` to the `C-terminal` region - Fig. 2.2) three residues are perturbed simultaneously to generate a new trial conformation. This generated conformation is discarded when one of its amino acid residues overlap. Otherwise, the energy of the conformation is calculated by adding interactions between all residues separated in sequence by no more than the current interval. The energy is then evaluated; if not rejected, this conformation is defined as the new current conformation. A complete progression from `N` to `C` is a cycle. For each interval interaction 6000 cycles are performed (1000 equilibrium steps followed by 5000 trial structures are generated). The trial conformations are retained. Chain segments in the trial ensembles that adopt a persisting conformation in an interval are constrained, and remain in that conformation during subsequent intervals. The energy of a current conformation decreases over the course of each interval and from each interval to the next. At the end, the predicted structure is the conformation with the lowest energy in the last interval. A `Monte Carlo` procedure is used by algorithm to escape energy local minimal.

Each simulation package and protein structure prediction method make use of specific computational strategies in order to search the conformational space and find the native structure of the target polypeptide. Table B.2 summarizes the main computational strategies implemented and used in the described molecular packages and `ab initio` protein structure prediction methods. Molecular modelling packages implement many potential energy functions (as presented in the second Column of Table B.2) and their application and use depend on the type of the molecular simulation problem and simulation parameters used as solvent, temperature, pressure, etc. Usually, `ab initio` prediction methods use only one potential energy or scoring functions that analyze specific features. For example, `BHAGEERATH` uses an empirical energy function that considers the non-bonded energy of a protein, expressed as a sum of three energy terms: electrostatic, `van der Waals`, and hydrophobic (ARORA; JAYARAM, 1998); `LINUS` is based on steric and conformational entropy and the terms used in the scoring functions are the hydrogen bonds and hydrophobic interactions; `ASTROFOLD` uses the `ECEPP/3` force-field.

There are several other `ab initio` simulation packages and `ab initio` algorithms that are also used in the context of the `PSP` problem. These packages and algorithms are similar in some aspects to the ones previously described. The structure of these packages and algorithms are basically the same. Here, we list the other commonly used ab initio packages and prediction algorithms: ABALONE[12], GROMOS (SCOTT et al., 1999), MACROMODEL[13], MOIL (ELBER et al., 1995; ELBER, 2005), MOE[14], NAB (Nucleic Acid Builder) (MACKE; CASE, 1998), ADUN (JOHNSTON; FERNÁNDEZ-GALVÁN; VILLÀ-FREIRA, 2005), ACEMD (HARVEY; GIUPPONI; FABRITIIS, 2009), SPARTAN[15], PLOP (JACOBSON; FRIESNER; HONIG, 2002; JACOBSON et al., 1968B, 2004), BOSS (JORGENSEN; TIRADO-RIVES, 2005B), HOOMD (ANDERSON; TRAVESSET, 2008), LAMMPS (PLIMPTON, 1995), ITAP (STADLER; MIKULLA; TREBIN, 1997), CPMD (ANDREONI; CURIONI, 2000; HUTTER; CURIONI, 2005), SMMP (EISENMENGER et al., 2001, 2006; MEINKE et al., 2008), MOLDY (REFSON, 2000), MACSIMUS[16], DL POLY (SMITH; FORESTER, 1996; SMITH; YONG; RODGER, 2002), ESPRESSO (LIMBACH et al., 2006), MDYNAMIX (LYUBARTSEV; LAAKSONEN, 2000), MCPRO (JORGENSEN; TIRADO-RIVES, 2005B), OPENMD (KUANG et al., 2009), ORAC (MARSILI et al., 2010; PROCACCI et al., 1997), PACKMOL (MARTÍNEZ et al., 2009), PINYMD (TUCKERMAN et al., 2000), Q (MARELIUS et al., 1999), SIESTA (Spanish Initiative for Electronic Simulations with Thousands of Atoms) (SOLER et al., 2001), VASP (KRESSE; MARSMAN; FURTHMULLER, 2009), SAGEMD (SELEZENEV et al., 2003), NAMD (PHILLIPS et al., 2005), MOSCITO (PASCHEK; GEIGER, 2003), MCCCS TOWHEE (MARTIN; SIEPMANN, 1999). Table B.2 lists the simulation packages most widely used in the protein folding and PSP problems. The main computational strategies offered by each package are also listed.

ASTROFOLD (KLEPEIS; FLOUDAS, 2003), LINUS (SRINIVASAN; ROSE, 2002, 1995) are examples of methods based on ab initio protein structure prediction concepts. However, there are other prediction methods that are based on the same concepts and computational techniques. A number of these methods are based on Genetic Algorithms (GA). Hoque (HOQUE; CHETTY; SATTAR, 2009) presents a recent comprehensive review of the application of GA in the protein folding problem. Methods that use GA concepts are presented by many authors (DANDEKAR; ARGOS, 1992, 1994; HOQUE; CHETTY; DOOLEY, 2005, 2006; LE GRAND; MERZ JR., 1993; PEDERSEN; MOULT, 1997; SUN, 1995; UNGER; MOULT, 1993,a). Other methods use MC based procedures to search the folding pathway of proteins. Gibbs (GIBBS; CLARKE; SESSIONS, 2001) is an ab initio prediction method that is based on backbone torsion angles and fixed side-chains torsions angles and on a MC algorithm used to search the conformational space and find the native energy minima only from primary sequence. Similar MC approaches applied to the PSP problem can be found in the works of Derreumaux (DERREUMAUX, 1999) and Abagyan (ABAGYAN; TOTROV, 1994). In Pokarowski (POKAROWSKI; KOLINSKI; SKOLNICKZ, 2003) a Replica Exchange Monte Carlo method (SWENDSEN; WANG, 1986) is used to reproduce a cooperative all-or-none folding transition and cooperative formation of secondary structures upon the folding process. Their method includes the interactions between the hydrophobic residues, repulsive interaction between hydrophobic and polar residues and the orientation-dependant

[12] Biomolecular simulations with Abalone. www.biomolecularmodeling.com/Abalone.

[13] MacroModel, Schrödinger, LLC, New York, NY - www.schrodinger.com.

[14] The Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada - www.chemcomp.com.

[15] Wavefunction, Inc., Irvine, California, USA - www.wavefun.com.

[16] MACromolecule SIMUlation Software. J. Kolafa, Prague Institute of Chemical Technology, Czech Republic - www.vscht.cz/fch/software/MACSIMUS.

polar-polar interactions. Similar strategies are presented by Thachuk (THACHUK; SHMYGELSKA; HOOS, 2007) and applied to a `Hydrophobic Polar` (`HP`) lattice model (DILL, 1985). Herges (HERGES et al., 2003) applies a stochastic tunneling algorithm to the protein folding problem. Schug (SCHUG et al., 2005) uses a Parallel Tempering approach to the `PSP` problem. Bahamish (BAHAMISH; ABDULLAH; SALAM, 2009) and Fonseca (FONSECA; PALUSZEWSKI; WINTER, 2010) apply swarm-based optimization algorithms to the `PSP` problem. Smith (SMITH, 2005) applies a Memetic algorithm to the `HP` protein model. Table B.2 lists the main protein structure prediction methods and their internal computational strategies.

4.3 First principle methods with database information

In first principle methods with database information general rules of protein structure are extracted from protein databases and used to build starting point `3-D` protein structures. These methods do not compare a target sequence to a known structure, but they compare fragments, i.e. short amino acid sub-sequences of a target fragment against fragments of known protein structures (FLOUDAS et al., 2006). This arises from the observation that when a new fold is discovered, it is composed of common structural motifs or fragments from super-secondary structures of proteins with known structures (TRAMONTANO, 2006). Thus, if there are protein fragments that fold into similar structures, then this information or these fragments can be used to construct `3-D` structural models of proteins. This is the essence of the methods based on fragments. The conformation of a protein is seen as a set of various fragments of amino acid sequences representing various structural motifs that are combined to form the `3-D` protein structure.

When homologue fragments are identified they are assembled into a structure through scoring functions and optimization algorithms. The fragments are assembled through a fragment assembly procedure (SIMONS et al., 1997; JONES, 1997) with the purpose of finding the structure with the lowest potential energy. When finding polypeptide structures with the lowest energy potential, these methods are similar to `ab initio` methods; however, they cannot be classified as `ab initio` methods because they use database information to predict the structure of polypeptides. `Fragment-based` methods are based on the premise that local interactions can define local structures in proteins. Local structures present in known protein structures are used in order to predict the structure of a target amino acid sequence. When appropriate fragments have been identified, compact structures can be assembled by randomly combining fragments using, for example, a simulated annealing approach (SIMONS et al., 1997; ROHL et al., 2004).

Similar local sequences do not always present the same `3-D` structure. This occurs because in a `3-D` structure a large number of physiochemical interactions are present; such interactions contribute not only to the stability of the global structure, but also to the configuration of the secondary structures. Thus, `fragment-based` methods cannot fragment the target amino acid sequence, search database template fragments, get their information and combine these fragments without any combination criterion. Non-covalent interactions between atoms of different regions of the molecule influence the formation of local structures (TRAMONTANO, 2006). `Fragment-based` methods need to establish a relationship criterion between the fragments so that they can determine the fragments with higher probability of in-

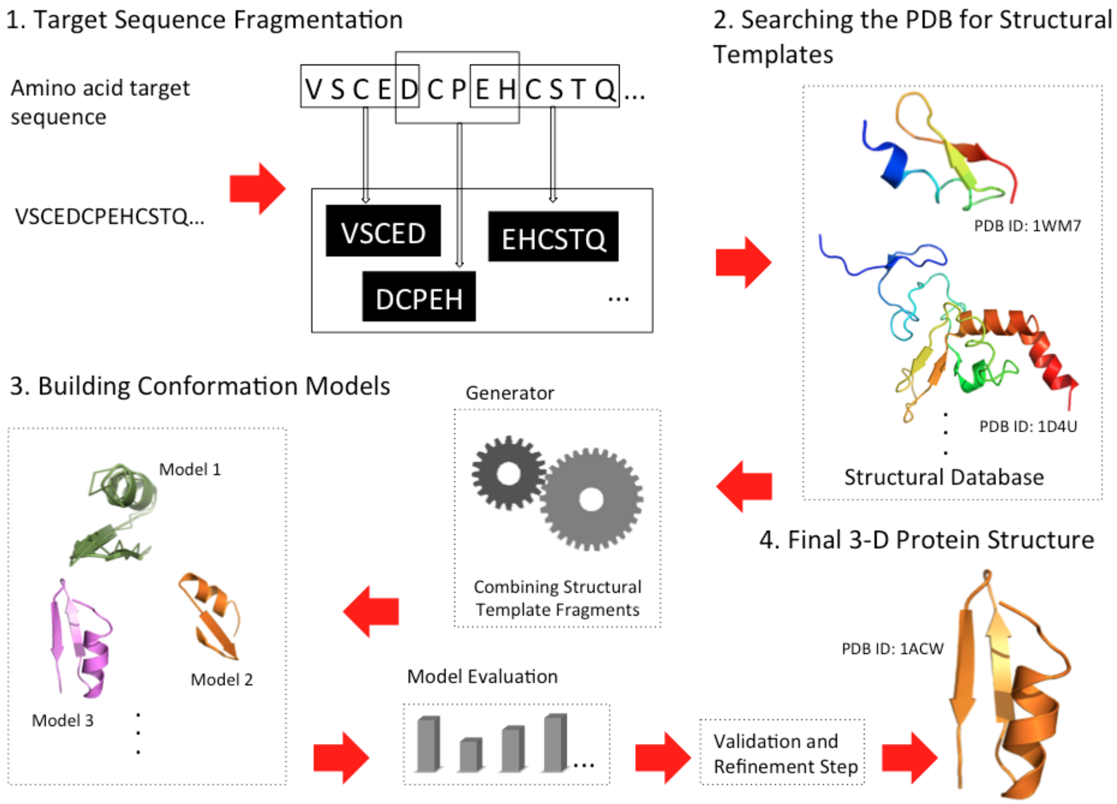

1. Target Sequence Fragmentation
Amino acid target sequence
VSCEDCPEHCSTQ...
V S C E D C P E H C S T Q ...
VSCED
EHCSTQ
DCPEH
...
2. Searching the PDB for Structural Templates
PDB ID: 1WM7
PDB ID: 1D4U
Structural Database
3. Building Conformation Models
Generator
Combining Structural Template Fragments
Model 1
Model 2
Model 3
Model Evaluation
Validation and Refinement Step
4. Final 3-D Protein Structure
PDB ID: 1ACW

Figure 4.1: General schematic representation of a `fragment-based` method for the `3-D PSP` problem: a target sequence is fragmented, templates are obtained from the `PDB`, the fragments are classified, the conformation is constructed and when appropriate, conformation is refined.

sertion during the prediction of the final structure. In this sense, scoring functions are frequently used. The fitness of a conformation can be assessed with scoring functions (ZHANG; SKOLNICK, 2004) derived from conformational statistics of known proteins (FLOUDAS et al., 2006). Additional information can be used in order to improve the scoring functions, for example, secondary structure information (SIMONS et al., 1999B). Figure 4.1 depicts a generic schematic representation of a fragment-based method.

Usually, given the complete sequence of amino acids in a protein, a `fragment-based` method is composed of five distinct stages where:

1. it divides the target sequence into fragments;

2. it carries out the search for similar sequences from each fragment, in a database of known structures;

3. it classifies the fragments (`scoring`);

4. it constructs the three-dimensional structure from the fragment template using a combination technique;

5. finally, it refines the conformation.

`Fragment-based` methods offer advantages over other prediction methods. The first advantage refers to the ability of predicting new folds, which cannot be achieved by methods based on comparative homology (Section 4.5). The second large advantage refers to the reduction of the conformational search space present in `ab initio` methods (Section 4.2). This reduction of conformational space is due to the fact that in a simple replacement of a fragment in the target protein, this fragment moves from one region of a protein which has a structure with minimum potential energy. However, despite reducing the conformational search space, the methods that use fragments still have two major limitations. The first is related to the challenge of dealing with large conformational search spaces caused by different combination of such fragments. The second refers to the challenge of reducing the potential energy in regions where combination of fragments occur. `Fragment-based` methods enjoyed very positive results in the `CASP` experiments. The methods that use database information can be classified as (FLOUDAS, 2007):

1. `Fragment-based recombination methods`: the fundamental idea is to use sequence-dependent local interactions to construct specific segments of the target sequence;

2. `Hybrid methods`: these methods combine multiple sequence comparison, threading methods, Monte Carlo optimization with scoring functions and clustering algorithms;

3. `Secondary structure information and restraint-based methods`: these methods combine information from secondary structure and select `3-D` restraints with `Monte Carlo` optimization and deterministic global optimization.

4.3.1 Overview of first principle methods with database information

I-TASSER (WU; SKOLNICK; ZHANG, 2007; ZHANG, 2007, 2008, 2009) is an interactive implementation of the TASSER method (ZHANG; SKOLNICK, 2004C,B). In the first stage the target sequence is threaded through the PDB to identify appropriate local fragments. Such fragments will incur further structural reassembly. The threading method used in I-TASSER is a simple profile-profile alignment (PPA) approach (OHSEN; SOMMER; ZIMMER, 2003). The frequency of amino acid residues obtained with a PDB PSI-BLAST (ALTSCHUL et al., 1990) search, the secondary structure prediction from PSI-PRED (JONES, 1999B) of the query sequence, and the secondary structure assignment by DSSP (KABSCH; SANDER, 1983) are used as terms in a score function. In I-TASSER the protein chain is divided into aligned and unaligned regions based on the PPA results and for a given target sequence the method proceeds as follows: (1) an initial model is built by connecting contiguous secondary structure fragments (I-TASSER considers a contiguous secondary structure a sequence with at least 5 residues) through a random walk of C_α-C_α bond vectors (WU; SKOLNICK; ZHANG, 2007); (2) this initial structure is submitted to a parallel Monte Carlo sampling for assembling/refinement (ZHANG; KIHARA; SKOLNICK, 2002). I-TASSER uses an energy function that includes predicted secondary structure propensities from PSI-PRED, hydrogen bonds (ZHANG et al., 2006), a variety of statistical short-range and long-range correlations (ZHANG; SKOLNICK, 2004C) and predicted accessible surface area through an artificial neural network approach (CHEN; ZHOU, 2005). Secondly, the trajectories obtained by the simulation in the first stage are clustered (ZHANG; SKOLNICK, 2004D); the cluster centroids are obtained and a Monte Carlo simulation is applied beginning with the cluster centroids conformation. Contact restraints are obtained from the combination of centroid structures and PDB structures searched by the structure alignment program TM-align (ZHANG; SKOLNICK, 2005) based on the cluster centroids. The conformation with the lowest energy is selected and the backbone atoms are added by PULCHRA (FEIG et al., 2000) and the side-chains are added and optimized by SCWRL (CANUTESCU; SHELENKOV; DUNBRACK JR., 2001).

FRAGFOLD (JONES, 2001): is based on the assembly of super-secondary structural fragments obtained from highly resolved protein structures using a simulated annealing approach. This method presents an objective function composed by a set of pairwise potentials of mean force, determined by a statistical analysis of highly resolved X-ray crystallized protein structures and the application of the inverse Boltzmann equation (O'TOOLE; DAHLER, 1960) with a solvation potential and a set of terms that describe the hydrogen network and the steric clashes between atoms of the protein. FRAGFOLD is composed by four basic steps: (1) favorable super-secondary structural fragments at each residue position along the target sequence are selected - the super-secondary structure classification model used by FRAGFOLD is defined as: α-hairpin (consecutive α-helices in a compact arrangement); α-corner (consecutive α-helices in a non-compact arrangement); β-hairpin (hydrogen-bonded consecutive β-strands); β-corner (non-hydrogen-bonded β-strands with intervening α-helix); split β-α-β unit (parallel non-hydrogen-bonded β-strands with intervening α-helix). The fragment selection involves also the summation of pairs of potential terms and solvation terms for the target sequence onto each super-secondary motif, at each position in the sequence; at the end, a sequence-specific list is generated; (2) a general fragment list is build from all tripeptide, tetrapeptide and pentapeptide fragments from the highly resolved structures; (3) A single folding simulation is executed: (a) a random sequence for the target sequence is generated by selecting fragments entirely randomly, (b) fragments are spliced by superposing the α-carbons and the main chain nitrogen and carboxyl-carbon atoms of the C-terminus of one fragment on the equivalent atoms of the N-terminus of the other fragment, (c) a random conformation is generated to each amino acid residue and a steric check is performed, (d) weights for the energy function are calculated, (e) after the weights and the random starting conformation have been determined a simulated annealing approach is used to minimize the energy function; (4) The final structure is selected - for each target sequence, twenty separated simulations using different random number seed values are carried out and the resulting structures are clustered (KELLEY; GARDNER; STUTCLIFFE, 1996). The five most representative populated clusters are assumed with the predicted final structure. There are some variations of the FRAGFOLD algorithm; genetic algorithms with Metropolis criterion are employed in searching the conformational space. FRAGFOLD uses an energy function composed by a pairwise, a solvation, a steric and hydrogen bonding terms.

ROBETTA (ROHL et al., 2004; SCHUELER-FURMAN et al., 2005; SIMONS et al., 1999B) is a fragment-based method for the PSP problem that makes use of an assembly strategy to combine native-like structures of fragments of unrelated protein structures with similar local sequences using Bayesian scoring functions. The main goal of the ROBETTA scoring function is to search for the most probable structure of a protein given the amino acid sequence and the large number of examples of sequences with known structure in the PDB. A Bayes-based theorem is used to describe the probability of a structure given an amino acid sequence (SIMONS et al., 1999). The use of this theorem includes some biological information such as radius gyration, solvation and residue pair interactions. The 3-D structures are generated by splicing together fragments of known structures with similar local sequences and evaluating them using the scoring function. ROBETTA represents a protein structure using a simplified model consisting of the heavy atoms of the main-chain and the C_β atom of the side-chain and the backbone torsional angles. All bond lengths are held constant. The low scoring conformations with distributions of residues of known proteins are identified through a simulated annealing approach in conjunction with the replacement of the torsion angles of segment in the polypeptide chain. ROBETTA uses both 3 and 9 amino acid residues as the fragment length. There are some variations of the ROBETTA method that implement a Monte Carlo procedure to search the protein conformation with the lowest energy (SCHUELER-FURMAN et al., 2005; BRADLEY; MISURA; BAKER, 2005).

ROBETTA@home (DAS et al., 2007) is a computing network based on the Berkeley Open Infrastructure Network Computing protocol (ANDERSON, 2004) that implements a ROBETTA-based algorithm on a Grid Computing platform. The all-atom energy function and the refinement procedure used by ROBETTA@home is the same used in Schueler-Furman (SCHUELER-FURMAN et al., 2005). ROBETTA@home uses three different template-based modelling strategies depending on the sequence size and sequence identity: (1) Loop modelling (targets with sequence identity with the closest template greater than 30% and targets longer than 200 residues with 20-30% sequence identity with the closest template); (2) Loop modelling with constrained all-atom refinement (targets longer than 200 residues with template sequence identity below 20%), and (3) Iterative segment rebuilding and all-atom refinement (targets shorter than 200 residues with template sequence identity below 30%). See Das (DAS et al., 2007) for a complete description of the three protocols.

SIMFOLD (CHIKENJIA; FUJITSUKAB; TAKADAC, 2003) is a fragment assembly algorithm for protein structure prediction. In the first stage of the basic SIMFOLD algorithm (CHIKENJIA; FUJITSUKAB; TAKADAC, 2003) fragment candidates are obtained from a given sequence of amino acid residues. These fragments are three-residues long. Contiguous three-residues fragment templates are searched for an exact match in a non-redundant database. SIMFOLD uses the CULLPDB database (WANG; DUNBRACK, 2003)) and the resulting fragment templates are stored in a fragment library. The homologous database is also searched to find protein templates for each target fragment, using BLOSUM62 (HENIKOFF; HENIKOFF, 1993) scoring over nine residues, which includes an additional three residues in the N-terminal region and in the C-terminal region around the central amino acid fragment). After obtaining the template fragments from the structural database, the algorithm starts a simulation with a random conformation. In this simulation, a move consists of substituting the torsional angles of a randomly chosen candidate in a randomly chosen three-residue fragment for those of the current configuration. Each movement is evaluated using a Metropolis criterion and this procedure is repeated with a decrease in the temperature. SIMFOLD uses a "replicated system" (CHIKENJIA; FUJITSUKAB; TAKADAC, 2003) to chose the contiguous fragments with high probability. A multi-canonical ensemble Monte Carlo (BERG; NEUHAUS, 1991) algorithm is used to search the conformational assembly space. SIMFOLD applies an energy function based on physical terms: $\mathtt{V}_{tot} = \mathtt{V}_{\omega} + \mathtt{V}_{\phi} + \mathtt{V}_{\psi} + \mathtt{V}_{vdw} + \mathtt{V}_{HB} + \mathtt{V}_{HP} + \mathtt{V}_{Rama} + \mathtt{V}_{pair}$, where $\mathtt{V}_{\omega}$, $\mathtt{V}_{\phi}$ and $\mathtt{V}_{\psi}$ are torsion angle potentials, $\mathtt{V}_{vdw}$ is the van der Waals interaction, $\mathtt{V}_{HB}$ is the Hydrogen bonding term, $\mathtt{V}_{HP}$ is the hydrophobic interaction, $\mathtt{V}_{Rama}$ represents the secondary structure propensity based on the entropy contribution of the side-chain, and $\mathtt{V}_{pair}$ denotes pairwise interaction (such as Coulomb interactions) (FUJITSUKA et al., 2004). Recent SIMFOLD versions present some optimizations in the energy function where the energetic parameters are optimized using a set of proteins with known X-ray crystal structures (FUJITSUKA; CHIKENJI; TAKADA, 2006).

PROFESY (LEE et al., 2004) (PROFile Enumerating SYstem) predicts 3-D protein structures that use secondary structure prediction information of the query sequence and the fragment assembly procedure based on global optimization. PROFESY uses the information obtained from the secondary structure prediction method PREDICT (PRofile Enumeration DICtionary) (JOO et al., 2004). For a given sequence of amino acid residues PREDICT, using PSI-BLAST (ALTSCHUL et al., 1997), defines patterns for its amino acid residues. Each pattern is a pair of fifteen amino acid residues. Each pattern is compared with those in the PDB and the patterns closest to the query sequence are selected to determine the secondary structure of the query residues. For each amino acid residue in consideration, a fragment library is built and is composed by the backbone dihedral angles of the patterns. The tertiary structure of a given sequence is generated by this library by fragment assembly. The random conformations are built from a N to C-terminal region that selects a random fragment from the fragment library that is related to an amino acid residue from the target sequence. The global energy minimization of the energy function is performed by the Conformational Space Annealing method (CSA) (LEE;

SCHERAGA; RACKOVSKY, 1997, 1998). PROFESY energy function includes the number of long-range hydrogen bonds, the radius gyration, the Lennard-Jones, van der Waals interactions of the CHARMM (BROOKS et al., 1983; FIELD et al., 1998) (available in the TINKER package), force field for avoiding steric clashes, and the accessible surface area solvation energy (OOI et al., 1987).

CREF (DORN; SOUZA, 2008, 2010) is a PSP method based on short fragments from the PDB. The main goal of CREF is to predict approximate 3-D protein structure that can then be refined through MM techniques. The main characteristic of CREF is that it does not use entire template fragments, but only the ϕ and ψ torsion angles of the main chain of the central amino acid residue of the template fragments obtained from the PDB. Clustering techniques are applied to the template information to identify the Ramachandran plot regions where the central amino acid residues of the template are more concentrated. CREF uses a fixed fragment length equal to five amino acid residues. The clusters identified in the clustering step are labeled with respect to the conformational state indicated by the regions in the Ramachandran plot. The secondary structure of the target sequence is predicted and the approximated conformation is built through a mapping function using the clustering results.

A3N (DORN; SOUZA, 2010B) is a fragment-based method to predict approximate native-like protein structure from primary sequences of amino acid residues. A3N fragments the target sequence in consecutive amino acid fragments. All fragments with five, seven, nine and eleven amino acid residues are generated. A search procedure in the PDB is performed for each target amino acid fragment. Only the information from the central amino acid residue from the templates is considered for analysis. The structural (torsion angles) information from protein templates is analyzed through a statistical function and the secondary structure of the target sequence is predicted. A clustering algorithm is applied in order to identify similar correlated templates in specific regions of the Ramachandran plot. Each Ramachandran region represents a class of conformational states and torsion angle values. A mapping function is used to create training patterns for each amino acid residue from the target sequence. The training patterns of one amino acid residue are learned using Back-propagation in Artificial Neural Networks (GARCEZ; LAMB; GABBAY, 2009; HAYKIN, 1998; RUMELHART; HINTON; WILLIAMS, 1986; GARCEZ; LAMB; GABBAY, 2007; GARCEZ; LAMB, 2006). The torsion angles ϕ and ψ are predicted for each amino acid residue of the target sequence. The polypeptide structure is then predicted.

QUARK [17]: is a algorithm for protein folding and protein structure prediction. Models are built from small fragments by replica-exchange Monte Carlo simulation under the guide of an atomic-level knowledge-based force field.

There are other prediction methods that use the concept of knowledge-fragments for predicting protein structures: CABS (KOLINSKI, 2004), UNDERTAKER (KARPLUS et al., 2003), ABLE (ISHIDA et al., 2003), Fragment-HMM (LI et al., 2008) and ANGLOR (WU; ZHANG, 2008a). Park (PARK, 2005) uses a genetic algorithm for fragment assembly to find low-energy conformations. Cutello et al. (CUTELLO; NARZISI; NICOSIA, 2006) use a genetic algorithm for solving a multi-objective representation of a protein structure. Table C.1 lists the main computational strategies used in the context of this class of methods.

[17] Quark. zhanglab.ccmb.med.umich.edu/QUARK

4.4 Fold recognition and threading methods

Fold recognition methods focus on predicting the three-dimensional folded structure of protein amino acid sequences for which comparative methods do not provide reliable predictions (FLOUDAS et al., 2006). These methods are motivated by the notion that structure is more evolutionary preserved than sequence, i.e., proteins with no apparent sequence similarity could have similar folds (FINKELSTEIN; PTITSYN, 1987; LEVITT; CHOTHIA, 1976; SETUBAL; MEIDANIS, 1997). Several studies in the last years have indicated that the number of protein structural folds in nature are limited (RICHARDSON, 1981; LI et al., 1996; WANG, 1998). Today, for example, there are approximately ten different folds in fifty percent of the proteins that have known structure (RUSSELL; BARTON, 1994). Based on SCOP classification the PDB currently presents 1195 distinct folds.

The general goal of 3-D protein structure prediction by threading methods is to fit a protein sequence correctly against a structural model. This involves two basic procedures: (i) choosing a structural model from a library of models and (ii) finding the correct alignment between the target sequence against the sequences of the structural models in the space of possible sequence-structure alignments. In the threading methodology the 3-D structure prediction problem can be potentially classified as a pattern recognition problem, where the objective is to identify the appropriate fold that represents the structure of the target amino acid sequence. Threading methods use structural information such as residue-residue contact patterns and solvent accessibility in a structure. After identifying the structural similarities, which cannot be detected solely by the similarities between the amino acid sequences, the predicted structural models are constructed. Frequently, threading methods make use of structural classification databases (described in Section 2.6) to construct the template library. Representative instances for each super-family or even each family in the template library are used in order to achieve better predictions.

In threading methods for the 3-D PSP problem it is necessary to solve the problem of sequence-structure alignment, where, given a solved structure T for a sequence $t = t_1, t_2, \ldots, t_n$ and a new sequence $s = s_1, s_2, \ldots, s_m$ the main goal is to find the best match between s and T. Like comparative homology modelling (CM), threading methods use known protein structures as templates for sequences of unknown structures. The main difference between comparative homology modelling and threading methods is that CM methods try to match proteins or clear evolutionary relations while threading methods try to identify templates of the similar fold with or without direct evolutionary relations. CM usually employs sequence-sequence comparison while threading usually exploits structure information to assist alignment (ZHANG, 2009-B). Compared to first principle methods without database information (ab initio), threading methods seek to optimize a potential energy function (an objective or scored function) measuring the fit quality of a sequence in a particular 3-D configuration. A threading method typically consists of four components (SMITH et al., 1997): (i) construction of a library of potential folds or structural templates; (ii) a scoring schema to evaluate any particular placement of a sequence into each fold; (iii) a method to search over the vast space of possible alignments between each sequence and each fold for the best set that gives the best total score; and (iv) a means of choosing the best fold of the set of all alignment of all possible folds to the target sequence. Next, we detail these four components.

`Library of potential folds or structural templates`: the library of folds is constructed from known native protein structures derived from the `PDB`. Usually, the `3-D` coordinates of a protein structure are reduced to more abstract representations. Structural core elements are defined by the secondary structure elements: β`-sheet`, α`-helix`, `left-handed helix`, `coil`, `strands`; frequently, side-chain information is removed. What remains is a backbone template of blank or empty amino acid positions (SMITH et al., 1997).

`A scoring schema to evaluate each placement of a sequence-fold`: the scoring functions are usually a list of statistical references of each amino acid residue to each structural or fold environment (SMITH et al., 1997). Energy functions can also be used. These functions describe how favorable an alignment between a query sequence and a template structure is (JIANG; XU; ZHANG, 2002). Most threading methods do not use physical full-atom free energy function as used by first principle methods without database information. Most threading objective energy functions are determined empirically by statistical analysis of `3-D` data obtained from the `PDB`. These functions are referred to in general as `Knowledge-based` functions. Different approaches for potential functions have been developed: `Boltzmann` statistics (SIPPL, 1995), hydrophobic contact potential (HUANG et al., 1996), probability model based on `Markov Random Fields` (WHITE; MUCHNIK; SMITH, 1994), logistic regression (BRYANT; LAWRENCE, 1993).

`A method to search over the vast space of possible alignments`: the use of an algorithm to identify the optimal sequence-structure alignment is essential in a threading method. The main task is to identify the global best score and the optimal alignment or threading. There are at least three approaches to the `sequence-structure` alignment: `(i)` to use protein sequence alignment strategies (SMITH, 1999; TAYLOR, 1996; MADHUSUDHAN et al., 2006; ALTSCHUL et al., 1997; THOMPSON; HIGGINS; GIBSON, 1994; NOTREDAME; HIGGINS; HERINGAL, 2000; HIGGINS; SHARP, 1988; HIGGINS; THOMPSON; GIBSON, 1996); `(ii)` `3-D` profile methods (BOWIE; LUTHY; EISENBERG, 1991; LUTHY; BOWIE; EISENBERG, 1992; ALEXANDROV; NUSSINOV; ZIMMER, 1996; KELLEY; MACCALLUM; STERNBERG, 2000; SHI; BLUNDELL; MIZUGUCHI, 2001); and `(iii)` contact potentials (CASARI; SIPPL, 1992; BRYANT; LAWRENCE, 1993; SIPPL; HENDLICH; LACKNER, 1992; HENDLICH et al., 1990). Today most threading methods fall into category `iii` above.

`Choosing the best fold of the set of all alignments`: choosing the best template based on alignments is also critical to the success of protein threading. This means that fold recognition requires a criterion to identify the best template for one target sequence. The `sequence-template` alignment score cannot be directly used to rank the templates due to the bias introduced by the residue composition and the number of alternative `sequence-template` alignments (BRYANT; ALTSCHUL, 1995). There exist two basic strategies: `(i)` recognition based on `Z`-scores (BRYANT; ALTSCHUL, 1995) and `(ii)` recognition by machine learning methods (JONES, 1999; XU; PENG; ZHAO, 2009; XU et al., 2003B).

In some aspects a threading algorithm (Fig. 4.2) is close to a sequence alignment method used in comparative modelling (described in Section 4.5). When compared with a comparative modelling method, threading methods present some particular-

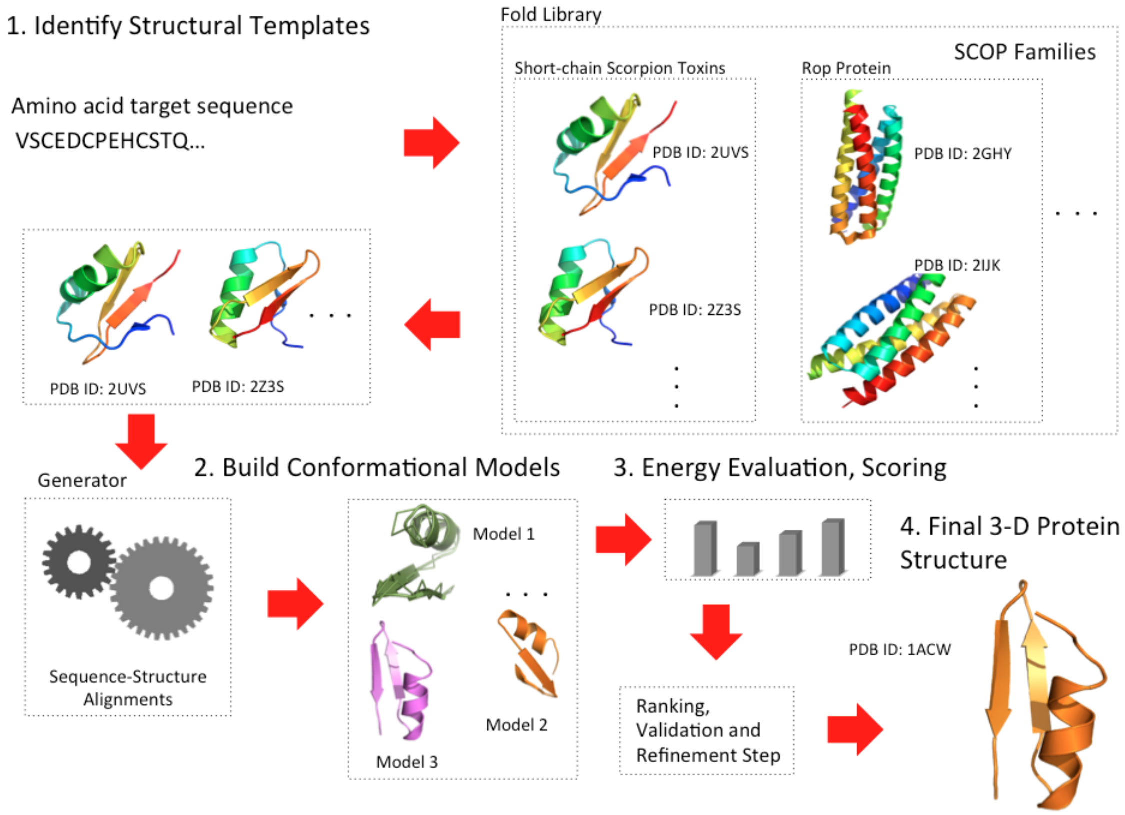

1. Identify Structural Templates
Amino acid target sequence
VSCEDCPEHCSTQ...
Fold Library
SCOP Families
Short-chain Scorpion Toxins
Rop Protein
PDB ID: 2UVS
PDB ID: 2GHY
PDB ID: 2Z3S
PDB ID: 2IJK
PDB ID: 2UVS
PDB ID: 2Z3S
Generator
2. Build Conformational Models
3. Energy Evaluation, Scoring
4. Final 3-D Protein Structure
Sequence-Structure Alignments
Model 1
Model 2
Model 3
Ranking, Validation and Refinement Step
PDB ID: 1ACW

Figure 4.2: General schematic representation of a threading procedure: template folds are selected from a library of protein structures, models for the target protein are constructed, the potential energy of the structures is calculated and the models are scored, the structures are ranked and validated and, when necessary, the best ranked structure is refined.

ities (LESK, 2002): (i) in homology modelling, first the homologue's are identified while in threading methods all possible parents are tried; (ii) in homology modelling the optimal is then determined, while in threading methods many possible alignments are tried; (iii) homology methods optimize one model, whereas threading methods evaluate many rough models.

4.4.1 Overview of fold recognition and threading methods

In order to estimate the quality of the predicted models normalized threading scores are commonly used. Threading methods use comparative homology methods as a basis to align target sequences with template sequences. Many threading methods have been developed and tested recently. The most commonly used methods are presented below.

GENTHREADER (JONES, 1999) performs the calculation of pairs of potential and solvation terms and uses an implementation of an artificial neural network in order to evaluate the alignment. The prediction method uses a traditional multi-sequence alignment algorithm (MULTAL (TAYLOR, 1988)). A sequence profile (GRIBSKOV; MCLACHLAN; EISENBERG, 1987; GRIBSKOV, 1994) is constructed using a BLOSUM matrix (HENIKOFF; HENIKOFF, 1992, 1993). GENTHREADER uses an evaluation function based on a set of pairwise potentials of mean forces (HENDLICH et al., 1990) to determine and select the conformation with lowest potential energy.

123D (ALEXANDROV; NUSSINOV; ZIMMER, 1996)] is a threading method that uses a single empirical potential function to map sequences onto structural positions of any of the proposed folds. The empirical scoring function is derived from an analysis of a non-redundant database of known structures by converting relative frequencies into pseudo-energies using a normalization according to the inverse Boltzmann law. After the sequence-structure alignment, the alignments are evaluated and ranked according to their potential and statistical significance. The best alignment is estimated in comparison with the other alignments. 123D uses a fast dynamic programming optimization procedure adapted to CCPs (Contact Capacity Potentials) (BOWIE; LUTHY; EISENBERG, 1991; OUZOUNIS et al., 1993), mostly for position and secondary structure dependent costs (specially gap costs) to identify similar structures that can be used to model the three-dimensional structure of the target sequence.

ORFEUS (GINALSKI et al., 2003) can be considered a hybrid threading approach. It combines predicted secondary structure information with the information about sequence conservation and variability. The secondary structure information is stored as profile of probabilities (it uses the FFAS strategy (RYCHLEWSKI et al., 2000)). The original algorithm uses the PSIPRED algorithm (JONES, 1999B) to predict the secondary structure. However, any other secondary prediction algorithm that produces estimated probabilities for local structures can be used. ORFEUS uses the SCOP classification database to extract sequence families and a genetic algorithm implementation to improve the parameters of FFAS.

PROSPECT (XU; XU, 2000) uses a scoring function for alignment composed by four terms: (i) mutation term, (ii) singleton fitness term, (iii) pairwise-contact potential term, and (iv) alignment gap penalties. The energy function has the following form: $\mathrm{E}_{\mathrm{total}} = \omega_{mutate}\ \mathrm{E}_{mutate} + \omega_{single}\ \mathrm{E}_{single} + \omega_{pair}\ \mathrm{E}_{pair} + \omega_{gap}\ \mathrm{E}_{gap}$. The mutation energy $\mathtt{E}_{mutate}$ is the sum of the compatibility measurements $\mathtt{e}_{mutate}$ $(\mathtt{a}_1, \mathtt{a}_2)$ for substituting the template amino acid $\mathtt{a}_1$ by the target amino acid $\mathtt{a}_2$. PROSPECT uses the PAM50 matrix (GONNET; COHEN; BENNER, 1992) for calculating $\mathtt{E}_{mutate}$. The singleton energy $\mathtt{E}_{single}$ represents the sum of the preferences $\mathtt{E}_{single}$ (a,s,t) for aligning amino acid a of the target sequence onto a template position with a structural environment defined by secondary structure s and solvent accessibility, or Accessible Surface Area (ASA) t. E_{pair} is the sum of pair-contact potentials $\mathtt{e}_{pair}$ $(\mathtt{a}_1, \mathtt{a}_2)$ between amino acids $\mathtt{a}_1$ and $\mathtt{a}_2$ of the target sequence when they are aligned to template positions that are spatially close. The $\mathtt{E}_{gap}$ is the sum of the penalties $\mathtt{e}_{gap}$ (g) for an alignment gap of length g (GONNET; COHEN; BENNER, 1992). All the ω terms are scaling factors, which are determined by optimizing the threading alignments of the training set against the Structure-Structure alignments. PROSPECT considers pair contacts only between core residues (α-helix or β-sheet) and alignment gaps only in loop regions. All statistics for estimating the terms in the above equation are collected from FSSP (HOLM et al., 1992). The algorithm employs a divide-and-conquer strategy to solve the optimal threading problem. The algorithm solves the entire optimal alignment problem by recursively solving a series of alignment problems between sub-structures and sub-sequences, under various constraints, and then combining these sub-alignments in a consistent and optimal way.

MUSTER (WU; ZHANG, 2008b) is used to identify template structures from the PDB library. It generates sequence-template alignments by combining sequence profile-profile alignment with multiple structural information.

Other threading methods can be found in the literature: SEGMER (WU; ZHANG, 2010), HHPRED (SODING, 2005; SODING; BIEGERT; LUPAS, 2005), THREADER 2 (JONES; MILLER; THORNTON, 1995), 3DPSSM (KELLEY; MACCALLUM; STERNBERG, 2000), FFAS (RYCHLEWSKI et al., 2000), ESYPRED3D (LAMBERT et al., 2002), FUNGUE (SHI; BLUNDELL; MIZUGUCHI, 2001), RAPTOR (XU et al., 2003,B), SAM-T99, SAM-T02 (KARPLUS et al., 2001), SAM-T99 (KARPLUS; BARRETT; HUGHEY, 1992), SAM-T02 (KARPLUS et al., 2001), TOPITS (ROST, 1995,b), LIBRA I (OTA; NISHIKAWA, 1997) and COTH[18]. Turcotte et al. (TURCOTTE; MUGGLETON; STERNBERG, 1998, 2001,B,C) apply Inductive Logic Programing (ILP) to discover rules that govern the three-dimensional topology of protein structure. Xu et al. (XU; XU; UBERBACHER, 1998) developed an algorithm that solves the globally optimal threading problem efficiently. Table D.1 summarizes the main computational strategies used in the context of the main threading methods.

One of the most recent advancements in the field of 3-D protein structure prediction and threading methods is the idea of meta-strategies or meta-serves (BUJNICKI et al., 2001). This idea is related to the concept of consensus-based approach (LUNDSTROM et al., 2001). As shown in the last 8 editions of CASP (MOULT et al., 2009, 2007, 2005; MOULT; FIDELIS; HUBBARD, 2003; MOULT et al., 2001, 1997, 1999, 1995) there is no method that is always the best in the predictions. This occurs because the quality of the predictions depends on many factors which are unknown when the prediction is run. In META-SERVERS all prediction methods are applied to a given sequence; a computational strategy, such as ANNs are used in 3D-Judge (JASKOWSKI et al., 2007), are then applied in order to choose the most realistic prediction. The META-SERVERS approach has many advantages: (i) as shown in the CASP experiments, META-SERVERS produce generally better results than individual servers; (ii) the prediction in meta-serves are more stable than those made when only a single prediction method is used.

META-SERVERS approaches represent one of the most significant advances in the field of protein structure prediction problem. Currently, 3D-Jury (GINALSKI et al., 2003) is one of the most popular META-SERVERS, it computes structure similarities between models using a MaxSub measure (SIEW1 et al., 2000) and chooses the most realistic one as the predicted final result. Other examples of META-SERVERS can be found in the literature: 3D-Judge (JASKOWSKI et al., 2007), LOMETS (WU; ZHANG, 2007), STRUCLA (SASIN; KUROWSKI; BUJNICKI, 2003), Pcons.net (WALLNER; LARSSON; ELOFSSON, 2007), ProCKSi (BARTHEL et al., 2007) and TASSER (ZHOU; SKOLNICK, 2007, 2009; ZHOU; PANDIT; SKOLNICK, 2009). A good review of META-SERVERS can be found in Fischer's work (FISCHER, 2006).

4.5 Comparative modeling methods and sequence alignment strategies

In comparative modelling by homology a target sequence of amino acid residues (target protein) is aligned against the amino acid sequence of another protein with known structure (template protein) and stored in the PDB (BERMAN et al., 2000).

[18] COTH: CO-THreader. zhanglab.ccmb.med.umich.edu/COTH.

If the target sequence is similar to the sequence of the template protein, the structural information obtained from the known structure is used for modelling the target protein (MCLACHLAN, 1992; BAJORATH; STENKAMP; ARUFFO, 1994; BLUNDELL et al., 1987; JOHNSON et al., 1994; SALI, 1995; SÁNCHEZ; SALI, 1997; PEITSCH, 1996). The main idea of this kind of method is to construct an atomic-resolution model of the target protein from its amino acid sequence and a experimental `3-D` structure of a related homologous protein.

Comparative modelling by homology can be applied whenever it is possible to detect an evolutionary relationship between the target protein and the template protein of which the `3-D` structure is known (MARTÍ-RENOM et al., 2000). The evolutionary relationship between proteins is a fundamental factor in methods of comparative modelling by homology and the target protein can be modelled from homologous proteins with `3-D` structures determined experimentally (STERNBERG, 1997). The structure of these proteins are similar in the sense that amino acid residues with identical physiochemical properties occupy the same position in homologous proteins. The torsion angles of the protein backbone preserve a certain regularity in their values.

The quality of the comparative modelling methods is dependent on the quality of the sequence alignment methods. the sequence alignment are used to produce a structural model of the target sequence. There are three main classes of methods used as sequence alignment strategies (MARTÍ-RENOM et al., 2000):

1. `Sequence-sequence comparison (pairwise)`: this class includes the methods that compare the target sequence with each candidate sequence in the database independently (APOSTOLICO; GIANCARLO, 1998). `FASTA` (PEARSON; LIPMAN, 1988) and `BLAST` (ALTSCHUL et al., 1990) are examples of systems used in `sequence-sequence` comparison.

2. `Multiple sequence comparison methods`: perform multiple sequence alignments (THOMPSON; PLEWNIAK; POCH, 1999; NOTREDAME, 2002, 2007; WALLACE; BLACKSHIELDS; HIGGINS, 2005) to improve the sensitivity of the search (GRIBSKOV, 1994; KROGH et al., 1994; ALTSCHUL et al., 1997; HENIKOFF; HENIKOFF, 1994). `CLUSTALW` (THOMPSON; HIGGINS; GIBSON, 1994), `PSI-BLAST` (ALTSCHUL et al., 1997) and `T-COFFEE` (NOTREDAME; HIGGINS; HERINGAL, 2000) are examples of multiple sequence alignment methods.

3. `Threading or 3-D template matching pairwise comparison`: these methods rely on pairwise comparison of a protein sequence and a protein of known structure (JONES; TAYLOR; THORNTON, 1992; BOWIE; LUTHY; EISENBERG, 1991).

Martí-Renom and Sanchez (MARTÍ-RENOM et al., 2000; SÁNCHEZ; SALI, 1997) enumerate four basic steps of a comparative modelling procedure: (`i`) fold assignment and template selection, (`ii`) template target alignment, (`iii`) model building, and (`iv`) model evaluation. Initially, sequences similar to the target sequence are collected using search engines over a database (fold assignment and template selection). Templates can be found searching in structural databases such as `PDB` (BERMAN et al., 2000), `CATH` (ORENGO et al., 1999), and `SCOP` (LO CONTE et al., 1999). The four basic steps are detailed as follows.

1. `Fold assignment and template selection`: the starting point in a Comparative Modelling method is to identify all protein structures with sequences related to the target sequence, then to select templates that will be used as templates. There are a numerous protein sequence and structure databases and database scanning software (ALTSCHUL et al., 1994; HOLM et al., 1992). Templates can be found using the target sequence as a query for searching structure databases such as the `PDB` (BERMAN et al., 2000), `SCOP` (LO CONTE et al., 1999) and `CATH` (ORENGO et al., 1997, 1999).

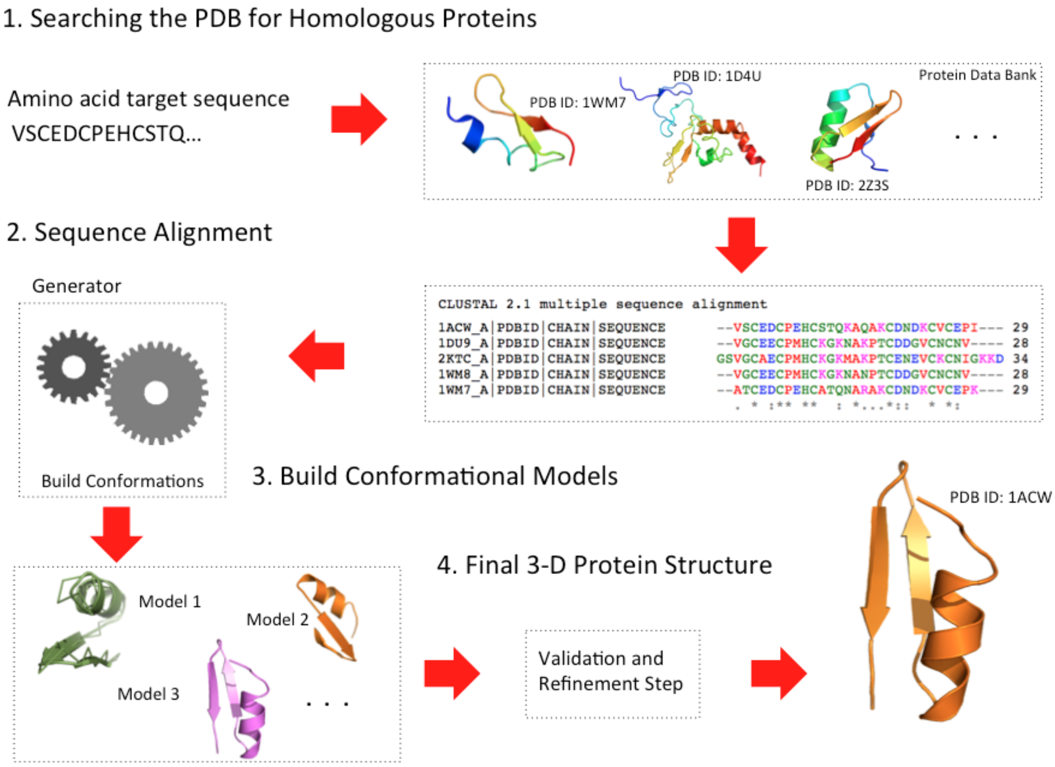
2. `Template target alignment`: in the alignment step the sequence of the target protein is aligned with sequence(s) of protein(s) with known structure(s). It forms the base model. There are other methods that are usually tuned for detection of remote relationships (MARTÍ-RENOM et al., 2000; BAXEVANIS, 1998; BRIFFEUIL et al., 1998; HOLM; SANDER, 1996; SMITH, 1999; TAYLOR, 1996). In these methods, non-optimal alignments are exploited. Pairwise sequence alignment methods (APOSTOLICO; GIANCARLO, 1998) are used to find the best-matching local or global alignments of two sequences. Pairwise alignments can only be used between two sequences at a time. The three primary methods of producing pairwise alignments are dot-matrix methods and dynamic programming.

   Multiple sequence alignment (LIPMAN; ALTSCHUL; KECECIOGLU, 1989; HIROSAWA et al., 1995; GRASSO; LEE, 2004; KIM; PRAMANIK; CHUNG, 1994; EDGAR, 2004; BRUDNO et al., 2003; NOTREDAME, 2002; THOMPSON; PLEWNIAK; POCH, 1999; WALLACE; BLACKSHIELDS; HIGGINS, 2005; NOTREDAME, 2007) is an extension of pairwise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all of the sequences in a given query set.

3. `Model building`: when building the model of the protein, it is common that, first, all the backbone from the homologous regions is constructed, then the different regions, loop regions and, finally, the side-chains (VÁSQUEZ, 1996). A variety of methods can be used to construct the `3-D` model of the target protein. These methods are divided into three groups: modelling by assembly of rigid bodies (BLUNDELL et al., 1987; GREER, 1990), modelling by segment matching or coordinate reconstruction (LEVITT, 1992; CLAESSENS et al., 1989; JONES; THORNTON, 1997B), and modelling by satisfaction of spatial restraints (SALI; BLUNDELL, 1993; HAVEL; SNOW, 1991; SRINIVASAN; MARCH; SUDARSANAM, 1993; ASZÓDI; TAYLOR, 1996).

4. `Model evaluation`: the evaluation of the final model takes into account all available information of the target protein (TRAMONTANO, 2006). According to Baxevanis (BAXEVANIS; QUELLETTE, 1990), the most critical step in homology modelling is the alignment. A misalignment can have a distorting effect on the other steps, generating a distorted and incorrect final structural model.

Comparative modeling by homology is the most used method in protein structure prediction for two main reasons (TRAMONTANO, 2006): `(i)` the quality of the predicted models - when a reasonable evolutionary relationship is present then



```
CLUSTAL 2.1 multiple sequence alignment

1ACW_A|PDBID|CHAIN|SEQUENCE      --VSCEDCPEHCSTQKAQAKCDNDKCVCEPI--- 29
1DU9_A|PDBID|CHAIN|SEQUENCE      --VGCEECPMHCKGKNAKPTCDDGVCNCNV---- 28
2KTC_A|PDBID|CHAIN|SEQUENCE      GSVGCAECPMHCKGKMAKPTCENEVCKCNIGKKD 34
1WM8_A|PDBID|CHAIN|SEQUENCE      --VGCEECPMHCKGKNANPTCDDGVCNCNV---- 28
1WM7_A|PDBID|CHAIN|SEQUENCE      --ATCEDCPEHCATQNARAKCDNDKCVCEPK--- 29
                                   . * :** **  : *...*::  * *:
```

Figure 4.3: Schematic representation of a typical process of comparative modelling by homology. Initially, template proteins are identified. Then the sequence of the target protein is aligned against the sequence of the `protein-templates`, and then a model is built and validated, obtaining in the end, the `3-D` structure of the target protein. If necessary, the final structure may undergo a refinement process.

the accuracy of the predicted models is greater than those produced with other techniques; (ii) the reliability of the model can be estimated a priori and the quality of the predicted structures can be estimated.

In the last years considerable progress has been made in ab initio protein structure prediction methods; however, comparative modeling is a very precise and accurate prediction method (KOEHL; LEVITT, 1999; MARTÍ-RENOM et al., 2000). Despite the high quality predictions, comparative modelling by homology has some limitations. The first limitation concerns the inability to perform prediction of new folds. This is explained by the fact that this methodology can only predict structures of protein sequences which are similar or nearly identical to other protein sequences of known structures in the PDB. The second limitation is that it is not possible to study the folding process of the protein, i.e., the path that an unfolded protein traverses to the functional state (native state).

4.5.1 Overview of comparative modeling methods and sequence alignment strategies

SWISS-MODEL (ARNOLD et al., 2006; KIEFER et al., 2009) is a web-based integrated service dedicated to protein structure homology modelling. It employs an automated, knowledge-based protein modelling tool: ProMod (PEITSCH; JONGENEEL, 1993; PEITSCH, 1996). SWISS-MODEL presents three types of modelling modules: (i) automated mode, (ii) alignment mode and (iii) project mode. The first is computed by the SWISS-MODEL server homology pipeline (SCHWEDE et al., 2003). This module is used in cases where the target sequence similarity is sufficiently high to allow for fully automated mode. In alignment mode, the submitted alignment is matched against the sequence of the template structure extracted from the SWISS-MODEL template library [19]. A rigid fragment assembly modelling and heuristics are used to improve the placement of insertions and deletions based on the structural context. In the project mode, the correct alignment between target and template cannot be clearly determined by sequence-based methods. Further, visual inspection and manual manipulation of the alignment are used (BATES et al., 2001). SWISS-MODEL provides access to a set of increasingly complex and computationally demanding methods for templates searching: BLAST (ALTSCHUL et al., 1997), Interactive profile BLAST (ALTSCHUL et al., 1997) that uses information from the NR (non-redundant) database (WHEELER et al., 2005), HMM-based template library search that uses a library of Hidden Markov Models, where each model of the library was created from multiple sequence alignment generated by iterative search of NR databases using SAM-T2K (HUGHEY; KROGH, 1996).

MODELLER (ESWAR et al., 2006; MARTÍ-RENOM et al., 2000) is a computing system for comparative protein structure modelling that incorporates a large set of functionalities. In the most simple case, MODELLER is used to calculate a model containing all non-hydrogen atoms with only the input of an alignment of a sequence with the template structures and the atomic coordinates of the templates. In other cases, MODELLER can perform fold assignment alignment of two protein sequences (ESWAR et al., 2003) or their profiles (MARTÍ-RENOM; MADHUSUDHAN; SALI, 2004), comparative structure modelling by satisfaction of spatial restraints, multi alignment of sequences and/or structures (MADHUSUDHAN et al.,

[19] The template structure database is derived from the PDB.

2006), calculation of phylogenetic trees (FITCH; MARGOLIASH, 1967), and de novo modelling of loops in proteins (FISER; DO; SALI, 2000). A 3-D model is obtained by optimization of a molecular Probability Density Function (PDF). In order to optimize this function methods of conjugate gradient and Molecular Dynamics with simulated annealing are employed.

T-COFFEE (NOTREDAME; HOLM; HIGGINS, 1998; NOTREDAME; HIGGINS; HERINGAL, 2000) is a multiple sequence alignment package. It provides a simple and flexible means of generating multiple alignments using heterogeneous data sources through a library of pairwise alignments that use a structure similar as the one presented in Notredame et.al (NOTREDAME; HOLM; HIGGINS, 1998). T-COFFEE uses a progressive strategy (FENG; DOOLITTLE, 1987; TAYLOR, 1988; THOMPSON; HIGGINS; GIBSON, 1994) (dynamic programming) to find the best multi-alignment. It uses the information in the library of the pairwise alignments to carry out progressive alignment in a way that considers the alignments between all pairwise alignments, while each step of the progressive multi-alignment is executed. In the progressive alignment, pairwise alignments are made to produce a distance matrix between all the sequences which in turn is used to produce a guide tree using the neighbor-joining method of Saitou and Nei (SAITOU; NEI, 1987). CLUSTALW (LARKIN et al., 2007; HIGGINS; SHARP, 1988; THOMPSON; HIGGINS; GIBSON, 1994), TIP-STRUCTFAST (DEBE et al., 2006), MULTALIN (CORPET, 1988), COMPASS (SADREYEV; GRISHIN, 2003), PSI-BLAST (ALTSCHUL et al., 1997), FASTA (LIPMAN; PEARSON, 1985) and BLAST (ALTSCHUL et al., 1997) a are examples of other comparative modelling methods found in the literature. Table E.1 lists the main computational strategies used in the context of comparative modeling methods for the PSP problem. These methods are also classified into three groups according to the type of structural information used and the strategy used to build the polypeptide structures: modelling by assembly of rigid body; modelling by segment matching or coordinate reconstruction and modelling by satisfaction of spatial restraints.

4.6 Chapter conclusions

Experimentally, the generation of a protein sequence is considerably easier than the determination of its 3-D structure. However, the knowledge of the 3-D structure of the polypeptide gives researchers very important information about the function of the protein in the cell. The difficulty in determining and finding out the 3-D structure of proteins has generated a large discrepancy between the volume of data (sequences of amino acid residues) generated by the GENOME projects[20] and the number of 3-D structures of proteins which are known nowadays. These figures not only clearly illustrate the need for, but also motivate further research in Computational Protein Structure Prediction Methods. In addition, the analysis presented in this chapter demonstrates the importance of the development of accurate computational methods that can compute and predict the 3-D structure of proteins when only their amino acid sequence is known. This Chapter have presented several computational techniques that have been widely applied in the context of the PSP problem. Next Chapter (Chapter 5) presents a new computational strategy to predict 3-D structures of polypeptides.

[20] DOE Genomic Science. genomics.energy.gov.

5 MOIRAE: REDUCING THE CONFORMATIONAL SEARCH SPACE OF 3-D PROTEIN STRUCTURES USING INFORMATION OF EXPERIMENTALLY DETERMINED PROTEINS

5.1 Introduction

The classification of the prediction methods into four classes, (i) first principle methods without database information; (ii) first principle methods with database information; (iii) fold recognition and threading methods and (iv) comparative modelling methods - gives a more general view about which computational methods can be used in 3-D protein structure prediction, how experimental data can be used in the prediction tasks, and how a protein conformation can be represented in terms of physical and chemical laws (in the protein folding process). Knowledge-based methods are limited to experimental data, e.g., comparative homology modelling can only predict structures of protein sequences which are similar or nearly identical to other protein sequences of known structure. Fold recognition via threading is limited to the fold library derived from the PDB structure database. Ab initio methods can obtain new structures with novel folds. However, the complexity and high dimensionality of the conformational search space even for a small protein molecule still makes the problem intractable considering methods of any of the four classes presented in Chapter 4.

Over the last years, probably the most important results in this field were produced by hybrid methods such as the ones based on first principles with database information. Such hybrid methods combine the accuracy of knowledge-based methods with a more realistic, force field-based, physiochemical description of a protein. The last results presented in the CASP[1] competition corroborate this statement. ROBETTA (ROHL et al., 2004; SIMONS et al., 1999B), FRAGFOLD (JONES, 2001), I-TASSER (ZHANG, 2007) and LINUS (SRINIVASAN; ROSE, 1995) all belong to this class of methods. ROBETTA and I-TASSER have been the most successful predictors over the last years according to data from the biannual CASP experiments.

Protein Structure Prediction is a challenging problem and further research remains to be done. As revealed by last CASP, the development of new strategies, the adaptation and investigation of new methods and the combination of existing and state-of-the-art computational methods and techniques to the PSP problem is clearly needed. Understanding how experimental data can be better used in combination

[1] CASP. predictioncenter.org

with ab initio techniques is another open research question. In summary, there are several research opportunities and avenues to be explored in this field, with relevant multidisciplinary applications in computer science, bioinformatics, chemistry, biochemistry, and the medical sciences.

In tertiary protein structure prediction methods that use database information, there are basically four main problems that must be solved: (i) find a better way to represent computationally the polypeptide structure; (ii) built a strategy to acquire structural information from experimental templates; (iii) find a way to better distinguish between a good and a bad candidate solution, i.e, the development of an energy function that can discriminate between the set o native-like protein conformations and all other conformations; and (iv) develop a search strategy that can find the protein structure with the lowest energy in the conformational space. In the proposed strategy, called MOIRAE, the main efforts are concentrated in order to: (i) built a new strategy to acquire useful structural information from experimental templates and (ii) develop an efficient and effective strategy to search the polypeptide conformational space. The proposed method is composed of two main stages: (i) structural templates acquisition and torsion angles constraints generation and; (ii) prediction of approximate 3-D protein structures using a genetic algorithm combined with a local-search strategy. This Chapter describes the developed strategy to acquire structural template information from experimentally determined 3-D structures of proteins. The developed strategy to search the protein conformational space will be presented in Chapter 6.

5.2 Proposed method

5.2.1 Polypeptide representation

As described in Chapter 3, there are two main usually forms to represent a polypeptide structure. The first considers the Cartesian positions of each atom of the protein structure. The second model represents the polypeptide structure by its torsions angles. The use of dihedral angles has a great advantage over the Cartesian model because the degree of freedom is reduced. Nevertheless, this model presents the drawback that every small change in one dihedral angle causes drastic changes in the polypeptide structure. When dihedral angles are used, computational strategies to search the conformational space should be adapted in order to deal with this kind of problem.

In this thesis the dihedral angles model is used to represent and manipulate the polypeptide conformation as described in Chapter 3. A protein conformation C is represented as a vector $C = [d_1, d_2, \ldots, d_n]$, where d_i represents a set of main-chain and side-chain dihedral angles of an amino acid residue i. The set of consecutive dihedral angles represents the internal rotations of the polypeptide conformation. In order to compute the dihedral angles values for PHI , PSI and CHI of the polypeptide 3-D structure and also compute the position of an atom after a rotation, several routines were developed using the NAB[2] language. These computational routines were developed based on the concepts described in Chapter 3. Rotations over the torsion angles χ were performed according the group of atoms presented in Appendix A.

[2] Nucleic Acid Build. casegroup.rutgers.edu/casegr-sh-2.2.html

5.2.2 Structural templates

The acquirement of structural templates is done in ten steps: (`i`) target amino acid fragmentation; (`ii`) searching for protein templates; (`iii`) calculate secondary structure and torsion angles information; (`iv`) filter structural templates; (`v` and `vi`) compute the main-chain torsion angles of the central amino acid residue; (`vii`) clustering templates torsion angles; (`viii`) build structural patterns; (`ix`) pattern recognition using artificial neural and (`x`) construct intervals of main-chain torsion angles. Figure 5.1 shows the schematic representation of the developed strategy to acquire structural information from experimentally determined proteins from the `PDB`. `Step i` fragments the target amino acid sequence into short and consecutive amino acid fragments. `Step ii` searches the `PDB` for near exact match of protein templates[3]. `Step iii` computes the secondary structure information of each template fragment identified in `step ii`. In `step iv` a filter for secondary structure information is applied in order to remove possible structural distortions in the templates. `Step v` computes the dihedral angles of the main-chain of the central amino acid residue of the templates fragments returned by the filter process in `step iv`. If the target fragment have structural templates then built main-chain torsion angles intervals. For target fragments that does not have structural templates after the filter process, a clustering algorithm is applied into all templates obtained by `step ii` (`step viii`). `Step vii` computes the torsion angles of the central amino acid residue of all templates obtained from the `PDB`. `Step viii` constructs structural patterns that are used to train artificial neural networks (`step ix`) and predicts constrain torsion angles intervals (`step x`). Each of these steps are detailed in the next sub-sections.

5.2.2.1 Target protein sequence fragmentation

Let $X = [a_1, a_2, \ldots, a_n]$ be a vector representing the target amino acid sequence, where a_1 and a_n represent, respectively, the first and the last amino acid residues of sequence X; and L a fragment length. The set of all fragments (target fragments) with size L of sequence X can be represented as $S = \{s_1, s_2, \ldots, s_m\}$, where s_1 and s_m represent, respectively, the first and the last fragments. Two consecutive fragments s_i and s_{i+1} overlap the last $L-1$ amino acid residues from fragment s_i. Thus, $m = |X| - (L+1)$ fragments. Figure 5.2 illustrates the fragmentation process for $L = 5$. The length L is chosen to be odd because only the structural information of the central amino acid residue is further considered. The central amino acid residue is influenced by its amino acid neighbors (for example, α`-helices` have hydrogen pattern and specific values in their torsion angles ϕ and ψ). This type of fragmentation has been used successful by (DORN; BREDA; SOUZA, 2008; DORN; SOUZA, 2010). A size of five amino acid residues for each fragment was chosen because in a polypeptide structure at least four residues are needed to form the most basic structures. Each amino acid fragment consists of a vector of consecutive torsion angles $[(\phi,\psi)_1, (\phi,\psi)_2, \ldots, (\phi,\psi)_{L=5}]$, which describes the internal rotations

[3] As described by `BLAST` documentation (`www.ncbi.nlm.nih.gov/blast`): *"Short sequences (less than 20 bases) will often not find any significant matches to the database entries under the standard nucleotide-nucleotide `BLAST` settings. The usual reasons for this are that the significance threshold governed by the expect value parameter is set too stringently and the default word size parameter is set too high"*. In order to search short sequences we adopt a word size = 2 with a Expect Value = 200000.

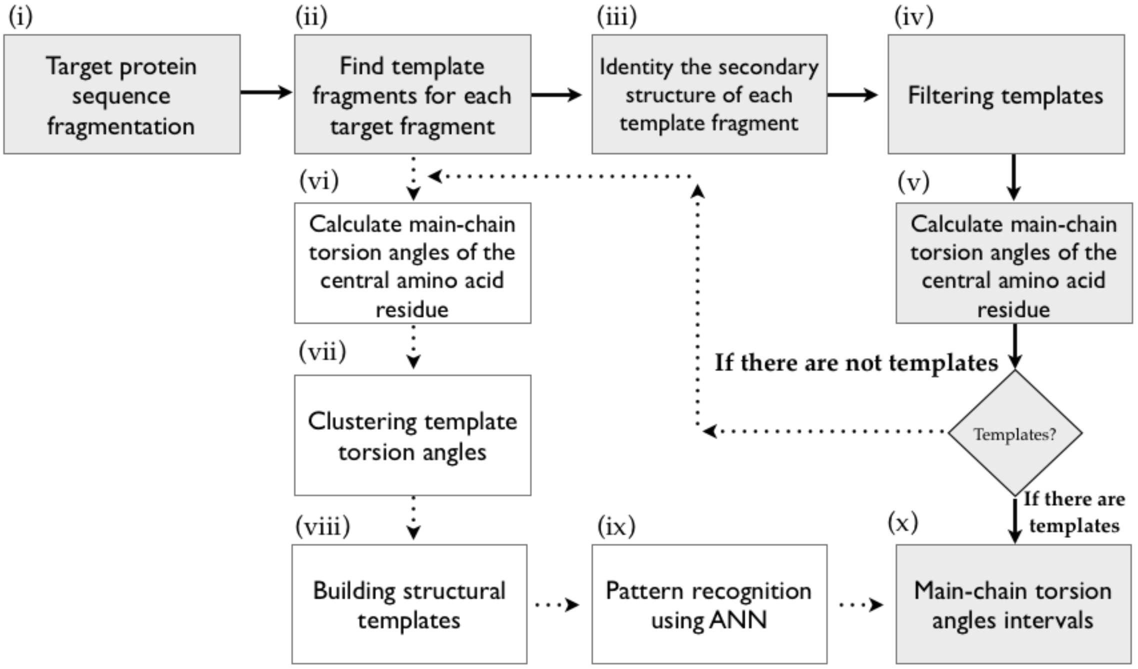

(i)
Target protein sequence fragmentation
(ii)
Find template fragments for each target fragment
(iii)
Identity the secondary structure of each template fragment
(iv)
Filtering templates
(v)
Calculate main-chain torsion angles of the central amino acid residue
(vi)
Calculate main-chain torsion angles of the central amino acid residue
(vii)
Clustering template torsion angles
If there are not templates
Templates?
If there are templates
(viii)
Building structural templates
(ix)
Pattern recognition using ANN
(x)
Main-chain torsion angles intervals

Figure 5.1: Schematic representation of the structural templates acquisition process. Steps highlighted represent the main flow of the process. White boxes represent the steps executed when no templates were present after the filtering step.

of the main-chain of the fragment.

Fragments	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12
s1	a1	a2	a3	a4	a5							
s2		a2	a3	a4	a5	a6						
s3			a3	a4	a5	a6	a7					
s4				a4	a5	a6	a7	a8				
s5					a5	a6	a7	a8	a9			
s6						a6	a7	a8	a9	a10		
s7							a7	a8	a9	a10	a11	
s8								a8	a9	a10	a11	a12

Figure 5.2: Schematic representation of the target amino acid sequence fragmentation. Central amino acid residues of each consecutive fragment of 5 residues are highlighted.

5.2.2.2 Find template fragments

Template amino acid fragments[4] for each s_k, for $k = 1, \ldots, m$, amino acid target fragment are obtained from the Protein Data Bank (PDB) (BERMAN et al., 2000) by sequence alignments using the BLASTp (Basic Local Alignment Search Tool)[5] algorithm (ALTSCHUL et al., 1997) (Algorithm 1, line 3). Let be t_i a template

[4] Template fragments are short sub-sequences of amino acid residues of proteins with known 3-D structure.

[5] BLAST blast.ncbi.nlm.nih.gov

fragment retrieved when a s_k amino acid target fragment is given as input, than let be $T_k = \{t_1, t_2, \dots, t_n\}$ represents the set of all templates fragments belonging to a target fragment s_k. Each s_k target fragment obtained from the fragmentation step is a very short sequence of amino acid residues and BLASTp can often not find significant matches in the PDB. In order to deal with this problem BLASTp documentation recommends the use of small word sizes and a large expected cut-off. This parameters values were adjusted to 2 and 200000, respectively.

Alignment substitution matrices are tailored to detect similarities among sequences with different levels of divergence (KOONIN; GALPERIN, 2002). For short sequences, BLAST documentation recommends the use of BLOSUM80 (HENIKOFF; HENIKOFF, 1993), PAM30 or the PAM70 (DAYHOFF; SCHWARTZ; ORCUTT, 1978) matrices. In this step the PAM70 substitution matrix is adopted. Algorithm 1 gives a general overview of the template fragment search step. Figure 5.3 shows a schematic representation of the set T_k of template amino acid fragments returned by the search procedure for a target fragment s_k.

```
input  : target fragments $s_k \in S$.
output: template proteins.
forall the $s_k \in S$ do
    Search the Protein Data Bank for templates;
    $T_k \leftarrow$ BLASTp($s_k$) For each target fragment a set of template fragments
    are obtained from the PDB;
end
```

Algorithm 1: Search the Protein Data Bank for template proteins.

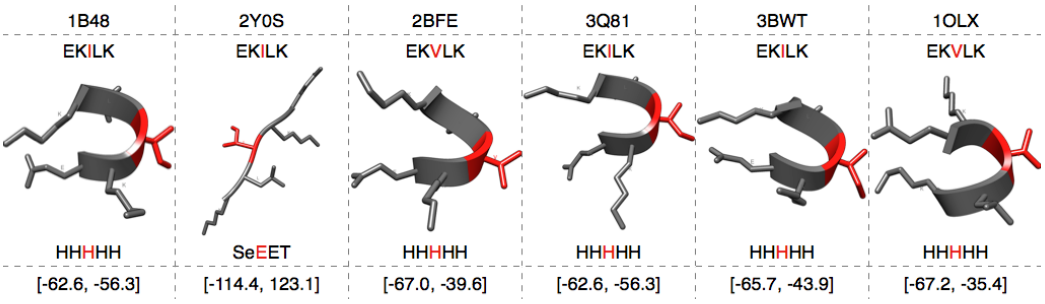


Figure 5.3: Example of templates fragments obtained from the PDB for a target fragment EKILK. Six templates fragments were returned: 1B48 (EKILK), 2Y0S (EKILK), 2BFE (EKVLK), 3Q81 (EKILK), 3BWT (EKILK), 1OLX (EKVLK). Central amino acid residues of each template fragment are highlighted. For each template fragment the secondary structure of its amino acid residues is calculated. The phi and psi torsion angles of the central amino acid residue for each template fragment is calculated. Conformational state of the template fragments were calculated by PROMOTIF (HUTCHINSON; THORNTON, 1996). Molecular graphics and analyses were performed with the UCSF Chimera package - www.cgl.ucsf.edu/chimera.

5.2.2.3 Identify secondary structures

A single polypeptide may contain multiple secondary structures. α-helix and β-sheet are the most stable secondary structures and can be considered as the principal elements present in 3-D structures of proteins. As described in Section 2.4, in a polypeptide chain, the NH group of the backbone can form a hydrogen bond with the CO group of the fourth nearest amino acid residue. These repetitions define a regular pattern known as α-helix (LESK, 2010; PAULING; COREY; BRANSON, 1951). α-helices have on average 3.6 amino acids residues by turn. This structure is stabilized by one hydrogen bond between the nitrogen (N) atom of a peptide bond and the oxygen (O) atom of the carbonyl group in the fourth amino acid residue of the region N-terminal (LEHNINGER; NELSON; COX, 2005). Each successive turn of helix is held with the adjacent turns by three or four hydrogen bonds. There is another type of helices: the 3^{10} helices. In this type of helix, hydrogen bonds are formed between residues i and $i + 3$ (VENKATACHALAM, 1968; RICHARDSON, 1981). 3^{10} helices have an average of 3 residues per turn and are narrower when compared with α-helix (LILJAS et al., 2001). The β-strands consist of extended polypeptide chains with neighboring chains extending parallel or anti-parallel to each other. The amine and carboxyl groups of peptide bonds point toward each other and in the same plane, so hydrogen bonding can occur between adjacent polypeptide chains. Adjacent β-strands can form hydrogen bonds in anti-parallel or parallel arrangements. In the first the successive β-strands alternate directions so that the N-terminal of one sheet is adjacent to he C-terminal of he next. In a parallel arrangement, all successive N-terminals are oriented in the same direction.

There is other type of regular secondary structure: β-turn. This type of secondary structure does not occur so frequently as α-helices and β-strands. β-turns are short segments and often connect two β-strands. These structures have an hydrogen bond between the CO of residue i and the NH of residue $i + 3$. There is an another type of turn where a hydrogen bond is formed between the CO group of residue i and the NH hydrogen of residue $i + 2$. This type of turn is known as γ-turn (MILNER-WHITE et al., 1988; ROSE; GIERASCH; SMITH, 1985). Finally, there are Irregular structures (or coils) that are formed in regions where the polypeptide changes their directions, i.e., after a regular secondary structure in α-helix, 3^{10}-helix, β-sheet, γ-turn state and β-turn. The irregular structures are structural elements that join successive regular secondary structures.

In this step the secondary structure state of amino acid residues of each template fragment $t_i \in T_k$ is calculated (Fig. 5.3). The set of all secondary structures templates belonging to s_k is represented as $SS_k = \{ss_1, ss_2, \ldots, ss_m\}$ (Algorithm 2). The computation of the secondary structure of each template fragment is performed by PROMOTIF (HUTCHINSON; THORNTON, 1996) that uses a local implementation of the DSSP (KABSCH; SANDER, 1983) algorithm. Eleven conformational states are used to represent the secondary structure of the template fragments:

- α-helix (H): are formed from a single consecutive set of residues in the amino acid sequence. The hydrogen-bonding patterns links the C=O group of amino acid residue i to the H-N group of residue $i + 4$ (LESK, 2010);

- α-helix (h): occurs when a α-helix secondary structure begins changing its conformational state;

- β-strand (E): are formed by the lateral interaction of independent set of residues (LESK, 2010). The amino and carboxyl groups of peptide bonds point toward each other and in the same plane, so hydrogen bonding can occur between adjacent polypeptide chains;

- β-strand (e): occurs when a β-strand secondary structure begins changing its conformational state;
- β-turn (T): is defined for 4 consecutive residues (denoted by i, i+1, i+2 and i+ 3) if the distance between the C_α atom of residue i and the C_α atom of residue $i+3$ is less than 7Å and if the central two residues are not helical (KABSCH; SANDER, 1983);
- β-turn (t): occurs when a β-turn secondary structure begins changing its conformational state;
- 3^{10}-helix (G): the hydrogen-bonding patterns links the C=O group of residue i to the H-N group of residue $i - i + 3$ (LESK, 2010);
- 3^{10}-helix (g): occurs when a 3^{10}-helix secondary structure begins changing its conformational state;
- γ-turn (Bends B and S): is represented by regions with high curvature (KABSCH; SANDER, 1983). This type of secondary structure occurs when the polypeptide chain folds back on itself to form an anti-parallel β-sheet;
- γ-turn (Bend b): occurs when a Bend secondary structure begins changing its conformational state.

```
input  : template set $T_k$.
output: secondary structure templates $ss_i \in SS_k$.
forall the $s_k \in S$ do
    $SS_k \leftarrow 0$;
    forall the $t \in T_k$ do
        compute the secondary structure information;
        $SS_k^t \leftarrow$PROMOTIF($t$);
        $SS_k \leftarrow SS_k \cup SS_k^t$;
    end
end
```

Algorithm 2: Secondary structure calculation for each template fragment.

5.2.2.4 Calculate main-chain torsion angles

Amino acid residues in a secondary structure usually adopt particular backbone torsion angles (PHI and PSI) (HOVMOLLER; OHLSON, 2002). Residues in α-helix generally adopt torsion angles that range between 180.0° < ϕ < 0.0°, 100.0° < ψ < 45.0°. 3^{10}-helix usually adopt values between -85.0°< ϕ < -65.0°, -15.0° < ψ < 10.0°. β-strands usually present values between -90.0° < ϕ < -70.0°, 140.0°< ψ < 160.0°. β-turns have an hydrogen bond between the CO of residue i and the NH of residue $i+3$. This impose strong restrictions on the conformational torsional angles of residue $i+1$ and $i+2$ (LILJAS et al., 2001). Turns are important for the preservation of the particular fold of the protein structure and are classified according to the separation between the two end residues (NÉMETHY; PRINTZ, 1972; LEWIS; MOMANY; SCHERAGA, 1973). Table F.1 (Appendix) shows the conformational angles for most common types of β-turns. Irregular structures as coils and loops, patterns of hydrogen bonds and in the same way pairs of torsion angles (ϕ and ψ) are not regular. Let $AA_k = \{aa_1, aa_2, \ldots, aa_m\}_t$ be the set of template torsion angles of the central amino acid residue retrieved when set T_k is given as input. All $aa_i \in AA_k$ torsion angles templates are computed as described by Algorithm 3. In this step the main-chain torsion angles of the central amino acid residue of each template fragments is computed. Chapter 3 describe the procedure adopted to compute the torsion angles of each template fragments. The program `Torsions` (by Dr. Andrew C.R. Martin, UCL, London) is used to automatize this step. For each template fragment $t \in T_k$ a pair of torsion angles ϕ, ψ is calculated. Each pair of torsion angles is represented as one 2-tuple aa_i=(ϕ, ψ).

```
input  : templates $t_i \in s_k$.
output: torsion angles $aa_i \in AA_k$
forall the $s_k \in S$ do
    $AA_k \leftarrow 0$;
    forall the $t \in T_k$ do
        compute $\phi$ and $\psi$ torsion angles for the central amino acid residue;
        $AA_k^t \leftarrow$TORSIONS($t$);
        $AA_k \leftarrow AA_k \cup AA_K^t$;
    end
end
```

Algorithm 3: Torsion angles calculation of the central amino acid residues.

5.2.2.5 Filtering structural templates

Searching the `PDB` using `BLAST` for structural templates means that sequence alignments were performed in order to identify protein templates with homologous sequences. Similar to comparative modelling methods it is assumed that, if the target fragment sequence is in someway similar to the sequence of the fragment template protein, the structural information obtained from the known structure can be used to model the target protein fragment (MCLACHLAN, 1992; BAJORATH; STENKAMP; ARUFFO, 1994; BLUNDELL et al., 1987). This is partially true because not always identical sequences assume the same conformation. This problem is worse when only short fragments are used to search the Protein Data Bank. Figure 5.3 shows some structural templates obtained for a target fragment s_k=`EKILK` after the template search procedure. In this example, the template fragment from protein with `PDB ID = 2YOS` present a distinct fold when compared with other template fragments despite it sequence similarities.

In order to handle with this problem a filtering strategy was developed. Let $RSS_k = \{ss_1, ss_2, \ldots, aa_m\}$ be the reduced sets of secondary structure templates, and $RAA_k = \{aa_1, aa_2, \ldots, aa_m\}$ the reduced set of torsion angles of the central amino acid residue of template fragment $t \in T_k$ after the filtering process. Algorithm 4 shows the developed filtering procedure. Initially, the secondary structure of each target fragment s_k is computed using `PROMOTIF`. The templates of secondary structures $ss_i \in SS_k$ are compared against the secondary structure of its related target fragment s_k. Only the templates $ss_i \in SS_k$ that present the central amino acid residue i and its neighbors $i+1$ and $i-1$ with the same secondary structure and the same amino acid residue type of the central amino acid residue i of its related amino acid target fragment s_k are considered for further analysis.

```
input : templates $t_i \in T_k$.
input : secondary structure templates $ss_i \in SS_k$.
input : torsion angles templates $aa_i \in AA_k$.
input : target fragments $s_K \in S$.
output: a reduced set $RSS_k$ of secondary structure templates .
output: a reduced set $RAA_k$ torsion angles templates.
forall the $s_k \in S$ do
    $RSS_k \leftarrow 0$;
    $RAA_k \leftarrow 0$;
    if secondary structure of the three central amino acid residues of
    $ss_i \in SS_k$ is equal to the secondary structure of the target sequence $s_k$
    then
        if the central amino acid residue type of $s_k$ = the central amino
        acid residue of the template fragment $t_i \in T_K$ then
            $RSS_k^t \leftarrow ss_i$;
            $RAA_k^t \leftarrow aa_i$;
            $RSS_k \leftarrow RSS_k \cup RSS_k^t$;
            $RAA_k \leftarrow RAA_k \cup RAA_k^t$;
        end
    end
end
```

Algorithm 4: Filtering template fragments.

5.2.2.6 Clustering torsion angles templates

In this step clustering techniques are applied in order to identify cluster(s) in the `Ramachandran` plot between the set of templates obtained from `PDB` (Fig. 5.4). For each s_k fragment which does not have related templates after the filtering step an clustering strategy is applied over all torsion angles templates $aa_i \in AA_k$ (Algorithm 5). Let be CLU_k a set of p clusters when a target fragment s_k is given as input, than all 2-tuples $aa_i \in AA_k$ are clustered into $c_i \in CLU_k$ clusters using the `K-means` method (LLOYD, 1982; MITRA; ACHARYA, 2005). A $c_i \in CLU_k$ represents a cluster of $aa_i \in AA_k$ pairs of torsion angles ϕ and ψ. `K-means` considers the different probabilities of distribution for each individual cluster in order to identify which set of clusters is more favorable for a given set of data. `K-means` minimizes a function E of quadratic error (Eq. 5.1), in which p clusters are present.

$$E = \sum_{j=1}^{p} \sum_{aa_i \in AA_K} |aa_i - m(c_j)|^2, \tag{5.1}$$

$$m(c_j) = \frac{1}{n} \sum_{i=1}^{n} aa_i, \tag{5.2}$$

where n describes the number of $aa_i \in c_j$ and $m(c_j)$ (Eq. 5.2) computes the mean value of all $aa_j \in c_j$. The mean value (Eq. 5.3 and Eq. 5.4) and the estimate standard deviation (Eq. 5.5 and Eq. 5.6) for each cluster $c_j \in CLU_k$ are calculated individually for ϕ and ψ.

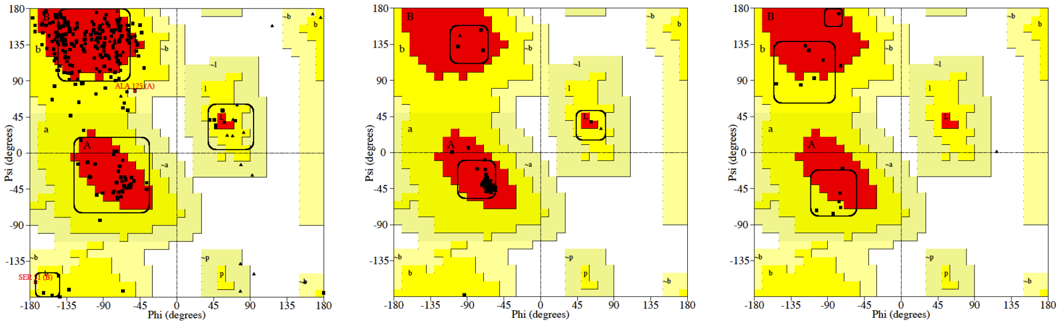
$$mphi_j = \frac{1}{n}\sum_{i=1}^{n} aa_i.\phi \tag{5.3}$$

$$mpsi_j = \frac{1}{n}\sum_{i=1}^{n} aa_i.\psi \tag{5.4}$$

$$ephi_j = \sum_{i=1}^{n} |aa_i.\phi - mphi_j|^2 \tag{5.5}$$

$$epsi_j = \sum_{i=1}^{n} |aa_i.\psi - mpsi_j|^2 \tag{5.6}$$

Empirical observations reveal that template fragments obtained from the PDB present certain regularity in the torsion angles of the central amino acid residue. This occurs, first, because in the step of collecting protein templates from the PDB homologous sequences are identified. Homologous sequences that present comparable 3-D conformations can have similar values in their main-chain torsion angles. A second issue is related with the fact that only the information of the central amino acid residue of a template is considered. This means that torsion angles values with few variations occur mainly because the central amino acid is influenced by its neighboring amino acid residues.



(a) PDB ID: 1AIL (b) PDB ID: 3CRE (c) PDB ID: 1NIL

Figure 5.4: Clustering procedure of template fragments in the Sasisekharan-Ramakrishnan-Ramachandran plot of protein with PDB ID: 1AIL (a), PDB ID: 3CRE (b) and PDB ID: 1NIL (c). Outlined areas represent identified clusters. Illustrations were prepared with PROCHECK (LASKOWSKI et al., 1996).

5.2.2.7 Built structural patterns

The template secondary structure information $ss_i \in SS_k$, the template amino acid sequence information $t \in T_k$ and the cluster information $c \in CLU_k$ are used to build structural training patterns. Let be $TP_k = \{tp_1, tp_2, \ldots, tp_m\}$ a set of training patterns when a $s_k \in S$ target fragment is given, a training pattern tp_i has

```
input  : template torsion angles $aa_i \in AA_K$.
output: mean value and estimate standard deviation for $c_j \in CLU_k$
Clusterize torsion angles using K-means.;
Minimize(E) where;
$E = \sum_{j=1}^{p} \sum_{aa_i \in AA_K} |aa_i - m(c_j)|^2$
$m(c_j) = \frac{1}{n} \sum_{i=1}^{n} aa_i$
Calculate the mean value and the standard deviation of each cluster.;
for $j \leftarrow 1$ to $p$ do
    $mphi_j = \frac{1}{n} \sum_{i=1}^{n} aa_i.\phi$
    $mpsi_j = \frac{1}{n} \sum_{i=1}^{n} aa_i.\psi$
    $ephi_j = \sum_{i=1}^{n} |aa_i.\phi - mphi_j|^2$
    $epsi_j = \sum_{i=1}^{n} |aa_i.\psi - mpsi_j|^2$
end
```

Algorithm 5: Clustering template torsion angles.

the form: $tp_i = ss_1, ss_2, ss_3, ss_4, ss_5, aa_1, aa_2, aa_3, aa_4, aa_5 : c_j$, where ss_i represents the secondary structure state of the ith amino acid residue of a template fragment $t \in T_K$, aa_i represents the ith amino acid residue of a template fragment $t_i \in T_k$ and c_j represents jth cluster $c_j \in CLU_k$ on which the template fragment belongs after the clustering step. The secondary structure $(ss_1, ss_2, ss_3, ss_4, ss_5)$ and the type of each amino acid $(aa_1, aa_2, aa_3, aa_4, aa_5)$ are used as the input object and cluster identification (c_j) as a desired output. All template fragment obtained from the PDB are used. Figure 5.5 illustrates the structural patterns building.

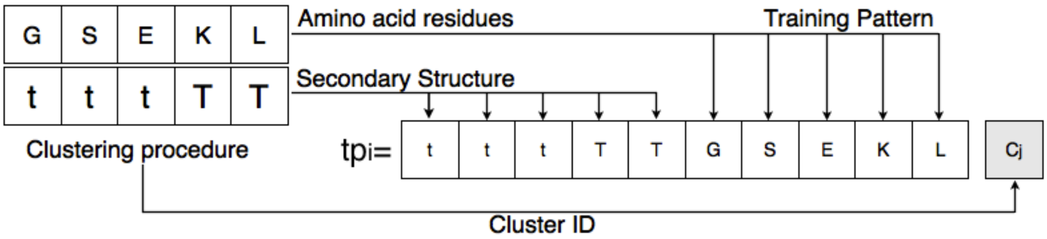


Figure 5.5: Building conformational training patterns. The secondary structure information, the amino acid residue information and its relate cluster are used to built structural patterns.

5.2.2.8 Built artificial neural networks

In this step, for each s_k target fragment which does not have related templates after the filtering procedure, artificial neural networks are built. The main goal in using ANNs is to learn how the secondary structure information combined with the amino acid sequence of a template fragment influences the torsions of the central amino acid residue. The architecture of each artificial neural network consists of three layers:

- `Layer 1`: input layer with 10 neurons corresponding to the conformational pattern size.

- `Layer 2`: hidden layer with 5 hidden neurons. The number of hidden neurons in each hidden layer was defined empirically and represents a trade off between performance and risk of over-fitting.

- `Layer 3`: output layer, the number of output neurons is equal of the number of clusters identified in the clustering step.

The artificial neural networks are trained using all conformation patterns $tp_i \in TP_k$. For the learning task a back-propagation algorithm is used. The same training parameters were used to each neural networks: `"learning rate"` = 0.02, `"max epochs"` = 50,000, `"epochs between reports"` = 100. Weights are randomly selected. These parameters where selected empirically and represented the set which return the best results of the experiments.

Prediction samples are assembled with the information obtained from the target s_k fragments (secondary structure and amino acid sequence information) and submitted to its related trained artificial neural networks. The network outputs represent the cluster of torsion angles ϕ and ψ presented by the amino acid residue in the center of A s_k fragment. Figure 5.6 illustrates this process.

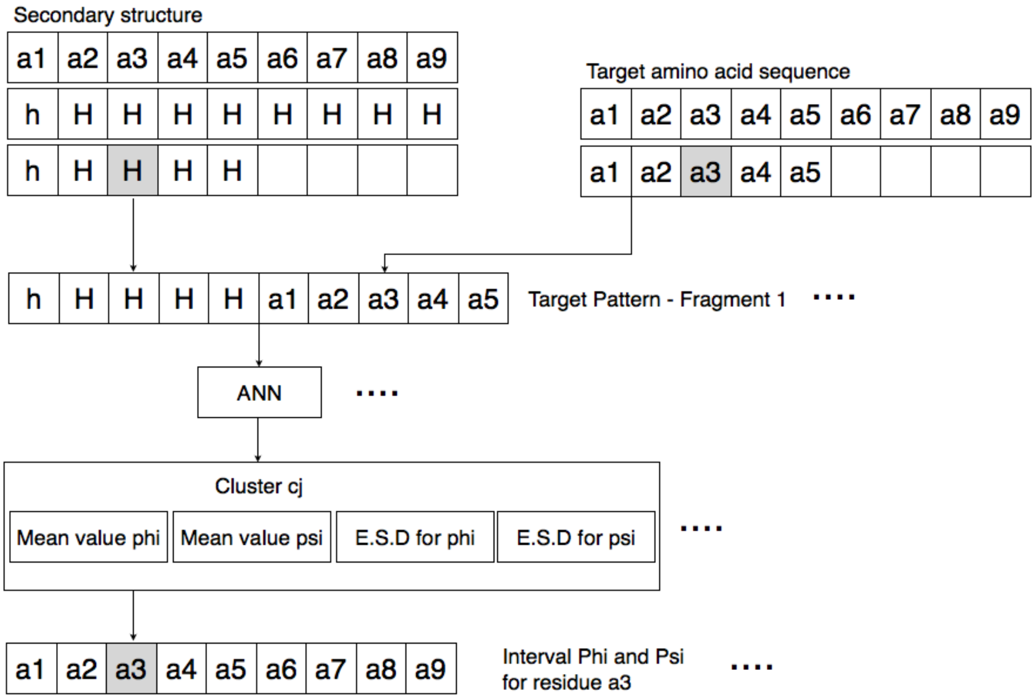

Secondary structure
a1 a2 a3 a4 a5 a6 a7 a8 a9
h H H H H H H H H
h H H H H
Target amino acid sequence
a1 a2 a3 a4 a5 a6 a7 a8 a9
a1 a2 a3 a4 a5
h H H H H a1 a2 a3 a4 a5
Target Pattern - Fragment 1
....
ANN
....
Cluster cj
Mean value phi
Mean value psi
E.S.D for phi
E.S.D for psi
....
a1 a2 a3 a4 a5 a6 a7 a8 a9
Interval Phi and Psi for residue a3
....

Figure 5.6: Schematic representation of the use of `ANN` to predict main-chain torsion angles intervals.

Artificial neural networks (`ANNs`): are computational models, which replicate the function of the biological neural network (HAYKIN, 1998) and are used to solve complex functions in various applications, for example: pattern recognition (MARQUES, 1999; SA, 2001) and secondary protein structure prediction (KARCI; DEMIR, 2009).

Artificial neural networks methods are widely used to recognize patterns and making simple rules for complex problems. They also have excellent training capabilities and are good at generalizing from a set of training data.

An ANN is composed by a set of processor units called neurons. In the usual way the type of processing of a single neuron is described as the linear combination of entries with weights ($w_i x_i$) (Eq. 5.7), followed by the passage of its values of an activation function $g(.)$ (Eq. 5.8). Depending on the type of problem to be solved there are some restrictions on the types of networks and of learning algorithms possible to use. A single neuron is defined by

$$u_k = \sum_{i=0}^{n} w_{ki} x_i, \tag{5.7}$$

$$y_k = g(u_k + b_k), \tag{5.8}$$

where $x_1, x_2, \ldots, x_n$ are input signals; $w_{k_1}, w_{k_2}, \ldots, w_{kn}$ are weights of a neuron k; u_k is the output of the linear combination obtained through the input signals (x_i); b_k is the bias; $g(.)$ is an activation function; and y_k is the output signal. The use of a bias b_k apply transformations in the output u_k of the linear combinator (Eq. 5.9).

$$v_k = u_k + b_k, \tag{5.9}$$

An important stage of an ANN is the training step. In this phase the ANN is trained to return a specific output when a specific input is given; this is done by continued training on a set of training data. Initially the weights and bias are chosen randomly. When training an ANN with a set of input and output data, weights are adjusted in order to give the same outputs as seen in the training data. The weights, after training, contain meaningful information. When a satisfactory level of performance is reached, the training phase stops, and the ANN uses the weights to make decisions about unknown inputs. The goal of any training algorithm is to minimize the global error (e_k). When one tries to minimize this error using gradient descent for the class of neural networks called Multi-Layer Perceptrons. There are two major learning paradigms, each corresponding to a particular abstract learning task. The first is the `supervised learning`, where we have a set of example pairs and the aim is to find a function in the allowed class of functions that matches the examples. Tasks that fall within the paradigm of `supervised learning` are pattern recognition (also known as classification) and regression (also known as function approximation). The second paradigm is the `unsupervised learning` where we have some data x, and the cost function to be minimized can be any function of the data x and the networks output. Many training algorithms are available to `supervised learning`, one example is the `back-propagation` algorithm (HAYKIN, 1998).

In the back-propagation algorithm, after propagating an input through the network, the error is calculated and then is propagated back through the network while the weights are adjusted in order to make the error smaller. First the input is propagated through the ANN to the output. After this, the error e_k on a single output neuron k is calculated (Eq. 5.10)

$$e_k = d_k + y_k, \tag{5.10}$$

where y_k is the calculated output and d_k is the desired output of a neuron k. This error value is used to calculate a δ_k value, which is again used for adjusting the weights (Eq. 5.11).

$$\delta_k = e_k g(y_k), \tag{5.11}$$

where $g(.)$ is the derived activation function. When the δ_k value is calculated, we can calculate the δ_j value for preceding layers. The δ_j values of the previous layer is calculated from the δ_k values of this layer (Eq. 5.12)

$$\delta_j = \eta g(y_j) \sum_{k=0}^{k} \delta_k w_{jk} \tag{5.12}$$

where K is the number of neurons in this layer and η is the learning rate parameter, which determines how much the weight should be adjusted. Using these δ values, the Δ_w values that the weights should be adjusted by, can be calculated (Eq. 5.13)

$$\Delta W_{jk} = \delta_j k \tag{5.13}$$

The ΔW_{jk} value is used to adjust the weight w_{jk}, by $w_{jk} = w_{jk} + \Delta W_{jk}$ and the back-propagation algorithm moves onto the next input and adjusts the weights according to the output. This process goes on until a certain stop criteria is reached. The stop criteria is typically determined by measuring the mean square error of the training data while training with the data, when this mean square error reaches a certain limit, the training is stopped.

5.2.2.9 Construct main-chain torsion angles intervals

In this step main-chain torsion angles intervals for ϕ and ψ are built for each amino acid residue of target sequence X. A constrain interval $I \in \mathbb{R}$ is represented as $I = [\underline{i}, \bar{i}]$, where $\underline{i}$ and $\bar{i}$ represent, respectively, the lower and the upper bound of an interval I. Algorithm 6 schematizes the steps to construct main-chain torsion angles intervals. There are two main flows: (i) built torsion angles intervals for s_k target fragments that present structural templates after the filtering process and (ii) when structural templates are not present after the filtering process then apply artificial neural networks to predict the torsion angles intervals (clusters).

For each s_k target fragment that have templates fragments ($aa_i \in RAA_k$) after the filtering process (first flow) the mean value for ϕ (Eq. 5.14) and ψ (Eq. 5.15) are computed. Additionally, the standard deviation estimate for ϕ (Eq. 5.16) and ψ (Eq. 5.17) are also computed.

$$mphi_k = \frac{1}{m} \sum_{aa_i \in RAA_k} aa_i.\phi \tag{5.14}$$

$$mpsi_k = \frac{1}{m} \sum_{aa_i \in RAA_k} aa_i.\psi \tag{5.15}$$

$$ephi_k = \sum_{aa_i \in RAA_k} |aa_i.\phi - mphi_k|^2 \tag{5.16}$$

$$epsi_k = \sum_{aa_i \in RAA_k} |aa_i.\psi - mpsi_k|^2 \tag{5.17}$$

For s_k target fragment that does not present template fragments after the filtering procedure, artificial neural networks are built in order to predict the torsion angles intervals (flow 2). After predict the cluster of a target fragment s_k the mean and estimate standard deviation values are obtained. Let θ be a torsion angle ϕ or ψ. Intervals of torsion angles are built through the mean and the estimated standard deviation values. A closed interval of torsion angles is represented as $[\theta] = [\underline{\theta}, \overline{\theta}]$. The lower bound $\underline{\theta}$ of an interval $[\theta]$ is built from the difference between the mean value $m(.)$ and the estimated standard deviation $\sigma(.)$. The upper bound $\overline{\theta}$ of an interval $[\theta]$ is obtained trough the mean $m(.)$ and the estimated standard deviation $\sigma(.)$ sum (Eq. 5.19).

$$\underline{\theta} = m(\theta) - \sigma(\theta) \tag{5.18}$$

$$\overline{\theta} = m(\theta) + \sigma(\theta) \tag{5.19}$$

```
input  : torsion angles $aa_i \in RAA_k$.
input  : clusters $c_j \in CLU_k$.
output: Main-chain torsion angle interval.
forall the $s_k \in S$ do
    if there are templates after the filtering step. then
        forall the $aa_i \in RAA_k$ do
            $mphi_k = \frac{1}{n} \sum_{i=1}^{n} aa_i.\phi$
            $mpsi_k = \frac{1}{n} \sum_{i=1}^{n} aa_i.\psi$
            $ephi_k = \sum_{i=1}^{n} |aa_i.\phi - mphi_k|^2$
            $\sigma_{psi} = \sum_{i=1}^{n} |aa_i.\psi - mpsi_k|^2$
        end
    end
    else
        Use $mphi_k$, $mpsi_k$, $ephi_k$, $epsi_k$ from ANNs predictions.
    end
end
```

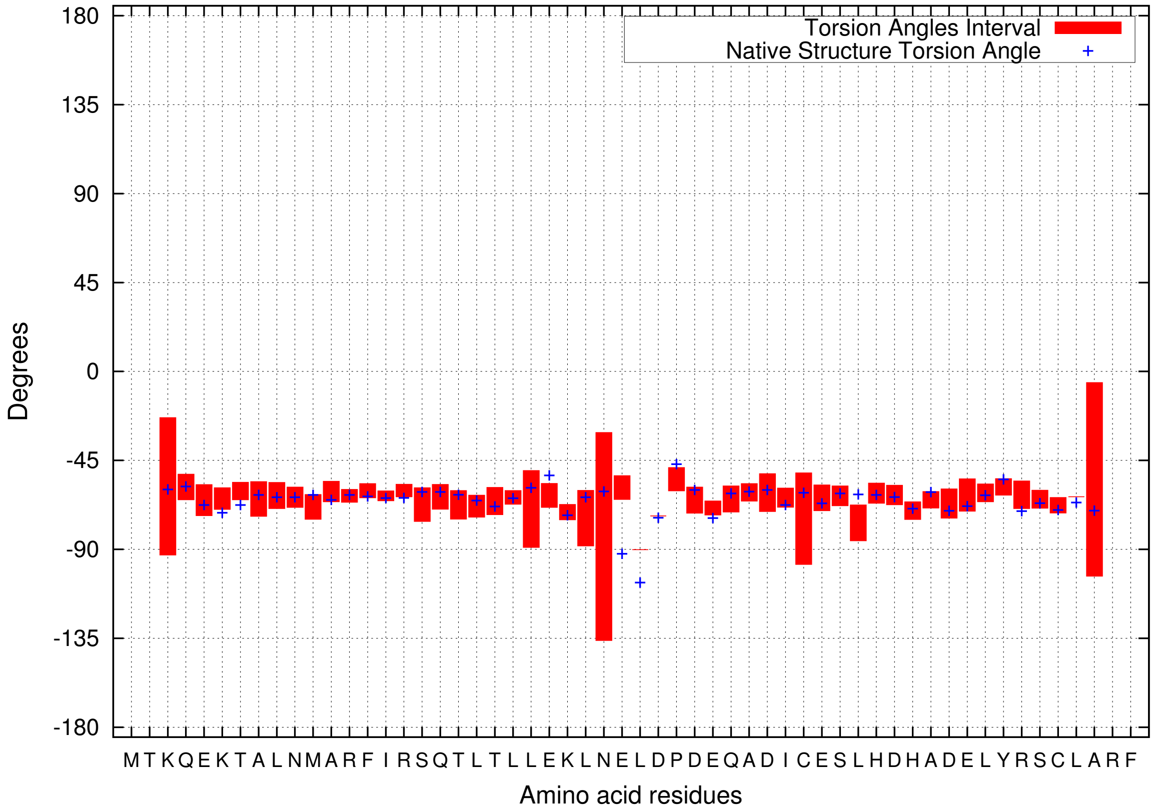
Algorithm 6: Compute main-chain torsion angles intervals.

Intervals of ϕ, ψ torsion angles are built for each s_k fragment. From this point, each s_k fragment is represented as a set of ϕ, ψ torsion angles intervals: $S = s_1$ = $(\underline{\phi}, \overline{\phi}, \underline{\psi}, \overline{\psi})$, $s_2 = (\underline{\phi}, \overline{\phi}, \underline{\psi}, \overline{\psi})$, ..., $s_m = (\underline{\phi}, \overline{\phi}, \underline{\psi}, \overline{\psi})$, where $(\underline{\phi}, \overline{\phi})$ and $(\underline{\psi}, \overline{\psi})$ are, respectively, the lower limit and the upper limit for ϕ and ψ of the central amino acid residue of a target fragment s_k. The lower and upper bounds of each dihedral angle represent a limited area of variation of the template torsion angles in the Ramachandran plot.

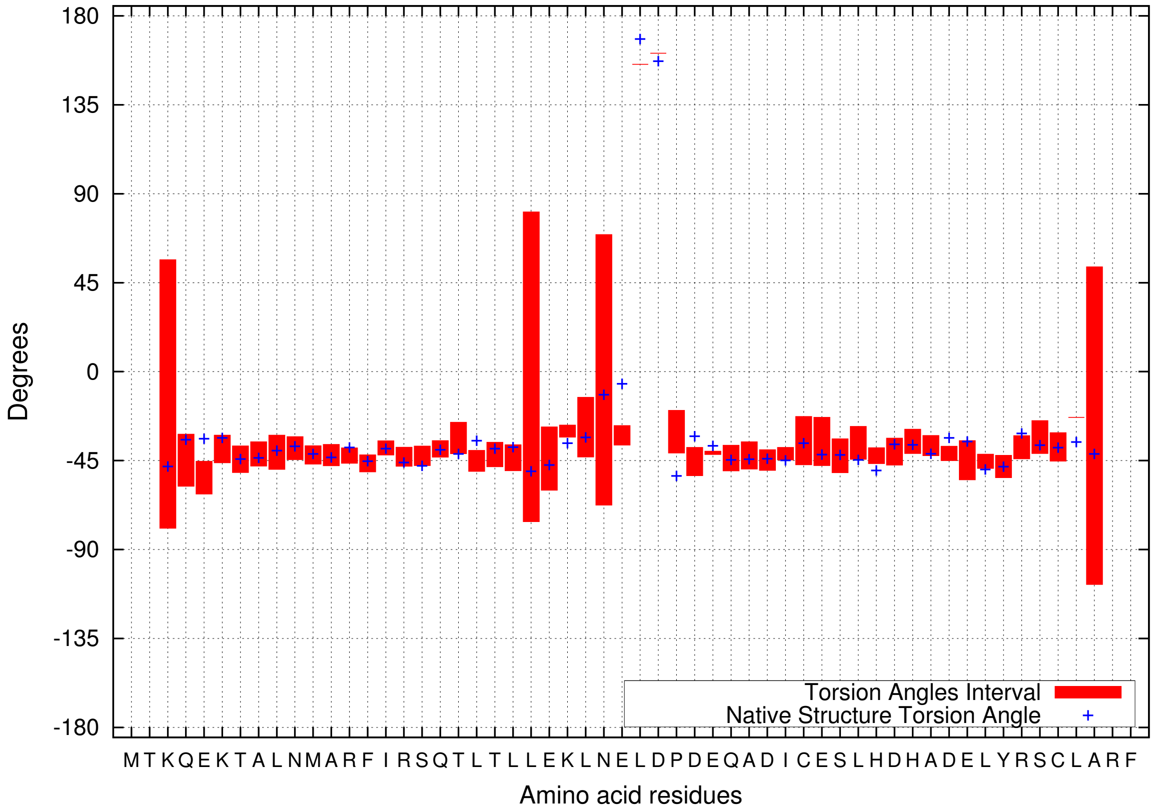
Conformational space reduction: we can estimate the proportion of reduction of the conformational space when torsion angle intervals are computed. Let n be the number of amino acid residues in a target protein sequence; let p be the number of torsion angles of each i amino acid residue of the target sequence (angles phi and psi); let k be the number of possible values that each p angle can assume; let $X = [\underline{x}, \overline{x}]$ be an interval of torsion angles and let ϵ be the size of an interval, than the number λ of possible combinations is of torsion angles is computed by Equation 5.20.

$$\lambda = \prod_{i=1}^{n} (k_i)^p = [k_i^p \times k_{i+1}^p \times \ldots \times k_{i=n}^p] \tag{5.20}$$

where k_i is obtained as $\epsilon = ||\underline{x} - \overline{x}|_\Phi - |\underline{x} - \overline{x}|_\Psi|$, where if $\epsilon = 0$, then $k_i = |\underline{x} - \overline{x}|_\Phi$ or $k_i = |\underline{x} - \overline{x}|_\Psi$, elif $\epsilon \neg 0$, then if $|\underline{x} - \overline{x}|_\Phi > |\underline{x} - \overline{x}|_\Psi$ then $k_i = \lfloor \frac{\epsilon}{2} + |\underline{x} - \overline{x}|_\Psi \rfloor$, or if $|\underline{x} - \overline{x}|_\Phi < |\underline{x} - \overline{x}|_\Psi$ then $k_i = \lfloor \frac{\epsilon}{2} + |\underline{x} - \overline{x}|_\Phi \rfloor$. If we assume the torsion angles space as discrete, then the number of all possible combination of phi and psi torsion in a polypeptide with 56 amino acid residues (`PDB ID = 1B6Q`) is equal to $\approx 2.0e^{+286}$ when the interval of each torsion angle is equal to `-180.0°  180.0°`. If we compute the torsion angles interval of each amino acid residue the number of combinations is equal to $\approx 9.3e^{+127}$. Figure 5.7 present the torsion angles intervals (phi and psi) for each amino acid residue of the protein with `PDB ID = 1B6Q`.



(a) Torsion angles PHI.



(b) Torsion angles PSI.

Figure 5.7: Torsion angles interval for the protein with PDB ID = 1B6Q. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.

5.3 Chapter conclusions

This chapter presented a computational strategy to acquire structural information from experimental `3-D` determined protein structures. Template information has been used to construct main-chain torsion angles intervals for each amino acid residue of the target sequence. These torsion angles intervals allows a considerable reduction in the conformational space of the protein structure. Chapter 6 presents a new computational strategy to search the conformation space using the structural information acquired from experimentall `3-D` determined protein structures in order to find native-like protein structures.

6 MOIRAE: A LOCAL-SEARCH-BASED GENETIC ALGORITHM TO SEARCH THE 3-D PROTEIN CONFORMATIONAL SPACE

6.1 Introduction

As described in Chapter 5, in order to develop a first principle method that uses database information we have to deal with four main steps: (i) to represent computationally the polypeptide structure; (ii) to acquire structural information of protein templates; (iii) to find a way to distinguish between good and bad solutions; and (iv) to provide a strategy to search the protein conformational space for the native-like 3-D protein structure. In Chapter 5, a representation of a polypeptide conformation (torsion angles of the main-chain and side-chains) has been described. Additionally, a new strategy to acquire useful structural information of protein templates was presented. This Chapter presents the developed strategies to deal with the two last steps. Firstly, a potential energy function and an implicit solvent model to evaluate protein conformations are described. Finally, a procedure to speed up the search of protein structure conformations by improving candidate solutions locally is presented.

6.2 Proposed method

6.2.1 Potential energy function and implicit solvation model

Commonly, a potential energy function incorporates two types of terms (Eq. 6.1) (MACKERREL, 2010): bonded and non-bonded. The bonded terms (bonds, angles and torsions) are covalently linked. The bonded terms constrain bond lengths and angles near to their equilibrium values. The bonded terms also include a torsional potential (torsion) that models the periodic energy barriers encountered during bond rotation. The non-bonded potential includes: ionic bonds, hydrophobic interactions, hydrogen bonds, van der Waals forces, and Dipole-dipole bonds. The potential energy function describes the interactions between the atoms in the system. The energy function is used to evaluate the quality of any given conformation defined by the Cartesian coordinate vector. The lower the energy value, the better should be the conformation. Equation 6.1 describes the AMBER[1] potential energy function (WEINER et al., 1984; CORNELL et al., 1995) used in this work. AMBER is one of the most commonly used potential energy functions to study protein struc-

[1] AMBER. ambermd.org

tures. The AMBER potential energy function is built-in in the programming language (NAB[2]) used in this work to develop the routines to manipulate the polypeptide structures.

$$\begin{aligned} E_{\text{total}} = & \sum_{\text{bonds}} K_b(r_{ij} - r_{eq})^2 \\ & + \sum_{\text{angle}} K_\theta(\theta - \theta_{eq})^2 \\ & + \sum_{\text{dihedrals}} \sum_{\text{n}} \frac{V_n}{2}[1 + cos(n\phi - \gamma)] \\ & + \sum_{\text{i}} \sum_{\text{j>i}} \left[4\varepsilon \left(\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right) + \frac{q_i q_j}{\epsilon r_{ij}} \right], \end{aligned} \tag{6.1}$$

where $\mathtt{r}_{ij} = \mathtt{r}_i$-$\mathtt{r}_j$ represents the bond lenght, $\mathtt{k}_b$ is the bond stretching constant, $\mathtt{r}_{eq}$ is the equilibrium bond distance, $\mathtt{k}_\theta$ is the bond angle constant, θ_{eq} is the equilibrium bond angle, ϕ is the torsion angle, γ is the phase angle and $\mathtt{v}_n$ is the torsional barrier. The last two terms represent the Lennard-Jones potential and the Coulumb interaction, where, ε is the van der Waals well depth, σ the van der Waals diameter, q is the charge of each atom, and ϵ is the dielectric constant. Each member of the family of AMBER force fields provides values for these parameters. In this work the AMBER ff99SB protein force field (HORNAK et al., 2006) was used. The AMBER ff99SB force field is an improved version of the widely used AMBER ff94 force field to study of proteins and nucleic acids. AMBER ff99SB uses the Cornel (CORNELL et al., 1995) electrostatic model but changes several geometrics values in order to improve this parameters.

Protein 3-D structures in the nature are greatly influenced by the aqueous environment in which they exist. The native structure of most proteins are only marginally stable, and achieve stability only within narrow ranges of conditions of solvent (LESK, 2010). Implicit solvation models treat the protein environment as continuum and reduces drastically the computational time. Implicit solvation models are useful when extensive sampling of conformational space is required, as for example in protein folding simulation. The most common implicit solvation model is the Generalized Born (GB) model (STILL; YEO H.C. AMD KOLATKAR; CLARKE, 1990; TSUI; CASE, 2001). In this work the OBC (Onufriev, Bashford, Case) variant of the GB model (ONUFRIEV; BASHFORD; CASE, 2000, 2004) provided by the AMBERTOOLS[3] package was used.

6.2.2 Search strategy

Along the last years, many search techniques have been applied to the 3-D protein structure prediction problem: Genetic Algorithms (PEDERSEN; MOULT, 1997; TUFFERY et al., 1991), Monte Carlo simulations (SIMONS et al., 1997), Molecular Dynamics simulations (GUNSTEREN; BERENDSEN, 1990; RAPAPORT, 2004)

[2] Nucleic Acid Build. casegroup.rutgers.edu/casegr-sh-2.2.html

[3] casegroup.rutgers.edu

(see Chapter 4). These methods changes the orientation of atoms of the protein structure in order to minimize an energy function (DESJARLAIS; CLARKEB, 1998). Genetic algorithms (HOLLAND, 1975) has been applied successfully to a large numbers of problems (LANGDON; POLI, 2010; FLOREANO; MATTIUSSI, 2008) and commonly, in order to save computational time, potential energy function as describe in Section 6.2.1 is used to evaluate the polypeptide structures (individuals) along the genetic algorithm simulations.

A `GA` is a population-based meta-heuristic that runs for many iterations, called `generations` (LUKE, 2009). During each generation, individuals are combined through a `crossover` procedure for generating new individuals for composing the next generation (LANGDON; POLI, 2010). In the `GA` context, a set of individuals form a population, a problem solution is an `individual`, and each element of the solution is a `gene`. Each individual from the population is evaluated, through an objective function. A `GA` selects well evaluated individuals to `crossover`, aiming at improving the quality of the population from one generation to the next.

A `GA` was proposed in this thesis to search the protein `3-D` conformational space. A procedure to speed up the search by improving candidate solutions locally was also developed. This procedure is called `local-search` and explores the neighbors of a solution aiming at finding a better solution than the current one. The genetic algorithm is combined with a structured population (BURIOL et al., 2005), and it is hybridized with the `local-search` procedure. The population is structured in `castes`. The fittest `20%` of the individuals compose caste `A`, the `50%` least fit ones compose caste `B`, and the remaining `30%` compose caste `C`. The crossover operator is a random key scheme that prioritizes (given `70%` of chances) genes originated from solution originated from set A. A `local-search` procedure is applied to each `individual` obtained after a crossover procedure. In the `local-search` procedure small perturbations are applied to genes. When selecting an individual for crossover, we know how it is classified in comparison with the other solutions from the population. Moreover, caste A is maintained as an elite set. Figure 6.1 schematizes the developed genetic algorithm. Following the main operators and strategies of the proposed `GA` algorithm are presented in details.

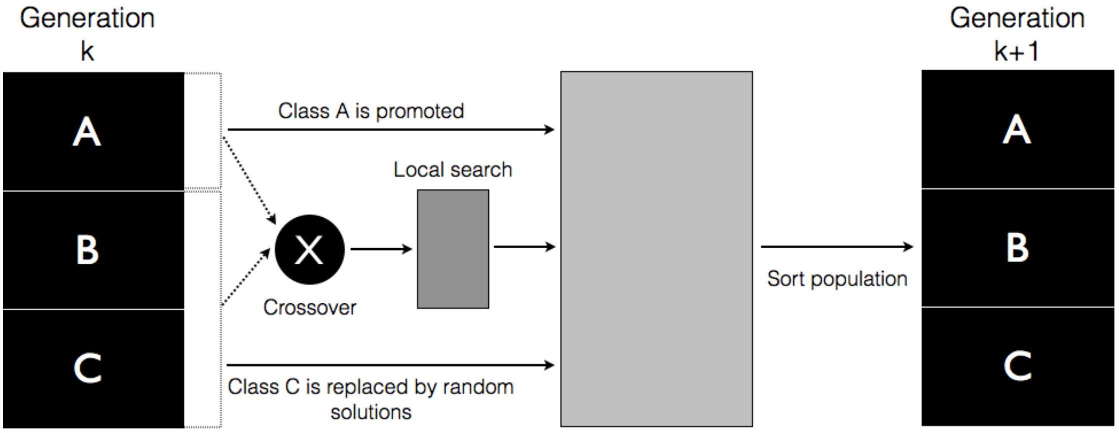
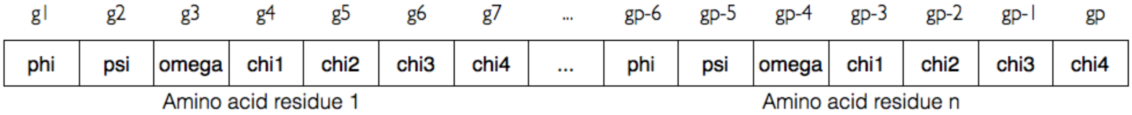


Figure 6.1: Genetic algorithm with the `local-search` operator. `Local-search` is used to improve a candidate solution.

`INDIVIDUAL REPRESENTATION:` as described in Chapter 3, a small change in one dihedral torsion angle causes drastic changes in the `3-D` protein structure. Therefore, we can define the problem of adjusting the torsion angles of a protein structure as a high-precision optimization problem. Thus, we represent an individual as a vector of n numbers belonging to the domain of real numbers (floating point numbers). Each position of this vector of torsion angles (backbone and side-chain) represents a `gene`. Each amino acid residue is comprised of at least by two `genes` representing the two dihedral angles (ϕ, ψ) from the protein backbone and a number of `genes` representing the χ angles that varies according with the type of amino acid residue. Figure 6.2 illustrates the representation of an individual.



g1	g2	g3	g4	g5	g6	g7	...	gp-6	gp-5	gp-4	gp-3	gp-2	gp-1	gp
phi	psi	omega	chi1	chi2	chi3	chi4	...	phi	psi	omega	chi1	chi2	chi3	chi4
		Amino acid residue 1								Amino acid residue n				

Figure 6.2: `Individuals` representation. Each `individual` is represented as a vector of main-chain and side-chain torsion angles.

`INITIAL POPULATION:` in a GA implementation the initial population should be sufficiently large and diverse to ensure that individuals display different fitness values (FLOREANO; MATTIUSSI, 2008). The size of the population depends on two factors: (i) the properties of the search space and (ii) the computational cost of evaluating all individuals of several generations. Taking into account these two factors we set a population size of 100 individuals. The initial population was randomly generated (`GENERATE` function in Algorithm 10). As described in Chapter 5, after the structural template acquisition each s_k target fragment is represented as a set of ϕ, ψ torsion angles intervals: $s_k = (\underline{\phi}, \overline{\phi}, \underline{\psi}, \overline{\psi})$. This means that for each amino acid residue of the target sequence we have associate an interval of torsion angles for phi and psi. Let be sol_i = [phi, psi, omega, chi1, chi2, chi3, chi4, ..., phi, psi, omega, chi1, chi2, chi3, chi4] a vector containing the main-chain and side-chain torsion angles of an individual. Each main-chain torsion angles (ϕ and ψ) were randomly selected from its corresponding intervals. The torsion angle *omega* is fixed in `180.0`°. `MOIRAE` computes intervals only for the main-chain torsion angles, side-chains torsion angles are selected randomly from intervals torsion angles obtained from the `Dunbrack` rotamers library (DUNBRACK JR.; COHEN, 1997; DUNBRACK JR.; KARPLUS, 2003) (see Tab. G.1).

`FITNESS FUNCTION:` the fitness function associates a value score to each individual (sol_i) of the population (function `EVALUATE` in Algorithm 9). Evaluating the fitness function of individuals is often the most time-consuming part of one evolutionary algorithm (FLOREANO; MATTIUSSI, 2008). In order to deal with this problem, the `AMBER` potential function (HORNAK et al., 2006) described in Section 6.2.1 and the `Generalized Born` implicit solvent model (ONUFRIEV; BASHFORD; CASE, 2000, 2004) were used for the evaluation of individuals of a population. The main advantage of using a `potential` energy function and an implicit solvation model is the reduction of the computational time to evaluate each individual.

`SORT OPERATOR:` let $SOL = \{sol_1, sol_2, \ldots, sol_{100}\}$ be a set of solutions (`individuals`) of a `population`. All $sol \in SOL$ are ranked according to its potential energy using a `QUICK-SORT` algorithm.

`CROSSOVER OPERATOR:` genetic operators introduce diversity in the population and allow the exploration of novel solutions. The crossover (or combination) operator combines inherit characteristics from two parents solutions by creating pairwise recombination of `genes`. Algorithm 7 presents a general structure of the crossover operator. The developed crossover operator is a random key scheme (line 2) that prioritizes (given `70%` of chances) genes originated from solution originated from class A (line 3). All the individuals of class B are generated using the crossover operator (Fig. 6.1).

```
  input : a $sol_A$ from class A
  input : a $sol_{BC}$ from class B+C
  output: a new $sol_{new}$
1 for $i \in \{1, ..., NumGenes\}$ do
2     if RAND(0,1)<=0.7 then
3         $sol_{new}$[i] = $sol_A$[i];
4     end
5     else
6         $sol_{new}$[i] = $sol_{BC}$[i];
7     end
8 end
9 Return $sol_{new}$.
```

Algorithm 7: Crossover operator.

`DIVERSITY CONTROL OPERATOR:` if the population of a `GA` simulation has lost most of its diversity[4], the fitness values may not grow further or may take a lot of generations to display some improvement. In order to control the diversity of the individuals of class A (class with the `30%` best individuals) during the `GA` simulation we developed a `Diversity Control Operator`. Algorithm 8 presents the basic structure of the `Diversity Control Operator`. Let sol_{i-1} and sol_i be two solutions of a sorted class A, than the diversity control operator computes the c_α `RMSD` (root mean square deviation) value between these two solutions (Eq. 6.2). If the `RMSD` value is less then a `Threshold` (line 3) then the solution sol_{i-1} is maintained (line 7) and the solution sol_i is repleaced by a new randomly generated solution (line 4).

$$\mathbf{RMSD}(a,b) = \sqrt{\left(\sum_{i=1}^{n} \|r_{ai} - r_{bi}\|^2\right)/n}, \tag{6.2}$$

The population size is kept the same, since once this solution is inserted, after sorting the population, the worst solution is discarded.

`LOCAL-SEARCH OPERATOR:` after each `crossover` operation the `local-search` operator is applied to the new solution. Algorithm 9 presents a general structure of the `local-search` operator. Let λ be an adjustment value to be applied to a g

[4] Diversity is important in genetic algorithms because crossing over a homogeneous population does not yield new solutions.

```
input  : $SOL$ current population
input  : Threshold
output: an updated population $SOL$
$SOL_{new} = 0$;
for $i \in \{2, ..., |A|\}$ do
    if RMSD($sol_i$,$sol_{i-1}$)$<=$Threshold then
        $SOL_{new} = SOL_{new} \cup$ Generate a random new $sol_i$
    end
    else
        $SOL_{new} = SOL_{new} \cup sol_i$
    end
end
Return $SOL$=$SOL_{new}$.
```

Algorithm 8: Diversity control operator.

(gene) value than the `local-search` operator starts with a gene of and individual (*sol*) and proceeds as follow: (1) computes the energy of the current solution (line 3), generate an adjustment randomly (line 4), increase the value of gene i and computes the potential energy of the new solution (line 6). If the potential energy decreases after the adjustment then continue to increase the value of gene i until the energy begins to increase (lines 7-14). Save and apply `local-search` operator to gene $i + 1$; (2) If in the first attempt the energy of the new solution increases than start to decreases the value of gene i until the energy of the solution begins to increase (lines 16-30). Save and apply `local-search` to next gene $i + 1$. When all genes of the current solution were processed then return the solution (line 32). The `Local-search` operator is used so that the `GA` can escape from local minima. It changes each individual torsion angle from the protein backbone and side-chain so that a small change in the angles does not lead to stereo-chemical discharges which increase the potential energy of the molecules.

In Algorithm 10 a general structure of the developed `GA` is presented. Initially, in line 1, the population is generated. Initial solutions are generated at random, with angles selected from the intervals described in Chapter 5. The population is then sorted by increasing order of their objective function values, and classes A, B and C are defined (they are, respectively, the first `20%`, `50%` and `30%` of the solutions). Next, the loop in lines 3 to 14 iterates *NumGen* times. In each generation, $|B|$ solutions from the next generation are generated by crossover (lines 4-8). Each crossover is applied considering one solution selected at random from set A, and another selected at random from sets $B + C$ (Fig. 6.1). The generated solution is evaluated and inserted in the population of the next generation (line 8). To promote elite solutions, the `GA` adds all solutions belonging to set A directly to the next population (line 10). To complete the population solutions from the next generation, the `GA` adds $|C|$ new solutions generated at random with the same procedure used for generating the solutions from the initial population (line 11).

```
input  : an individual sol
output: an individual sol with minimized energy
for $i \in \{1, ..., NumGen\}$ do
    sol' = sol;
    $E_1$ = EVALUATE(sol);
    $\lambda$ = RAND(0,1);
    sol' = sol[i] + $\lambda$; the gene $i$ (torsion angle) is increased
    $E_2$ = EVALUATE(sol');
    if $E_2 < E_1$ then
        while $E_2 < E_1$ do
            sol = sol';
            $\lambda$ = RAND(0,1);
            sol'[i] = sol'[i] + $\lambda$; the gene $i$ (torsion angle) is
            increased
            $E_2$ = EVALUATE(sol');
            $E_1$ = EVALUATE(sol);
        end
    end
    else
        sol' = sol;
        $\lambda$ = RAND(0,1);
        sol'[i] = sol'[i] - $\lambda$; the gene $i$ (torsion angle) is
        decreased
        $E_2$ = EVALUATE(sol');
        if $E_2 < E_1$ then
            while $E_2 < E_1$ do
                g = g';
                $\lambda$ = RAND(0,1);
                sol'[i] = sol'[i] - $\lambda$; the gene $i$ (torsion angle) is
                decreased
                $E_2$ = EVALUATE(sol');
                $E_1$ = EVALUATE(sol);
            end
        end
    end
end
Return sol;
```

Algorithm 9: The `local-search` operator.

```
input  : a target sequence $X$ of amino acid residues.
output: Best solution.
GENERATE and EVALUATE the initial population.
SORT the solutions and define classes $A$, $B$ and $C$;
for $i \in \{1, ..., NumGen\}$ do
    for $j \in \{1, ..., |B|\}$ do
        Select solution $sol_1$ from solution set $A$;
        Select solution $sol_2$ from solution set $B + C$;
        Apply CROSSOVER($sol_1$, $sol_2$) operator;
        Apply LOCAL-SEARCH operator and add the
        solution to the next population;
    end
    Add all solutions from solution set $A$ to next
    population;
    GENERATE, EVALUATE and add $|C|$ random solutions
    to the next population;
    SORT the next population and define classes $A$, $B$ and
    $C$;
    Consider the next population as the current
    population;
end
Return the best solution from the current population.
```

Algorithm 10: Genetic algorithm with the `local-search` operator.

6.3 Chapter conclusions

In this chapter, a genetic-based algorithm to search the protein conformation space was presented . The algorithm combines a genetic algorithm with a structured population and it is hybridized with a `local-search` operator. The developed method allows efficient mechanisms for protein structure prediction. This is achieved by the use of `local-search` operator which allows the `GA` to escape from local minima. Next chapter (Chapter 7) presents the experiments and obtained results with the application of the techniques described in this chapter and in Chapter 5.

7 EXPERIMENTAL RESULTS

7.1 Introduction

This chapter presents and discusses the results obtained with the application of the computational strategy MOIRAE described in Chapters 5 and 6 to predict the 3-D structure of 20 target protein sequences. The proposed method was tested with protein sequences whose sizes vary from 14-70 amino acid residues. The results show that the predicted tertiary structures adopt a fold comparable to the experimental structures. Structural analysis reveals that the proposed method presents satisfactory results in their predictions. All MOIRAE routines described in Chapters 5 and 6 were implemented in C and NAB languages. The evaluation function (potential energy function) was executed on a shared memory scheme using OPENMP. Each prediction reported in this chapter takes about 24-48 hours of CPU time in a Linux environment of a PC Intel Core i7 3.07 GHz 8MB of Cache and 5GB RAM. This time depends on the length of the amino acid sequence of the target protein. Section 7.2 presents the target protein sequences used to test the MOIRAE strategy. Section 7.3 shows the obtained results with the acquisition of templates from the PDB and the torsion angles intervals construction. Section 7.4 presents the obtained results with the application of the developed search strategy in order to find the native-like 3-D structure of the target protein sequences. Finally, section 7.5 shows the structural analysis of the predicted 3-D structures.

7.2 Model and target proteins

The amino acid sequences of 20 proteins were obtained from the PDB (BERMAN et al., 2000) and used as study cases in our experiments: 1AB1 (Fig. 7.1a - Black), 1ACW (Fig. 7.1b - Black), 1AIL (Fig. 7.1c - Black), 1BO3 (Fig. 7.1d - Black), 1B6Q (Fig. 7.1e - Black), 1BDC (Fig. 7.1f - Black), 1BGK (Fig. 7.1g - Black), 1BHI (Fig. 7.1h - Black), 1DFN (Fig. 7.1i - Black), 1DVO (Fig. 7.1j - Black), 1EOQ (Fig. 7.1k - Black), 1ENH (Fig. 7.1l - Black), 1FME (Fig. 7.1m - Black), 1K43 (Fig. 7.1n - Black), 1OVX (Fig. 7.1o - Black), 1Q2K (Fig. 7.1p - Black), 1QR8 (Fig. 7.1q - Black), 1ROO (Fig. 7.1r - Black), 1ROP (Fig. 7.1s - Black), 1WQC (Fig. 7.1t - Black). These study cases were selected in order to test our method with different classes of polypeptides with different folding patterns (LILJAS et al., 2001). Table 7.1 presents details of the target proteins. Column 2 presents the target amino acid sequences, Column 3 presents the number of amino acid residues of each target protein and Column 4 shows the SCOP classification of each target protein. As described in Section 2.5, proteins can be classified into groups according on their structural motifs and evolutionary relationships: all α-helical; all β-sheet; α+β; α/β; Membrane and cell; Multi-domain and Small proteins. For the experiments protein sequences were selected from five different groups: Small proteins (1AB1 - Fig. 7.14a, 1ACW - Fig. 7.14b, 1BGK - Fig. 7.2g, 1BHI - Fig. 7.2h, 1DFN - Fig. 7.2i, 1OVX - Fig. 7.2o, 1Q2K - Fig. 7.2p and 1ROO - Fig. 7.2r); all α-helical proteins (1AIL - Fig. 7.14c, 1B6Q - Fig. 7.14e, 1BDC - Fig. 7.2f, 1DV0 - Fig. 7.2j, 1ENH - Fig. 7.2l and 1ROP - Fig. 7.2s); Designed proteins (1FME - Fig. 7.2m and 1K43 - Fig. 7.2n); Peptides (1B03 - Fig. 7.14d and 1E0Q - Fig. 7.2k) and Coiled coil (1QR8 - Fig. 7.2q). Designed proteins are experimental structures of protein with essentially non-natural sequences. The class of Small proteins represents the proteins with only some secondary structures maintained by disulphide bonds or ligands. The class of all alpha protein represents the proteins with its secondary structure formed exclusively by α-helices. Coiled coil is a structural motif in which 2-7 α-helices are coiled together (LIU et al., 2006). The class peptides represents the structures of peptides and fragments.

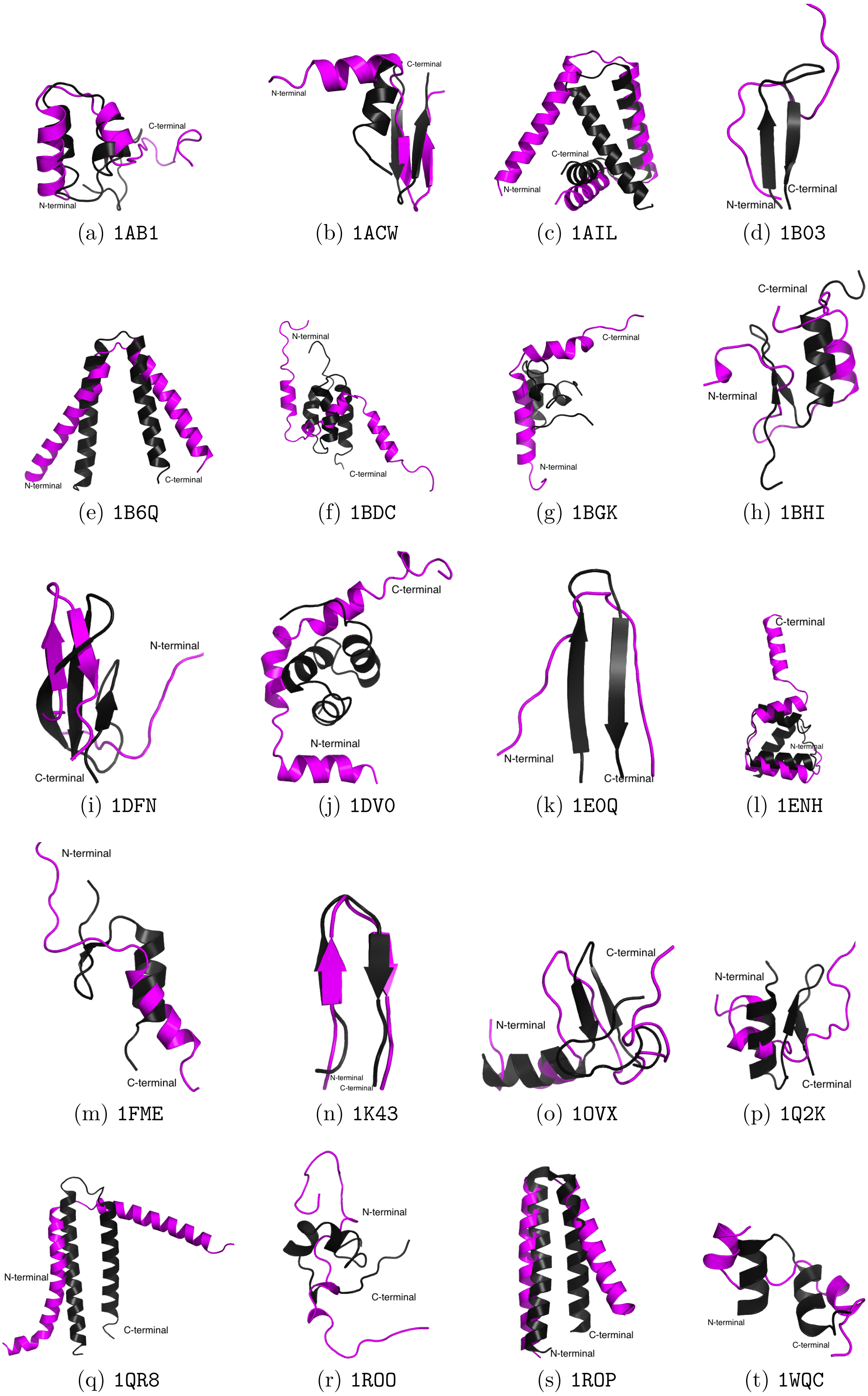
One way to characterize the fold of a protein structure is by the arrangement of these secondary structures as they pack together. The protein topology can be defined as the relationship between the sequential ordering of secondary structures and their spatial organization. Figure 7.2 shows the topology arrangement of the twenty target proteins an its secondary structure composition. Helices are shown in red, strands in pink. Proteins with PDB ID 1K43 (Fig. 7.2n), 1E0Q (Fig. 7.2k) and 1B03 (Fig. 7.14d) are composed only by β-strands organized in a structural motif known as beta hairpin[1] (sometimes also called beta-ribbon or beta-beta unit). The protein with PDB ID 1DFN (Fig. 7.2i) is composed by three consecutive antiparallel β-strands linked together by hairpin loops forming a structural motif known as β-meander. Proteins with PDB ID 1AIL (Fig. 7.14c), 1B6Q (Fig. 7.14e), 1BDC (Fig. 7.2f), 1BGK (Fig. 7.2g), 1DV0 (Fig. 7.2j), 1ENH (Fig. 7.2l), 1QR8 (Fig. 7.2q), 1ROO (Fig. 7.2r), 1ROP (Fig. 7.2s), 1WQC (Fig. 7.2t) are composed by α-helices joined by short strands of amino acid residues in a structural motif known as Helix-turn-Helix. Proteins with PDB ID 1ACW (Fig. 7.14b), 1OVX (Fig. 7.2o) , 1BHI (Fig. 7.2h) and 1Q2K (Fig. 7.2p) presents two beta strands with an alpha helix end folded presenting a motif known as Zinc Finger. For each study case we remove all protein templates whose sequences are equal to the full sequence of the target protein.

The 20 target protein sequences were submitted to MOIRAE in order to predict their 3-D structures. In Section 7.3 we analyse the construction of the main-chain torsion angles intervals for each target protein and show the benefices of using this procedure to reduce the protein 3-D conformational space. The main-chain torsion angles intervals of each target protein ware used as input for the GA-based search strategy described in Chapter 6. In Section 7.4 we show and analyse the time costs of the developed search strategy. Fitness graphs are presented in order to show the convergence of the GA strategy. Structural analysis of the predicted 3-D structures are presented in Section 7.5. We analyse the root mean square deviation (RSMD) of the predicted 3-D structures when compared with its corresponding native structure, the stereo chemical quality of the predicted secondary structures and the topology of the predicted structures.

[1] Two antiparallel beta strands connected by a tight turn of a few amino acids between them.

Table 7.1: Target protein sequences and its SCOP classification. Case studies were selected in order to test the MOIRAE strategy with different protein folding patterns.

PDB ID	Target Sequence	Res.	SCOP Class
1AB1 (YAMANO; HEO; TEETER, 1997)	TTCCPSIVARSNFNVCRLPGTSEAICATYTGCIIIPGATCPGDYAN	46	Small
1ACW (BLANC et al., 1996)	VSCEDCPEHCSTQKAQAKCDNDKCVCEPI	29	Small
1AIL (LIU et al., 1997)	MDSNTVSSFQVDCFLWHVRKQVVDQELGDAPFLDRLRRDQKSLRGRGSTLGLNIEAATHVGKQIVEKILK	70	All alpha
1B03 (TUGARINOV; ZVI; LEVY, 1999)	RKSIRIQRGPGRAFVTIG	18	Peptides
1B6Q (GLYKOS; CESARENI, 1999)	MTKQEKTALNMARFIRSQTLTLLEKLNELDPDEQADICESLHDHADELYRSCLARF	56	All alpha
1BDC (GOUDA et al., 1992)	TADNKFNKEQQNAFYEILHLPNLNEEQRNGFIQSLKDDPSQSANLLAEAKKLNDAQAPKA	60	All alpha
1BGK (DAUPLAIS et al., 1997)	VCRDWFKETACRHAKSLGNCRTSQKYRANCAKTCELC	37	Small
1BHI (NAGADOI et al., 1999)	MSDDKPFLCTAPGCGQRFTNEDHLAVHKHKHEMTLKFG	38	Small
1DFN (HILL et al., 1991)	DCYCRIPACIAGERRYGTCIYQGRLWAFCC	30	Small
1DV0 (WITHERS-WARD et al., 2000)	QEKEAIERLKALGFPESLVIQAYFACEKNENLAANFLLSQNFDDE	45	All alpha
1E0Q (ZERELLA et al., 2000)	MQIFVKTLDGKTITLEV	17	Peptides
1ENH (CLARKE et al., 1994)	RPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI	54	All alpha
1FME (SARISKY; MAYO, 2001)	EQYTAKYKGRTFRNEKELRDFIEKFKGR	28	Designed
1K43 (PASTOR et al., 2002)	RGKWTYNGITYEGR	14	Designed
1OVX (DONALDSON; WOJTYRA, 2003)	LLYCSFCGKSQHEVRKLIAGPSVYICDECVDLCNDIIR	38	Small
1Q2K (CAI et al., 2004)	AACYSSDCRVKCVAMGFSSGKCINSKCKCYK	31	Small
1QR8 (JI et al., 1999)	SGIVQQQNNLLRAIEAQQHLLQLTVRGIKQLQARSGGRGGWMEWDREINNYTSLIHSLIEESQNQQE	67	Coiled coil
1ROO (TUDOR et al., 1996)	RSCIDTIPKSRCTAFQCKHSMKYRLSFCRKTCGTC	35	Small
1ROP (BANNER; KOKKINIDIS, 1987)	MTKQEKTALNMARFIRSQTLTLLEKLNELDADEQADICESLHDHADELYRSCLARF	56	All alpha
1WQC (CHAGOT et al., 2005)	DPCYEVCLQQHGNVKECEEACKHPVE	26	absent



(a) 1AB1 (b) 1ACW (c) 1AIL (d) 1B03

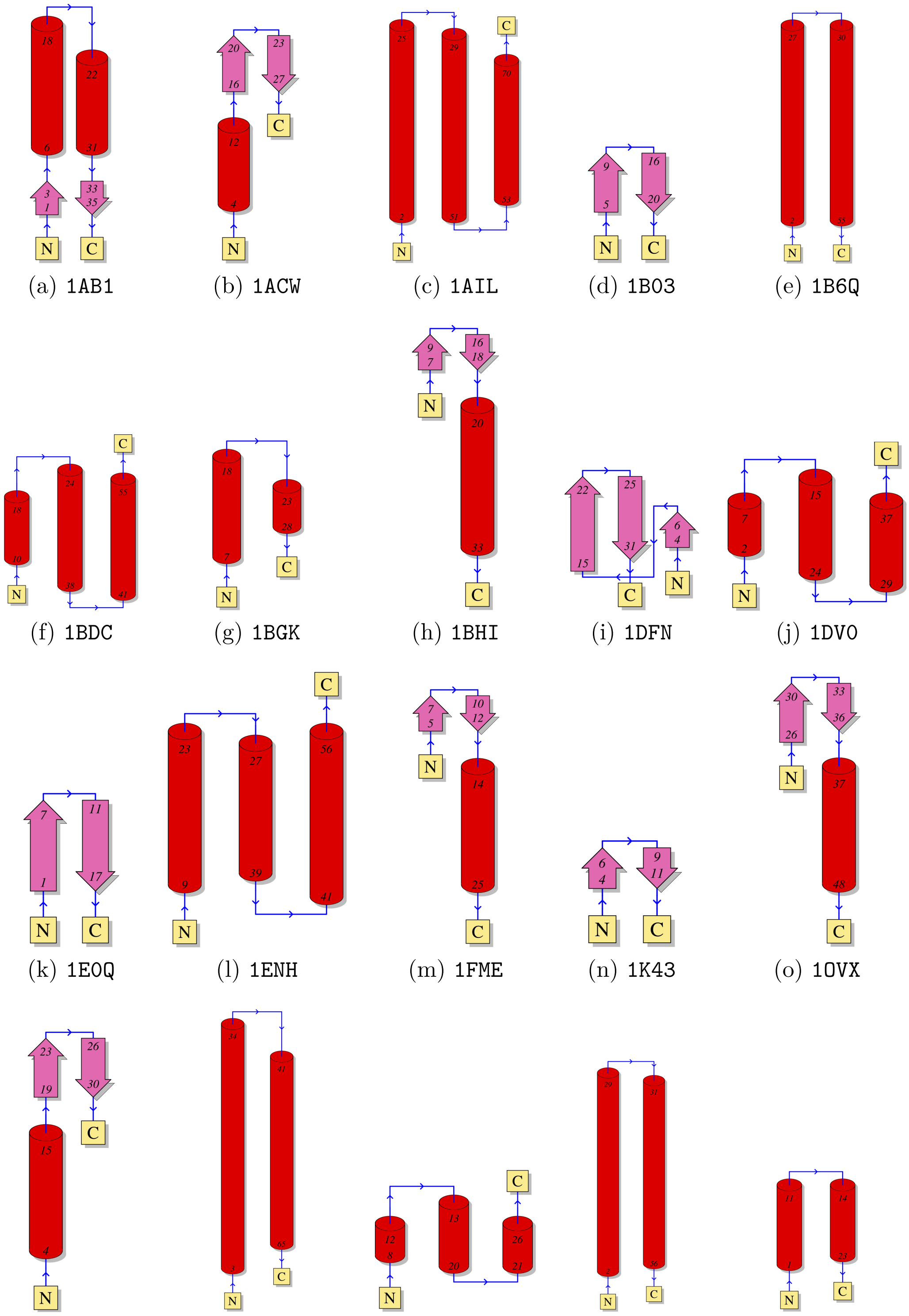
(e) 1B6Q (f) 1BDC (g) 1BGK (h) 1BHI

(i) 1DFN (j) 1DV0 (k) 1E0Q (l) 1ENH

(m) 1FME (n) 1K43 (o) 1OVX (p) 1Q2K

(q) 1QR8 (r) 1ROO (s) 1ROP (t) 1WQC

Figure 7.1: Ribbon representation of the experimental (black) and predicted by MOIRAE (magenta) 3-D structures. The C_α of the experimental and the predicted 3-D structure are fitted. Amino acid side chains are not shown for clarity. Graphic representation was prepared with PYMOL (www.pymol.org).



(a) 1AB1 (b) 1ACW (c) 1AIL (d) 1B03 (e) 1B6Q

(f) 1BDC (g) 1BGK (h) 1BHI (i) 1DFN (j) 1DV0

(k) 1E0Q (l) 1ENH (m) 1FME (n) 1K43 (o) 1OVX

(p) 1Q2K (q) 1QR8 (r) 1ROO (s) 1ROP (t) 1WQC

Figure 7.2: Diagram representing the topology of the target 3-D protein structures. N and C represent the N-terminal and the C-terminal regions, respectively. α-helices are showed in red, β-sheets are showed in pink and coil regions are showed in blue. Graphic representation was generated by PDBSUM (www.ebi.ac.uk/pdbsum).

7.3 Computing main-chain torsion angles intervals

MOIRAE was applied to all target amino acid sequences presented in Table 7.1 in order to obtain structural templates from the PDB. Table 7.2 summarizes the approximate number of combinations of the backbone torsion angles when a full interval and when the computed intervals (Eq. 5.20) are used. As described in Section 5.2.2.9 the approximate number of combinations presented in Table 7.2 was calculated in a discrete space. Column 3 presents the approximate number of combinations when full intervals are considered for ϕ and ψ and Column 4 presents the approximate number of combination when the computed intervals are considered. As can be observed, the number of combinations is greatly reduced. This number does not consider the real number space, however, gives us an idea about the search space reduction when the interval approach is used.

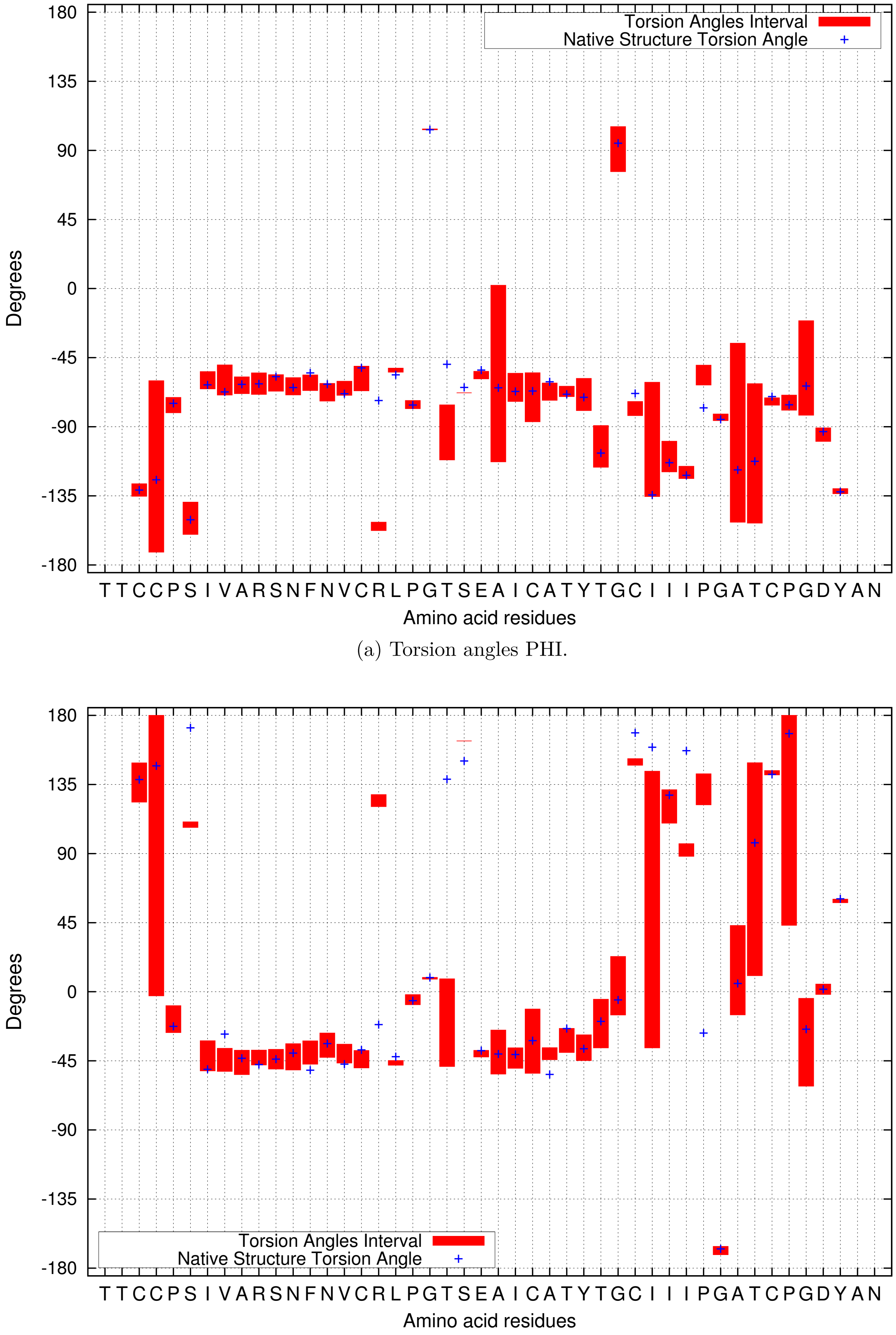
Table 7.2: Approximate number of combinations.

PDB ID	Number of residues	Full interval (-180°,180°)	Predicted Intervals
1AB1	46	≈ 1.0e+235	≈ 2.7e+101
1ACW	29	≈ 2.0e+148	≈ 3.1e+092
1AIL	70	≈ 1.0e+358	≈ 1.4e+124
1B03	18	≈ 1.0e+092	≈ 3.3e+042
1B6Q	56	≈ 2.0e+286	≈ 9.3e+127
1BDC	60	≈ 7.0e+306	≈ 1.9e+167
1BGK	37	≈ 1.0e+189	≈ 2.6e+115
1BHI	38	≈ 2.0e+194	≈ 6.1e+094
1DFN	30	≈ 2.0e+153	≈ 4.3e+068
1DV0	45	≈ 1.0e+230	≈ 1.9e+099
1E0Q	17	≈ 9.0e+086	≈ 1.4e+038
1ENH	54	≈ 1.0e+276	≈ 7.5e+122
1FME	28	≈ 1.0e+143	≈ 1.4e+080
1K43	14	≈ 4.0e+071	≈ 4.0e+007
1OVX	38	≈ 2.0e+194	≈ 3.0e+101
1Q2K	31	≈ 3.0e+158	≈ 2.4e+063
1QR8	67	≈ 6.0e+347	≈ 1.2e+143
1ROO	35	≈ 1.0e+179	≈ 1.5e+077
1ROP	56	≈ 2.0e+286	≈ 1.0e+117
1WQC	26	≈ 9.0e+132	≈ 4.1e+054

We estimate the quality of the predicted torsion angles intervals by analysing if there enclosures the torsion angle value present in the native structure of the protein. Figure 7.3 shows the torsion angles intervals for ϕ (7.3a) and ψ (7.3b) of the protein with PDB ID: 1AB1. Filled boxes represent the torsion angles intervals computed by MOIRAE. Blue dots identify the torsion angle values of the amino acid residue in the native state of the target protein. For ϕ (7.3a), from the 42 amino acid residues with computed intervals[2], 38 (91%) of them enclosures the torsion angle values of the protein native structure. For ψ (7.3b), 34 (81%) of them enclosures

[2] 1AB1 presents 46 amino acid residues. However, the fragmentation scheme adopted in MOIRAE makes that the first and the last two amino acid residues are lost. For these amino acid residues the torsion angles are fixed on 180.0°.

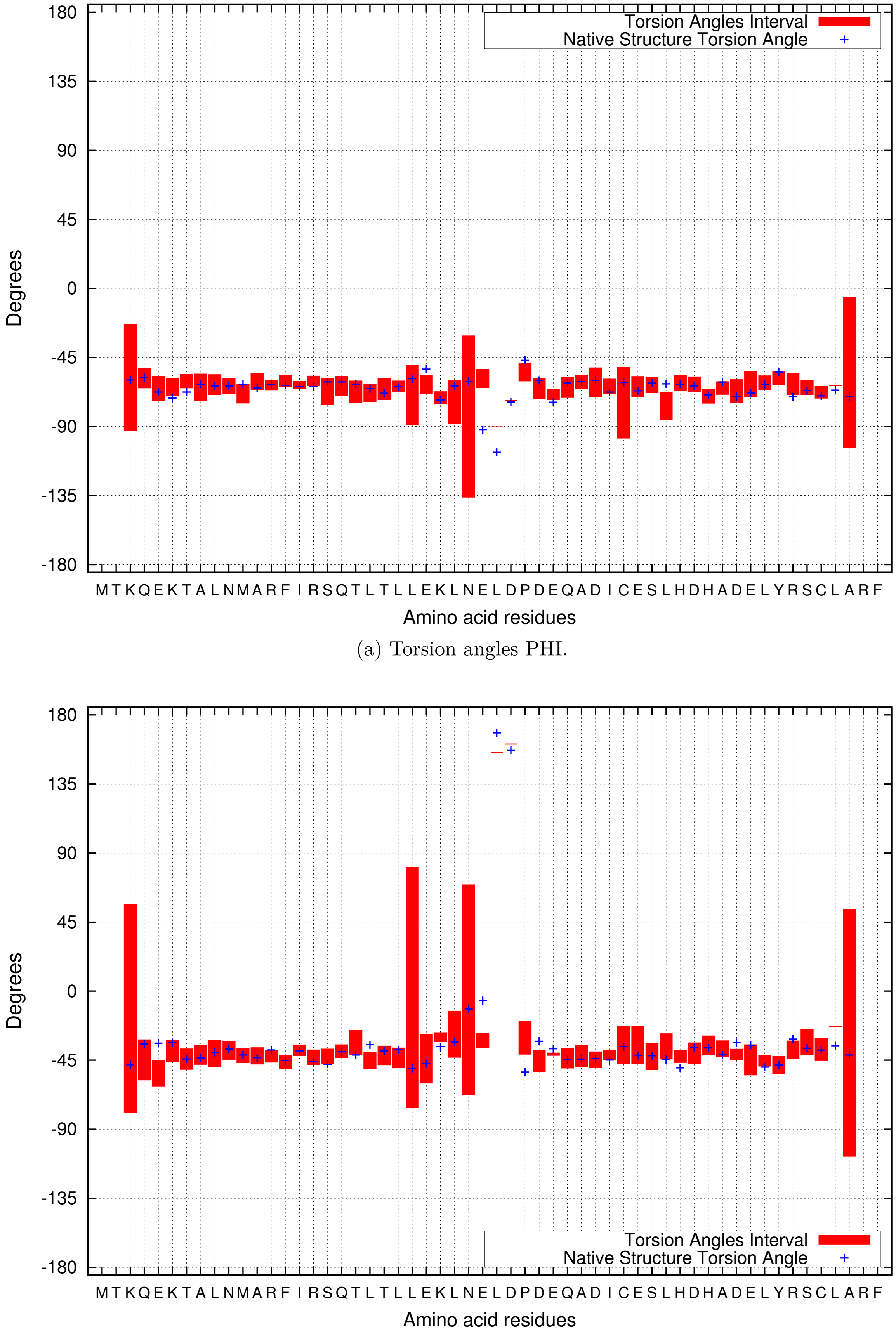
the torsion angle values of protein native structure. Although some of the torsion angles are not enclosed, they are very near to the interval limits. Figures 7.4, 7.6, 7.7 illustrate the torsion angles intervals for phi (a) and psi (b) of the proteins with 1B6Q, 1ROP and 1DFN, respectively. Figure 7.5 shows the torsion angles intervals for the protein with PDB ID 1K43. Figure 7.5a shows the torsion angles intervals for phi and Figure 7.5b shows the torsion angles intervals for psi. As can be observed the size of the torsion angles intervals of 1K43 are very small when compared with the torsion angles intervals of other study cases. This occurs, because the number of template fragments identified by MOIRAE for 1K43 is small. Appendix H presents the torsion angles interval for the proteins PDB ID: 1ACW (Fig. H.1), 1AIL (Fig. H.2), 1B03 (Fig. H.3), 1BDC (Fig. H.4), 1BGK (Fig. H.5), 1BHI (Fig. H.6), 1DV0 (Fig. H.7), 1E0Q (Fig. H.8), 1ENH (Fig. H.9), 1FME (Fig. H.10), 1OVX (Fig. H.11), 1Q2K (Fig. H.12), 1QR8 (Fig. H.13), 1ROO (Fig. H.14), 1WQC (Fig. H.15). As can be observed, in ∼90% of the computed torsion angles interval of each target protein enclosures the torsion angle values of the protein native structure.



(a) Torsion angles PHI.

(b) Torsion angles PSI.

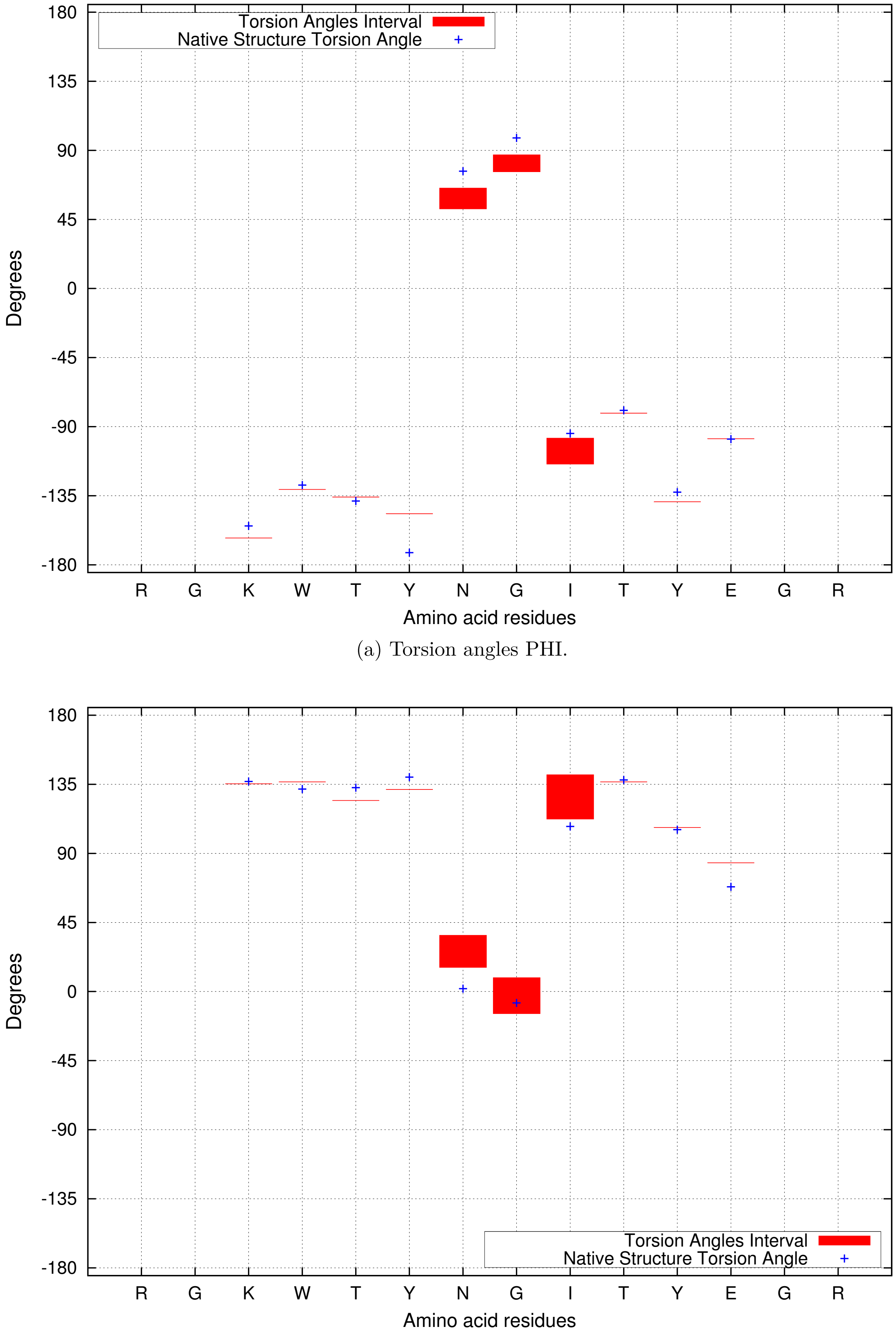
Figure 7.3: Torsion angles interval for the protein with PDB ID = 1AB1. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.



(a) Torsion angles PHI.

(b) Torsion angles PSI.

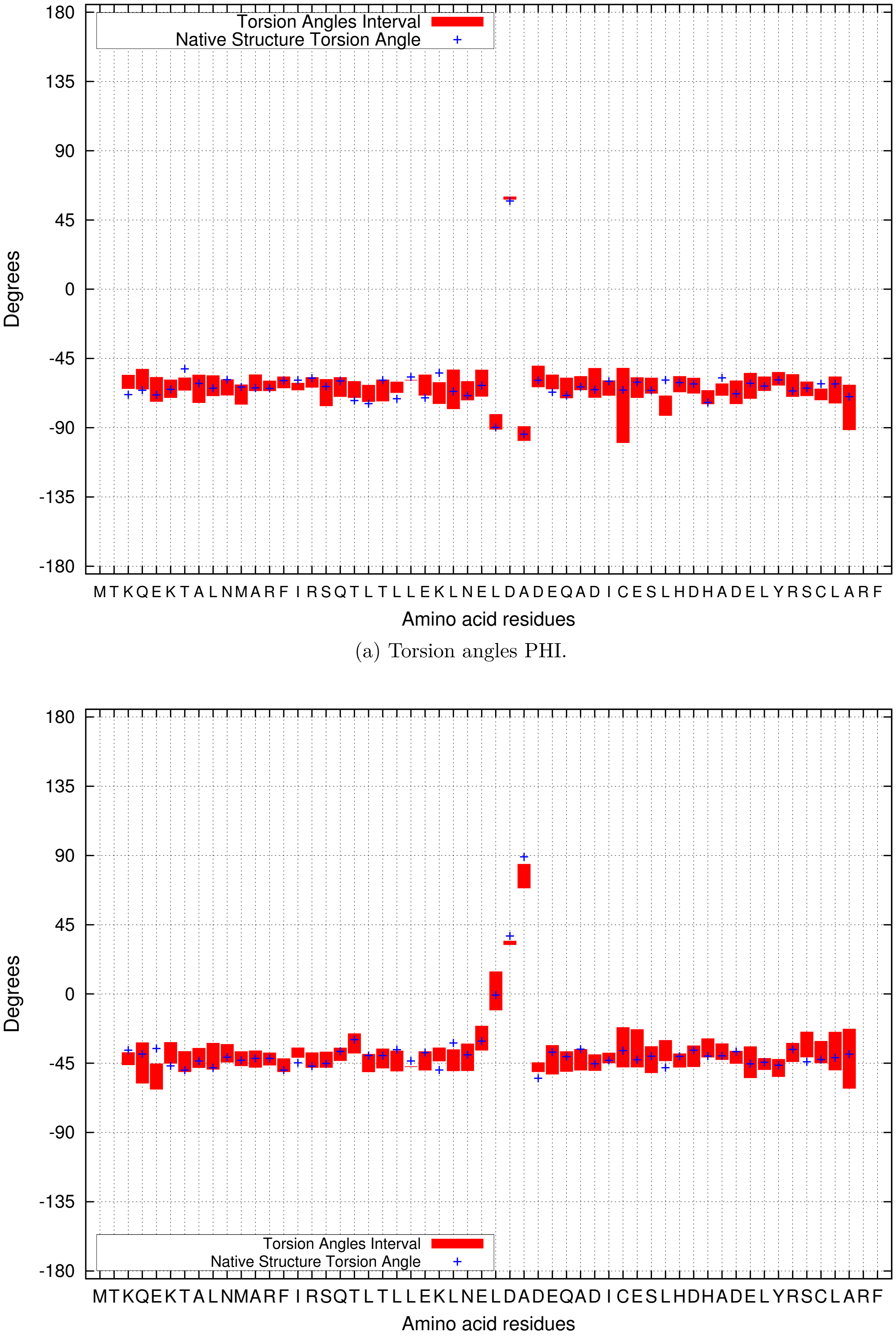
Figure 7.4: Torsion angles interval for the protein with `PDB ID` = `1B6Q`. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.



(a) Torsion angles PHI.

(b) Torsion angles PSI.

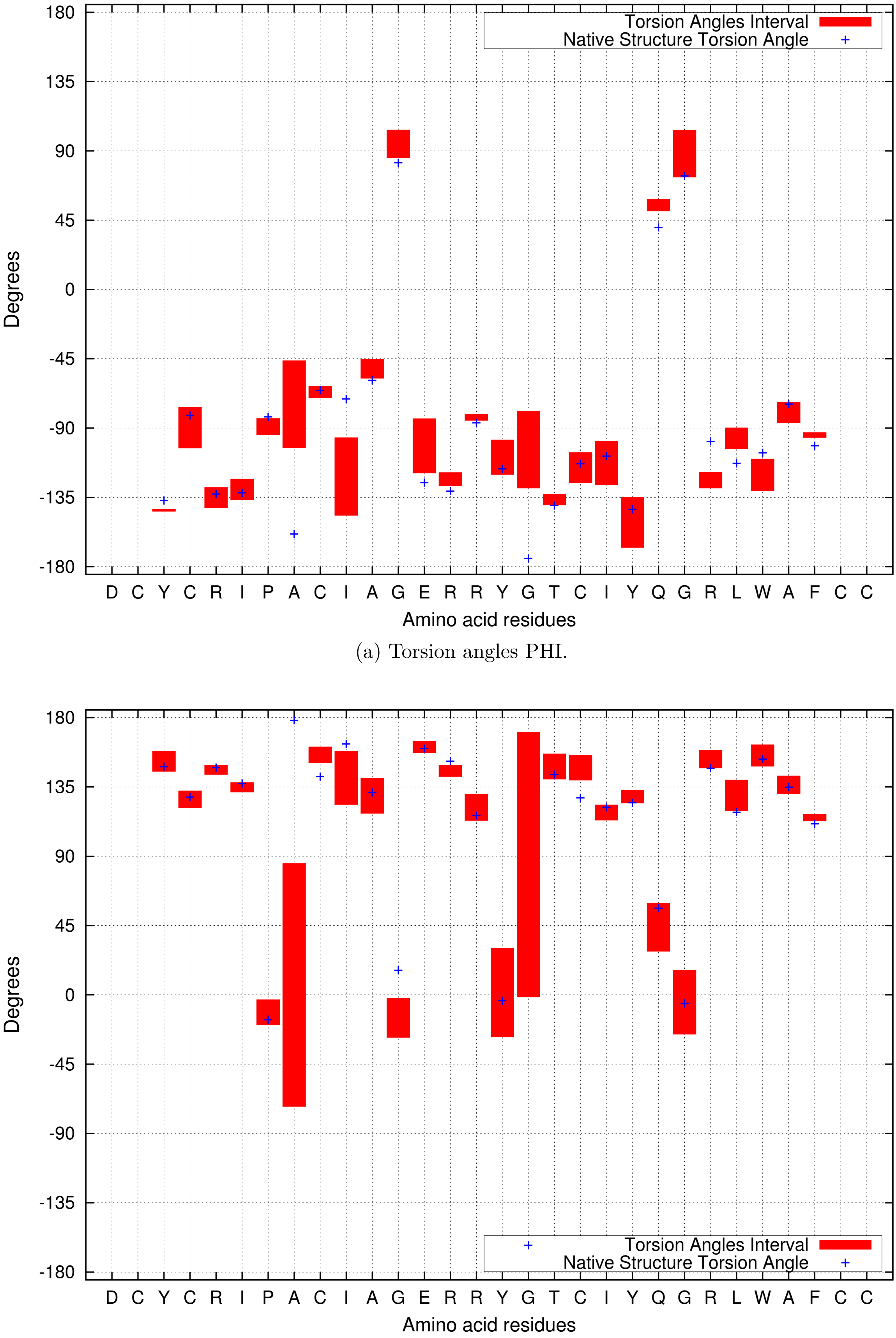
Figure 7.5: Torsion angles interval for the protein with `PDB ID` = `1K43`. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.



(a) Torsion angles PHI.

(b) Torsion angles PSI.

Figure 7.6: Torsion angles interval for the protein with `PDB ID` = `1ROP`. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.



(a) Torsion angles PHI.

(b) Torsion angles PSI.

Figure 7.7: Torsion angles interval for the protein with PDB ID = 1DFN. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.

7.4 Searching the native-like structures of polypeptides

Torsion angles intervals computed at the first stage of the MOIRAE strategy (Chapter 5) are used as input for the search strategy described in Chapter 6. The search strategy was run four times for each target protein and we obtain four classes of solutions. As described in Chapter 6 the GA population was structured in castes (fittest 20% of the individuals compose caste A, least 50% compose caste B, the remaining 30% compose caste C) and the population size was fixed on 100 individuals, the AMBER ff99SB (HORNAK et al., 2006) force field and the Geneneralized Born (OBC) (ONUFRIEV; BASHFORD; CASE, 2000, 2004) were used. We define as stop condition the time and the number of iterations, the search strategy stops when 2000 generations were computed or when the running time of 24 hours is reached. Table 7.4 summarizes the results obtained with the search procedure. In Table 7.4, Column 2 shows the number of generations computed in each run of the search strategy, Column 3 shows the initial lowest energy of cast A after the first generation and Column 4 shows the lowest energy in caste A at the last generation. Figure 7.8 illustrates the energy minimization procedure for the four runs of the GA. As can be observed the search strategy is effective in minimize the potential energy function. The local-search operator contributes in order to speed up the convergence of the simulation, as can be observed the potential energy of the polypeptide structure is greatly reduced already in the first 100 generations. The same can be observed in the other study cases: 1ACW (Fig. 7.9), 1B6Q (Fig. 7.10), 1DFN (Fig. 7.11), 1K43 (Fig. 7.12), 1ROP (Fig. 7.13), 1AIL (Fig. I.1), 1B03 (Fig. I.2), 1BDC (Fig. I.3), 1BGK (Fig. I.4), 1BHI (Fig. I.5), 1DV0 (Fig. I.6), 1E0Q (Fig. I.7), 1ENH (Fig. I.8), 1FME (Fig. I.9), 1OVX (Fig. I.10), 1Q2K (Fig. I.11), 1QR8 (Fig. I.12), 1ROO (Fig. I.13), 1WQC (Fig. I.14).

We analyse the time costs of the search strategy for each study case. Table 7.3 shows the simulation times. Column 2 presents the total time in seconds of the GA execution, Column 3 shows the average time of each generation, Column 4 shows the time costs of the local-search operator. As can be observed, the time costs of the local-search operator are superior at 90% of the GA total time. Column 5 shows the average time of the local-search operation at each GA generation. Column 6 presents the total time costs of the diversity control operator and Column 7 shows its average costs at each GA generation. The last three Columns shows respectively, the total time costs of the function that calculates the potential energy of the individuals, the average time of each call of this function and the total number of calls to the function along the GA simulation. As can be observed most of the time costs of the developed search strategy are associated with the fitness function.

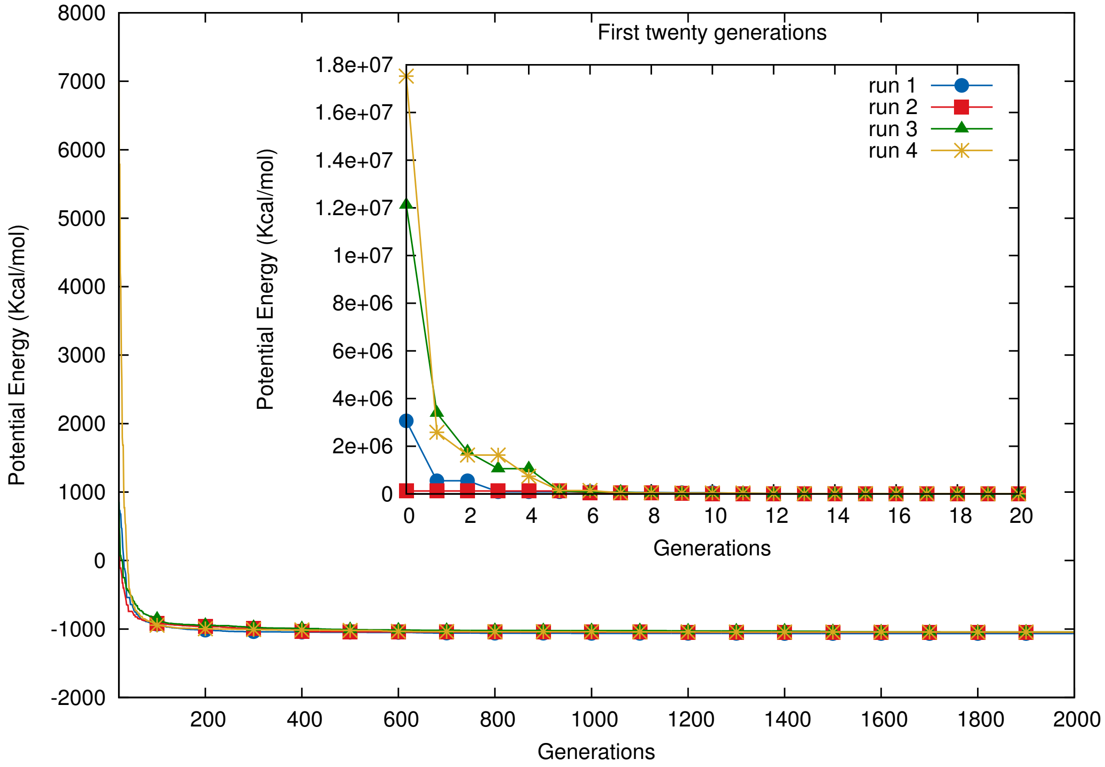


Figure 7.8: Potential energy minimization for protein with `PDB ID` = `1AB1`.

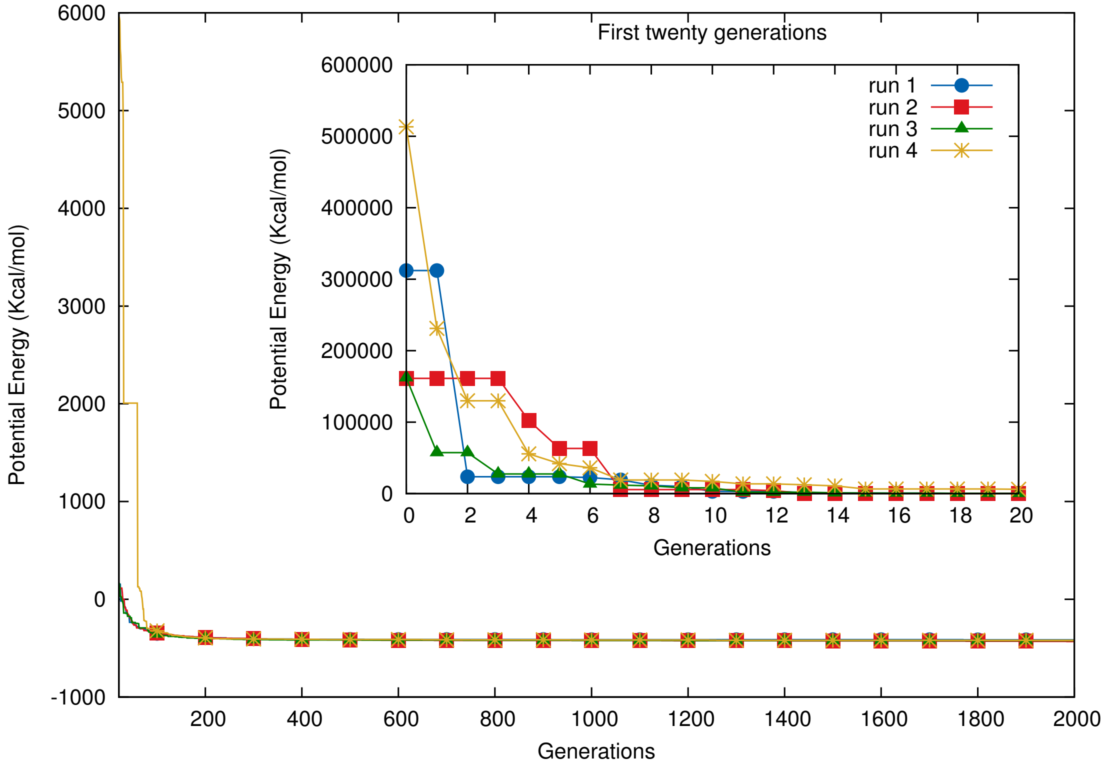


Figure 7.9: Potential energy minimization for protein with `PDB ID` = `1ACW`.

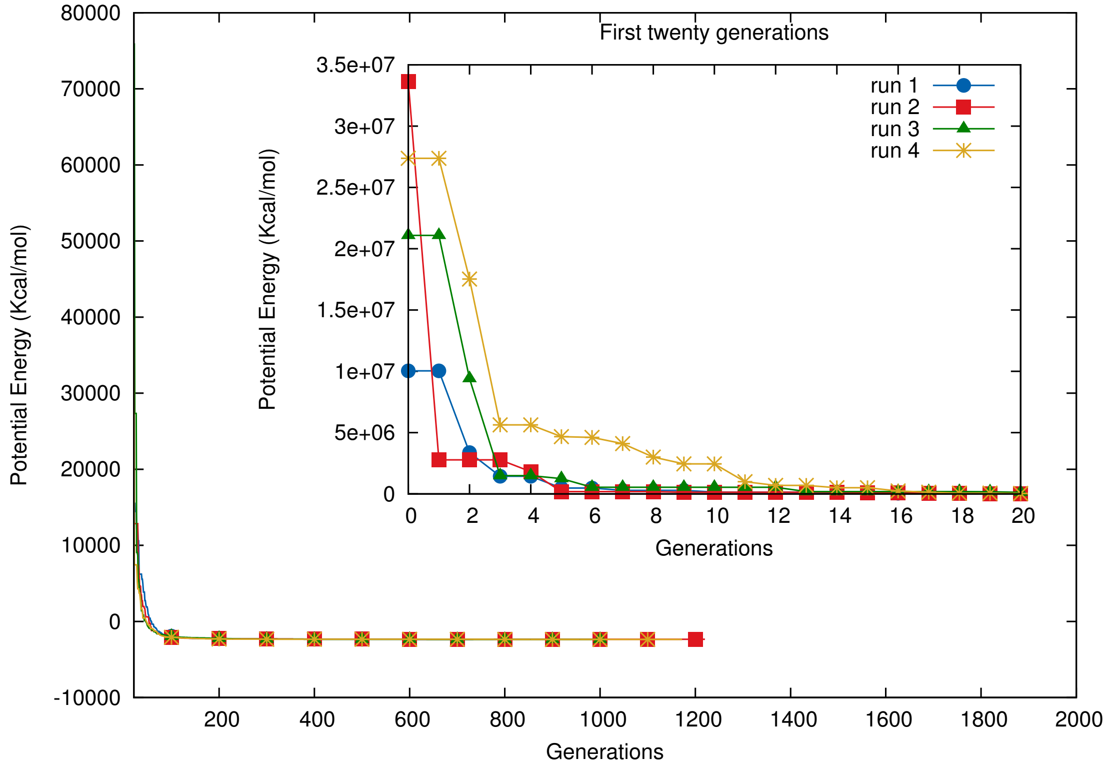


Figure 7.10: Potential energy minimization for protein with PDB ID = 1B6Q.

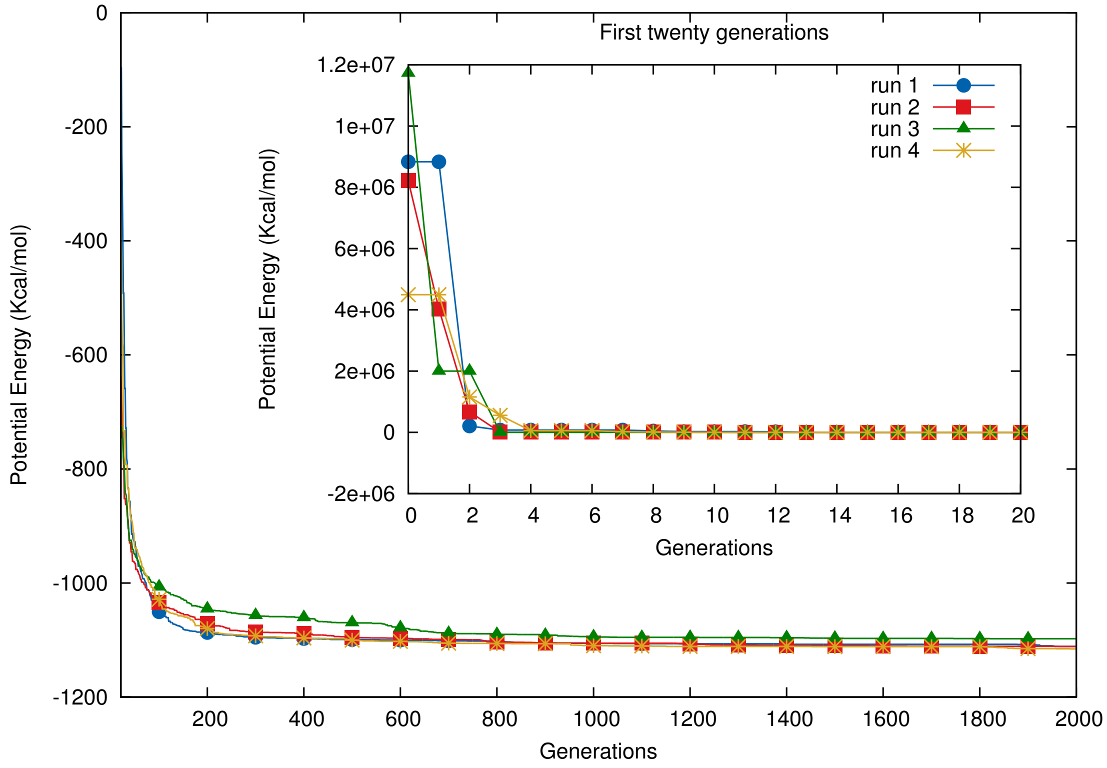


Figure 7.11: Potential energy minimization for protein with PDB ID = 1DFN.

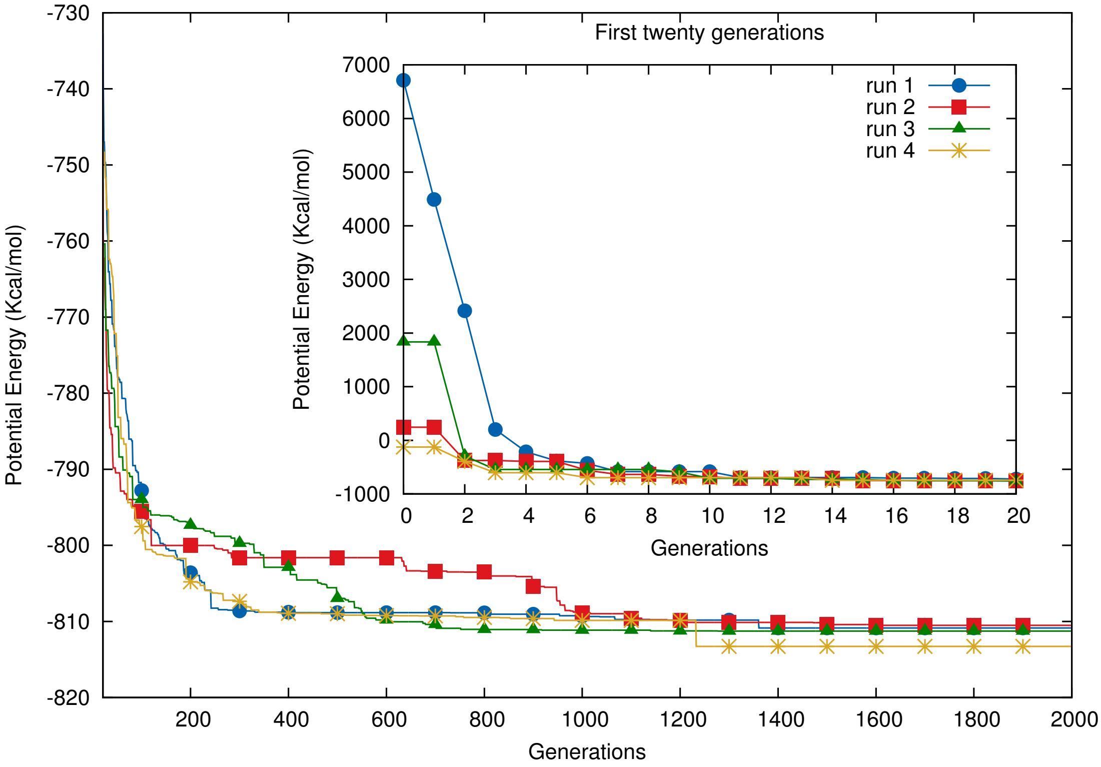


Figure 7.12: Potential energy minimization for protein with PDB ID = 1K43.

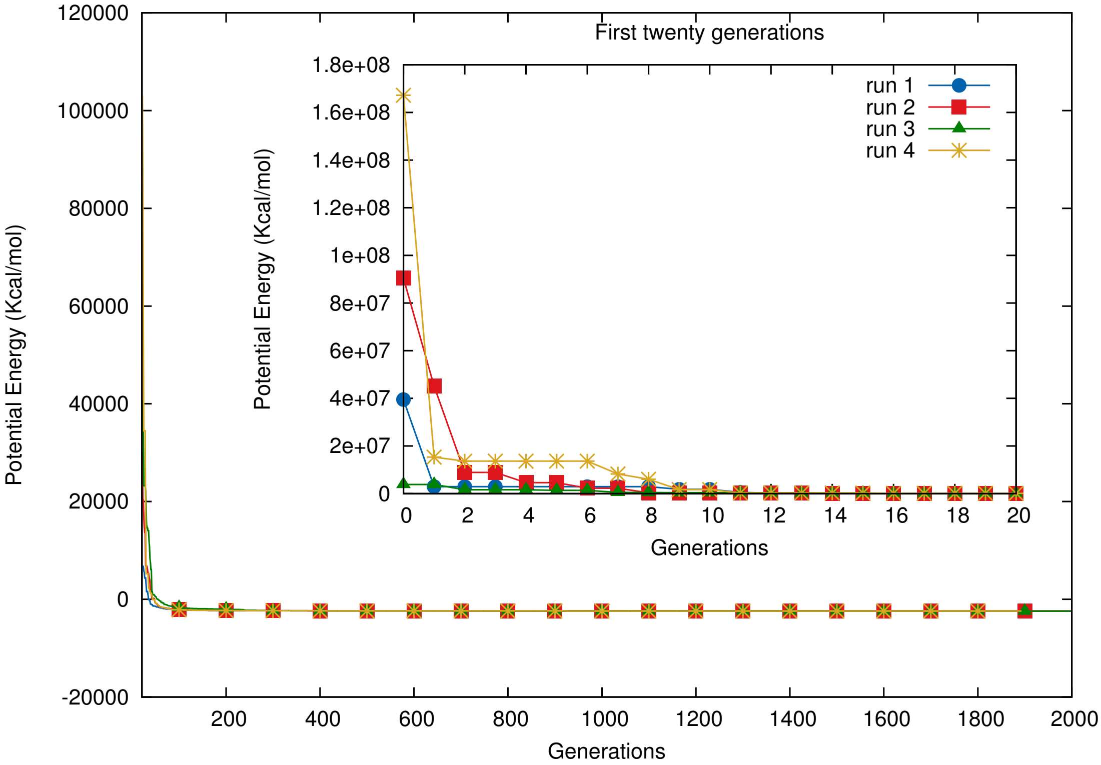


Figure 7.13: Potential energy minimization for protein with PDB ID = 1ROP.

Table 7.3: Time analysis of the GA search strategy. 2nd Column shows the total time of each GA execution. 3rd Column shows the mean time of each generation. 4th Column shows the total time of the local-search operator. 5th Column shows the mean time of the local-search operator by generation. 6th Column shows the total time of the diversity control operator. 7th Column shows the mean time of the diversity control operator by generation. 8th Column shows the total time of the potential energy function (MME). 9th Column shows the mean time of each call of the MME. 10th Column shows the number of calls to evaluation function (MME) for each run. Time values are in seconds.

PDB	Total Time	Mean Time	Total Time	Mean Time	Total Time	Mean Time	Total Time	Mean Time	Number
ID	of the GA	per Gen.	LS	LS per Gen.	DC	DC per Gen.	MME	MME calls	MME calls
1AB1-1	131,626	65.81	105,578	52.78	9,206	4.60	96,358	0.013	6,999,956
1AB1-2	132,226	66.11	106,003	53.00	9,334	4.66	96,803	0.013	6,954,703
1AB1-3	133,210	66.60	107,228	53.61	9,163	4.58	98,188	0.014	6,989,666
1AB1-4	128,906	64.45	102,902	51.45	9,176	4.58	93,916	0.013	6,951,318
AVERAGE	**131,492**	**65.74**	**105,427**	**52.71**	**9,219**	**4.60**	**96,316**	**0.013**	**6,973,910**
1ACW-1	44,259	22.12	30,860	15.43	5,975	1.81	12,071	0.004	2,889,158
1ACW-2	49,146	24.57	35,284	17.64	6,203	1.76	12,537	0.004	2,938,161
1ACW-3	43,506	21.75	30,059	15.02	6,018	1.76	12,574	0.004	2,944,855
1ACW-4	42,723	21.36	29,202	14.60	6,011	1.71	12,360	0.004	2,933,215
AVERAGE	**44,908**	**22.45**	**31,351**	**15.67**	**6,051**	**1.76**	**12,385**	**0.004**	**2,926,347**
1AIL-1	172,832	86.41	149,398	74.69	6,703	3.35	139,406	0.021	6,497,595
1AIL-2	172,836	86.41	147,463	73.73	7,379	3.68	137,461	0.020	6,810,213
1AIL-3	172,878	86.43	147,735	73.86	7,335	3.66	138,204	0.020	6,590,933
1AIL-4	172,843	86.42	148,171	74.08	7,131	3.56	138,607	0.020	6,627,110
AVERAGE	**172,847**	**86.42**	**148,191**	**74.09**	**7,137**	**3.56**	**138,419**	**0.020**	**6,631,462**
1B03-1	19,977	9.98	12,914	6.45	3,621	1.81	12,071	0.004	2,889,158
1B03-2	20,195	10.09	13,308	6.65	3,527	1.76	12,537	0.004	2,938,161
1B03-3	20,250	10.12	13,402	6.70	3,536	1.76	12,574	0.004	2,944,855
1B03-4	19,953	9.97	13,202	6.60	3,434	1.71	12,360	0.004	2,933,215

Continued on next page

Table 7.3 – continued from previous page

PDB ID	Total Time GA	Mean Time Gen.	Total Time LS	Mean Time LS Gen.	Total Time DC	Mean Time DC Gen.	Total Time MME	Mean Time MME calls	Number MME calls
AVERAGE	**20,093**	**10.04**	**13,206**	**6.60**	**3,529**	**1.76**	**12,385**	**0.004**	**2,926,347**
1B6Q-1	172,951	86.47	156,877	78.43	5,213	2.60	144,364	0.032	4,466,248
1B6Q-2	172,802	86.40	147,341	73.67	7,487	3.74	126,300	0.018	7,005,620
1B6Q-3	172,936	86.46	151,706	75.85	6,471	3.23	135,243	0.023	5,800,226
1B6Q-4	172,869	86.43	148,421	74.21	7,249	3.62	127,645	0.018	6,910,332
AVERAGE	**111,328**	**86.44**	**151,086**	**75.543**	**6,605**	**3.302**	**133,388**	**0.023**	**6,045,604**
1BDC-1	172,982	86.49	150,782	75.39	6,367	3.18	132,012	0.021	6,111,618
1BDC-2	172,807	86.40	150,517	75.25	6,322	3.16	131,126	0.021	6,202,221
1BDC-3	172,962	86.48	151,532	75.76	6,108	3.05	131,959	0.021	6,186,657
1BDC-4	172,904	86.45	152,487	76.24	5,838	2.91	133,182	0.021	6,056,441
AVERAGE	**172,913**	**86.45**	**151,329**	**76.66**	**6,158**	**3.93**	**132,069**	**0.021**	**6,139,234**
1BGK-1	125,022	62.51	103,860	51.93	7,876	3.93	90,075	0.010	8,293,071
1BGK-2	126,224	63.11	104,906	52.45	7,985	3.99	91,730	0.011	8,110,055
1BGK-3	121,309	60.65	100,266	50.13	7,836	3.91	86,865	0.010	8,127,542
1BGK-4	120,832	60.41	99,588	49.79	7,951	3.97	86,990	0.011	7,884,768
AVERAGE	**123,346**	**61.67**	**102,155**	**51.077**	**7,912**	**3.956**	**88,915**	**0.010**	**8,103,859**
1BHI-1	109,396	54.69	88,257	44.12	7,839	3.91	77,123	0.010	7,388,361
1BHI-2	115,958	57.97	94,674	47.33	7,880	3.94	83,561	0.011	7,536,264
1BHI-3	109,936	54.96	88,398	44.19	8,115	4.05	77,509	0.010	7,216,185
1BHI-4	110,023	55.01	88,684	44.34	8,029	4.01	78,016	0.010	7,241,735
AVERAGE	**111,328**	**55.66**	**90,003**	**45.00**	**7,965**	**3.98**	**79,052**	**0.010**	**7,345,636**
1DFN-1	70,288	35.14	55,017	27.50	6,377	3.18	50,328	0.009	5,469,280
1DFN-2	70,498	35.24	55,263	27.63	6,370	3.18	50,275	0.009	5,459,061
1DFN-3	70,512	35.25	55,318	27.65	6,369	3.18	50,900	0.009	5,365,601
1DFN-4	78,997	39.49	63,005	31.50	6,650	3.32	59,547	0.01	5,282,663

Continued on next page

Table 7.3 – continued from previous page

PDB ID	Total Time GA	Mean Time Gen.	Total Time LS	Mean Time LS Gen.	Total Time DC	Mean Time DC Gen.	Total Time MME	Mean Time MME calls	Number MME calls
AVERAGE	**72,573**	**36.28**	**57,150**	**28.57**	**6,441**	**3.22**	**52,762**	**0.009**	**5,394,151**
1DV0-1	173,035	86.51	162,232	81.11	4,502	2.25	164,231	0.088	1,862,889
1DV0-2	172,829	86.41	144,521	72.26	9,887	4.94	127,012	0.014	8,828,951
1DV0-3	168,920	84,46	140,584	70.29	9,553	4.77	121,935	0.013	9,044,259
1DV0-4	169,008	85.50	140,703	70.35	9,639	4.81	122,505	0.013	8,911,348
AVERAGE	**170,948**	**85.47**	**147,010**	**73.50**	**8,395**	**4.19**	**133,920**	**0.032**	**7,161,661**
1E0Q-1	20,111	10.05	13,510	6.75	3,379	1.68	12,687	0.003	3,210,312
1E0Q-2	20,793	10.39	14,103	7.05	3,453	1.72	13,257	0.004	3,223,405
1E0Q-3	20,345	10.17	13,739	6.86	3,378	1.68	12,857	0.003	3,237,368
1E0Q-4	20,118	10.05	13,523	6.76	3,408	1.70	12,600	0.003	3,225,513
AVERAGE	**20,341**	**10.17**	**13,718**	**6,85**	**3,404**	**1.70**	**12,850**	**0.004**	**3,224,149**
1ENH-1	172,826	86.41	148,820	74.41	8,016	4.00	139,847	0.018	7,525,623
1ENH-1	172,826	86.41	148,820	74.41	8,016	4.00	139,847	0.018	7,525,623
1ENH-1	172,819	86.40	147,972	73.98	8,332	4.16	138,694	0.018	7,691,619
1ENH-1	172,867	86.43	148,902	74.45	8,099	4.04	140,634	0.019	7,267,851
AVERAGE	**172,834**	**86.41**	**148,628**	**74.31**	**8,115**	**4.05**	**139,755**	**0.018**	**7,502,679**
1FME-1	53,862	26.93	44,127	22.06	4,166	2.08	40,388	0.007	5,297,203
1FME-2	51,761	25.88	42,195	21.09	4,082	2.04	38,329	0.007	5,239,789
1FME-3	52,896	26.44	43,149	21.57	4,185	2.09	39,491	0.007	5,164,875
1FME-4	55,201	27.60	45,411	22.70	4,185	2.09	41,832	0.007	5,270,706
AVERAGE	**53,430**	**26.71**	**43,720**	**21.86**	**4,154**	**2.07**	**40,010**	**0.007**	**5,243,143**
1K43-1	9,378	4.68	6,031	3.01	1,859	0.92	6,253	0.002	2,120,204
1K43-2	10,052	5.02	6,535	3.26	1,950	0.97	6,933	0.003	2,092,856
1K43-3	9,581	4.79	6,248	3.12	1,861	0.93	6,504	0.003	2,109,776
1K43-4	9,581	5.79	6,248	3.12	1,861	0.93	6,504	0.003	2,109,776

Continued on next page

Table 7.3 – continued from previous page

PDB ID	Total Time GA	Mean Time Gen.	Total Time LS	Mean Time LS Gen.	Total Time DC	Mean Time DC Gen.	Total Time MME	Mean Time MME calls	Number MME calls
AVERAGE	**9,648**	**4.82**	**6,265**	**3.13**	**1,882**	**0.93**	**6,548**	**0.003**	**2,108,153**
1OVX-1	65,097	32.54	52,537	26.26	4,904	2.45	48,635	0.008	5,544,459
1OVX-2	68,912	34.45	56,022	28.01	5,036	2.51	52,079	0.009	5,636,531
1OVX-3	69,728	34.86	56,936	28.46	5,036	2.51	52,885	0.009	5,649,453
1OVX-4	62,792	31.39	50,279	25.13	4,917	2.45	46,101	0.008	5,491,878
AVERAGE	**66,632**	**33.31**	**53,943**	**26.97**	**4,973**	**2.48**	**49,925**	**0.009**	**5,491,878**
1Q2K-1	46,661	23.33	37,144	18.57	4,056	2.02	35,643	0.008	4,439,858
1Q2K-2	49,749	24.87	39,913	19.96	4,252	2.12	38,798	0.008	4,342,349
1Q2K-3	40,912	20.45	31,882	15.94	3,845	1.92	29,933	0.006	4,343,362
1Q2K-4	41,825	20.91	32,770	16.38	3,862	1.93	30,803	0.007	4,398,557
AVERAGE	**44,786**	**22.39**	**35,427**	**17.71**	**4,003**	**2.00**	**33,794**	**0.007**	**4,381,031**
1QR8-1	172,827	86.41	146,265	73.13	7,792	3.89	131,327	0.015	8,402,283
1QR8-2	172,876	86.43	147,289	73.64	7,391	3.69	132,488	0.015	8,305,841
1QR8-3	172,843	86.42	145,169	72.58	8,232	4.11	133,449	0.017	7,690,786
1QR8-4	172,926	86.46	145,159	72.57	8,071	4.03	131,738	0.017	7,554,720
AVERAGE	**172,868**	**86.43**	**145,970**	**72.98**	**7,871**	**3.93**	**132,250**	**0.016**	**7,988,407**
1ROO-1	60,055	30.02	48,431	24.21	4,737	2.36	45,109	0.008	5,239,071
1ROO-2	58,002	29.00	46,477	23.23	4,681	2.34	43,109	0.008	5,164,875
1ROO-3	61,428	30.71	49,835	24.91	4,685	2.34	45,907	0.008	5,402,867
1ROO-4	58,548	29.27	46,964	23.48	4,710	2.35	43,537	0.008	5,141,247
AVERAGE	**59,508**	**29.75**	**47,926**	**23.96**	**4,703**	**2.35**	**44,415**	**0.008**	**5,237,015**
1ROP-1	172,647	86.32	147,641	73.82	8,171	4.08	138,670	0.017	8,018,314
1ROP-2	161,273	80.63	136,649	68.32	8,178	4.08	127,704	0.016	7,851,932
1ROP-3	169,780	84.89	144,768	72.38	8,341	4.17	136,275	0.017	7,861,341
1ROP-4	172,869	86.43	148,864	74.43	8,096	4.04	140,947	0.018	7,520,441

Continued on next page

Table 7.3 – continued from previous page

PDB ID	Total Time GA	Mean Time Gen.	Total Time LS	Mean Time LS Gen.	Total Time DC	Mean Time DC Gen.	Total Time MME	Mean Time MME calls	Number MME calls
AVERAGE	**169,142**	**84.57**	**144,480**	**72.24**	**8,196**	**4.09**	**135,899**	**0.017**	**7,813,007**
1WQC-1	49,760	24.88	40,677	20.33	4,232	2.11	41,282	0.010	3,781,489
1WQC-2	45,356	22.67	36,695	18.34	4,015	2.00	37,202	0.009	3,768,990
1WQC-3	29,468	14.73	22,392	11.19	3,241	1.62	21,610	0.005	3,680,657
1WQC-4	28,384	14.19	21,564	10.78	3,102	1.55	20,512	0.005	3,784,168
AVERAGE	**38,242**	**19.12**	**30,332**	**15.16**	**3,647**	**1.82**	**30,151**	**0.008**	**3,753,826**

7.5 Structural analysis

For biochemical and structural analysis we selected the class of solutions that at the last GA simulation presents the solution with the lowest potential energy. The quality of the predicted structures were evaluated by similarity comparisons with the structures of the experimental proteins obtained from the PDB (Eq.7.1). Quality measurements have been made in terms of the root mean square deviation (RMSD) between the position of the C_α atoms of the predicted and the experimental structures. The RMSD measure was calculated using PROFIT[3].

$$\mathbf{RMSD}(a,b) = \sqrt{\left(\sum_{i=1}^{n} \|r_{ai} - r_{bi}\|^2\right)/n}, \tag{7.1}$$

were r_{ai} and r_{bi} are vectors representing the positions of the same atom i in each of two structures, a and b respectively, and where the structures a and b are optimally superimposed. Table 7.4 (Column 5) shows the RMSD value of each GA run. The predicted 3-D protein structure with the lowest RMSD was the protein with PDB ID = 1K43 (0.59Å - Fig.7.1n) followed by 1ROP (5.27Å - Fig.7.1s), 1ACW (5.95Å - Fig.7.1b), 1OVX (6.41Å - Fig.7.1o), 1E0Q (6.42Å - Fig.7.1k), 1AB1 (6.65Å - Fig.7.1a), 1B6Q (7.25Å - Fig.7.1e), 1DFN (7.30Å - Fig.7.1i), 1WQC (7.53Å - Fig.7.1t), 1Q2K (7.93Å - Fig.7.1p), 1B03 (8.22Å - Fig.7.1d), 1FME (8.38Å - Fig.7.1m), 1AIL (10.01Å - Fig.7.1c), 1BHI (11.46Å - Fig.7.1h), 1BGK (11.92Å - Fig.7.1g), 1ENH (12.58Å - Fig.7.1l), 1ROO (12.77Å - Fig.7.1r), 1DV0 (12.87Å - Fig.7.1j), 1QR8 (16.74Å- Fig.7.1q), 1BDC (13.96Å - Fig.7.1f). Case studies 1BDC, 1QR8, 1DV0, 1ROO, 1ENH, 1BGK presents higher RMSD. This result is somewhat expected given that these case studies shows a more complex folding pattern when compared with the other test cases. By visual inspection (Fig. 7.1), it is noticeable that the individual helices and other secondary structures are well formed in most of the study cases.

Table 7.4: GA simulation results. Columns 3 and 4 shows, respectively, the initial and the final lowest potential energy (Kcal/mol) of each run of the GA. Last Column shows the RMSD (Å) value. † identifies the run that achieves the lowest potential energy at the end of the GA simulation. -NUMBER identifies the GA run(for example "-1" added to 1AB1).

PDB ID	Generations	Initial energy (Kcal/mol)	Final energy (Kcal/mol)	RMSD (Å)
1AB1-1†	2000	3,066,806.77	-1,066.16	6.65
1AB1-2	2000	122,397.62	-1,048.59	9.42
1AB1-3	2000	12,113,647.82	-1,043.96	6.84
1AB1-4	2000	17,523,290.73	-1,044.49	8.54
AVERAGE		**8,206,535.73**	**-1,050.80**	**7.86**
1ACW-1	2000	312,117.46	-414.00	6.61
1ACW-2†	2000	161,353.27	-431.21	5.95
1ACW-3	2000	162,098.19	-422.20	6.77
1ACW-4	2000	513,233.97	-419.78	6.95

Continued on next page

[3] www.bioinf.org.uk/software/profit

Table 7.4 – continued from previous page

PDB ID	Generations	Initial energy (Kcal/mol)	Final energy (Kcal/mol)	RMSD (Å)
AVERAGE		**287,200.72**	**-419.45**	**6.57**
1AIL-1	1347	9,983,986.95	-2,812.26	8.72
1AIL-2†	1481	789,119,606.39	-2,825.70	10.01
1AIL-3	1430	140,121,694.99	-2,805.06	10.52
1AIL-4	1427	42,688,499.04	-2,790.40	9.17
AVERAGE		**245,478,446.84**	**-2,809.11**	**9.60**
1B03-1	2000	-203.34	-1,002.47	8.03
1B03-2	2000	-586.57	-1,003.82	8.20
1B03-3†	2000	821.38	-1,003.83	8.22
1B03-4	2000	-679.69	-1,002.93	8.08
AVERAGE		**-164.55**	**-1003.26**	**8.13**
1B6Q-1	735	10,033,050.69	-2,359.25	9.54
1B6Q-2	1221	33,636,730.51	-2,351.91	15.22
1B6Q-3	1001	21,088,183.80	-2,374.58	10.70
1B6Q-4†	1173	27,371,678.93	-2,375.75	7.25
AVERAGE		**23,032,410.99**	**-2,365.37**	**10.67**
1BDC-1	962	88,275,695.61	-2,097.98	15.80
1BDC-2†	973	582,975,317.49	-2,188.71	13.96
1BDC-3	934	367,694,523.89	-2,141.62	16.45
1BDC-4	887	202,464,886.29	-2,144.55	10.59
AVERAGE		**310,352,605.82**	**-2,143.21**	**14.20**
1BGK-1	2000	1,063,694.25	-1,496.39	12.83
1BGK-2	2000	10,407,280.25	-1,508.32	8.43
1BGK-3†	2000	4,980,452.72	-1,510.63	11.92
1BGK-4	2000	23,996,230.00	-1,499.36	11.18
AVERAGE		**10,111,914.30**	**-1,503.68**	**11.09**
1BHI-1	2000	21,073.07	-731.01	10.70
1BHI-2†	2000	460,303.53	-735.86	11.46
1BHI-3	2000	45,902.65	-723.09	13.51
1BHI-4	2000	74,342.34	-717.61	12.88
AVERAGE		**150,405.39**	**-726.89**	**12.13**
1DFN-1	2000	8,834,947.17	-1,111.27	6.39
1DFN-2	2000	8,215,465.94	-1,111.20	6.20
1DFN-3	2000	11,726,689.62	-1,097.74	5.69
1DFN-4†	2000	4,494,194.28	-1,115.53	7.30
AVERAGE		**8,317,824.25**	**-1,106.00**	**6.39**
1DV0-1	1384	24,591,968.29	-1,571.46	12.67
1DV0-2	1968	32,77,304.67	-1,583.96	12.12
1DV0-3†	2000	28,792,732.35	-1,589.85	12.87
1DV0-4	2000	13,309,098.34	-1,570.27	12.65
AVERAGE		**17,492,775.91**	**-1,578.89**	**12.57**
1E0Q-1	2000	622.99	-562.28	6.52
1E0Q-2	2000	363.05	-562.75	6.50
1E0Q-3	2000	1,198.04	-563.37	6.44

Continued on next page

Table 7.4 – continued from previous page

PDB ID	Generations	Initial energy (Kcal/mol)	Final energy (Kcal/mol)	RMSD (Å)
1EOQ-4†	2000	336.06	-564.49	6.42
AVERAGE		**630.035**	**-563.30**	**6.47**
1ENH-1	1812	197,871,083.13	-3,126.19	10.16
1ENH-2	1812	650,034,687.73	-3,126.19	11.65
1ENH-3†	1884	134,312,786.73	-3,128.14	12.58
1ENH-4	1768	218,510,602.24	-3,103.59	12.36
AVERAGE		**300,182,289.95**	**-3,121.03**	**11.68**
1FME-1	2000	6,257,978.99	-1,572.96	9.40
1FME-2†	2000	8,568,173.92	-1,579.14	8.38
1FME-3	2000	20,481,864.27	-1,564.59	9.52
1FME-4	2000	5,507,919.32	-1,566.81	9.49
AVERAGE		**10,203,984.12**	**-1,570.87**	**9.19**
1K43-1	2000	6,712.22	-810.86	0.46
1K43-2	2000	242.48	-810.52	0.60
1K43-3†	2000	1,834.88	-811.27	0.59
1K43-4	2000	-126.51	-811.27	0.73
AVERAGE		**2,165.76**	**-810.98**	**0.59**
1OVX-1	2000	130,273.15	-1,187.95	8.43
1OVX-2	2000	19,273.47	-1,180.32	8.06
1OVX-3†	2000	2,444,433.26	-1,191.58	6.41
1OVX-4	2000	57,971.79	-1,188.12	7.47
AVERAGE		**662,987.91**	**-1,186.99**	**7.59**
1Q2K-1†	2000	89,972,512.66	-198.18	7.93
1Q2K-2	2000	26,584,233.19	-162.93	9.39
1Q2K-3	2000	2,720,169.95	-170.87	8.42
1Q2K-4	2000	2,065,060.57	-190.36	7.92
AVERAGE		**30,335,494.09**	**-180.58**	**8.41**
1QR8-1	1728	94,223,404.84	-2,158.21	20.46
1QR8-2	1767	91,301,035.23	-2,257.74	22.26
1QR8-3†	1738	79,272,399.75	-2,947.88	16.74
1QR8-4	1658	250,366,250.90	-2,913.46	16.63
AVERAGE		**128,790,772.68**	**-2,569.32**	**19.02**
1ROO-1†	2000	56,624,797.13	-1,212.66	12.77
1ROO-2	2000	15,344,219.92	-1,205.09	12.43
1ROO-3	2000	16,731,038.24	-1,203.91	14.13
1ROO-4	2000	105,951,353.13	-1,211.63	13.83
AVERAGE		**48,662,852.105**	**-1,208.32**	**13.29**
1ROP-1	2000	39,507,903.38	-2,388.35	4.03
1ROP-2	2000	90,470,808.37	-2,409.99	5.59
1ROP-3†	2000	3,810,470.83	-2,422.90	5.27
1ROP-4	2000	167,218,892.74	-2,415.93	4.02
AVERAGE		**75,252,018.83**	**-2,409.30**	**4.72**
1WQC-1	2000	538,324.01	-676.91	7.24
1WQC-2†	2000	1,347,566.13	-694.87	7.53

Continued on next page

Table 7.4 – continued from previous page

PDB ID	Generations	Initial energy (Kcal/mol)	Final energy (Kcal/mol)	RMSD (Å)
1WQC-3	2000	22,021.76	-678.07	8.04
1WQC-4	2000	145,939.46	-676.64	7.54
AVERAGE		**513,462.84**	**-681.62**	**7.58**

7.5.1 Secondary structure analysis

Secondary structure analysis were performed with PROMOTIF (HUTCHINSON; THORNTON, 1996). We run PROMOTIF in order to analyze the patterns of hydrogen bonds that define the secondary structure of the predicted structures. In this analysis we compare the secondary structure contents of the predicted 3-D protein structures against the secondary structure of the native structures. Table 7.5 summarizes the obtained results with PROMOTIF. This analysis reveals that the secondary structure of the structures predicted by MOIRAE are comparable to their experimental structures. This can be observed when we examine the predicted structure of 1ACW (GA run 2) which presents 24.10% (against 24.10% of the experimental 3-D structure (1ACW-E)) of the amino acid residues in a α-helix state, 27.60% (against 34.50% of the experimental) in a β-sheet state, and 48.30% (against 41.40% of the experimental 3-D structure) representing other irregular structures. The predicted structure of 1AB1 presents 39.10% (against 41.30% of the experimental 3-D structure) of the amino acid residues in a α-helix state, 0.00% (against 8.70% of the experimental) in a β-sheet state, and 60.90% (against 50.00% of the experimental 3-D structure) representing other irregular structures. The predicted structure of 1B6Q presents 87.50% (against 85.70% of the experimental 3-D structure) of the amino acid residues in a α-helix state, and 12.50% (against 14.30% of the experimental 3-D structure) representing other irregular structures. The secondary structure of the predicted 1DFN presents 26.70% (against 60.00% of the experimental 3-D structure) of the amino acid residues in a β-sheet state and 73.30% (against 40.00% of the experimental structure) of the amino acid residues are other irregular structures. The 3-D structure of 1K43 presents 28.60% (against 42.90% of the experimental structure) of their amino acid residues in a β-sheet conformational state, 71.40% (against 57.10% of the experimental structure) of the amino acid residues representing other irregular structures. The predicted structure of 1ROP presents 91.10% (against 89.30% present in the experimental 3-D structure) of their amino acid residues in a α-helix state and 8.90% (against 10.70% present in the experimental 3-D structure) as irregular structures. The secondary structure similarity between the predicted and experimental structures can be also observed in case studies 1AIL, 1BDC, 1BGK, 1BHI, 1DV0, 1ENH, 1FME, 1OVX, 1Q2K, 1QR8, 1WQC.

The largest difference between the secondary structure elements of the predicted and experimental structures is observed in case studies 1B03, 1EOQ, 1ROO. The 3-D structure of 1B03 presents 0.00% (against 55.60% of the experimental structure) of their amino acid residues in a β-sheet conformational state, 100.00% (against 44.40% of the experimental structure) of the amino acid residues representing other irregular structures. Through visual inspection of Figure 7.1d we can observe that β-sheets regions (Fig. 7.14d) are not well formed, this in turns occur because the presence of distortions in the coil region (Fig. 7.14d). The same occurs in the study case of protein 1EOQ that presents 0.00% (against 70.60% in the experimental) of the amino acid residues in a β-sheet state and 100% (against 29.40% of the experimental structure) of amino acid residues in a regular structure. The α-helices of case study 1ROO were not well formed (Fig. 7.1r). PROMOTIF shows that predicted 1ROO presents 0.00% (against 31.40% in the experimental structure) of the amino acid residues in a α-helix state, 8.60% (against 8.60% in the experimental structure) in a 3^{10}-helix state and 91.40% (against 60.00% in the experimental structure) in an irregular conformational state.

Table 7.5: Analysis of the secondary structure contents of the predicted and the native 3-D protein structures. Suffix -NUMBER denotes the predicted 3-D structures with the lowest energy among the four runs of the GA simulation (for example "-1" added to 1AB1). Suffix -E denotes the experimental structure.

PDB ID	Strand/β-sheet(%)	Alpha-helix(%)	3^{10}-helix(%)	Others(%)
1AB1-1	0.00	39.10	0.00	60.90
1AB1-E	8.70	41.30	0.00	50.00
1ACW-2	27.60	24.10	0.00	48.30
1ACW-E	34.50	24.10	0.00	41.40
1AIL-2	0.00	87.10	0.00	12.90
1AIL-E	0.00	84.30	0.00	15.70
1B03-3	0.00	0.00	0.00	100.00
1B03-E	55.60	0.00	0.00	44.40
1B6Q-4	0.00	87.50	0.00	12.50
1B6Q-E	0.00	85.70	0.00	14.30
1BDC-2	0.00	46.70	0.00	53.30
1BDC-E	0.00	55.00	0.00	45.00
1BGK-3	0.00	29.70	16.20	54.10
1BGK-E	0.00	37.80	0.00	62.20
1BHI-2	0.00	21.10	7.90	71.70
1BHI-E	10.50	31.60	0.00	57.90
1DFN-4	26.70	0.00	0.00	73.30
1DFN-E	60.00	0.00	0.00	40.00
1DV0-3	0.00	53.30	6.70	40.0
1DV0-E	0.00	42.20	0.00	57.80
1EOQ-4	0.00	0.00	0.00	100.00
1EOQ-E	70.60	0.00	0.00	29.40
1ENH-3	0.00	68.50	0.00	31.50
1ENH-E	0.00	70.40	0.00	29.60
1FME-2	0.00	35.70	0.00	64.30
1FME-E	14.30	35.70	0.00	50.00
1K43-4	28.60	0.00	0.00	71.40
1K43-E	42.90	0.00	0.00	57.10
1OVX-2	10.50	13.20	0.00	76.30
1OVX-E	15.80	28.90	0.00	55.30

Continued on next page

Table 7.5 – continued from previous page

PDB ID	Strand/β-sheet(%)	Alpha-helix(%)	3^{10}-helix(%)	Others(%)
1Q2K-1	0.00	19.40	0.00	80.60
1Q2K-E	19.40	32.30	0.00	48.40
1QR8-3	0.00	85.10	0.00	14.90
1QR8-E	0.00	77.90	0.00	22.10
1ROO-1	0.00	0.00	8.60	91.40
1ROO-E	0.00	31.40	8.60	60.00
1ROP-3	0.00	91.10	0.00	8.90
1ROP-E	0.00	89.30	0.00	10.70
1WQC-2	0.00	46.20	0.00	53.80
1WQC-E	0.00	65.40	0.00	34.60

When we compare the topology of the predicted against the experimental (Fig. 7.14) 3-D structures (Fig. 7.2) we observe that the topologies are comparable, except for 1B03 (Fig. 7.14d) and 1EOQ (Fig. 7.14d). In these two cases the β-strands are not well formed (Tab. 7.5).

7.5.2 Stereo-chemical analysis

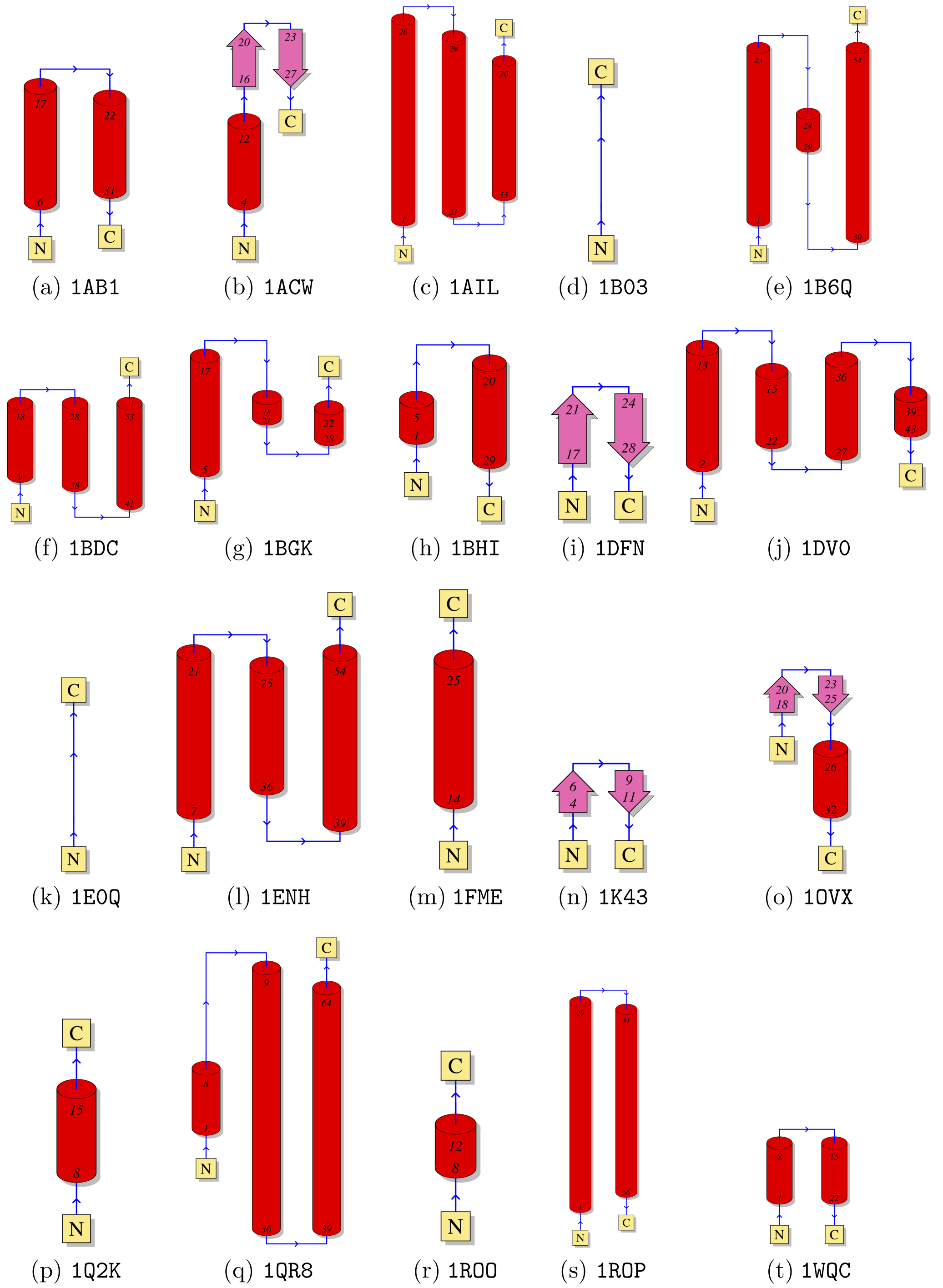
The distribution of the amino acid residues in the Ramachandran plot[4] and the stereo-chemical quality of the 3-D structures predicted by MOIRAE were analyzed by PROCHECK[5] (LASKOWSKI et al., 1993). Table 7.6 summarizes the numerical Ramachandran plot values for the experimental and predicted structures. We observe that in all of 3-D predicted structures, the amino acid residues are located in the most favorable regions of the map (favorable or additional allowed region) (Tab. 7.6): 1AB1 (Fig. 7.15b), 1ACW (Fig. 7.16b), 1B6Q (Fig. 7.17b), 1DFN (Fig. 7.18b), 1K43 (Fig. 7.19b), 1ROP (Fig. 7.20b), 1AIL (Fig. J.1b), 1B03 (Fig. J.2b), 1BDC (Fig. J.3b), 1BGK (Fig. J.4b), 1BHI (Fig. J.5b), 1DV0 (Fig. J.6b), 1EOQ (Fig. J.7b), 1ENH (Fig. J.8b), 1FME (Fig. J.9b), 1OVX (Fig. J.10b). The red, brown, and yellow regions in the Ramachadran plots represent the favored, allowed, and "generously allowed" regions of phi and torsion angles of amino acid residues, respectively. The percentage of residues in the "core" regions (most favorable regions) is one of the better guides to analyse the stereo-chemical quality of the predicted 3-D protein structures. When we compare the results obtained with the 3-D structure predicted by MOIRAE against the experimental structures we observe that these structures are comparable in terms of stereo-chemical quality.

[4] we use the Ramachandran plot to visualize backbone dihedral angles ϕ against ψ of amino acid residues in protein structure.

[5] www.ebi.ac.uk/thornton-srv/software/PROCHECK

Table 7.6: Numerical Ramachandran plot values for the experimental and predicted structures. -NUMBER denotes the predicted 3-D structures with the lowest energy (for example "-1" added to 1AB1). -E denotes the experimental structure.

PDB ID	Most favored region (%)	Most allowed region (%)	Generously allowed region (%)	Disallowed region (%)
1AB1-1	83.30	16.70	0.00	0.00
1AB1-E	94.40	5.60	0.00	0.00
1ACW-2	88.00	12.00	0.00	0.00
1ACW-E	84.00	16.00	0.00	0.00
1AIL-2	100.00	0.00	0.00	0.00
1AIL-E	98.40	1.60	0.00	0.00
1B03-3	100.00	0.00	0.00	0.00
1B03-E	100.00	0.00	0.00	0.00
1B6Q-4	98.10	1.90	0.00	0.00
1B6Q-E	100.00	0.00	0.00	0.00
1BDC-2	83.30	16.70	0.00	0.00
1BDC-E	70.40	29.60	0.00	0.00
1BGK-3	91.20	8.80	0.00	0.00
1BGK-E	73.50	23.50	2.90	0.00
1BHI-2	93.80	6.20	0.00	0.00
1BHI-E	78.10	21.90	0.00	0.00
1DFN-4	95.80	4.20	0.00	0.00
1DFN-E	95.80	4.20	0.00	0.00
1DV0-3	84.40	14.60	0.00	0.00
1DV0-E	82.90	14.60	2.40	0.00
1E0Q-4	100.00	0.00	0.00	0.00
1E0Q-E	100.00	0.00	0.00	0.00
1ENH-3	96.00	4.00	0.00	0.00
1ENH-E	100.00	0.00	0.00	0.00
1FME-2	91.70	8.30	0.00	0.00
1FME-E	62.50	37.50	0.00	0.00
1K43-4	66.70	33.30	0.00	0.00
1K43-E	66.70	33.30	0.00	0.00
1OVX-2	87.90	12.10	0.00	0.00
1OVX-E	90.90	9.10	0.00	0.00
1Q2K-1	88.90	7.40	3.70	0.00
1Q2K-E	81.50	11.10	7.40	0.00
1QR8-3	94.90	3.40	1.70	0.00
1QR8-E	91.70	6.70	1.70	0.00
1ROO-1	80.60	19.47	0.00	0.00
1ROO-E	71.00	25.80	3.20	0.00
1ROP-3	98.10	1.90	0.00	0.00
1ROP-E	98.10	1.90	0.00	0.00
1WQC-2	90.50	9.50	0.00	0.00
1WQC-E	90.50	0.00	9.50	0.00



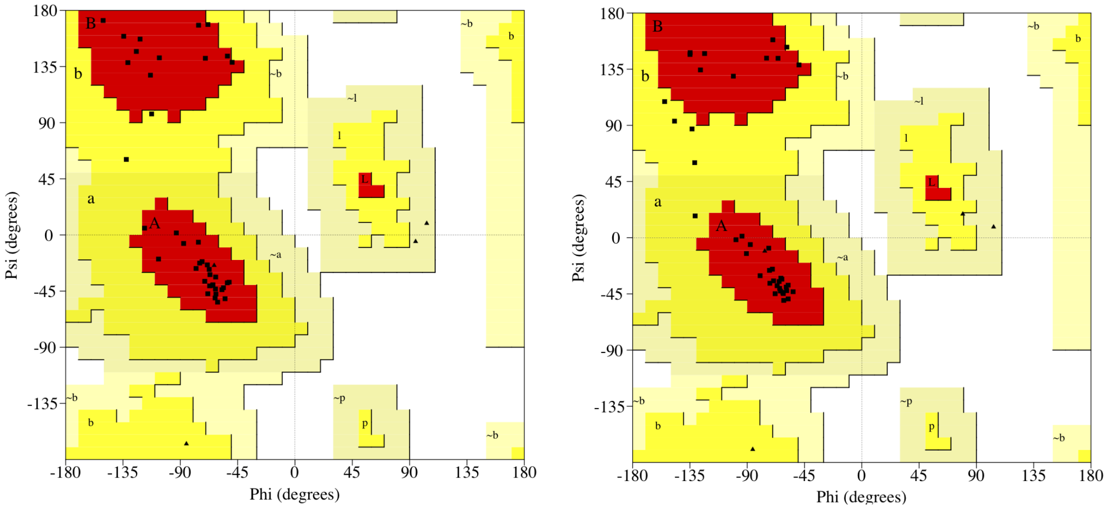
(a) 1AB1 (b) 1ACW (c) 1AIL (d) 1B03 (e) 1B6Q

(f) 1BDC (g) 1BGK (h) 1BHI (i) 1DFN (j) 1DV0

(k) 1E0Q (l) 1ENH (m) 1FME (n) 1K43 (o) 1OVX

(p) 1Q2K (q) 1QR8 (r) 1R00 (s) 1ROP (t) 1WQC

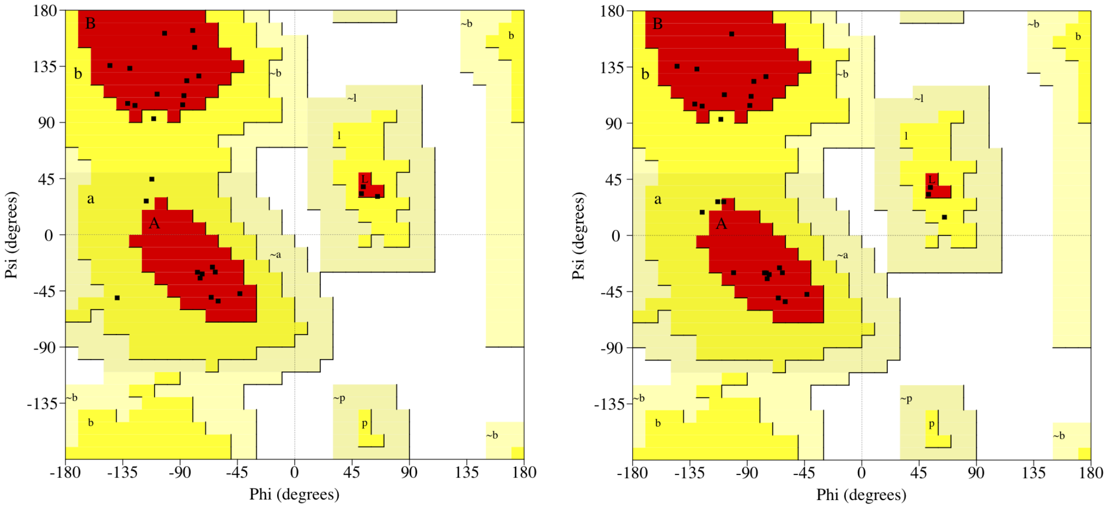
Figure 7.14: Diagram representing the topology of the predicted 3-D protein structures. N and C represents the N-terminal and the C-terminal regions, respectively. α-helices are showed in red, β-sheets are showed in pink and coil regions are showed in blue. Graphic representation was generated by PDBSUM (www.ebi.ac.uk/pdbsum).



(a) Experimental

(b) Predicted

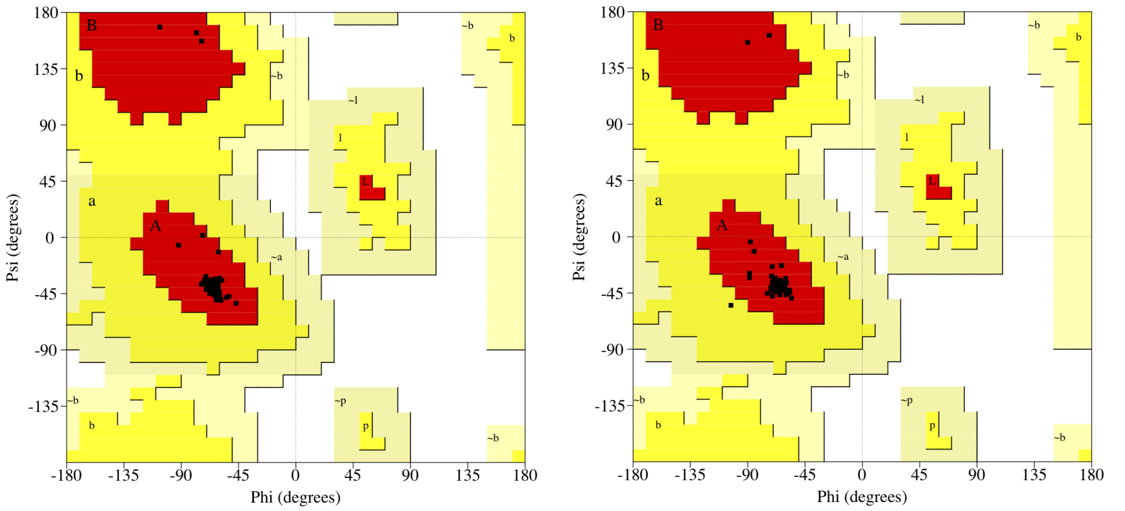
Figure 7.15: Ramachandran plot of the experimental and predicted structures. (a) Ramachandran plot of the experimental protein with PDB ID 1AB1. (b) Ramachandran plot of the predicted 3-D structure of the protein with PDB ID 1AB1.



(a) Experimental

(b) Predicted

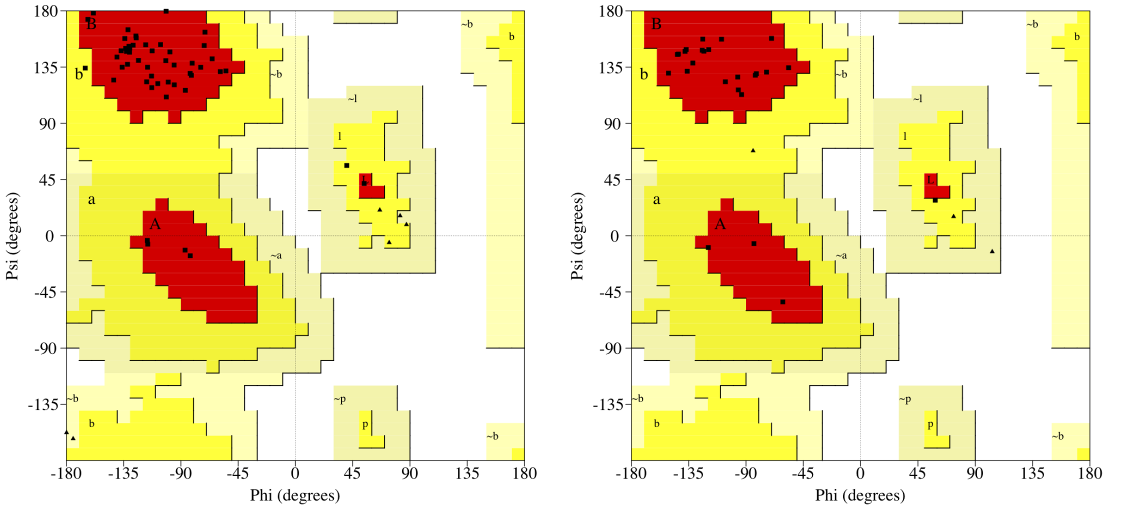
Figure 7.16: Ramachandran plot of the experimental and predicted structures. (a) Ramachandran plot of the experimental protein with PDB ID 1ACW. (b) Ramachandran plot of the predicted 3-D structure of the protein with PDB ID 1ACW.



(a) Experimental

(b) Predicted

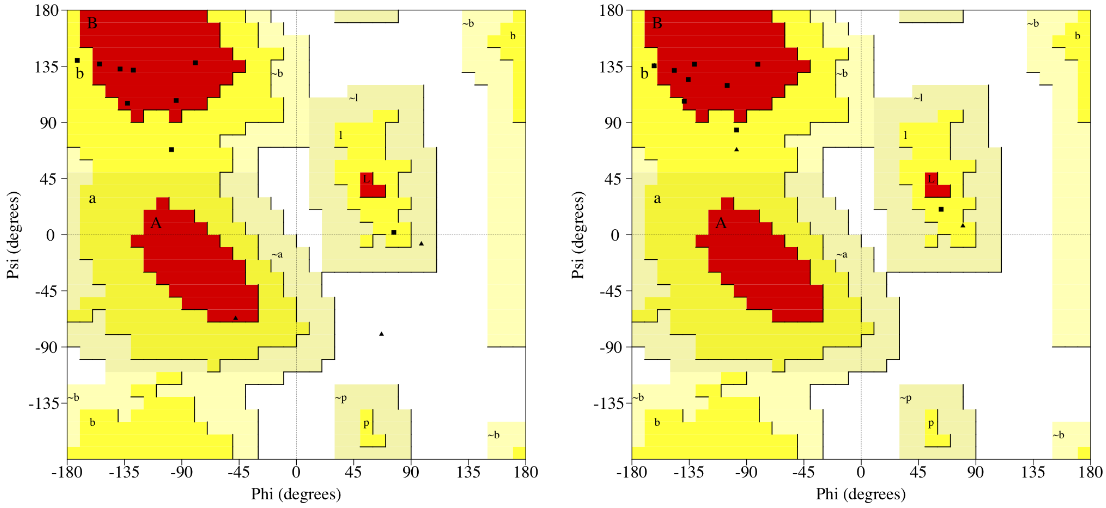
Figure 7.17: Ramachandran plot of the experimental and predicted structures. (a) Ramachandran plot of the experimental protein with PDB ID 1B6Q. (b) Ramachandran plot of the predicted 3-D structure of the protein with PDB ID 1B6Q.



(a) Experimental

(b) Predicted

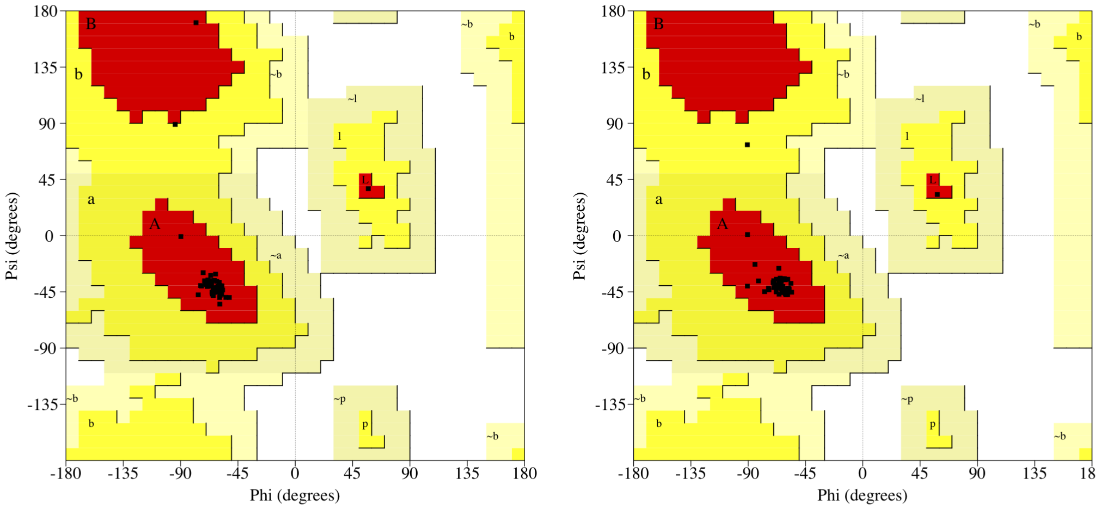
Figure 7.18: Ramachandran plot of the experimental and predicted structures. (a) Ramachandran plot of the experimental protein with PDB ID 1DFN. (b) Ramachandran plot of the predicted 3-D structure of the protein with PDB ID 1DFN.



(a) Experimental

(b) Predicted

Figure 7.19: `Ramachandran` plot of the experimental and predicted structures. (a) `Ramachandran` plot of the experimental protein with `PDB ID 1K43`. (b) `Ramachandran` plot of the predicted `3-D` structure of the protein with `PDB ID 1K43`.



(a) Experimental

(b) Predicted

Figure 7.20: `Ramachandran` plot of the experimental and predicted structures. (a) `Ramachandran` plot of the experimental protein with `PDB ID 1ROP`. (b) `Ramachandran` plot of the predicted `3-D` structure of the protein with `PDB ID 1ROP`.

7.6 Comparison of protein 3-D structure prediction methods

When compared with other prediction methods classified as first principle method that use database information, MOIRAE presents advantages in terms of demanded time to produced native-like approximate 3-D structures of proteins. Table 7.7 summarizes the root mean square deviations achieved by the homology modelling methods SWISS-MODEL (ARNOLD et al., 2006) and ESYPRED3D (LAMBERT et al., 2002); and the first principle methods that use database information BHAGEERATH (JAYARAM et al., 2006), PROTINFO (HUNG et al., 2005) and ROBETTA (ROHL et al., 2004). These results were computed by Jayaram et al. (JAYARAM et al., 2006). We compare the results obtained by Jayaram et al. with MOIRAE. As reported by Jayaram etal. (JAYARAM et al., 2006) the experiments using the computational strategy BHAGEERATH where executed on a cluster with 32 dedicated UltraSparc III 900 MHz processors and ROBETTA was configured and executed over only one dedicated processor. The processing time depends on the length of the sequence and number of secondary structure elements. BHAGEERATH methodology can process, over the computational architecture described above, around 4-5 normal size jobs per day. ROBETTA (ROHL et al., 2004), for example, took around 12 days to compute the 3-D structure of protein with PDB ID = 1EOQ. In Table 7.7, column 5 shows the mean RMSD of ten structures obtained when ROBETTA was executed for each target protein. In Table 7.7, column 6 shows the mean RMSD of five structures obtained by ProtInfo and Column 4 presents the mean RMSD of ten structures computed by BHAGEERATH for each target protein. Table 7.7, last column, shows the mean RMSD value of four computed 3-D protein structures for each target protein.

As could be observed, the best results were achieved by the homology modelling methods. However, this class of methods can only predict structures of protein sequences which are similar or nearly identical to other protein sequences of known structures. This means that when no sequence homologies are detected this class of methods can not predict accurate 3-D structures. The main goal of this thesis is to present a first principle method that can deal with this problem. Analysing the results achieved by MOIRAE we observe that the obtained RMSDs are comparable to the obtained by BHAGEERATH, ROBETTA and PROTINFO. ROBETTA was been the most successful predictor along the last years as revealed by the CASP experiments. Clearly, they present the best results, however, the proposed strategy is a novel idea to predict the 3-D structures of proteins with lower computational resources (see computational times in Table 7.3). MOIRAE was executed on an PC Intel Core i7 3.07 GHz and each prediction took around 24-48 hours of CPU time. Stereo-chemical analysis of predicted 3-D structures by BHAGEERATH, ROBETTA and PROTINFO were not reported by Jayaram et al. (JAYARAM et al., 2006) . Stereo-chemical analysis commonly highlight regions of the predicted 3-D proteins which appear to have unusual geometry. All the 3-D protein structures predicted by MOIRAE were analysed in terms of secondary structure arrangement (for example, hydrogen bonds formation necessary to stabilize the protein secondary structures) (Table 7.5) and by the phi/psi preferences using the Ramachandran plot (Table 7.6). As reported in last sections, the stereo-chemical quality of the predicted 3-D protein structure are comparable with the stereo-chemical quality of its experimental structure obtained from the PDB.

Table 7.7: A Comparison of protein tertiary structure prediction accuracy. Adapted from Jayaram et al. (JAYARAM et al., 2006). - denotes the cases where the 3-D structure could not be computed or was not available. The e.s.d columns shows the standard deviation.

PDB	SWISS-MODEL(Å)	ESYPRED3D(Å)	BHAGEERATH (Å)	e.s.d	ROBETTA(Å)	e.s.d	PROTINFO(Å)	e.s.d	MOIRAE(Å)	e.s.d
1AB1	2.80	0.40	4.33	0.86	2.86	0.52	4.64	1.22	7.86	1.34
1ACW	0.40	-	6.18	0.87	1.50	0.28	6.18	0.54	6.57	0.44
1AIL	0.73	0.46	6.72	2.20	5.20	1.33	8.84	0.99	9.61	0.81
1B03	-	3.50	6.32	1.71	2.85	0.21	4.34	0.29	8.13	0.09
1B6Q	5.00	2.70	6.56	2.90	9.04	1.46	10.22	0.23	10.68	3.35
1BDC	2.70	2.70	6.63	1.11	4.56	2.30	3.50	0.93	14.20	2.63
1BGK	0.50	-	5.49	0.62	4.38	1.35	6.24	0.17	11.09	1.90
1BHI	0.40	1.00	6.76	1.09	2.11	0.51	4.30	0.54	12.14	1.29
1DFN	0.40	1.30	6.31	0.70	5.23	1.46	6.36	0.46	6.40	0.67
1DV0	10.50	2.00	7.56	1.17	2.36	1.10	4.12	1.32	12.58	0.32
1E0Q	-	1.70	3.74	1.05	1.10	0.00	3.98	0.19	6.47	0.05
1ENH	0.80	0.90	7.32	1.99	3.22	1.32	5.36	1.89	11.69	1.09
1FME	0.90	-	4.82	0.90	3.58	0.61	2.26	0.42	9.20	0.55
1K43	-	-	-	-	-	-	-	-	0.60	0.11
1OVX	0.30	0.10	5.75	1.07	3.08	1.14	4.94	0.48	7.59	0.88
1Q2K	0.50	-	6.25	1.13	2.86	1.28	6.64	1.39	8.42	0.69
1QR8	0.50	1.10	8.79	2.83	9.63	1.68	10.44	1.04	19.02	2.80
1ROO	0.70	-	3.45	0.39	2.18	0.51	2.78	0.13	13.29	0.82
1ROP	0.60	4.70	8.52	3.05	9.42	2.38	11.26	0.86	4.73	0.82
1WQC	0.40	-	4.12	0.96	2.85	0.78	1.96	0.15	7.59	0.33

BHAGEERATH runs on a cluster with 32 dedicated UltraSparc III 900 MHz processors and took approximately 5 hours per job (JAYARAM et al., 2006).

ROBETTA was executed in a dedicate processor and took approximate 12 hours for the protein 1EOQ (JAYARAM et al., 2006).

PROTINFO was executed in a web server and processing time was not informed (JAYARAM et al., 2006).

7.7 Chapter conclusions

In this chapter, we present and discuss the results obtained with the application of MOIRAE to predict the 3-D structure of 20 target protein sequences. The sizes of the target sequences vary from 14-70 amino acid residues. The results show that the predicted tertiary structures adopt a fold comparable to the experimental structures. Stereo-chemical analysis has revealed that the secondary elements of the predicted 3-D structures are well formed.

8 CONCLUSIONS

The study of proteins and the prediction of their three-dimensional (3-D) structures is one of the key research problems in Structural Bioinformatics. Predicting the three-dimensional structure of a protein that has no templates in the Protein Data Bank is a very hard and sometimes virtually intractable task. Over the last years, many computational methods, systems and algorithms have been developed with the purpose of solving this complex problem. However, the problem still challenges computer scientists, biologists, chemists, bioinformaticians, and mathematicians because of the complexity and high dimensionality of the protein conformational search space. Experimentally, the generation of a protein sequence is considerably easier than the determination of its 3-D structure. However, the knowledge of the 3-D structure of the polypeptide gives researchers very important information about the function of the protein in the cell. The difficulty in determining and finding out the 3-D structure of proteins has generated a large discrepancy between the volume of data (sequences of amino acid residues) generated by the GENOME projects[1] and the number of 3-D structures of proteins which are known nowadays. These figures not only clearly illustrate the need for, but also motivate further research in Computational Protein Structure Prediction Methods.

Analysing the progress of CASP along the last years we can observe that it is still necessary the development of new strategies for extracting, representing and manipulating structural data from experimentally determined 3-D protein structures, as well the development of computational strategies to use this information in order to predict, from the amino acid sequence of a protein, its corresponding 3-D structure. In this work we present a new first principle computational strategy which uses database information to predict the 3-D structure of proteins. MOIRAE manipulates structural data from the PDB in order to generate main-chain torsion angles intervals. As could be observed by the experiments, the use of this strategy reduces the 3-D conformational space of the target protein. Torsion angles intervals computed at the first stage of MOIRAE are used as input for the search strategy based on a genetic algorithm (GA). The developed search strategy allows a efficient mechanisms for protein structure prediction. This is achieved by the use of local-search operator which allows the GA to scape from local minima. In the case in hand (the PSP problem) this occurs when torsion angles are modified by the GA. As corroborated by the experiments, the developed method can produce accurate predictions where the 3-D protein structures are comparable to their experimental structures.

When compared with other prediction methods classified as first principle method

[1] DOE Genomic Science. http://genomics.energy.gov.

that use database information, MOIRAE presents advantages in terms of demanded time to produced native-like 3-D structures of proteins. ROBETTA, FRAGFOLD, I-TASSER and LINUS has been the most successful predictors along the last years as revealed by the CASP experiments, however they make use of large high performance computing platforms. Clearly, they present the best results, however, the proposed strategy is a novel idea of computational strategy to predict 3-D structures of proteins.

The overall contribution of our work is threefold: First, the use of computational techniques and concepts to develop a new, effective algorithm for a relevant biological problem (the 3-D PSP problem) showing that the proposed strategy to manipulate templates from PDB is usefull to reduces the protein conformational search space. Second, the use of genetic algorithms with local-search operator shows that this combined techniques can lead to efficient applications in several domains. And finally, The combination of a fragment-based method with anab-initio method (GA to minimize the potential energy of the polypeptide structure).

Finally, Protein Structure Prediction is a very difficult problem and further research remains to be done. The development of new strategies, the adaptation and investigation of new methods and the combination of existing and state-of-the-art computational methods and techniques to the PSP problem is clearly needed. Understanding how experimental data can be better used in combination with *ab initio* techniques is another open research question. In summary, there are several research opportunities and avenues to be explored in this field, with relevant multidisciplinary applications in computer science, bioinformatics, chemistry, biochemistry, and the medical sciences. This work opens several interesting research avenues, with a range of applications in computational biology and bioinformatics. For instance, one could apply the developed method to other classes of proteins; second, one could think of using other search methods such as PSO, Simulated Annealing, GRASP or TABU search, which perhaps could lead to even more efficient algorithms for 3-D protein structure prediction. Predicted structures by MOIRAE could be also used as input structures in refinement methods based on molecular mechanics (MM), e.g. molecular dynamics (MD) simulations. The search space is expected to be greatly reduced and the ab initio methods can demand a much reduced computational time to achieve a more accurate polypeptide structure. This could in turn reduces the total time of *ab initio* methods which usually start from a fully extended conformation.

9 PUBLICATIONS

9.1 Published papers

- COSTA, A.L.P.; PAULI, I.; DORN, M.; SCHROEDER, E.K.; ZHAN, C-G; NORBERTO DE SOUZA, O. "Conformational changes in 2-trans-enoyl-ACP (CoA) Reductase (InhA) from M. tuberculosis induced by an inorganic complex: a molecular dynamics simulation study". `Journal of Molecular Modeling`, v.18, 1779-1790, 2012.
  `IMPACT FACTOR:` 1.797
  `QUALIS CAPES:` B2 (Computer Science), B2 (Biology)

- DORN, M.; BRAGA, A.S.; LLANOS, C.H.; COELHO, L.S.A GMDH polynomial neural network-based method to predict approximate three-dimensional structures of polypeptides, `Expert Systems With Applications`, Elsevier, V. 39, 12268-12279, 2012.
  `IMPACT FACTOR:` 2.203
  `QUALIS CAPES:` A1 (Computer Science), B2 (Biology)

- BRAGA, A.L.S.; ARIAS-GARCIA, J.; QUINTERO, C.H.L.; DORN, M.; FOLTRAN, A.; COELHO, L.S.. Hardware implementation of GMDH-type artificial neural networks and its use to predict approximate three-dimensional structures of proteins, `7th International Workshop` on `Reconfigurable Communication centric Systems-on-Chip`, IEEE, 2012.

- DORN, M.; BURIOL, L.S.; LAMB, L.C. "A hybrid genetic algorithm for the `3-D` protein structure prediction problem using a path-relinking strategy". `IEEE Congress on Evolutionary Computation - CEC"`, New Orleans, 2011. p. 2691-2698.
  `QUALIS CAPES:` A1 (Computer Science)

- DORN, M.; BURIOL, L.S.; LAMB, L.C. "Combining machine learning and optimization techniques to determine `3-D` structures of polypeptides" `IJCAI - Doctoral Mentoring Consortium`, 2011, Barcelona.
  `QUALIS CAPES:` A1 (Computer Science)

- GONÇALVES, W.W.; DORN, M.; BURIOL, L.S.; LAMB, L.C. "A Structured-Population Genetic Algorithm for the `3-D` Protein Structure Prediction Problem". `Brazilian Symposium on Bioinformatics`, Brasilia, 2011. v. 1. p. 17-24.
  `QUALIS CAPES:` B2 (Computer Science)

- ANDRADES, R.; DORN, M.; FARENZENA, D.S.; LAMB, L.C. "Aplicação de Técnicas de Inteligência Artificial e Mineração de Dados no Design de Proteínas. XXIII Salão de Iniciação Científica UFRGS, 2011, Porto Alegre.

- TABAJARA, L.M.; FARENZENA, D.S.; DORN, M.; LAMB, L.C. "Resolução de problemas através de computaç ão humana utilizando redes sociais". XXIII Salão de Iniciação Científica UFRGS, 2011, Porto Alegre.

- DORN, M.; LAMB, L.C.; BURIOL, L.S. "An artificial neural network based method for the prediction of approximated 3-D structures of mini-globular proteins". 6th International Conference of Brazilian Association for Bioinformatics and Computational Biology, 2010, Ouro Preto.

- DORN, M.; NORBERTO DE SOUZA, O. "Mining the Protein Data Bank with CReF to predict approximate 3-D structures of polypeptides". International Journal of Data Mining and Bioinformatics, v.4, p.281, 2010.
  IMPACT FACTOR: 0.681
  QUALIS CAPES: B2 (Computer Science) B4 (Biology)

- DORN, M.; NORBERTO DE SOUZA, O. "A3N: An artificial neural network n-gram-based method to approximate 3-D polypeptides structure prediction". Expert Systems with Applications, v.37, p.7497, 2010.
  IMPACT FACTOR: 2.203
  QUALIS CAPES: A1 (Computer Science) B2 (Biology)

- DORN, M.; NORBERTO DE SOUZA, O. "CReF: A central-residue-fragment-based method for predicting approximate 3-D polypeptides structures". Annual ACM Symposium on Applied Computing - SAC, Fortaleza, 2008. p. 1261-1267.
  QUALIS CAPES: A1 (Computer Science)

- DORN, M.; BREDA, A.E.; NORBERTO DE SOUZA, O. "A hybrid method for the protein structure prediction problem", Brazilian Symposium on Bioinformatics", Santo Andr, Lecture Notes in Computer Science, Advances in Bioinformatics and Computational Biology. Heidelberg, Germany : Springer, 2008. v. 5167. p. 47-56.
  QUALIS CAPES: B2 (Computer Science)

- DORN, M.; NOPRBERTO DE SOUZA, O. "A fragment-based clustering method for predicting approximate 3-D polypeptide structures". 3rd International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2007, So Paulo.

9.1.1 Under review

- DORN, M.; BURIOL, L.S.; LAMB, L.C. "A Molecular Dynamics and Knowledge-based Computational Strategy to Predict Native-like Structures of Polypeptides". Expert Systems with Applications, Elsevier, 2012.
  IMPACT FACTOR: 2.029
  QUALIS CAPES: A1 (Computer Science) B2 (Biology)

- DORN, M.; BURIOL, L.S.; LAMB, L.C. "Protein Tertiary Structure Prediction: Methods and Computational Strategies" `Chemical Reviews`, ACS, 2012.
  `IMPACT FACTOR:` 33.036
  `QUALIS CAPES:` A1 (Biology) A1 (Chemistry)
- DORN, M.; ANDRADES, R.; FARENZENA, D.S.; LAMB, L.C. "A Novel Cluster-DEE-based Strategy to Empower Protein Design" `Artificial Intelligence in Medicine`, Elsevier, 2012.
  `IMPACT FACTOR:` 1.568
  `QUALIS CAPES:` A1 (Engeneering) A2 (Computer Science)

APPENDIX A PROTEIN KINEMATICS: ROTATION OF SIDE-CHAIN TORSION ANGLES

Table A.1: List of atoms necessary to perform a rotation in the χ_1 angle.

CHI$_1$			
Residue	Axis for rotation	Atoms Used to Define Angle	zero value
ARG	CA-CB	N-CA-CB-CG	CG cis to N
GLU	CA-CB	N-CA-CB-CG	CG cis to N
GLN	CA-CB	N-CA-CB-CG	CG cis to N
CYS	CA-CB	N-CA-CB-SG	SG cis to N
ASP	CA-CB	N-CA-CB-CG	CG cis to N
ASN	CA-CB	N-CA-CB-CG	CG cis to N
HIS	CA-CB	N-CA-CB-CG	CG cis to N
ILE	CA-CB	N-CA-CB-CG1	CG1 cis to N
LEU	CA-CB	N-CA-CB-CG	CG cis to N
LYS	CA-CB	N-CA-CB-CG	CG cis to N
MET	CA-CB	N-CA-CB-CG	CG cis to N
PHE	CA-CB	N-CA-CB-CG	CG cis to N
PRO	CA-CB	N-CA-CB-CG	CG cis to N
VAL	CA-CB	N-CA-CB-CG1	CG1 cis to N
TYR	CA-CB	N-CA-CB-CG	CG cis to N
TRP	CA-CB	N-CA-CB-CG	CG cis to N
THR	CA-CB	N-CA-CB-OG1	OG1 cis to N
SER	CA-CB	N-CA-CB-OG	OG cis to N

Table A.2: List of atoms necessary to perform a rotation in the χ_2 angle.

CHI$_2$			
Side-Chain	Axis for rotation	Atoms Used to Define Angle	zero value
ARG	CB-CG	CA-CB-CG-CD	CD cis to CA
GLU	CB-CG	CA-CB-CG-CD	CD cis to CA
GLN	CB-CG	CA-CB-CG-CD	CD cis to CA
ASP	CB-CG	CA-CB-CG-OD1	OD1 cis to CA
ASN	CB-CG	CA-CB-CG-OD1	OD1 cis to CA
HIS	CB-CG	CA-CB-CG-ND1	ND1 cis to CA
ILE	CB-CG1	CA-CB-CG1-CD1	CD1 cis to CA
LEU	CB-CG	CA-CB-CG-CD1	CD1 cis to CA
LYS	CB-CG	CA-CB-CG-CD	CD cis to CA
MET	CB-CG	CA-CB-CG-SD	SD cis to CA
PHE	CB-CG	CA-CB-CG-CD1	CD1 cis to CA
PRO	CB-CG	CA-CB-CG-CD	CD cis to CA
TYR	CB-CG	CA-CB-CG-CD1	CD1 cis to CA
TRP	CB-CG	CA-CB-CG-CD1	CD1 cis to CA

Table A.3: List of atoms necessary to perform a rotation in the χ_3 angle.

		CHI$_3$	
Side-Chain	Axis for rotation	Atoms Used to Define Angle	zero value
ARG	CG-CD	CB-CG-CD-NE	NE cis to CB
GLU	CG-CD	CB-CG-CD-OE1	OE1 cis to CB
GLN	CG-CD	CB-CG-CD-OE1	OE1 cis to CB
LYS	CG-CD	CB-CG-CD-CE	CE cis to CB
MET	CG-SD	CB-CG-SD-CE	CE cis to CB

Table A.4: List of atoms necessary to perform a rotation in the χ_4 angle.

		CHI$_4$	
Side-Chain	Axis for rotation	Atoms Used to Define Angle	zero value
ARG	CD-NE	CG-CD-NE-cz	CZ cis to CG
LYS	CD-CE	CG-CD-CE-NZ	NZ cis to CG

APPENDIX B FIRST PRINCIPLE METHODS WITHOUT DATABASE INFORMATION

Table B.1: First principle methods without database information: simulation packages. MD (Molecular Dynamics); EM (Energy Minimization), CG (Conjugate Gradient), SD (Step Descendants), MM (Molecular Mechanics), MC (Monte Carlo), MO (semi-empirical Molecular Orbital) , QM (Quantum Mechanics), QM/MM (Quantum and Molecular Mechanics), DFT (Density Functional Theory), TN (Truncated Newton) and LBFGS (Limited memory Broyden Fletcher Goldfarb Shanno).

Simulation Packages	MD	EM	SD	CG	L-BFGS	TN	MC	MM	QM/MM	DFT	QM	MO
AMBER (CASE et al., 2005)	●	●	●	●								
UNRES (LIWO et al., 1998)							●	●				
GROMACS (HESS et al., 2008)	●	●	●	●	●							
TINKER (KUNDROT; PONDER; RICHARDS, 1991)	●	●		●		●						
CHARMM (BROOKS et al., 1983)	●	●					●					
MOIL (ELBER et al., 1995)	●						●					
MOE (www.chemcomp.com)	●	●										
NAB (MACKE; CASE, 1998)		●										
ADUN (JOHNSTON; FERNÁNDEZ-GALVÁN; VILLÀ-FREIRA, 2005)	●	●									●	
ACEMD (HARVEY; GIUPPONI; FABRITIIS, 2009)	●	●										
SPARTAN (www.wavefun.com)		●					●				●	
PLOP (JACOBSON; FRIESNER; HONIG, 2002)	●											
BOSS (JORGENSEN; TIRADO-RIVES, 2005B)							●	●	●			●
HOOMD (ANDERSON; TRAVESSET, 2008)	●											
LAMMPS (PLIMPTON, 1995)	●											
ITAP (STADLER; MIKULLA; TREBIN, 1997)	●											
SMMP (EISENMENGER et al., 2001)	●	●										
MACSIMUS (www.vscht.cz/fch/software/MACSIMUS)	●											
MOLDY (REFSON, 2000)	●											
DLPOLY (SMITH; FORESTER, 1996)	●											
ESPRESSO (LIMBACH et al., 2006)	●						●					
MDYNAMIX (LYUBARTSEV; LAAKSONEN, 2000)	●											

Continued on next page

Table B.1 – continued from previous page

Simulation Packages	MD	EM	SD	CG	L-BFGS	TN	MC	MM	QM/MM	DFT	QM	MO
MCPRO (JORGENSEN; TIRADO-RIVES, 2005B)							●					
OPENMD (KUANG et al., 2009)	●											
ORAC (MARSILI et al., 2010)	●											
PINYMD (TUCKERMAN et al., 2000)	●											
Q (MARELIUS et al., 1999)	●											
SIESTA (SOLER et al., 2001)	●											
VASP (KRESSE; MARSMAN; FURTHMULLER, 2009)	●										●	
SAGEMD (SELEZENEV et al., 2003)	●						●			●		
NAMD (PHILLIPS et al., 2005)	●											
MOSCITO (PASCHEK; GEIGER, 2003)	●											
MCCST. (MARTIN; SIEPMANN, 1999)							●					
CPMD (ANDREONI; CURIONI, 2000)										●		
PACKMOL (MARTÍNEZ et al., 2009)	●											

Table B.2: First principle methods without database information: computational `ab initio` protein structure prediction methods. `BB` (Branch and Bound), `CSA` (Conformational Space Annealing), `MC` (Monte Carlo), `ST` (Stochastic Tunneling), `PT` (Parallel Tempering), `SB` (Swarm-based optimization), `MME` (Memetic algorithm), `GA` (Genetic Algorithm) and `REMC` (Replica Exchange Monte Carlo).

FIRST PRINCIPLE METHODS	MC	BB	CSA	GA	REMC	ST	PT	SB	MME
ASTROFOLD (KLEPEIS; FLOUDAS, 2003)	●	●	●						
BHAGEERATH (JAYARAM et al., 2006)	●								
LINUS (SRINIVASAN; ROSE, 2002)	●								
Unger and Moult (UNGER; MOULT, 1993)				●					
Hoque (HOQUE; CHETTY; DOOLEY, 2005)				●					
Dandekar and Argos (DANDEKAR; ARGOS, 1992)				●					
Grand and Merz (LE GRAND; MERZ JR., 1993)				●					
Pedersen and Moult (PEDERSEN; MOULT, 1997)				●					
Gibbs et al. (GIBBS; CLARKE; SESSIONS, 2001)	●								
Sun (SUN, 1995)				●					
Derreumaux (DERREUMAUX, 1999)	●								
Abagyan (ABAGYAN; TOTROV, 1994)	●								
Thachuk (THACHUK; SHMYGELSKA; HOOS, 2007)					●				
Pokarowski (POKAROWSKI; KOLINSKI; SKOLNICKZ, 2003)					●				
Herges et al. (HERGES et al., 2003)						●			
Schug et al. (SCHUG et al., 2005)							●		
Bahamisch et al. (BAHAMISH; ABDULLAH; SALAM, 2009)								●	
Fonseca et al. (FONSECA; PALUSZEWSKI; WINTER, 2010)								●	
Smith (SMITH, 2005)									●

APPENDIX C FIRST PRINCIPLE METHODS WITH DATABASE INFORMATION

Table C.1: First principle methods with database information. CA (Clustering Algorithms), MC (Monte Carlo), PPA (Simple Profile Alignment Method), SA (Simulated Annealing), GA (Genetic Algorithms), MEMC (Multi-canonical Ensemble Monte Carlo) with Metropolis criterium, CSA (Conformational Space Annealing), REMC (Replica Exchange Monte Carlo), HMMS (Hidden Markov Model), ANN (Artificial Neural Networks).

FIRST PRINCIPLE METHODS WITH DATABASE	MC	CA	PPA	SA	GA	REMC	CSA	ANN	HMMS	MEMC
I-TASSER (WU; SKOLNICK; ZHANG, 2007)	●	●	●							
ANGLOR (WU; ZHANG, 2008a)								●		
FRAGFOLD (JONES, 2001)		●		●	●					
SIMFOLD (CHIKENJIA; FUJITSUKAB; TAKADAC, 2003)										●
PROFESY (LEE et al., 2004)							●			
ROBETTA (ROHL et al., 2004)	●			●						
ROBETTA@ (DAS et al., 2007)	●			●						
CREF (DORN; SOUZA, 2010)		●								
A3N (DORN; SOUZA, 2010B)		●						●		
UNDERTAKER (KARPLUS et al., 2003)					●				●	
ABLE (ISHIDA et al., 2003)	●	●		●						
CABS (KOLINSKI, 2004)	●	●								
Park (PARK, 2005)					●					
FRAGMENT-HMM (LI et al., 2008)	●								●	
Cutello et al. (CUTELLO; NARZISI; NICOSIA, 2006)					●					

APPENDIX D THREADING METHODS

Table D.1: Threading methods. LP (Linear Programming), GA (Genetic Algorithms), HMMS (Hidden Markov Models), ANN (Artificial Neural Networks), DP (Dynamic Programming), DC (Divide and Conquer), PA (Profile-analysis), ST (Screening Techniques), SEB (Sequence-based), STU (Structure-based), and ILP (Inductive Logic Programing).

Method	ANN	DP	GA	HMMS	LP	DC	PA	ST	SEB	STU	ILP
GENTHREADER (JONES, 1999)	●									●	
ORFEUS (GINALSKI et al., 2003)			●						●		
PROSPECT (XU; XU, 2000)						●				●	
123D (ALEXANDROV; NUSSINOV; ZIMMER, 1996)		●								●	
FFAS (RYCHLEWSKI et al., 2000)							●		●		
ESYPRED3D (LAMBERT et al., 2002)								●	●		
3DPSSM (KELLEY; MACCALLUM; STERNBERG, 2000)		●								●	
FUNGUE (SHI; BLUNDELL; MIZUGUCHI, 2001)							●			●	
RAPTOR (XU et al., 2003)					●					●	
SAM-T99 (KARPLUS; BARRETT; HUGHEY, 1992)				●					●		
SAM-T02 (KARPLUS et al., 2001)				●						●	
LIBRA I (OTA; NISHIKAWA, 1997)							●			●	
TOPITS (ROST, 1995)	●	●							●		
Turcotte et.al. (TURCOTTE; MUGGLETON; STERNBERG, 1998)											●
MUSTER (WU; ZHANG, 2008b)		●									
HHPRED (SODING, 2005)				●							
COTH (zhanglab.ccmb.med.umich.edu/COTH)		●									
THREADER2 (JONES; MILLER; THORNTON, 1995)	●										
SEGMER (WU; ZHANG, 2010)				●							

APPENDIX E COMPARATIVE MODELING METHODS SUMMARY

Table E.1: Comparative Modeling Methods Summary. CG (Conjugate Gradient), MD (Molecular Dynamics), SA (Simulated Annealing), HMMS (Hidden Markov Model), CA (Clustering Algorithm), GA (Genetic Algorithms), DP (Dynamic Programming), RB (Modeling by Assembly of Rigid Body), SM (Modelling by Segment Matching or Coordinate Reconstruction), SR (Modeling by Satisfaction of Spatial Restraints).

Method	CG	MD	SA	HMMS	CA	DP	RB	SM	SR	Package
SWISS-MODEL (ARNOLD et al., 2006)				●			●	●		yes
MODELLER (ESWAR et al., 2006)	●	●	●					●	●	yes
T-COFFEE (NOTREDAME; HOLM; HIGGINS, 1998)						●		●		yes
BLAST (ALTSCHUL et al., 1997)						●		●		no
PSI-BLAST (ALTSCHUL et al., 1997)						●		●		no
FASTA (LIPMAN; PEARSON, 1985)						●		●		no
CLUSTALW (LARKIN et al., 2007)						●				yes
TIP-STRUCTFAST (DEBE et al., 2006)						●		●		no
MULTALIN (CORPET, 1988)					●			●		no
COMPASS (SADREYEV; GRISHIN, 2003)						●		●		no

APPENDIX F TORSION ANGLES OF β-TURNS

Table F.1: Conformational angles for most common types of β-turns. Adapted from (LILJAS et al., 2001).

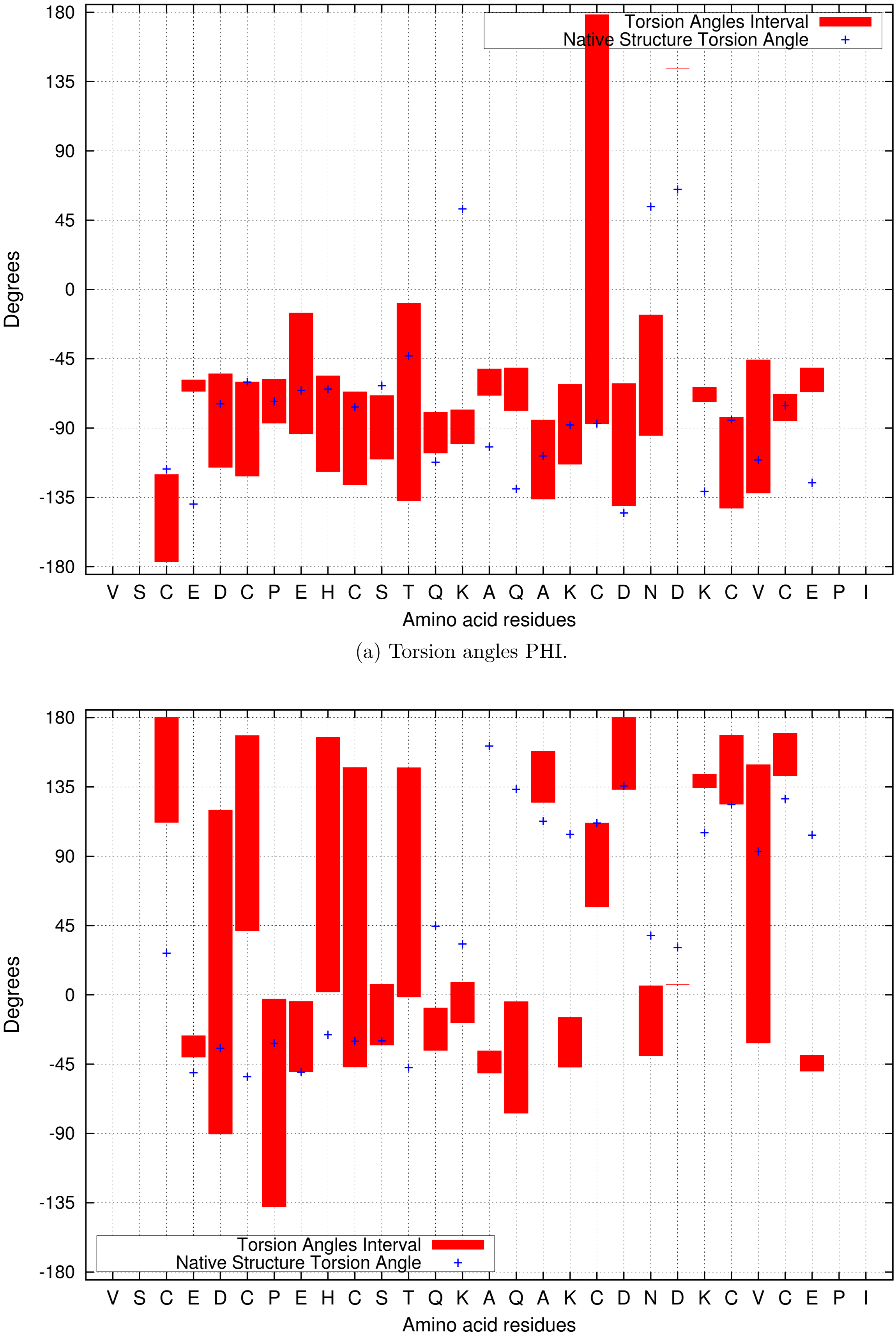
	Residue n+1		Residue n+2	
Type	PHI	PSI	PHI	PSI
I	-60.0	-30.0	90.0	0.0
I'	60.0	30.0	90.0	0.0
II	-60.0	120.0	80.0	0.0
II'	60.0	-120.0	-80.0	0.0
VIa	-60.0	120.0	-90.0	0.0
VIb	-135.0	135.0	-75.0	0.0
VIII	-60.0	-30.0	-120.0	120.0

APPENDIX G SIDE-CHAIN TORSION ANGLES

Table G.1: Side-chain torsion angle intervals obtained from the Dunbrack `Rotamers` library.

Amino acid	χ_1		χ_2		χ_3		χ_4	
residue	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX
Proline	-	-	-	-	-	-	-	-
Alanine	-	-	-	-	-	-	-	-
Cysteine	-	-	-	-	-	-	-	-
Serine	-39.09	159.36	-	-	-	-	-	-
Cysteine	-157.75	39.15	-	-	-	-	-	-
Threonine	-153.24	38.18	-	-	-	-	-	-
Valine	-37.18	156.98	-	-	-	-	-	-
Isoleucine	-150.62	36.96	-41.95	158.55	-	-	-	-
Leucine	-158.45	34.56	-41.16	153.69	-	-	-	-
Aspartic acid	-154.75	35.11	-46.01	46.96	-	-	-	-
Asparagine	-154.99	36.44	-78.68	102.88	-	-	-	-
Histidine	-156.43	39.01	-96.58	134..60	-	-	-	-
Phenylalanine	-158.22	37.19	-1.89	90.42	-	-	-	-
Tyrosine	-158.51	40.68	-3.56	90.33	-	-	-	-
Tryptophan	-159.33	37.20	-77.50	76.16	-	-	-	-
Methionine	-159.53	38.61	-95.98	136.93	-113.61	127.81	-	-
Glutamic acid	-154.33	36.01	-114.38	127.16	-46.26	45.90	-	-
Glutamine	-156.10	38.25	-130.31	110.02	-107.20	78.35	-	-
Lysine	-155.04	45.98	-131.52	107.88	-131.38	109.22	-131.38	109.22
Arginine	-151.83	55.45	-105.20	135.48	-136.74	95.51	-136.74	95.51

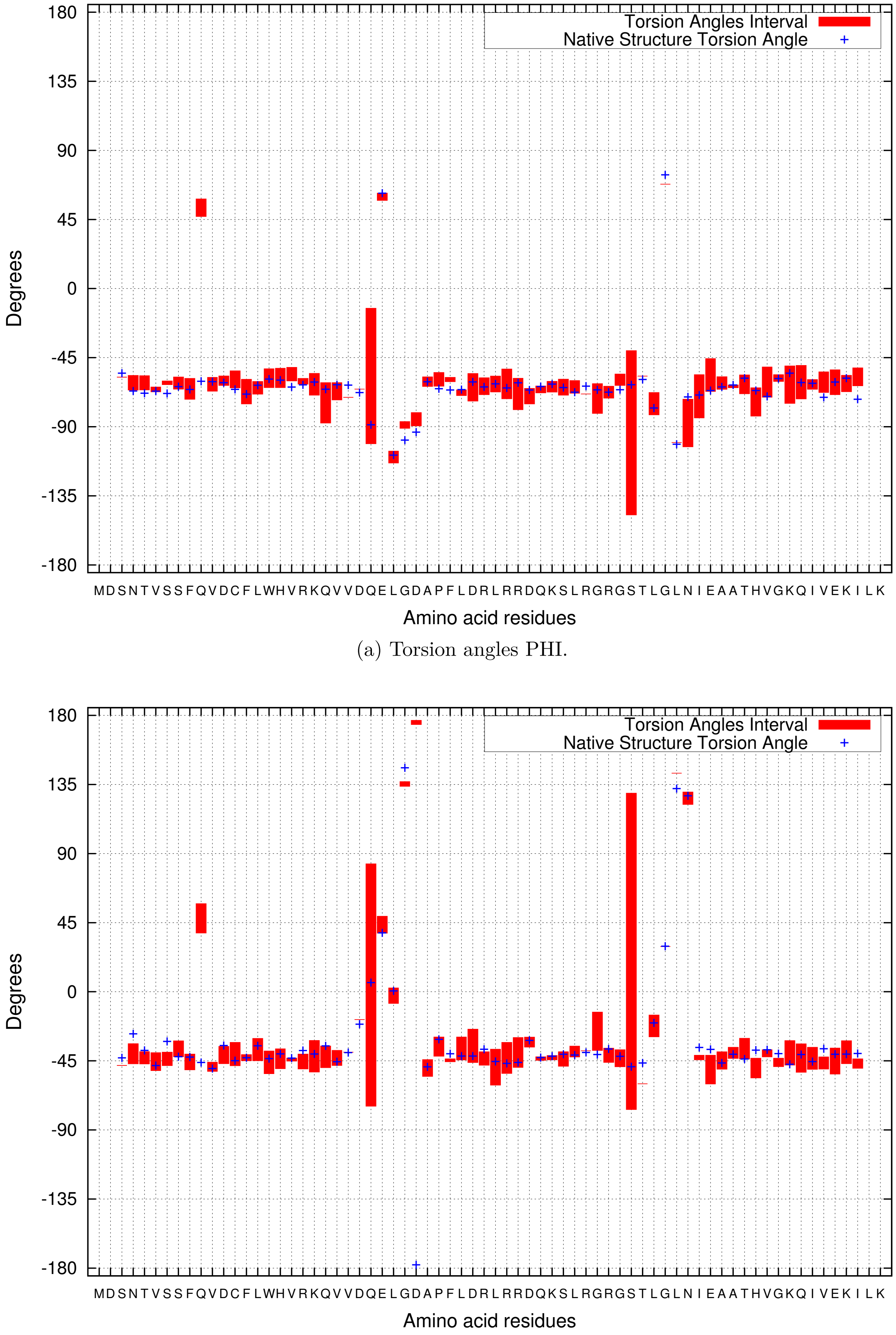
APPENDIX H TORSION ANGLES INTERVALS



(a) Torsion angles PHI.

(b) Torsion angles PSI.

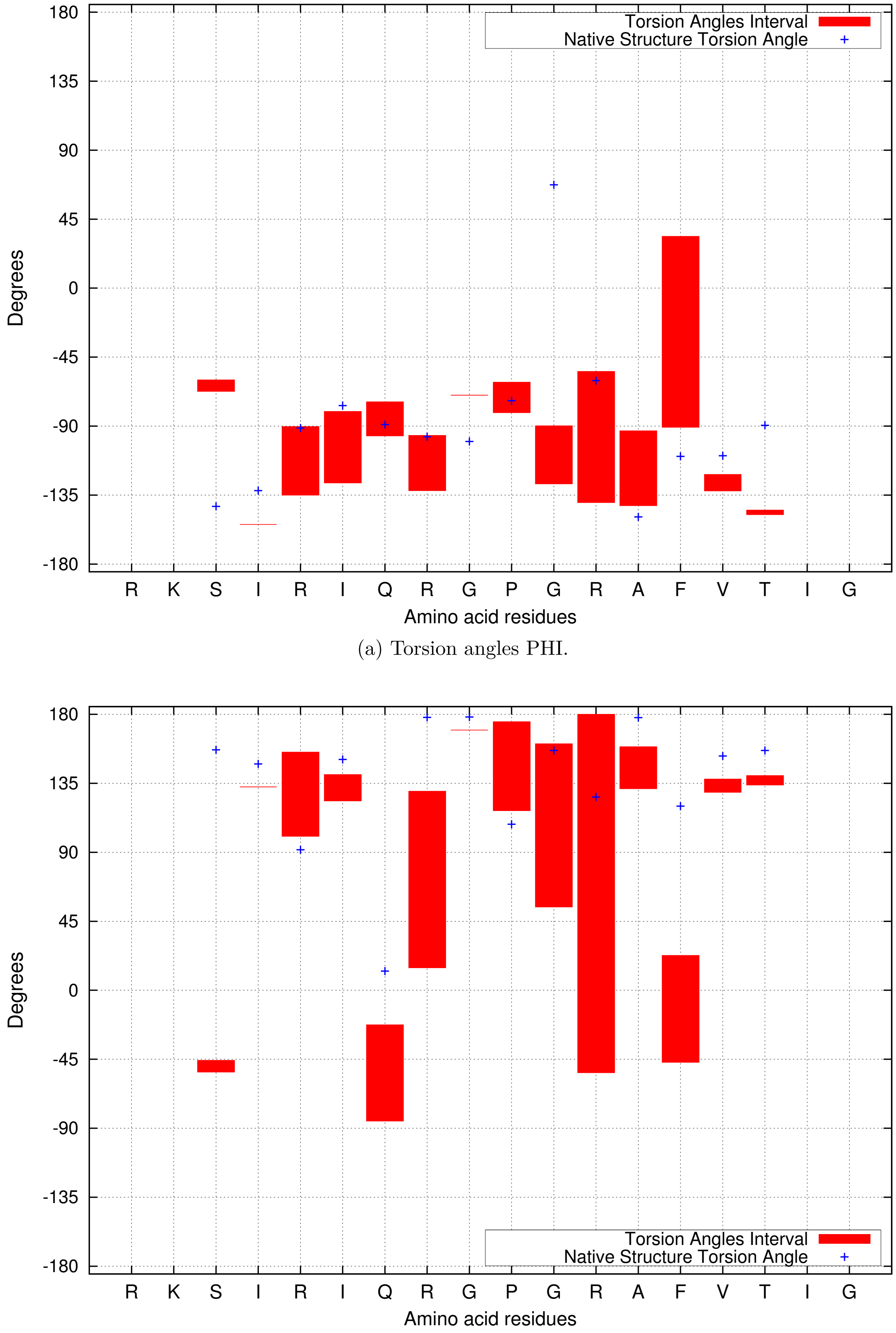
Figure H.1: Torsion angles interval for the protein with `PDB ID` = `1ACW`. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.



(a) Torsion angles PHI.

(b) Torsion angles PSI.

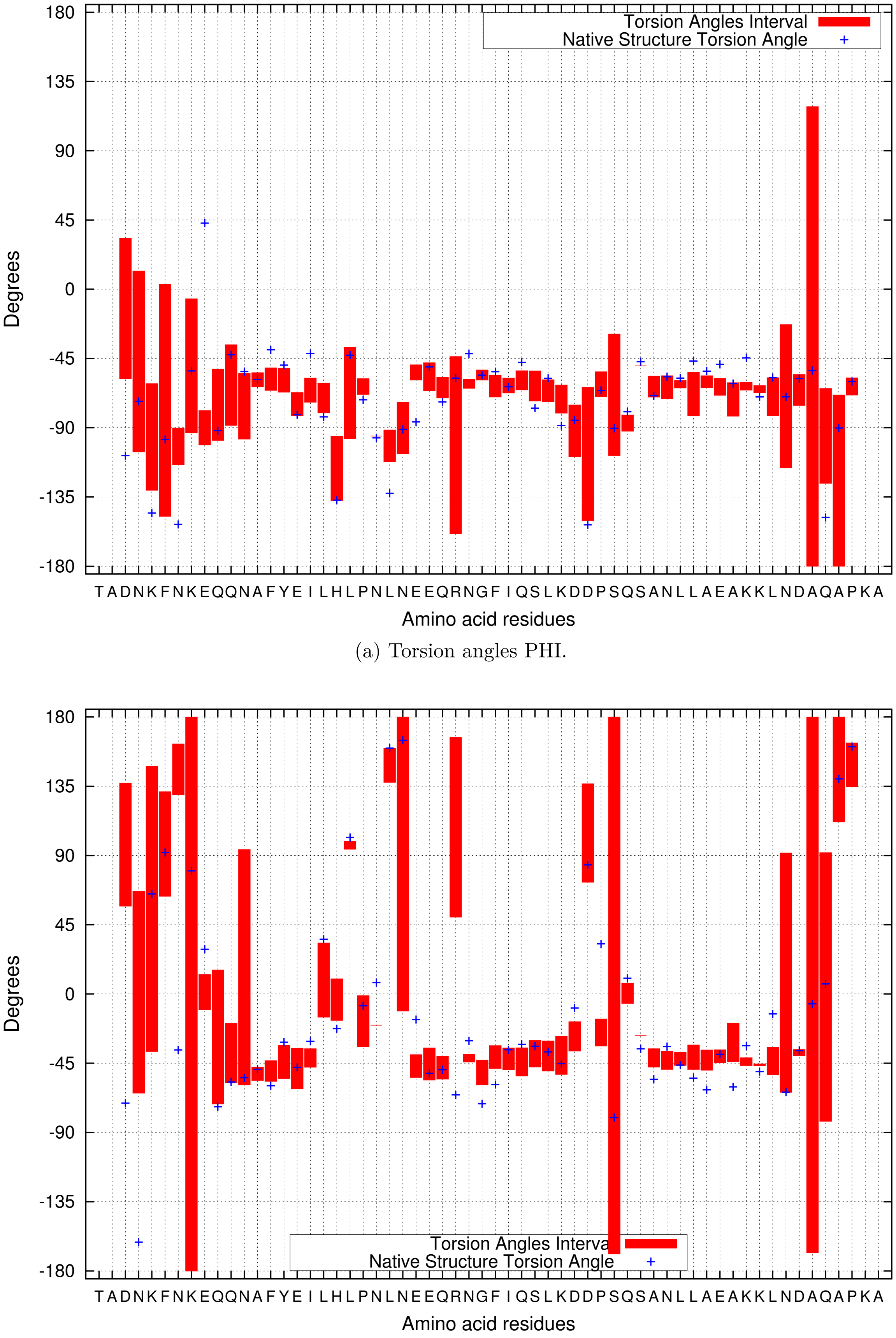
Figure H.2: Torsion angles interval for the protein with PDB ID = 1AIL. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.



(a) Torsion angles PHI.

(b) Torsion angles PSI.

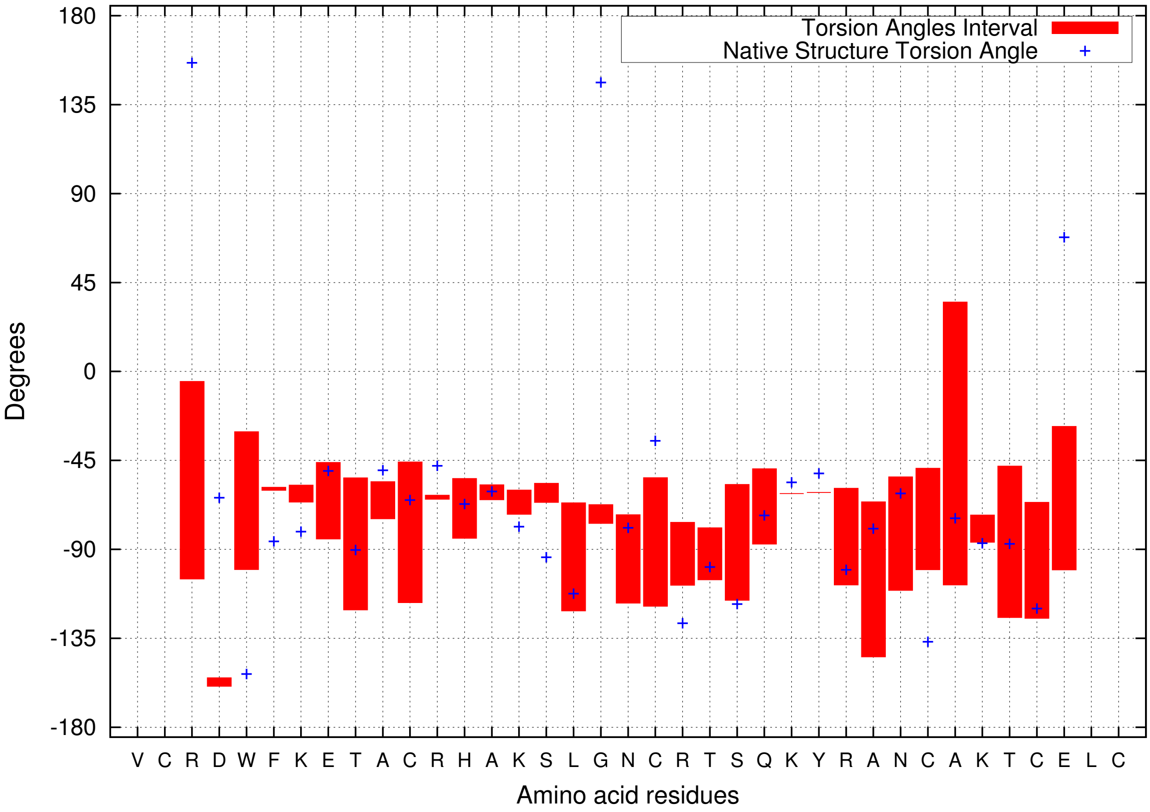
Figure H.3: Torsion angles interval for the protein with `PDB ID` = `1B03`. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.



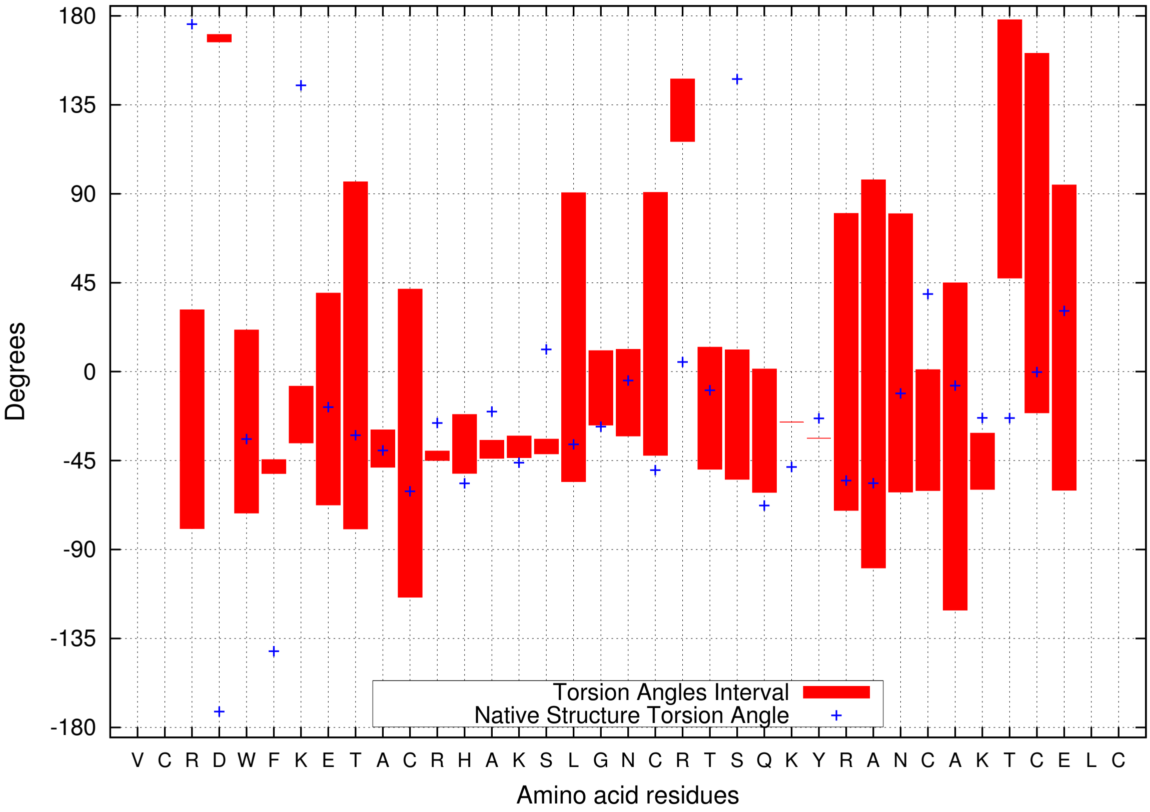
(a) Torsion angles PHI.

(b) Torsion angles PSI.

Figure H.4: Torsion angles interval for the protein with PDB ID = 1BDC. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.

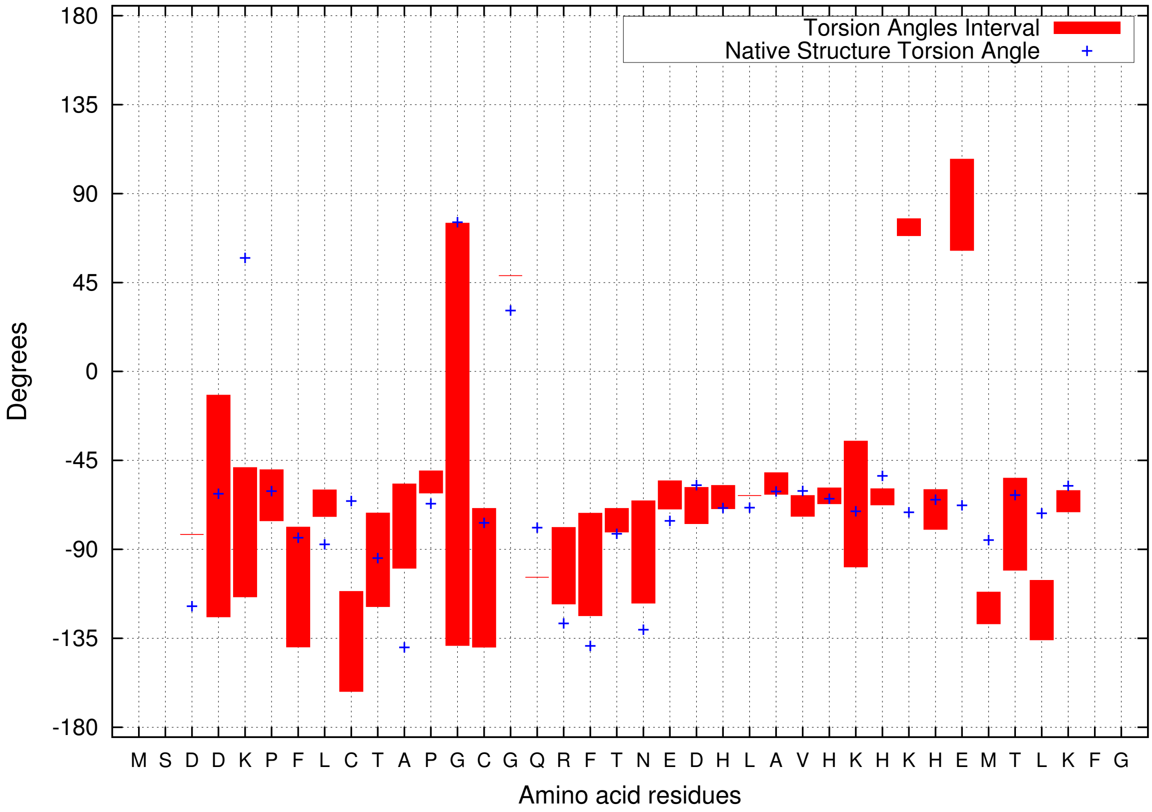


(a) Torsion angles PHI.

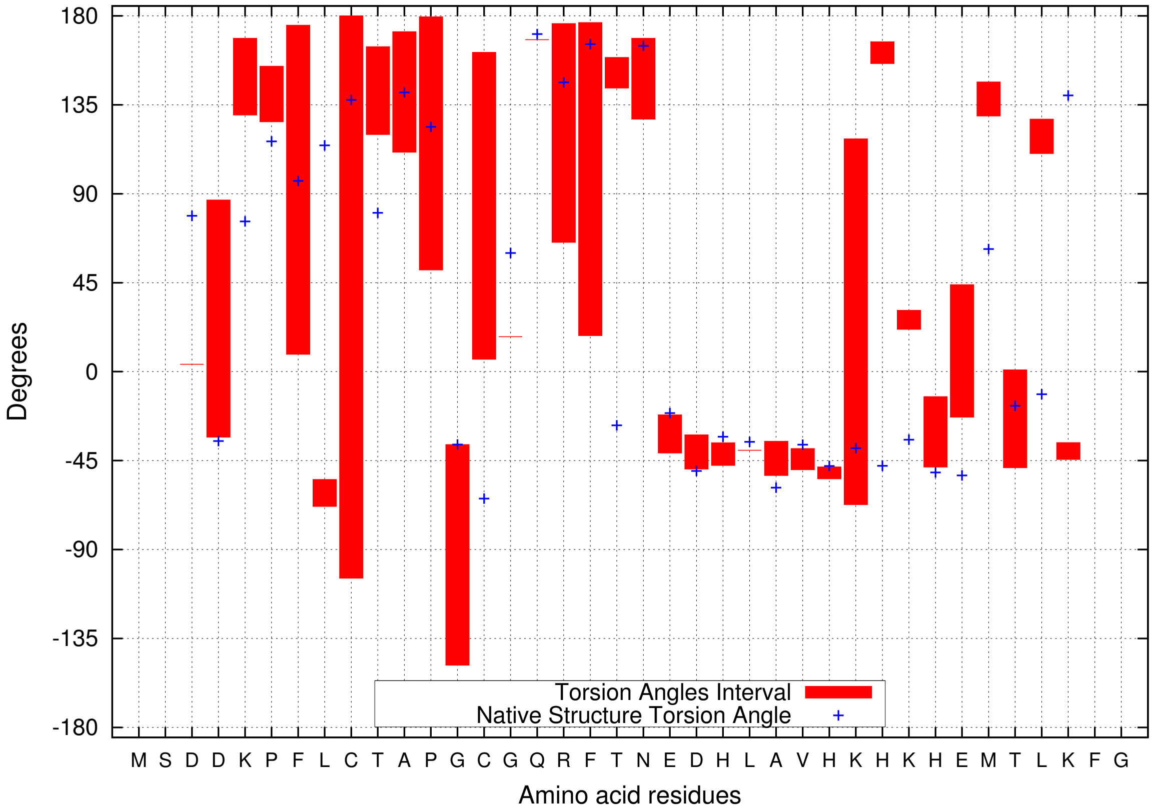


(b) Torsion angles PSI.

Figure H.5: Torsion angles interval for the protein with PDB ID = 1BGK. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.

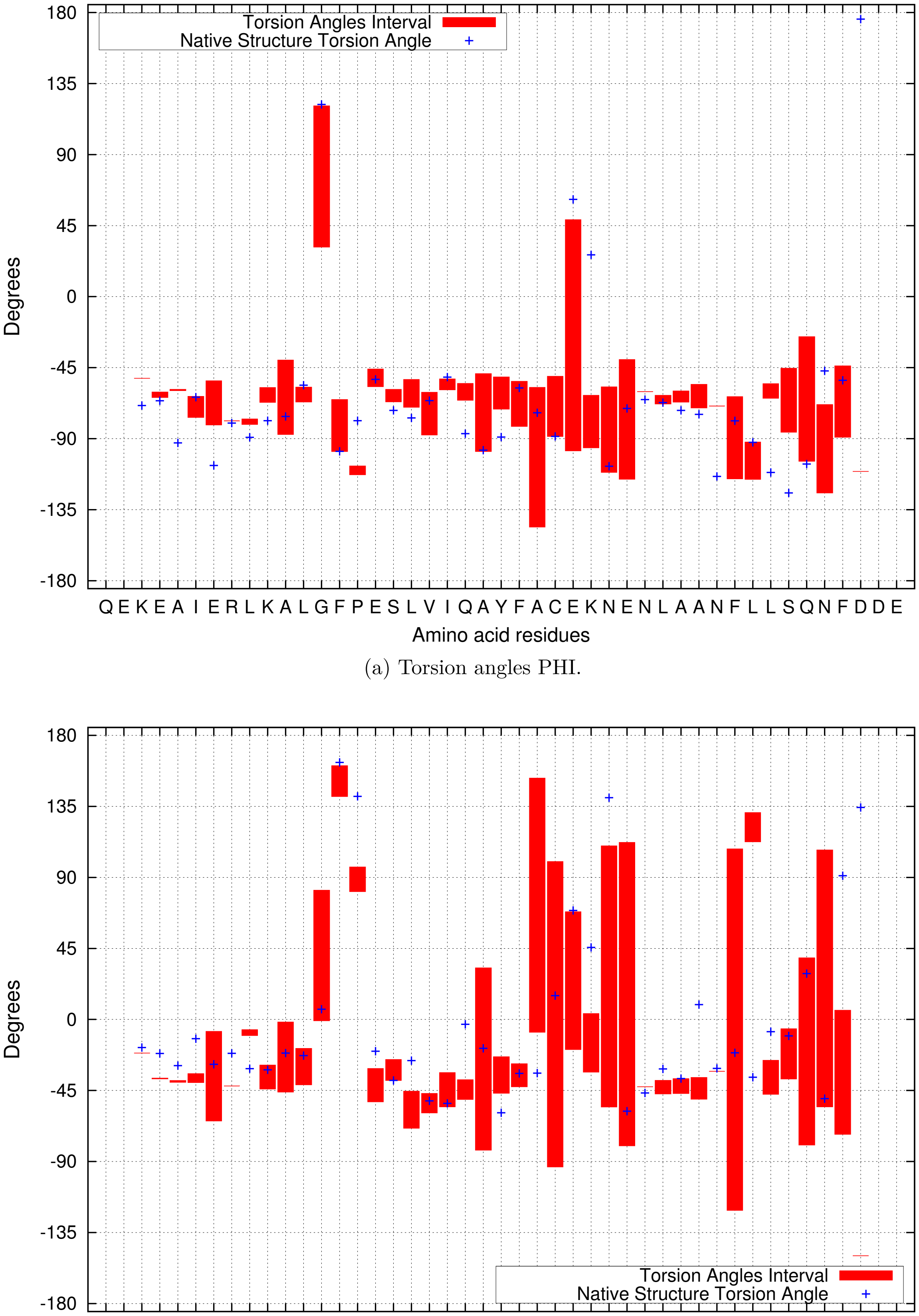


(a) Torsion angles PHI.



(b) Torsion angles PSI.

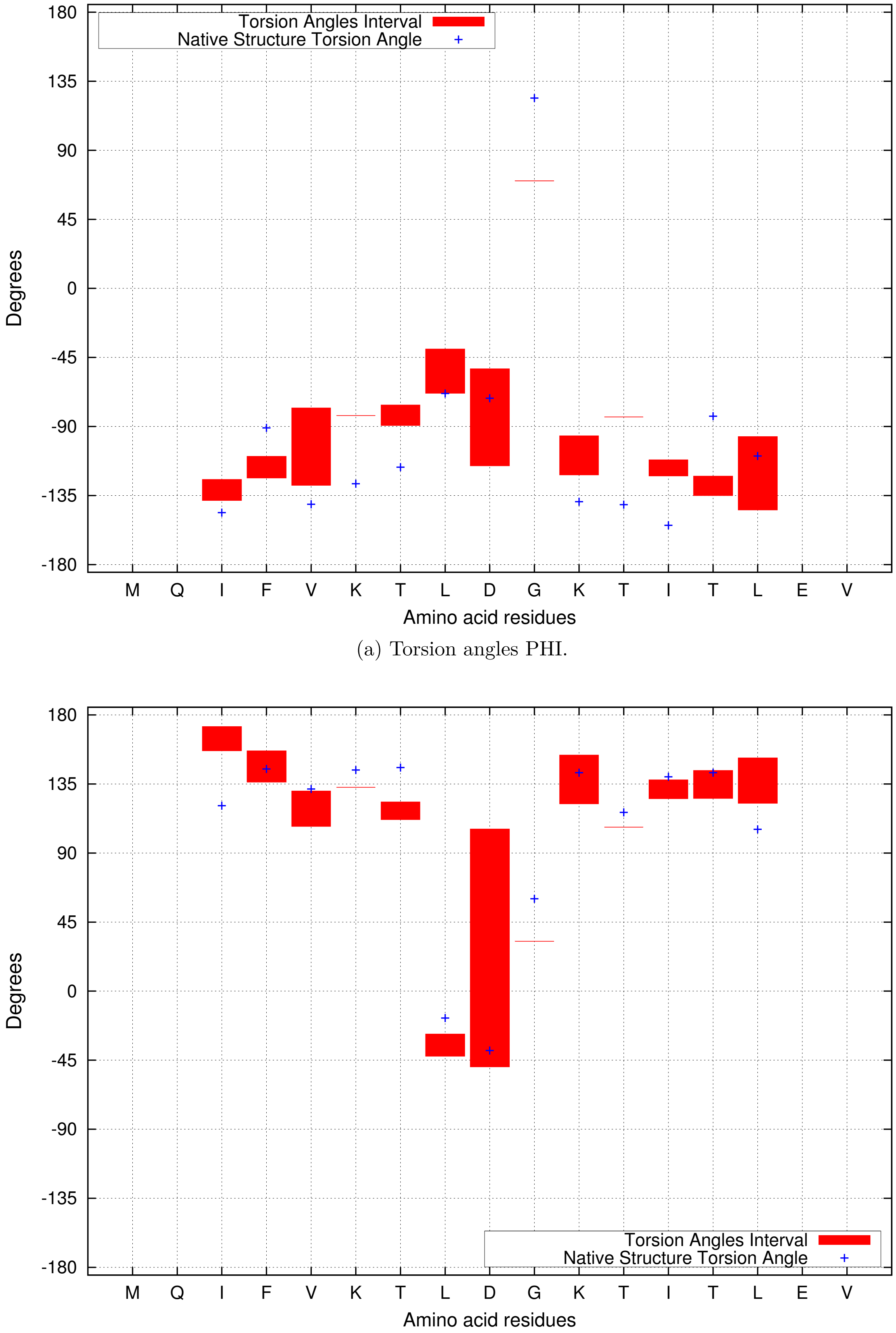
Figure H.6: Torsion angles interval for the protein with PDB ID = 1BHI. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.



(a) Torsion angles PHI.

(b) Torsion angles PSI.

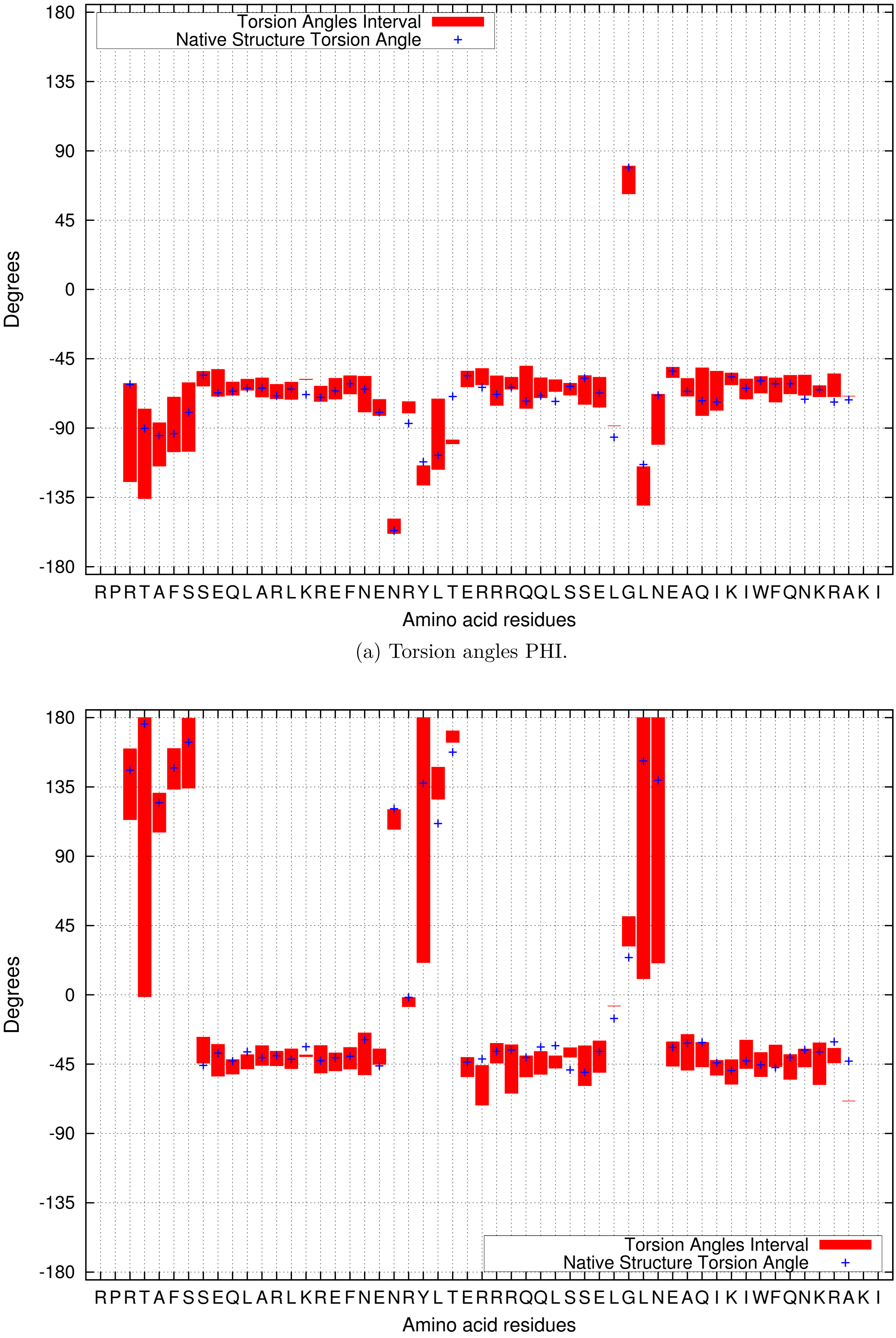
Figure H.7: Torsion angles interval for the protein with `PDB ID` = `1DV0`. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.



(a) Torsion angles PHI.

(b) Torsion angles PSI.

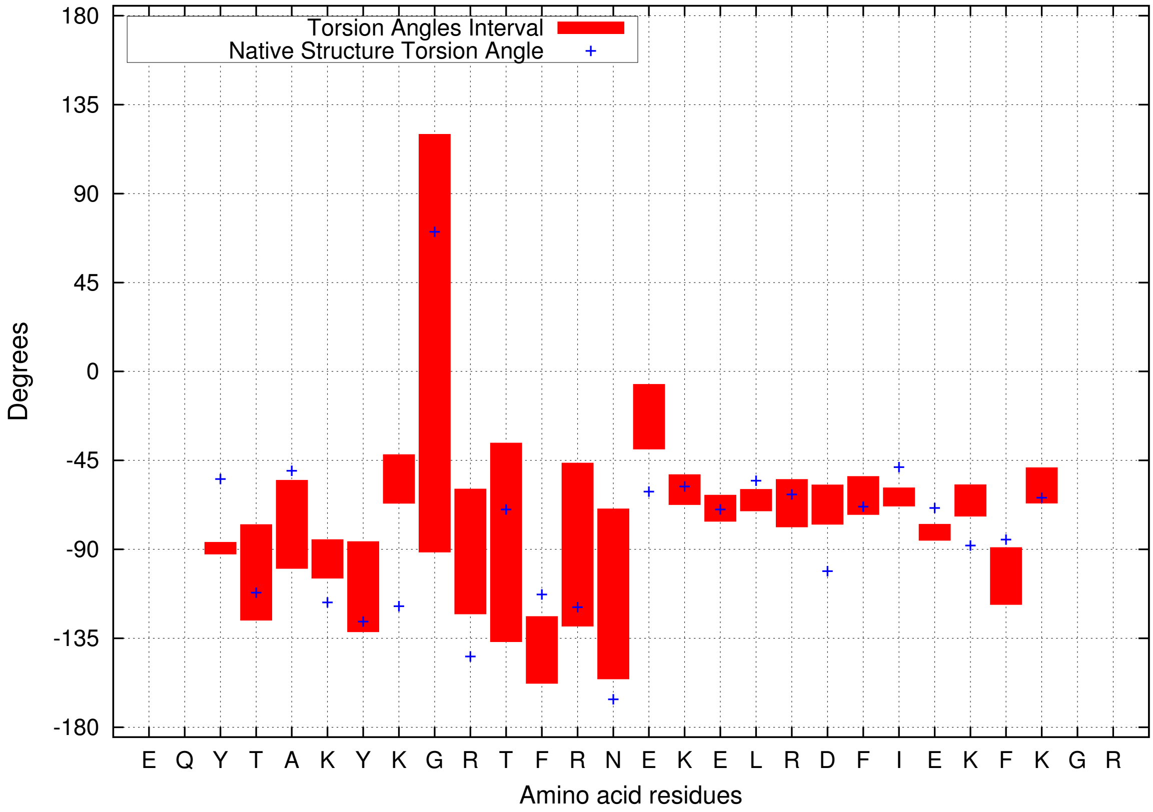
Figure H.8: Torsion angles interval for the protein with `PDB ID` = `1EOQ`. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.



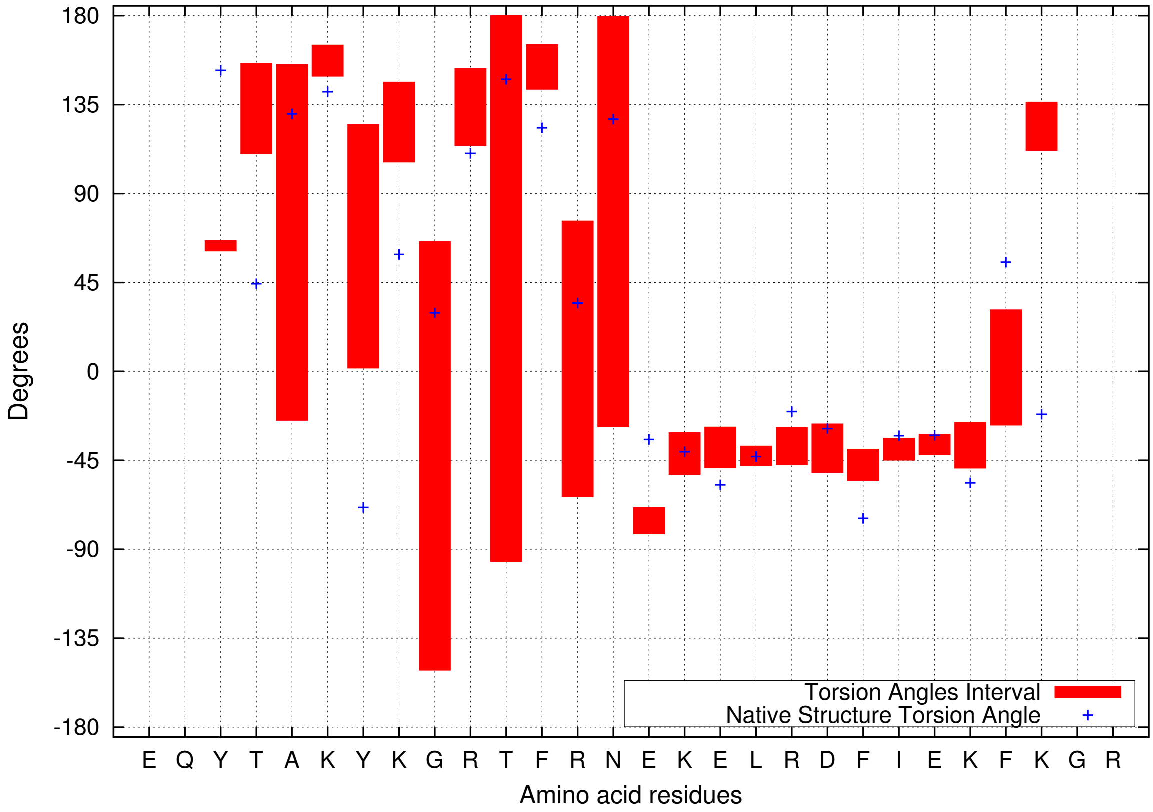
(a) Torsion angles PHI.

(b) Torsion angles PSI.

Figure H.9: Torsion angles interval for the protein with `PDB ID` = `1ENH`. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.

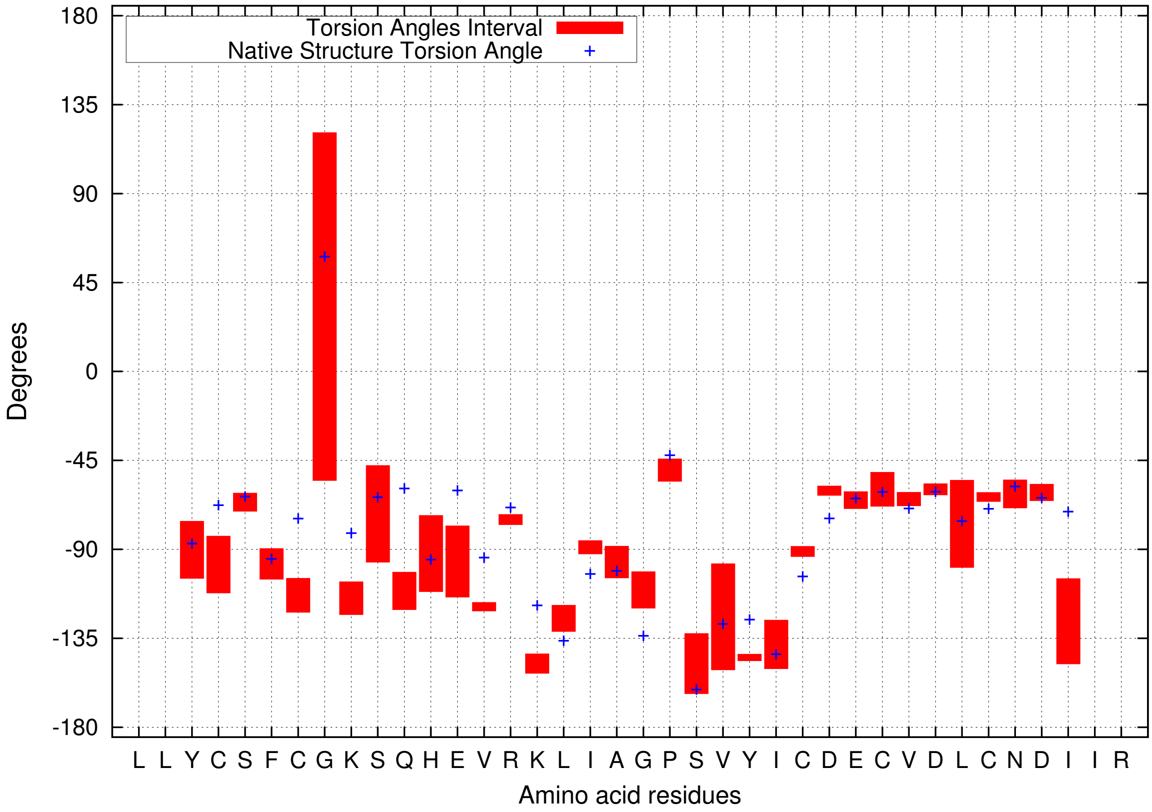


(a) Torsion angles PHI.

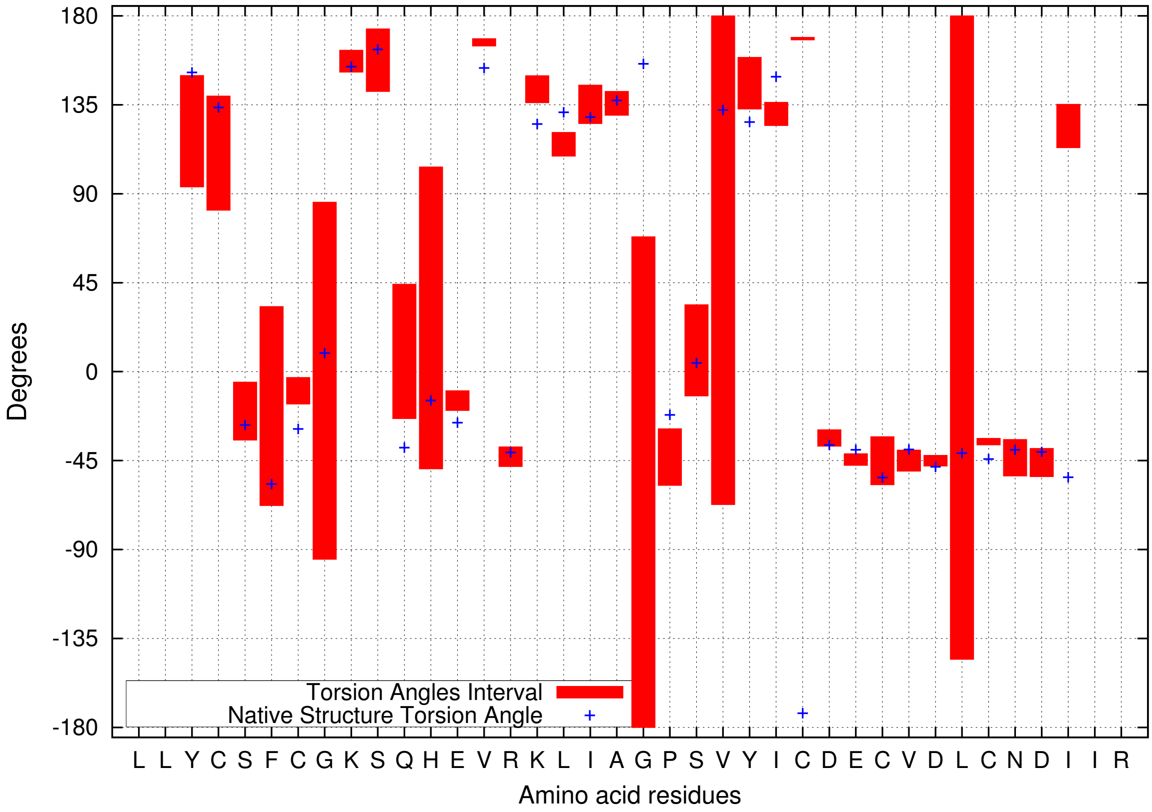


(b) Torsion angles PSI.

Figure H.10: Torsion angles interval for the protein with `PDB ID` = `1FME`. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.

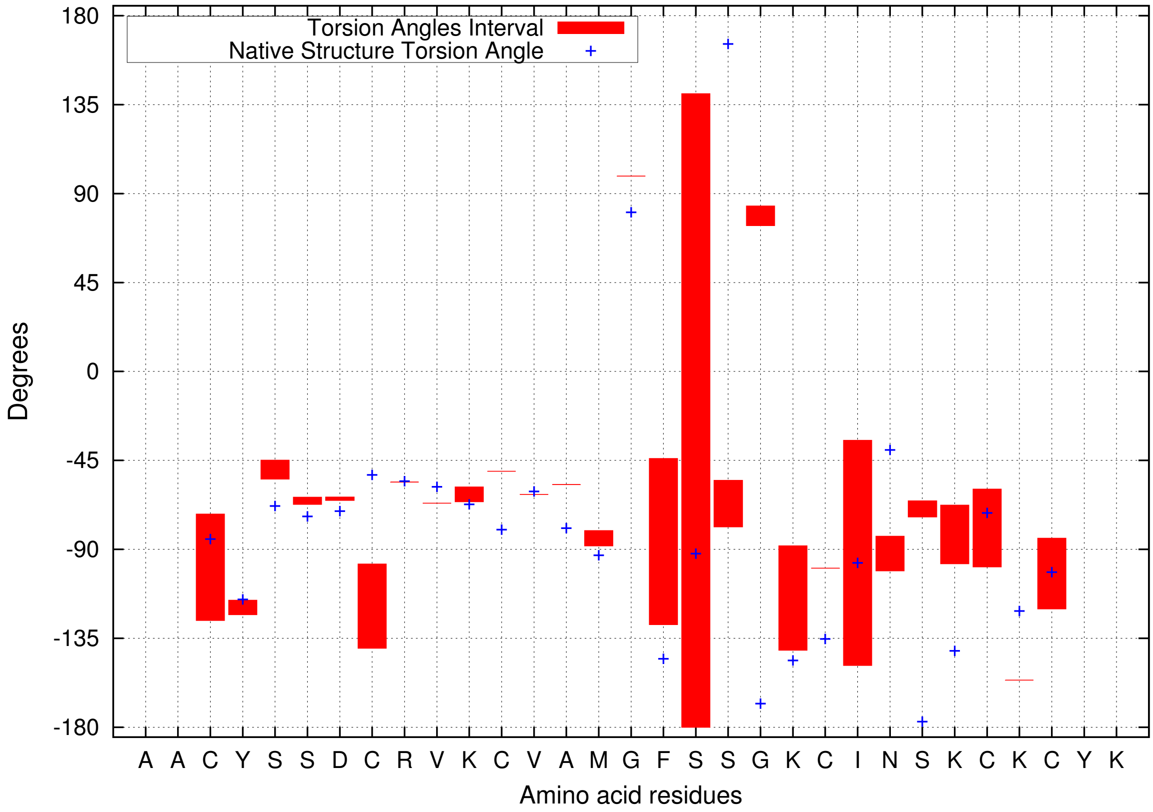


(a) Torsion angles PHI.

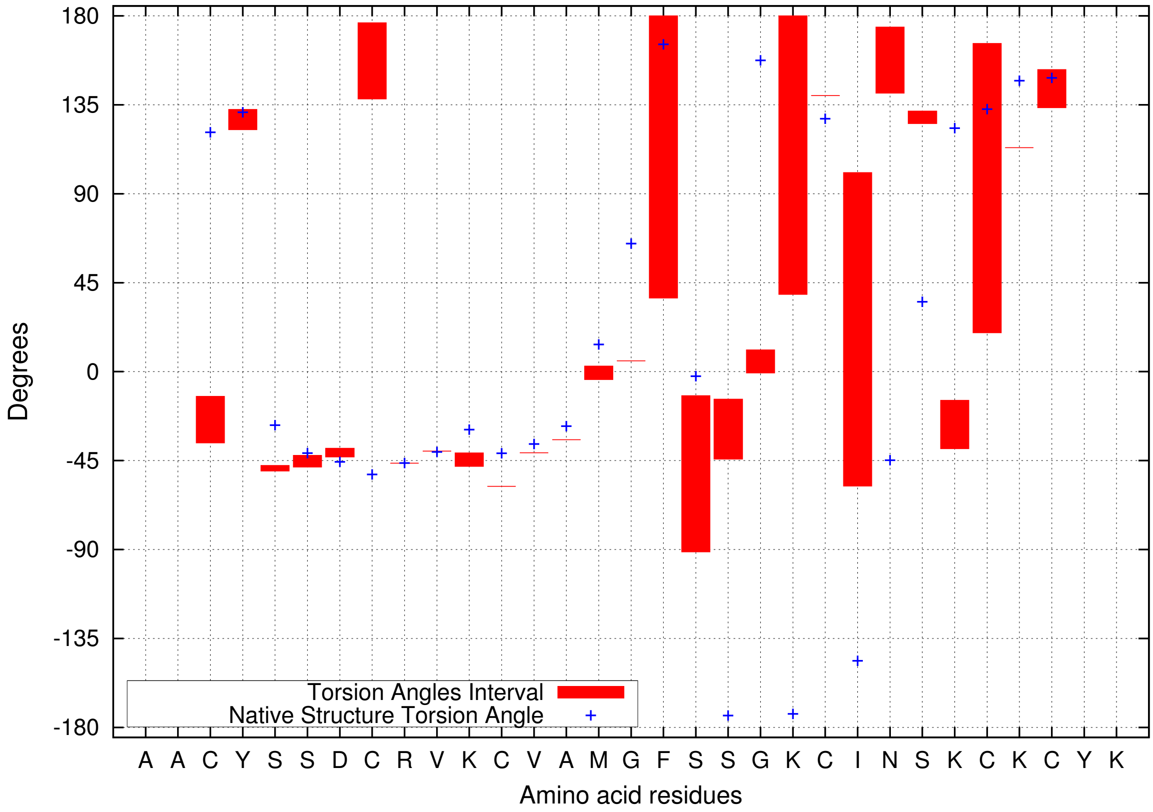


(b) Torsion angles PSI.

Figure H.11: Torsion angles interval for the protein with `PDB ID` = `1OVX`. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.

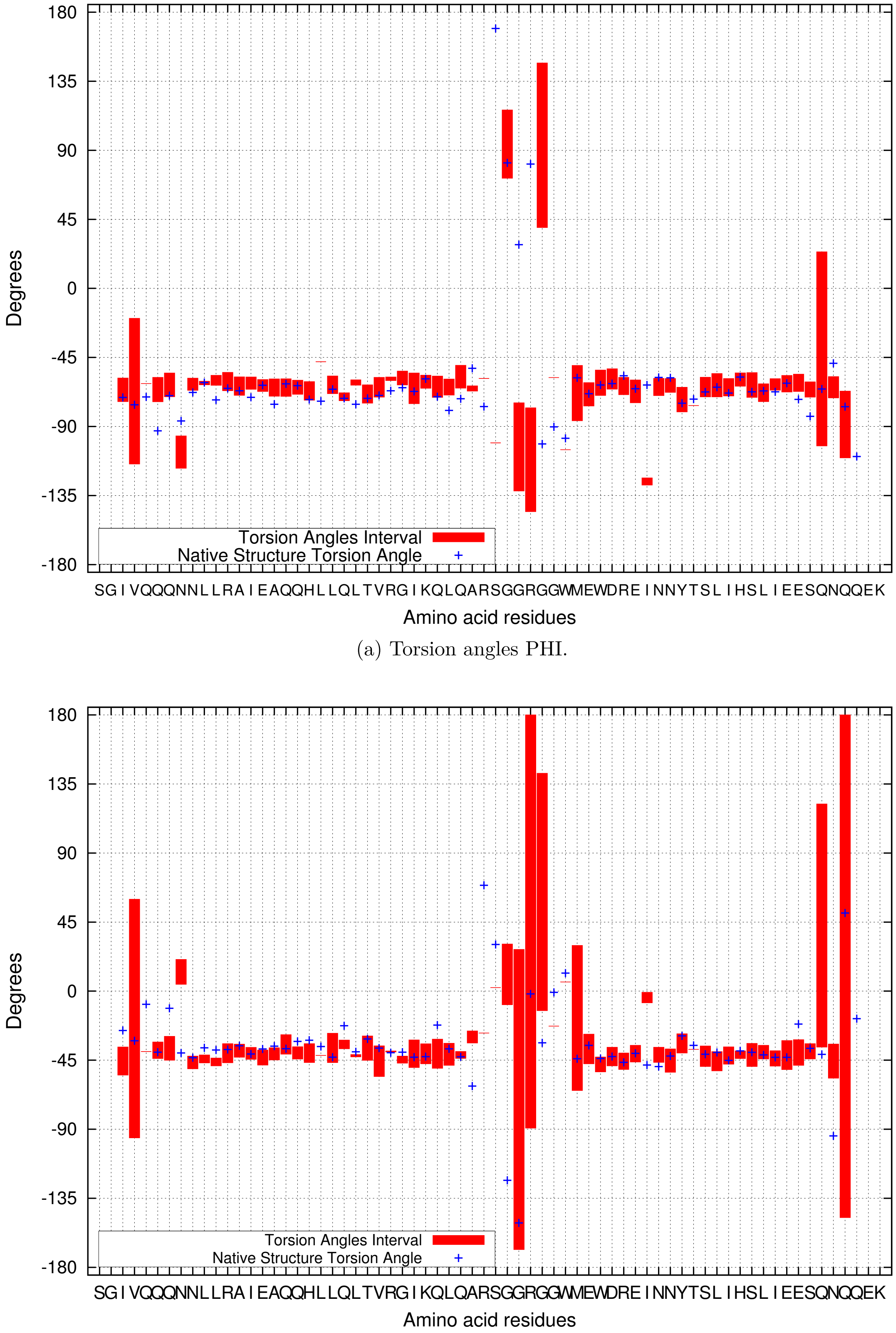


(a) Torsion angles PHI.



(b) Torsion angles PSI.

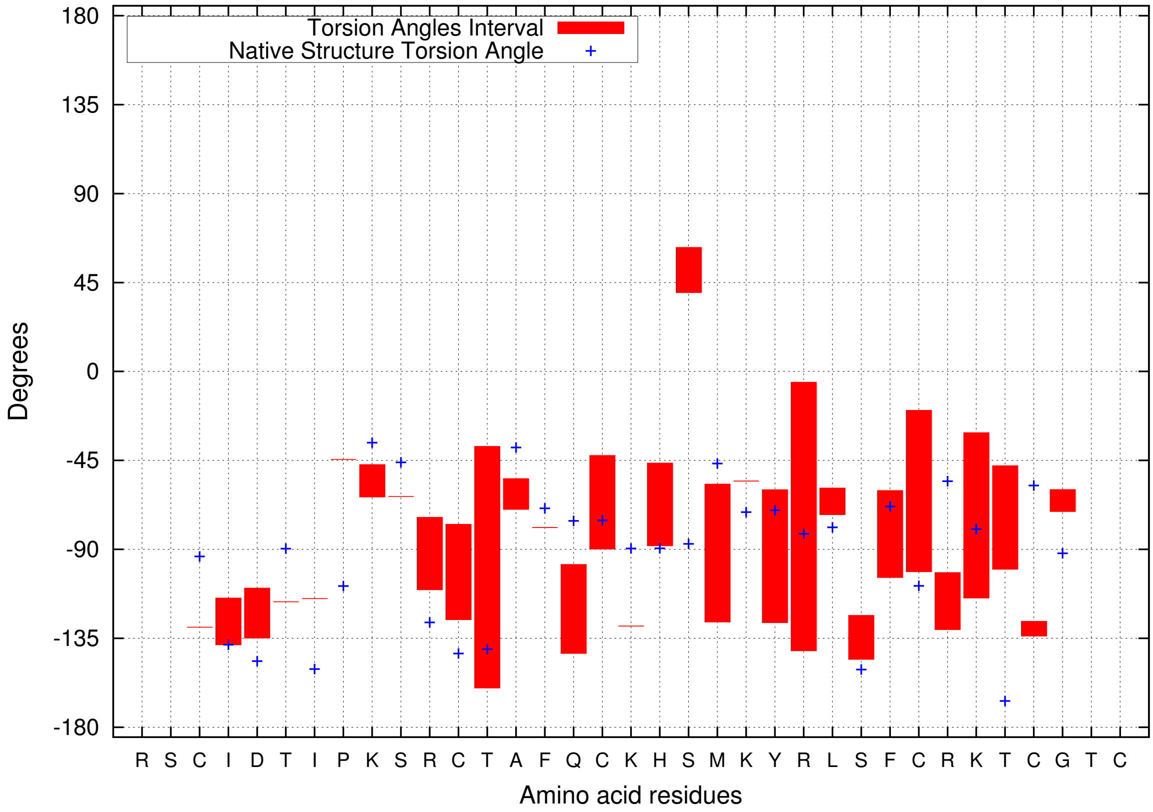
Figure H.12: Torsion angles interval for the protein with PDB ID = 1Q2K. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.



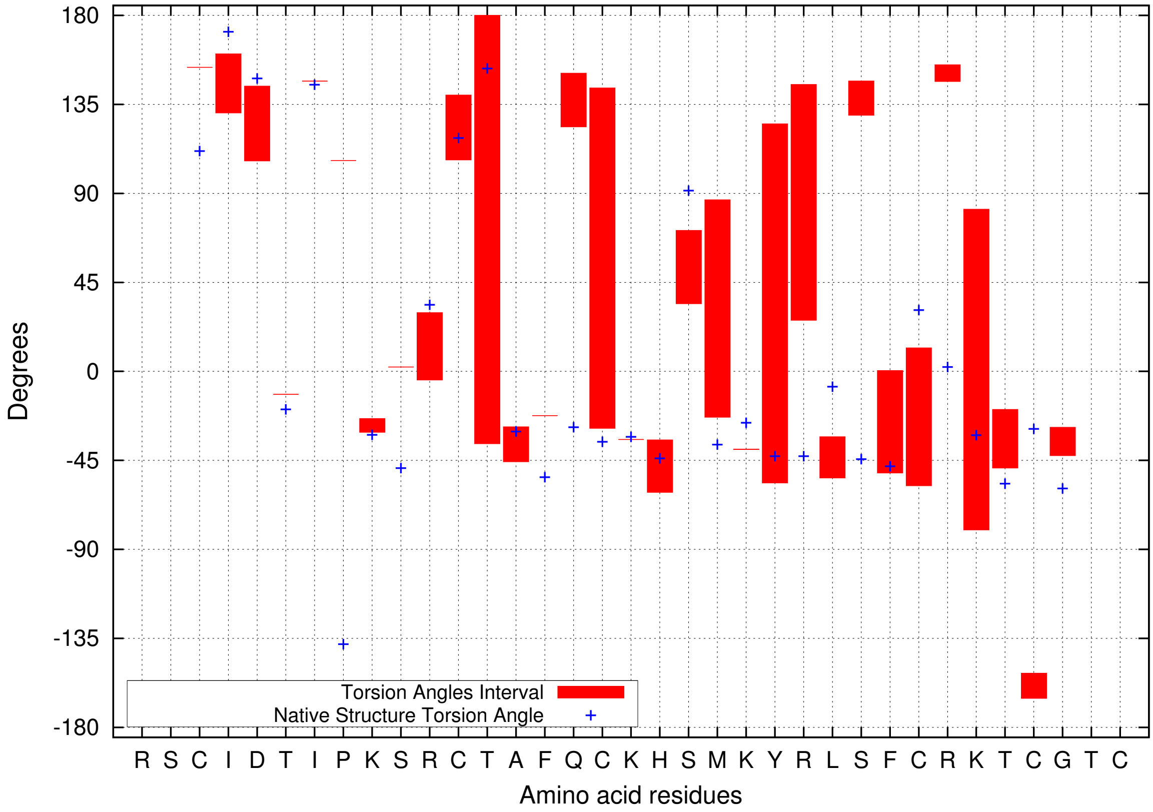
(a) Torsion angles PHI.

(b) Torsion angles PSI.

Figure H.13: Torsion angles interval for the protein with `PDB ID` = `1QR8`. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.

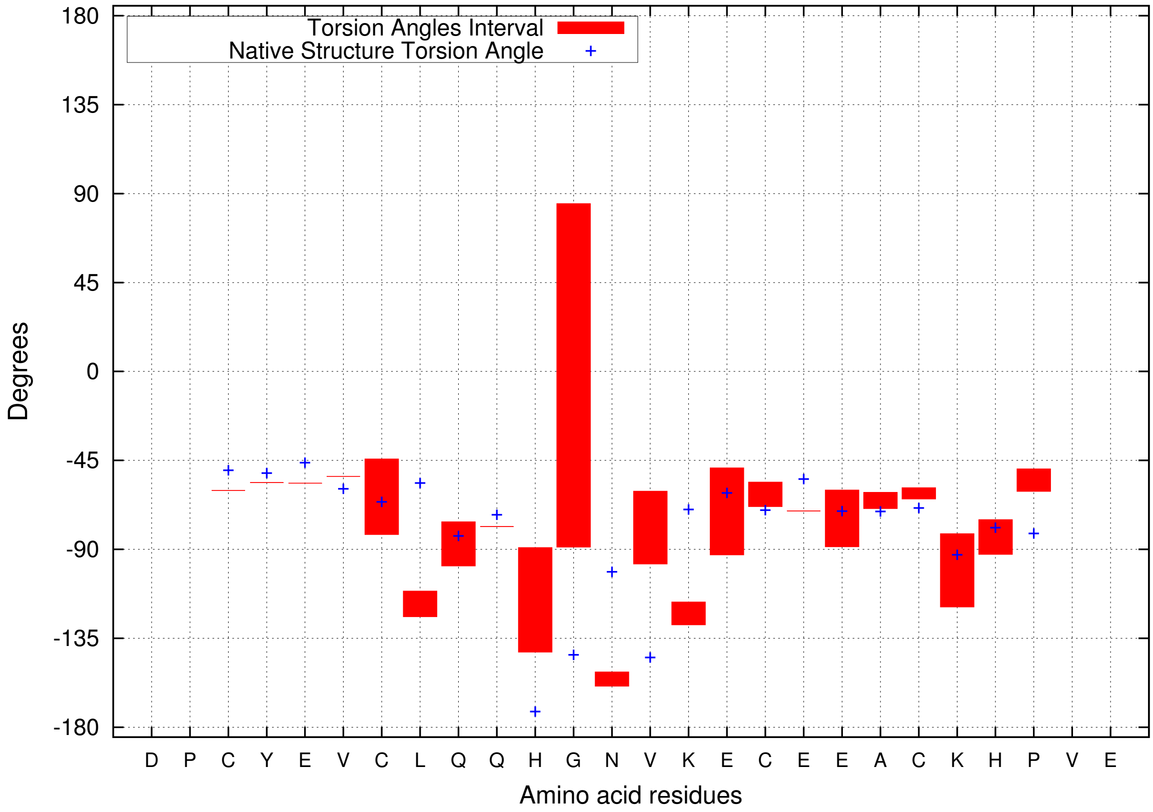


(a) Torsion angles PHI.

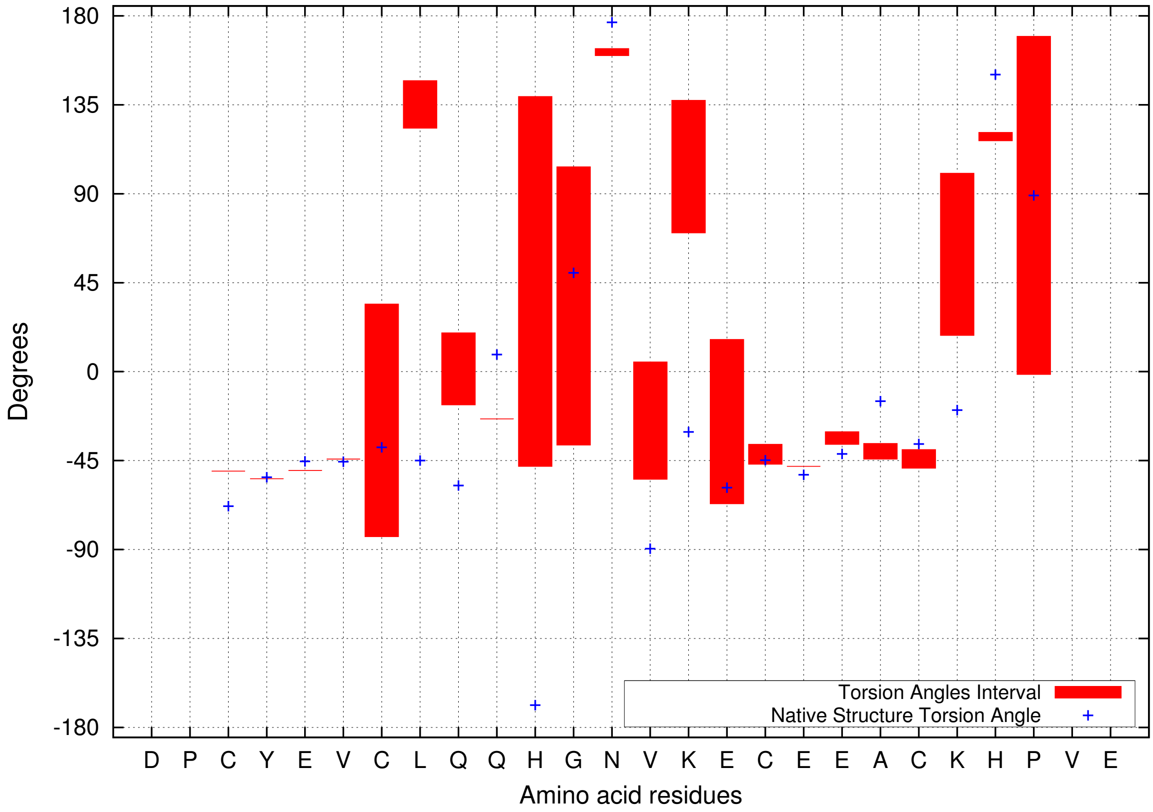


(b) Torsion angles PSI.

Figure H.14: Torsion angles interval for the protein with `PDB ID` = `1R00`. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.



(a) Torsion angles PHI.



(b) Torsion angles PSI.

Figure H.15: Torsion angles interval for the protein with `PDB ID` = `1WQC`. Filled boxes represent the constrained torsion angle interval. Blue dots identify the torsion angle value of the amino acid residue in the native state of the target protein.

APPENDIX I EXPERIMENTS: FITNESS GRAPHS

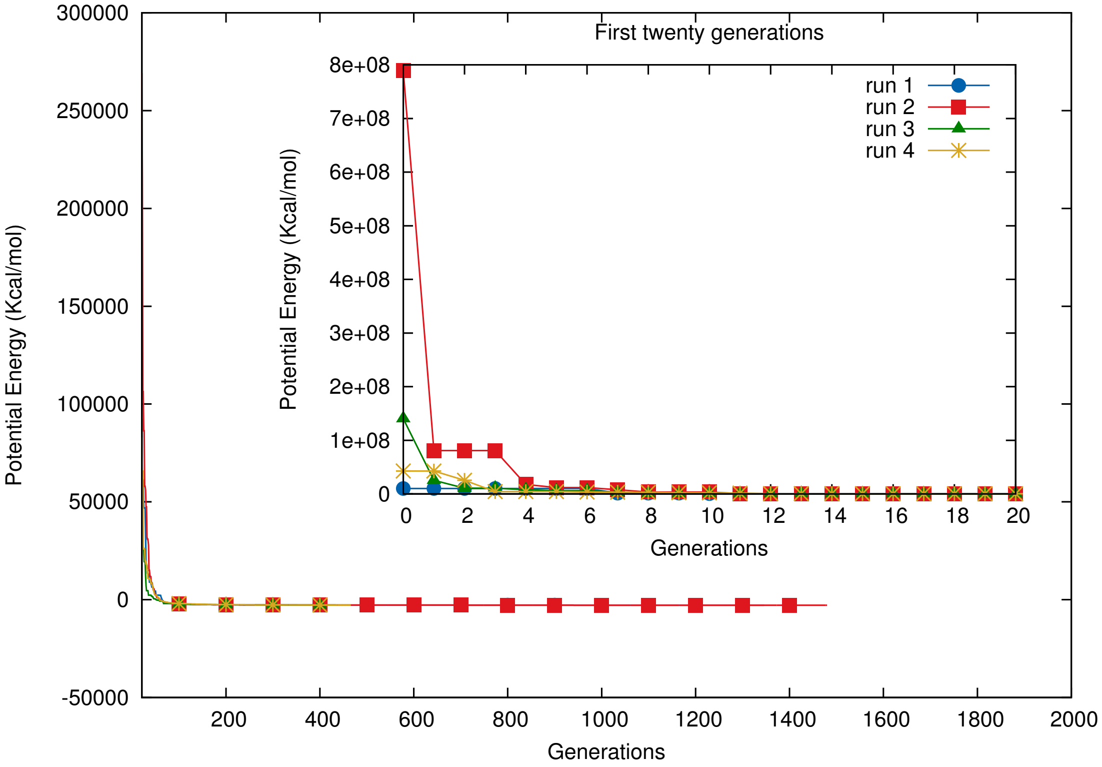


Figure I.1: Potential energy minimization for protein with `PDB ID` = `1AIL`.

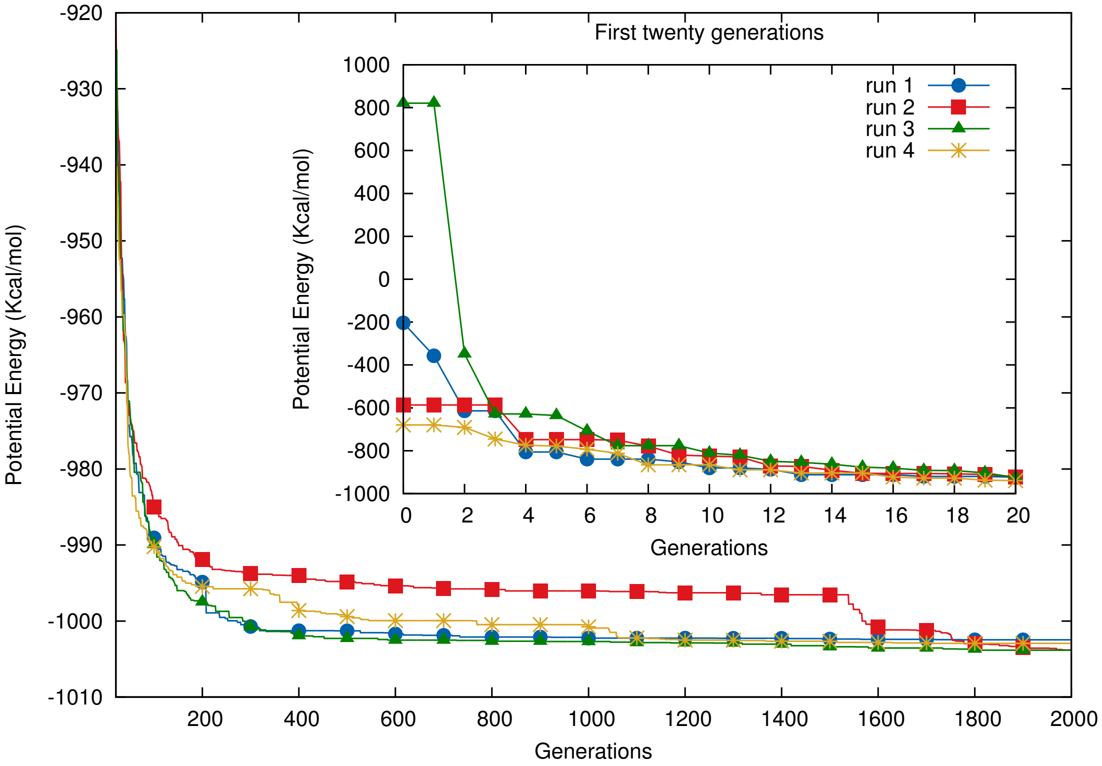


Figure I.2: Potential energy minimization for protein with `PDB ID` = `1B03`.

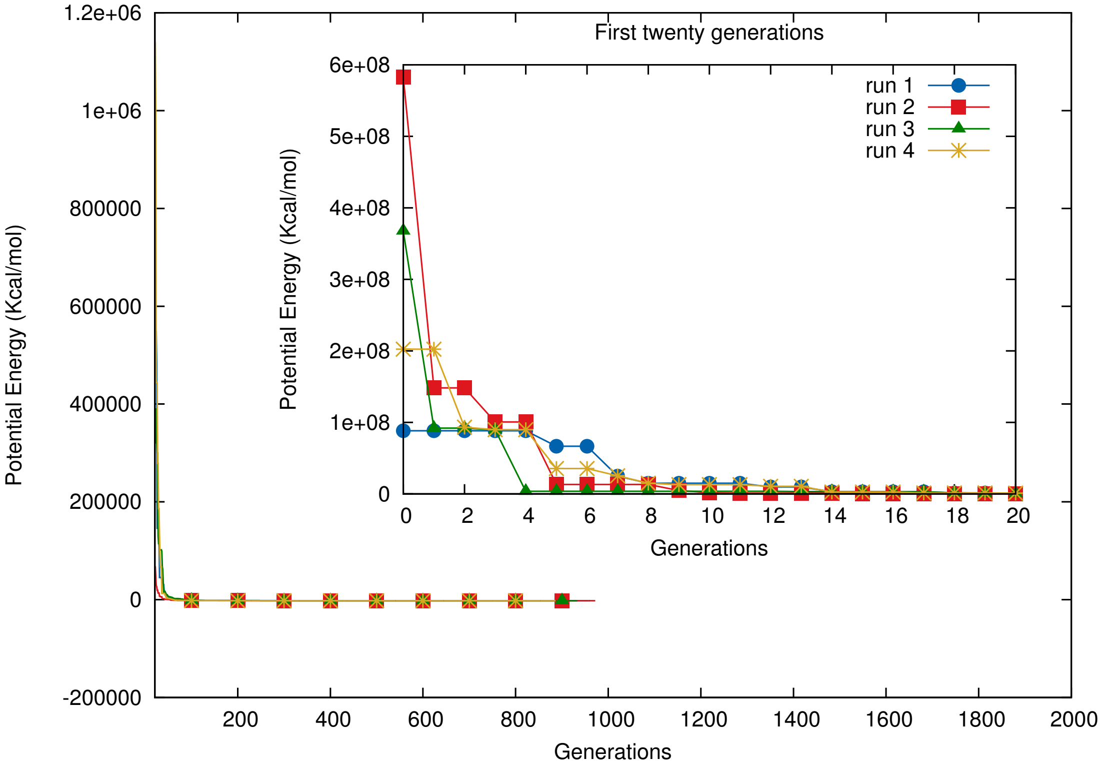


Figure I.3: Potential energy minimization for protein with PDB ID = 1BDC.

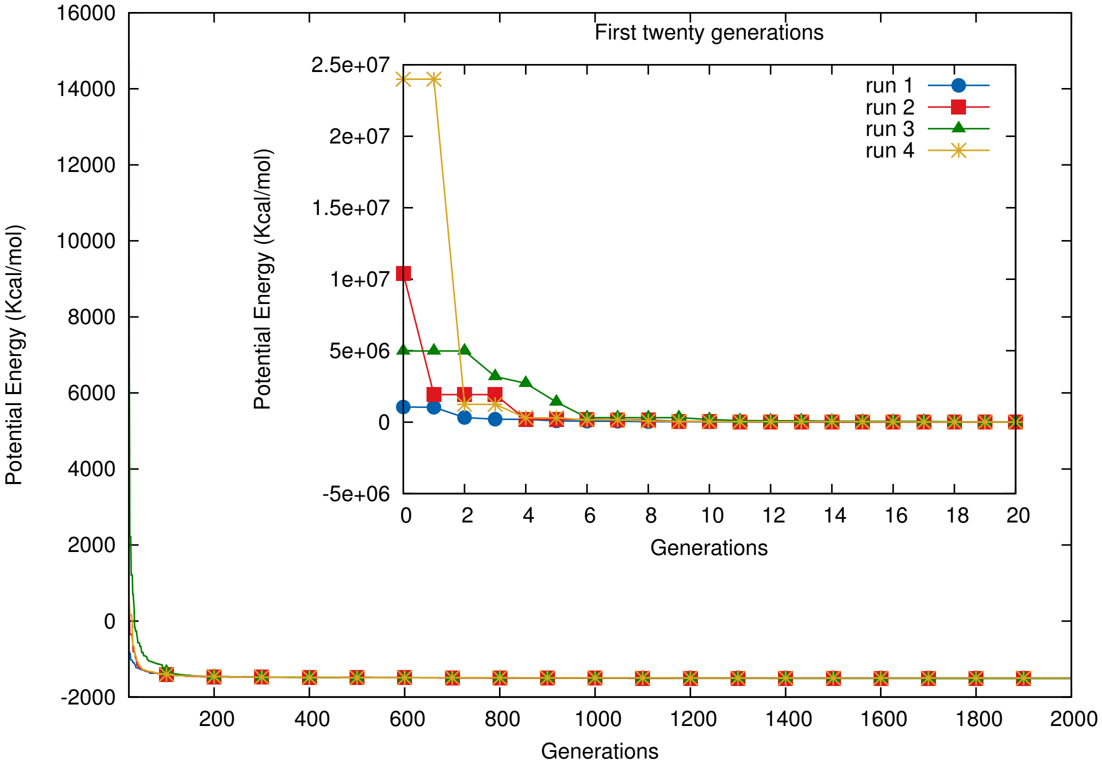


Figure I.4: Potential energy minimization for protein with PDB ID = 1BGK.

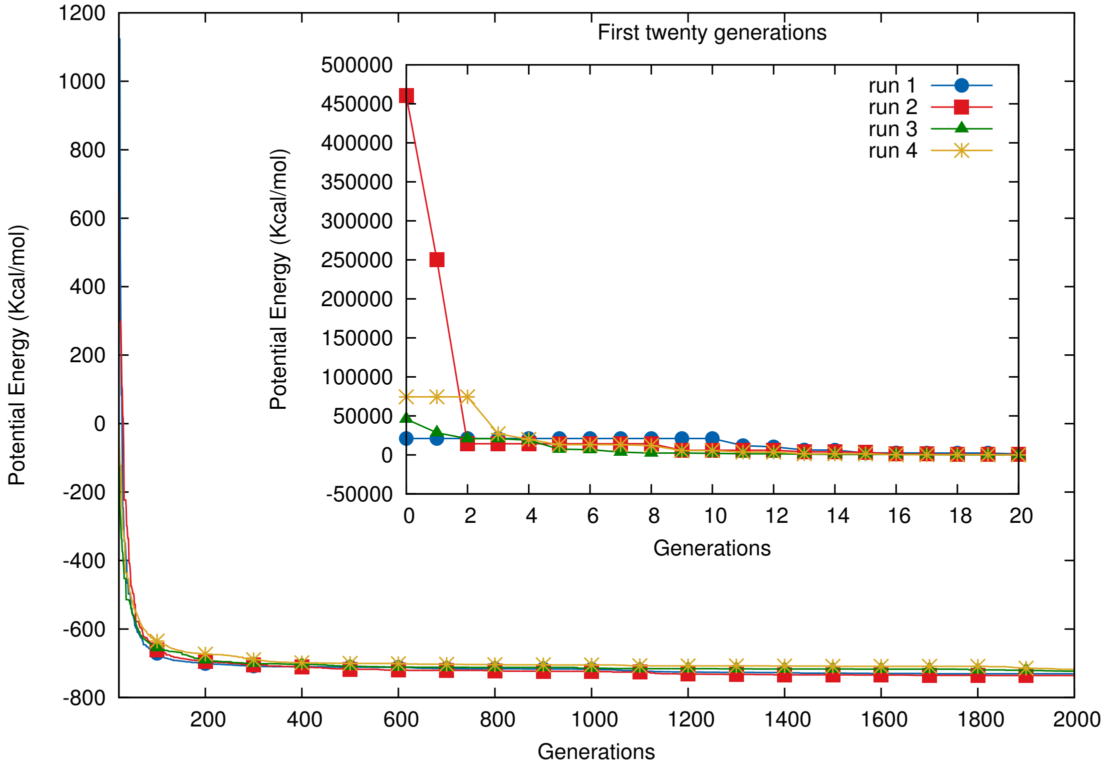


Figure I.5: Potential energy minimization for protein with PDB ID = 1BHI.

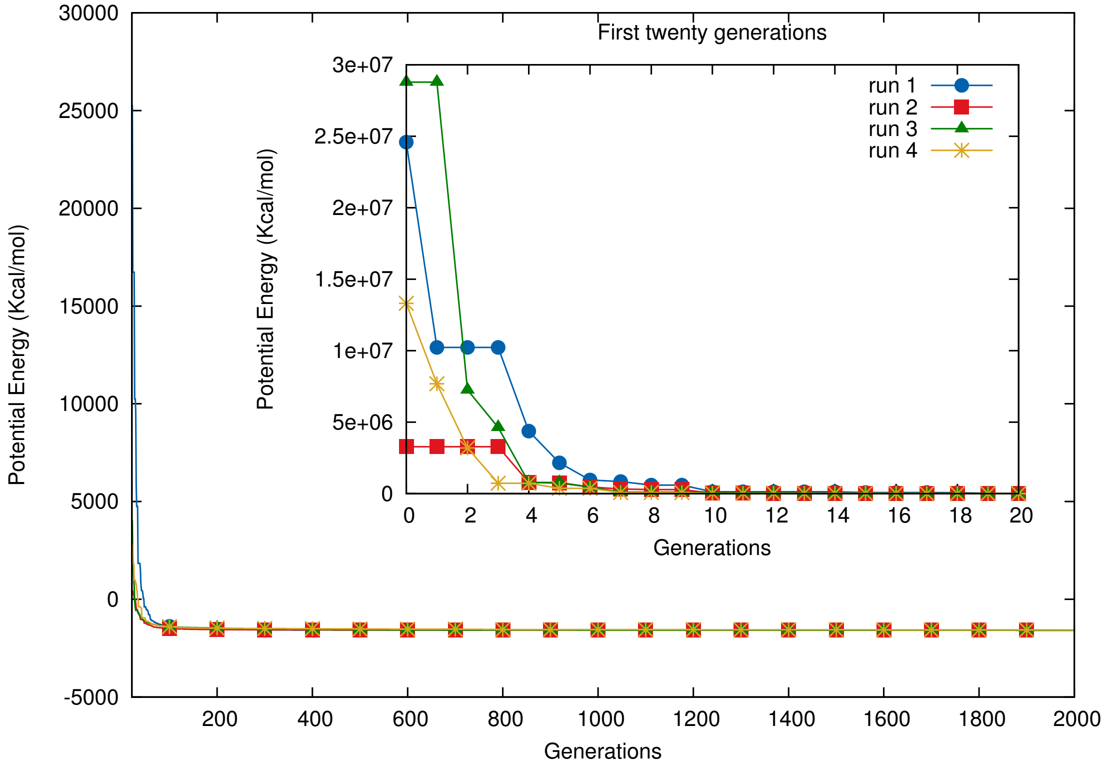


Figure I.6: Potential energy minimization for protein with PDB ID = 1DVO.

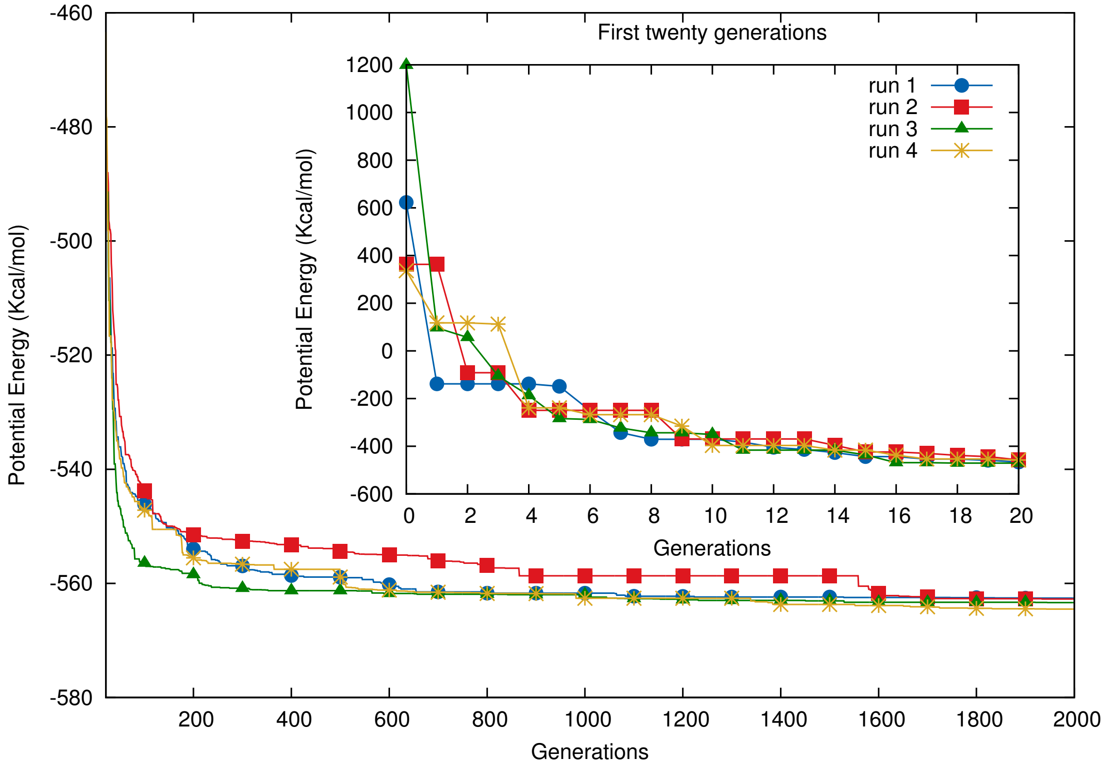


Figure I.7: Potential energy minimization for protein with `PDB ID` = `1EOQ`.

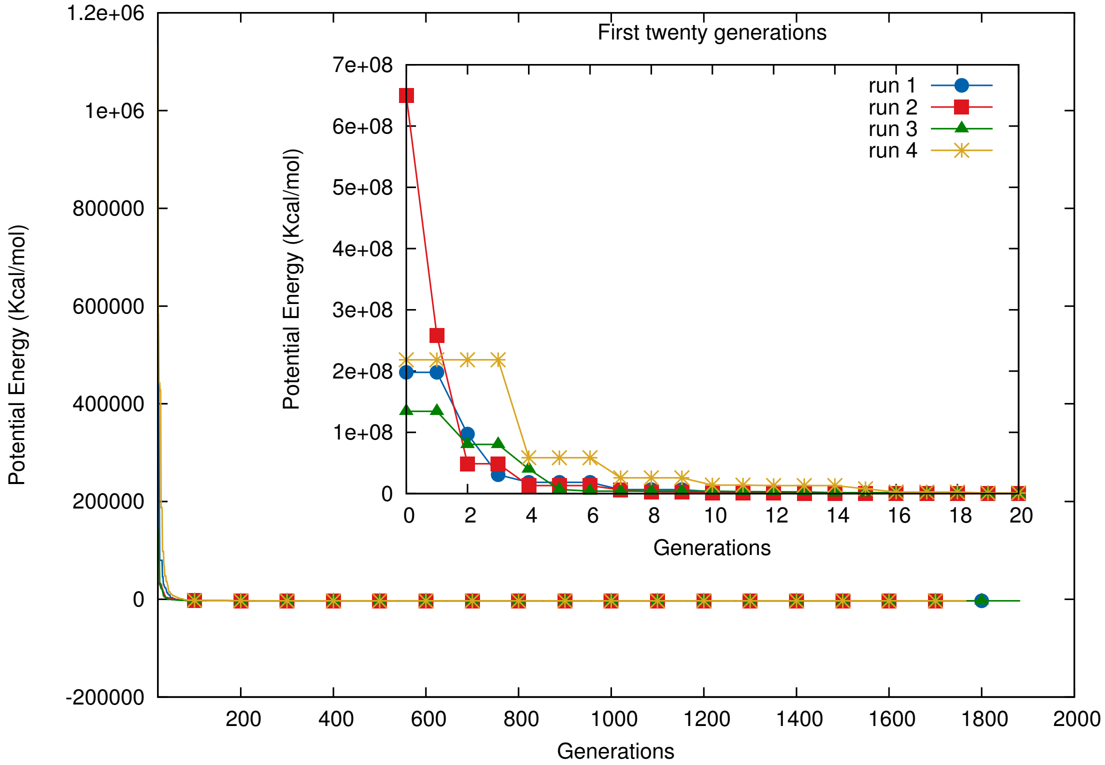


Figure I.8: Potential energy minimization for protein with `PDB ID` = `1ENH`.

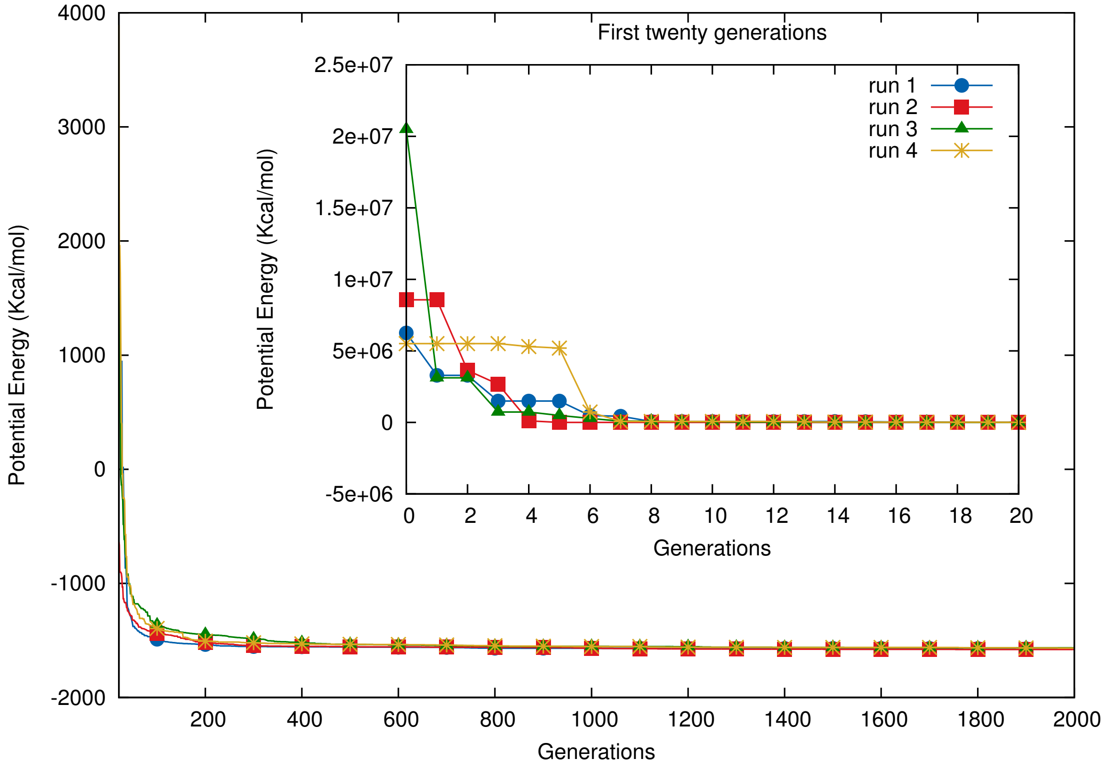


Figure I.9: Potential energy minimization for protein with PDB ID = 1FME.

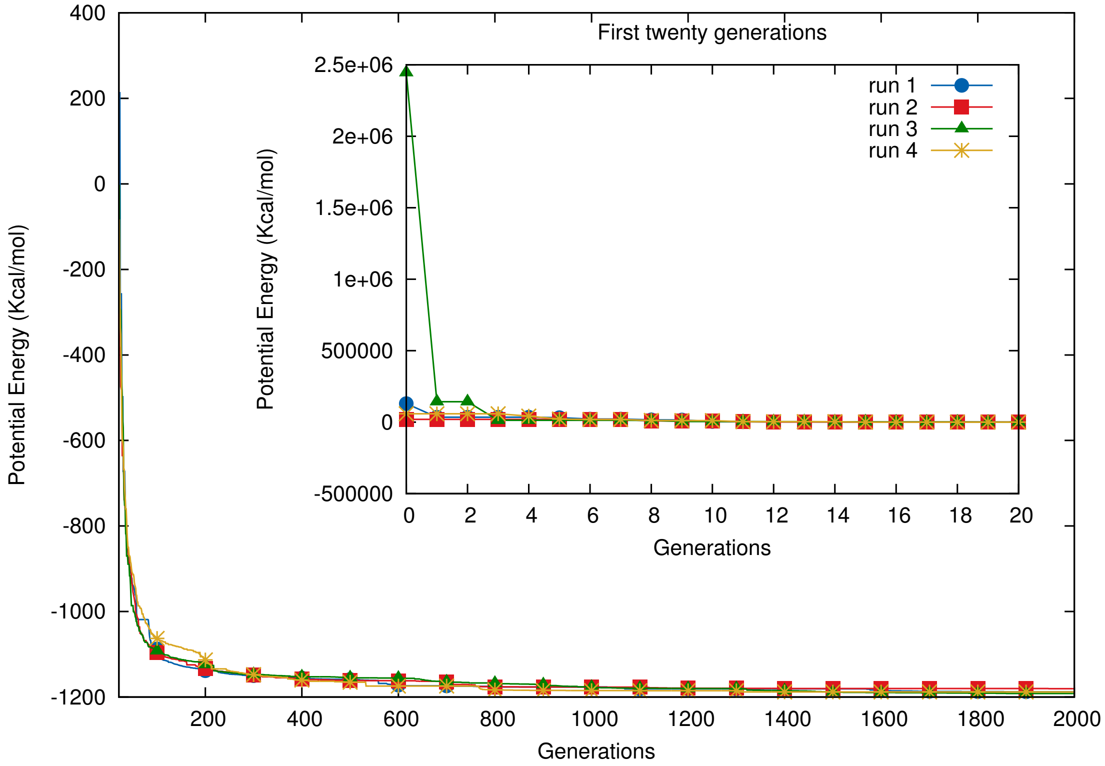


Figure I.10: Potential energy minimization for protein with PDB ID = 1OVX.

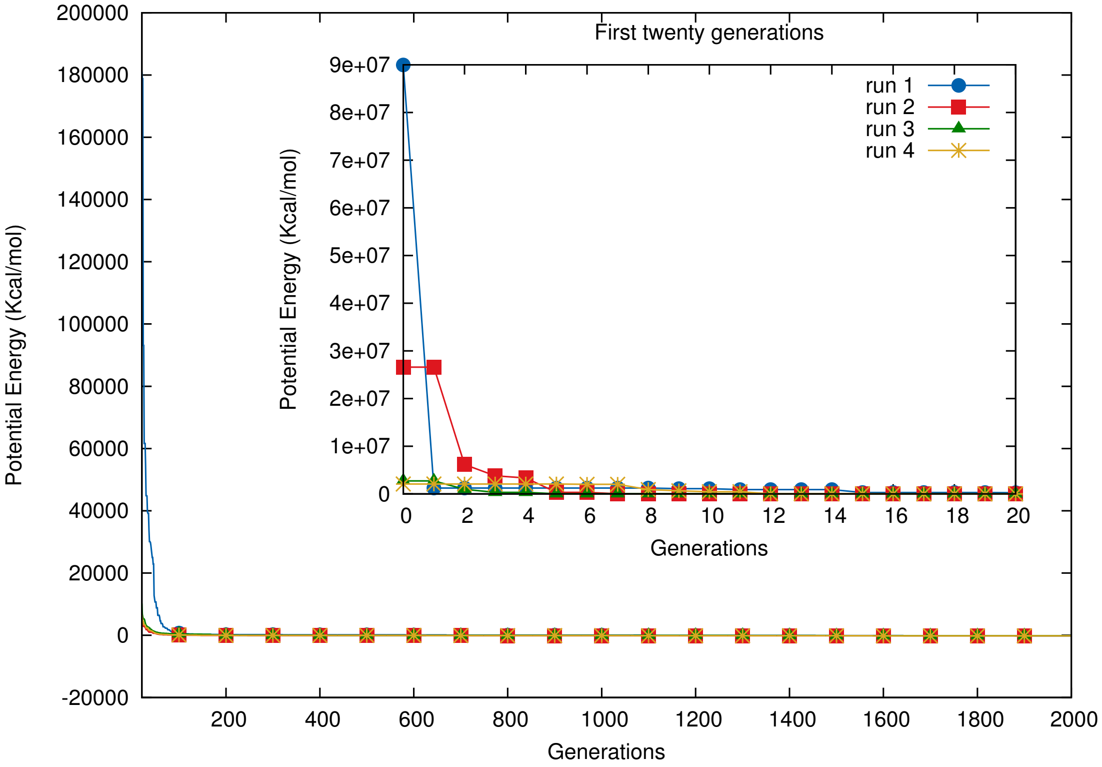


Figure I.11: Potential energy minimization for protein with PDB ID = 1Q2K.

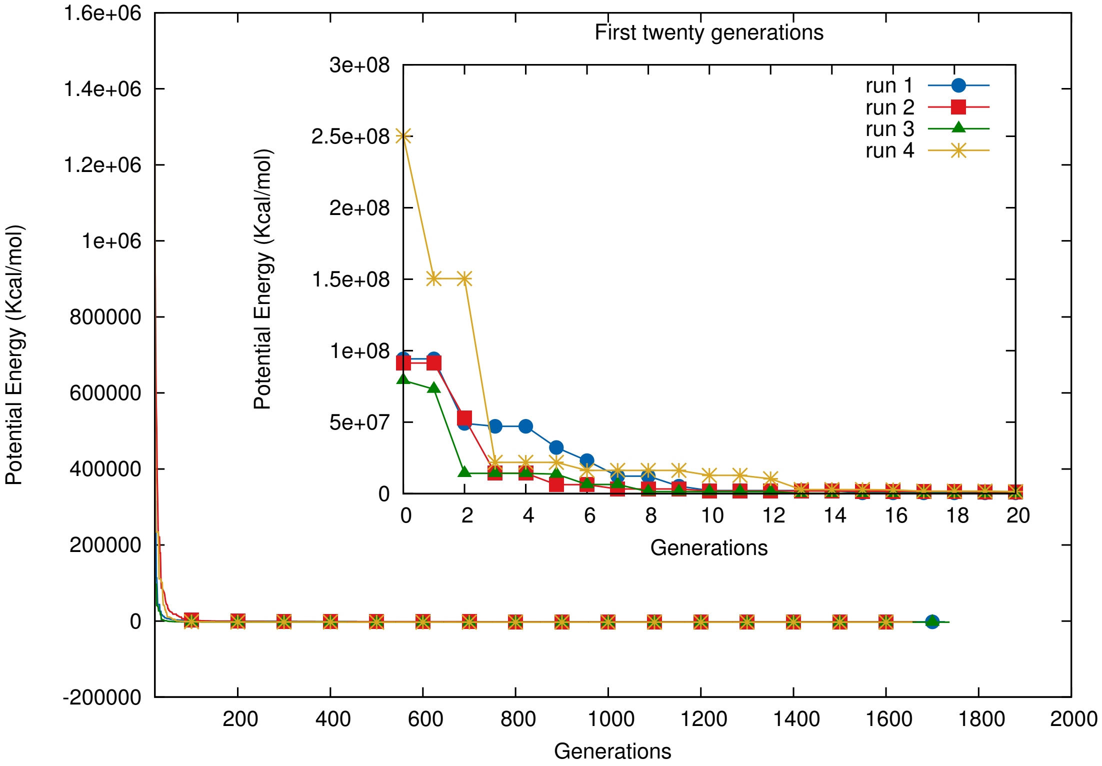


Figure I.12: Potential energy minimization for protein with PDB ID = 1QR8.

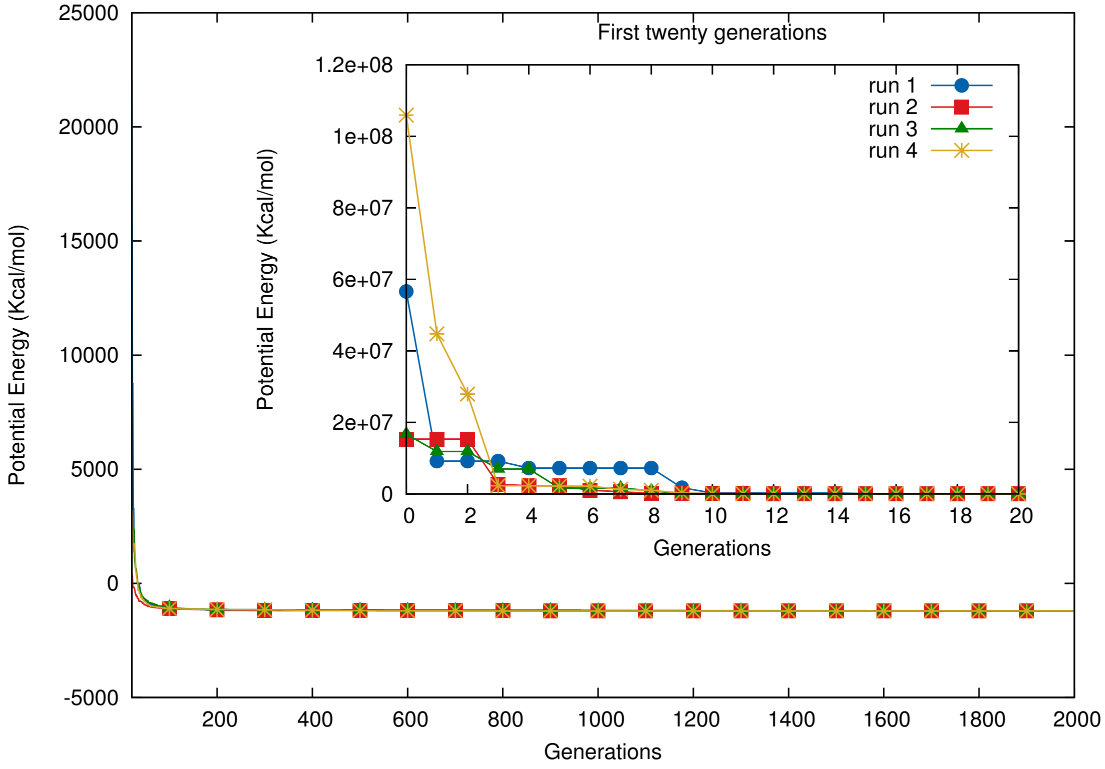


Figure I.13: Potential energy minimization for protein with `PDB ID` = `1ROO`.

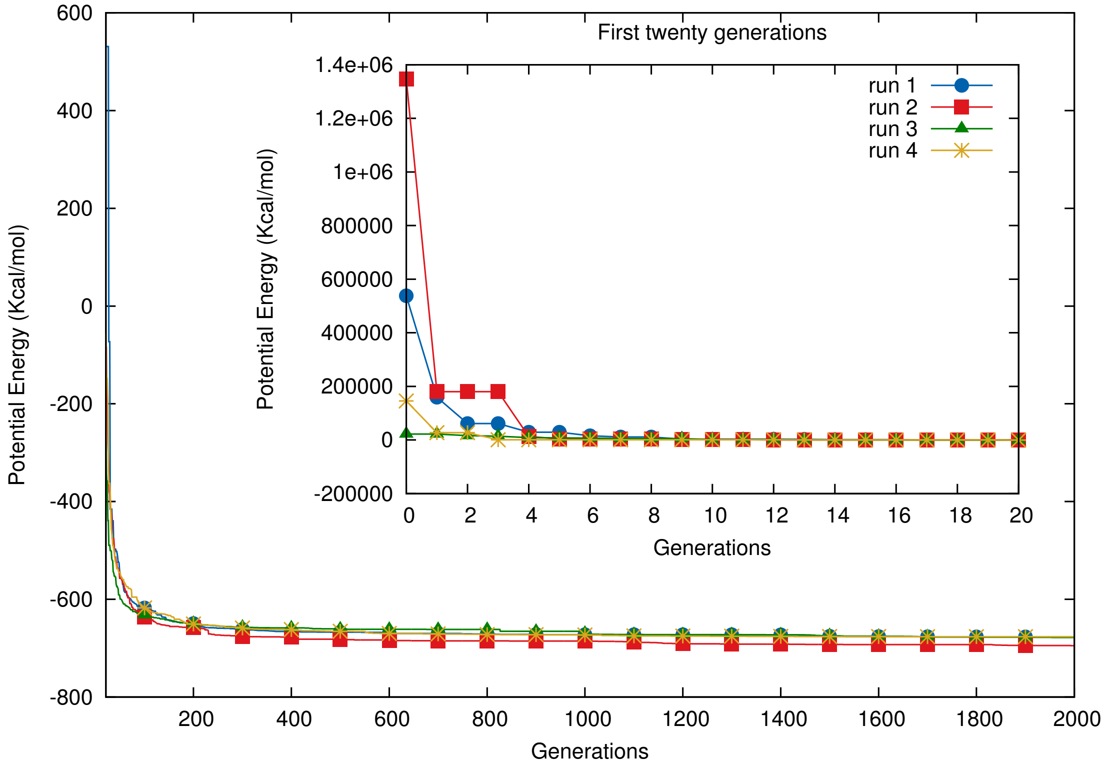
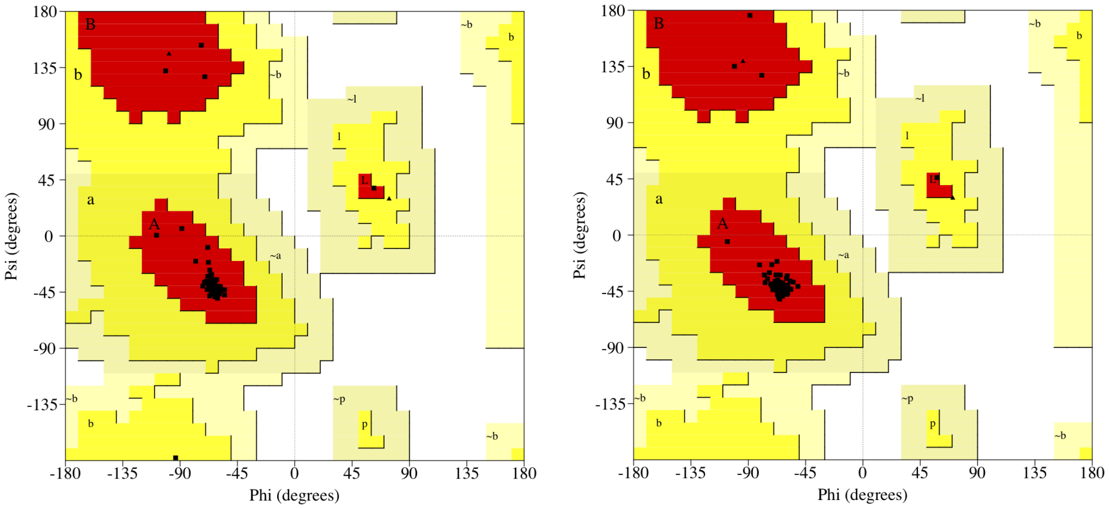


Figure I.14: Potential energy minimization for protein with `PDB ID` = `1WQC`.

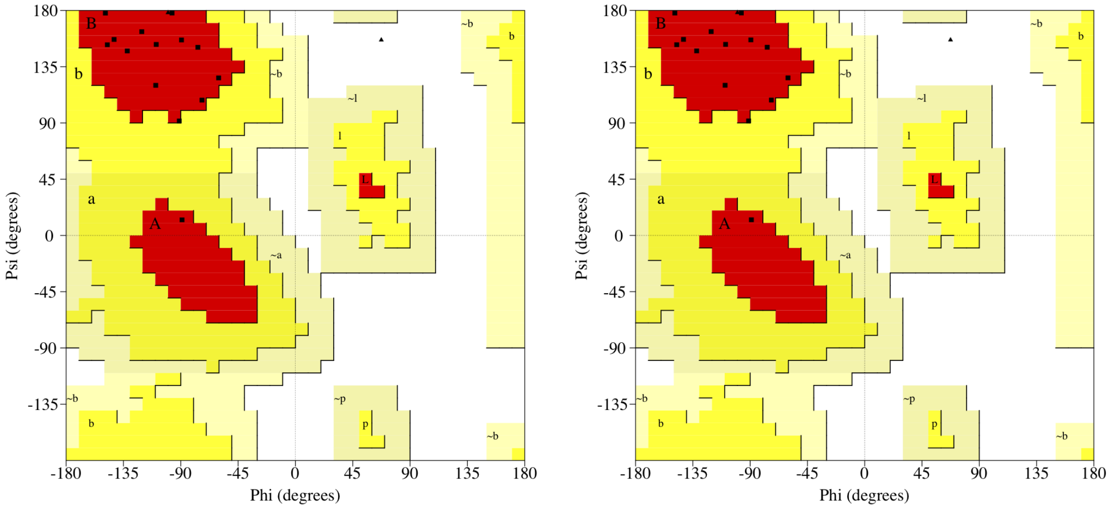
APPENDIX J EXPERIMENTS: RAMACHANDRAN PLOTS



(a) Experimental

(b) Predicted

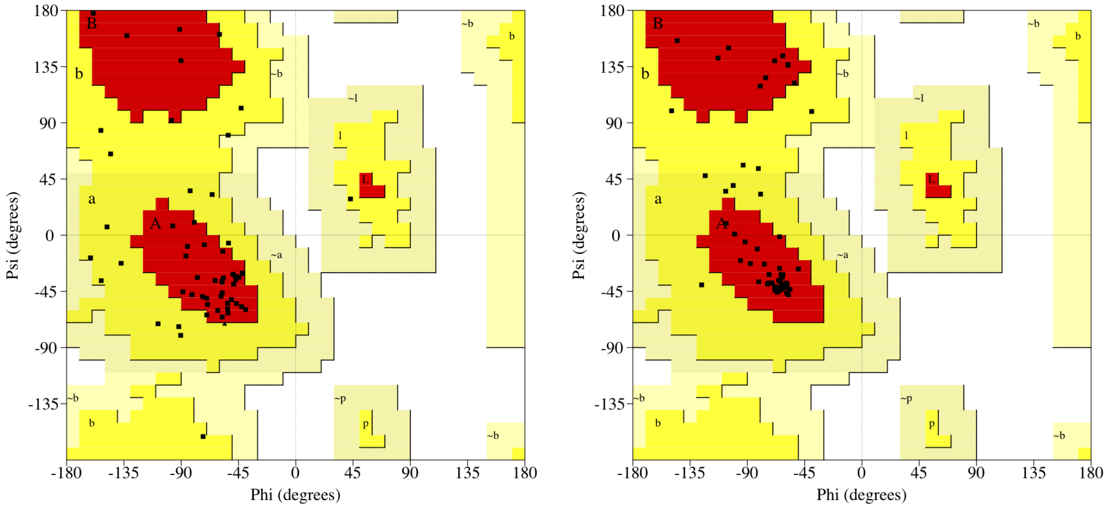
Figure J.1: `Ramachandran` plot of the experimental and predicted structures. (a) `Ramachandran` plot of the experimental protein with `PDB ID 1AIL`. (b) `Ramachandran` plot of the predicted `3-D` structure of the protein with `PDB ID 1AIL`.



(a) Experimental

(b) Predicted

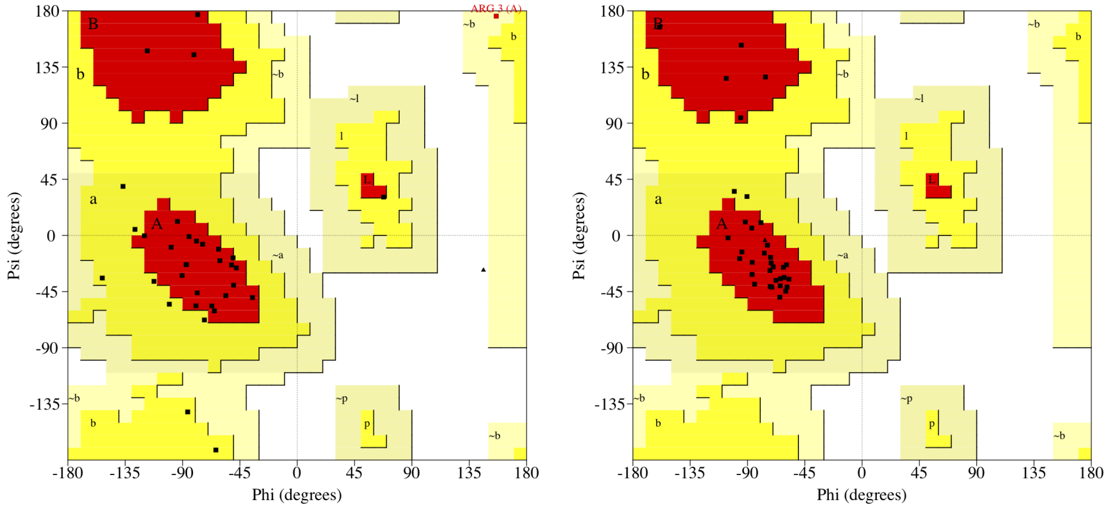
Figure J.2: `Ramachandran` plot of the experimental and predicted structures. (a) `Ramachandran` plot of the experimental protein with `PDB ID 1B03`. (b) `Ramachandran` plot of the predicted `3-D` structure of the protein with `PDB ID 1B03`.



(a) Experimental

(b) Predicted

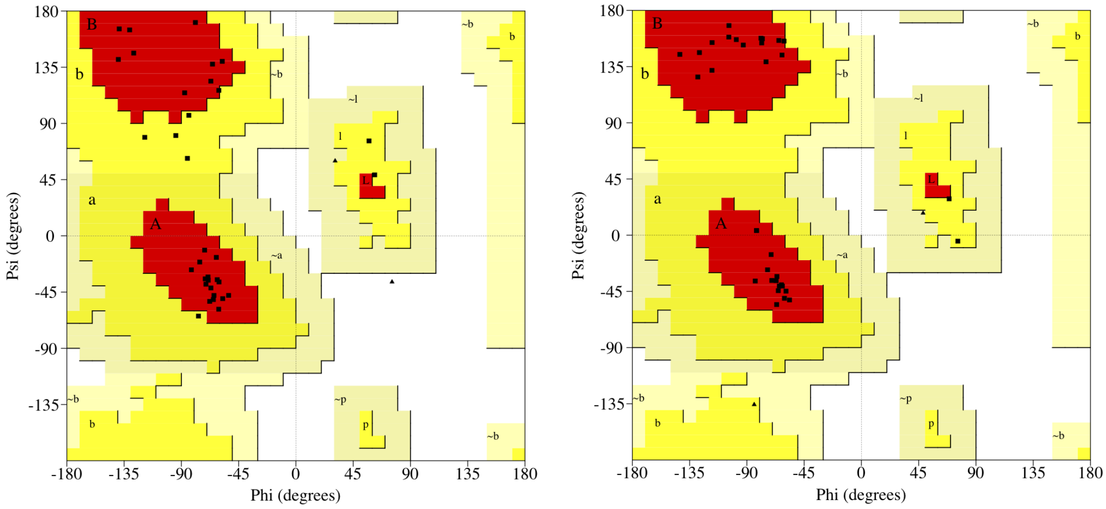
Figure J.3: Ramachandran plot of the experimental and predicted structures. (a) Ramachandran plot of the experimental protein with PDB ID 1BDC. (b) Ramachandran plot of the predicted 3-D structure of the protein with PDB ID 1BDC.



(a) Experimental

(b) Predicted

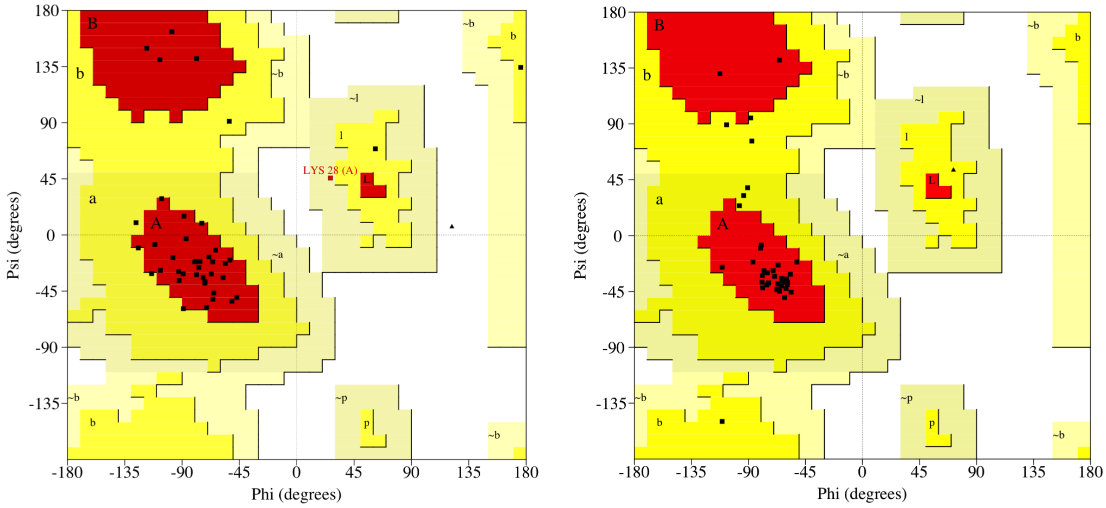
Figure J.4: Ramachandran plot of the experimental and predicted structures. (a) Ramachandran plot of the experimental protein with PDB ID 1BGK. (b) Ramachandran plot of the predicted 3-D structure of the protein with PDB ID 1BGK.



(a) Experimental

(b) Predicted

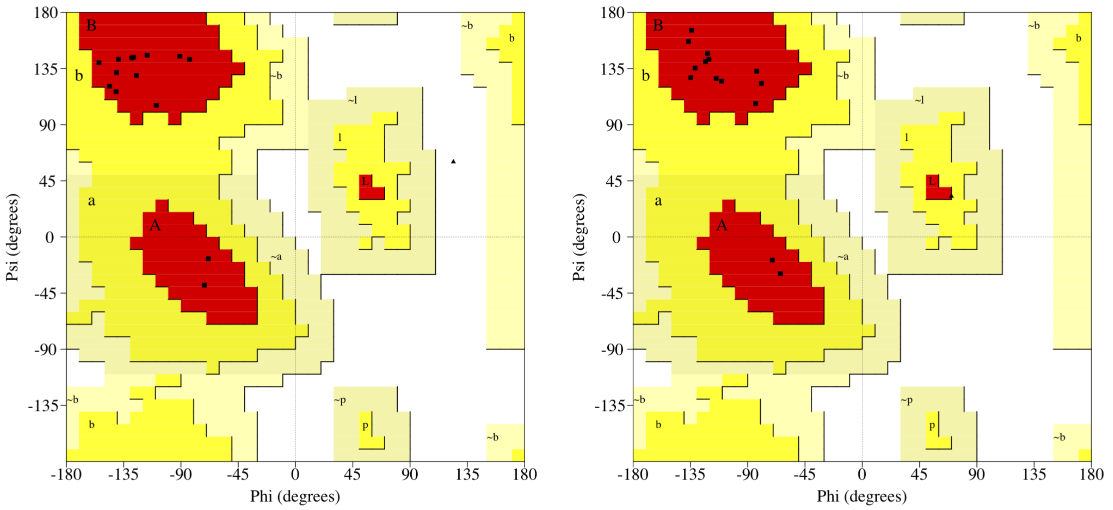
Figure J.5: Ramachandran plot of the experimental and predicted structures. (a) Ramachandran plot of the experimental protein with PDB ID 1BHI. (b) Ramachandran plot of the predicted 3-D structure of the protein with PDB ID 1BHI.



(a) Experimental

(b) Predicted

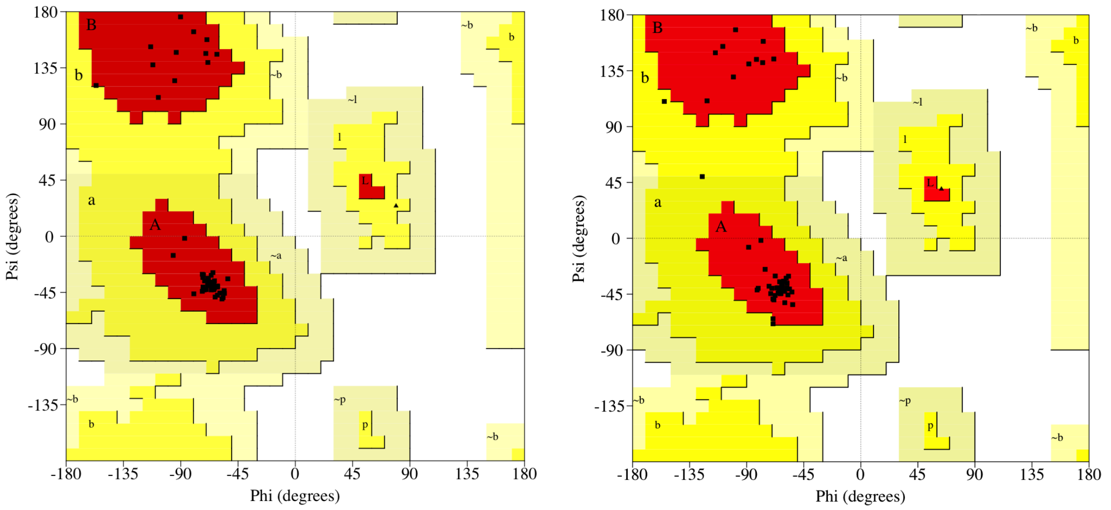
Figure J.6: Ramachandran plot of the experimental and predicted structures. (a) Ramachandran plot of the experimental protein with PDB ID 1DVO. (b) Ramachandran plot of the predicted 3-D structure of the protein with PDB ID 1DVO.



(a) Experimental

(b) Predicted

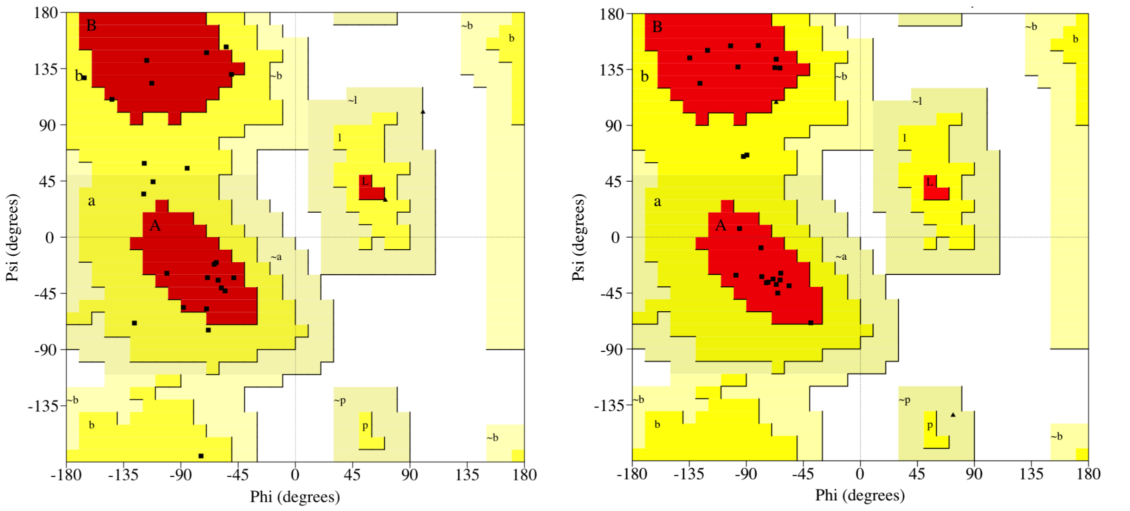
Figure J.7: `Ramachandran` plot of the experimental and predicted structures. (a) `Ramachandran` plot of the experimental protein with `PDB ID 1EOQ`. (b) `Ramachandran` plot of the predicted `3-D` structure of the protein with `PDB ID 1EOQ`.



(a) Experimental

(b) Predicted

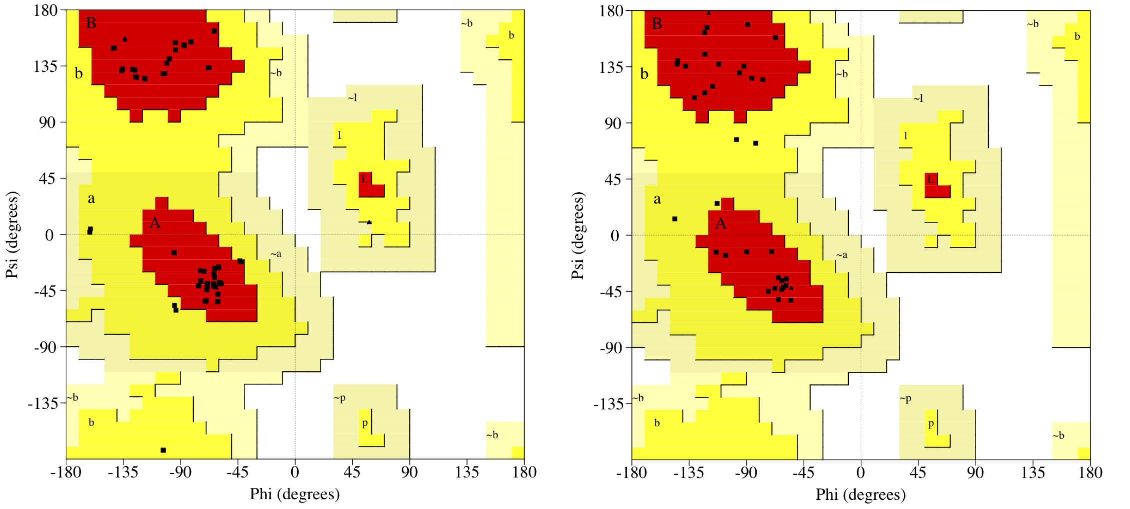
Figure J.8: `Ramachandran` plot of the experimental and predicted structures. (a) `Ramachandran` plot of the experimental protein with `PDB ID 1ENNH`. (b) `Ramachandran` plot of the predicted `3-D` structure of the protein with `PDB ID 1ENH`.



(a) Experimental

(b) Predicted

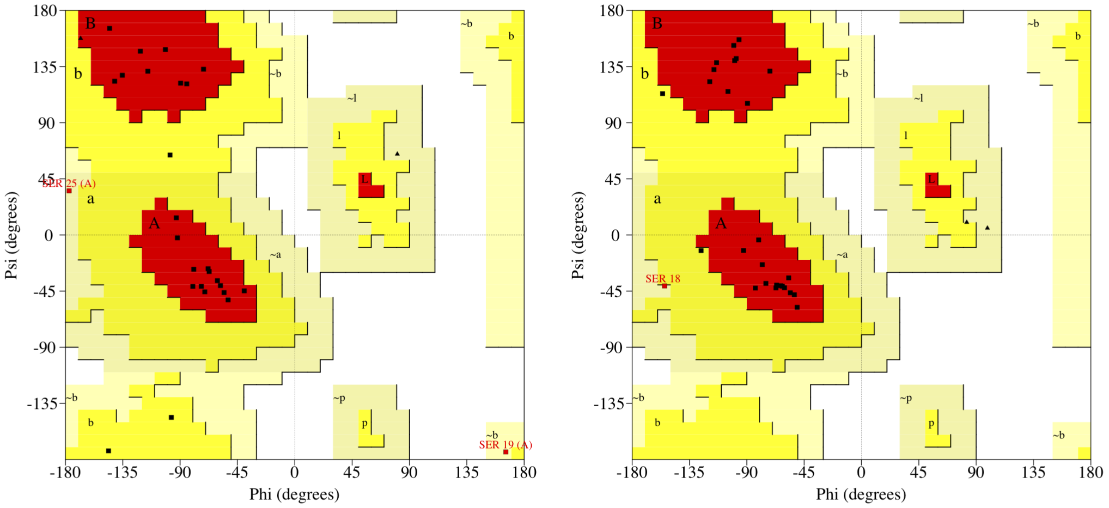
Figure J.9: `Ramachandran` plot of the experimental and predicted structures. (a) `Ramachandran` plot of the experimental protein with `PDB ID 1FME`. (b) `Ramachandran` plot of the predicted `3-D` structure of the protein with `PDB ID 1FME`.



(a) Experimental

(b) Predicted

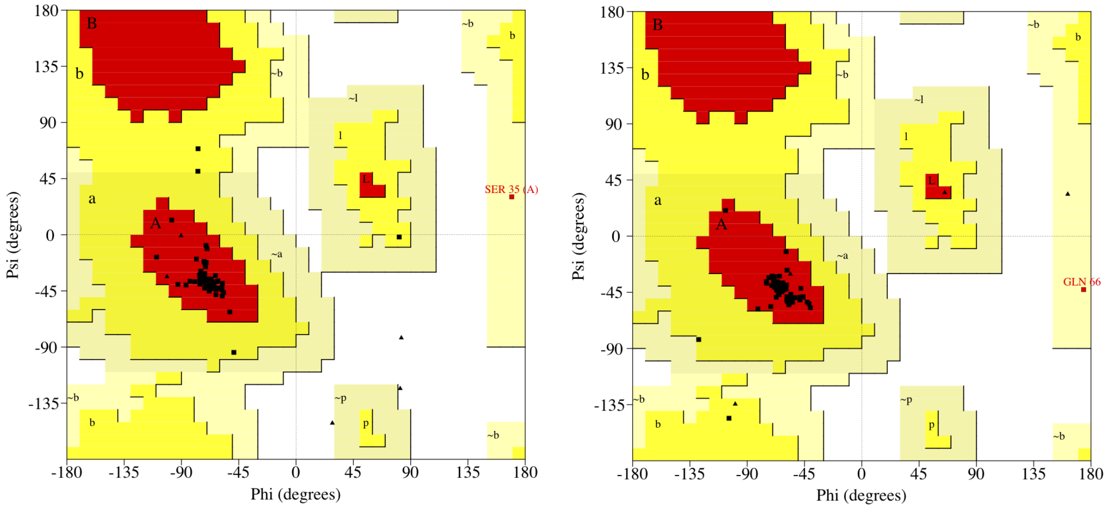
Figure J.10: `Ramachandran` plot of the experimental and predicted structures. (a) `Ramachandran` plot of the experimental protein with `PDB ID 1OVX`. (b) `Ramachandran` plot of the predicted `3-D` structure of the protein with `PDB ID 1OVX`.



(a) Experimental

(b) Predicted

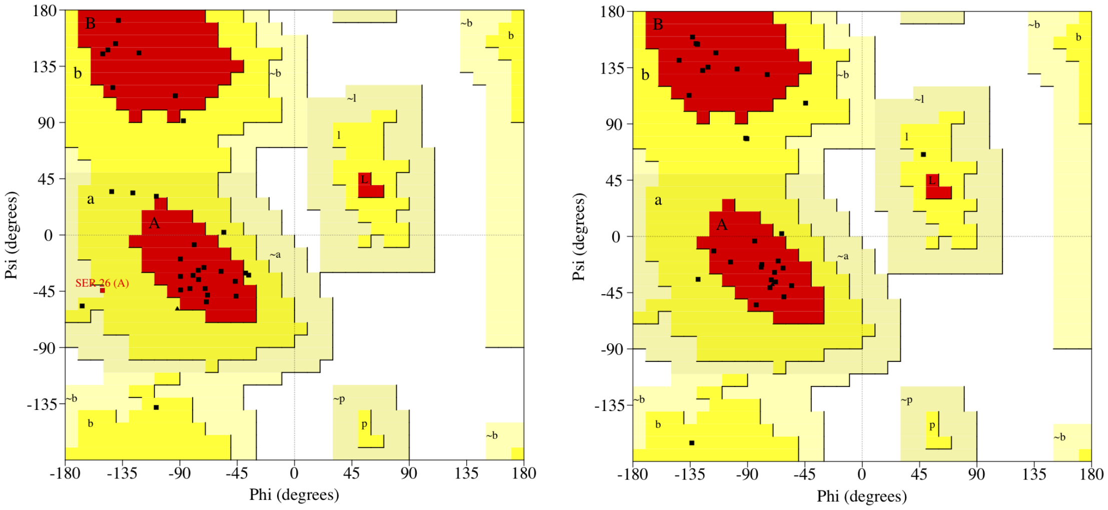
Figure J.11: `Ramachandran` plot of the experimental and predicted structures. (a) `Ramachandran` plot of the experimental protein with `PDB ID 1Q2k`. (b) `Ramachandran` plot of the predicted `3-D` structure of the protein with `PDB ID 1Q2K`.



(a) Experimental

(b) Predicted

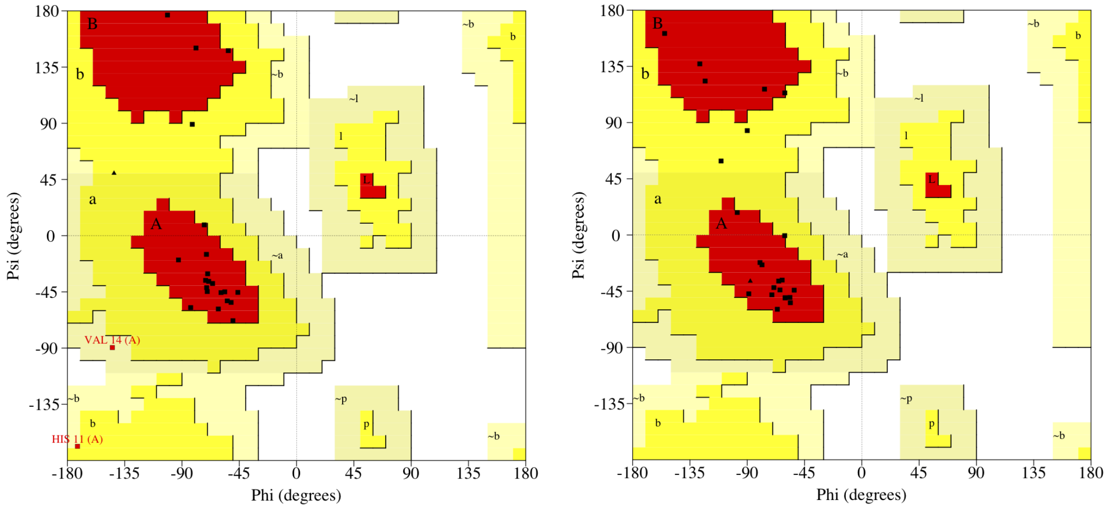
Figure J.12: `Ramachandran` plot of the experimental and predicted structures. (a) `Ramachandran` plot of the experimental protein with `PDB ID 1QR8`. (b) `Ramachandran` plot of the predicted `3-D` structure of the protein with `PDB ID 1QR8`.



(a) Experimental

(b) Predicted

Figure J.13: `Ramachandran` plot of the experimental and predicted structures. (a) `Ramachandran` plot of the experimental protein with `PDB ID 1ROO`. (b) `Ramachandran` plot of the predicted `3-D` structure of the protein with `PDB ID 1ROO`.



(a) Experimental

(b) Predicted

Figure J.14: `Ramachandran` plot of the experimental and predicted structures. (A) `Ramachandran` plot of the experimental protein with `PDB ID 1WQC`. (B) `Ramachandran` plot of the predicted `3-D` structure of the protein with `PDB ID 1WQC`.

APPENDIX K RESUMO ESTENDIDO

K.1 Desenvolvimento da Pesquisa

Atualmente, um dos mais importantes problemas de pesquisa na área da Bioinformática Estrutural trata da predição da estrutura tridimensional (3D) de proteínas. O conhecimento a respeito desta estrutura tridimensional nos permite investigar os processos biológicos de forma mais direta e em detalhes. Proteínas ou Polipeptídeos são polímeros constituídos de 20 diferentes tipos de resíduos de aminoácidos. Cada proteína é definida pela sua sequência única de resíduos de aminoácidos que em condições fisiológicas adequadas se enovela assumindo uma forma específica conhecida como estado nativo da proteína (ANFINSEN, 1973).

Cada resíduo de aminoácido é constituído por um carbono α (C_α) ligado covalentemente a um grupo amino (NH), a um grupo carboxílico (COOH) e a uma cadeira lateral (R) que representa as propriedades fisico-químicas específicas de cada resíduo de aminoácido. Um peptídeo é uma molécula composta por dois ou mais resíduos de aminoácidos ligados através de uma ligação peptídica, chamada de ligação peptídica. Esta ligação é formada quando o grupo carboxílico de um resíduo de aminoácido reage com o grupo amino de outro resíduo de aminoácido ocorrendo a liberação de uma molécula de água (H_2O) (LEHNINGER; NELSON; COX, 2005; TRAMONTANO, 2006; LESK, 2010; LILJAS et al., 2001). O conjunto de átomos formados pelo carbono α, oxigênio e nitrogênio formam a cadeia principal da proteína. Dois ou mais resíduos de aminoácidos ligados por meio de uma ligação peptídica são conhecidos como peptídeos e grandes peptídeos são geralmente conhecidos como polipeptídeos e proteínas (CREIGHTON, 1990; LESK, 2002).

Os projetos GENOMA, iniciados na década de 90, resultaram em um grande aumento no número de sequências de proteínas. Infelizmente, o número de estruturas tridimensionais de proteínas não cresceram no mesmo ritmo. Atualmente, o número de sequências de proteínas é muito maior que o número de estruturas tridimensionais conhecidas. Ao compararmos o número de sequências de proteínas que não são redundantes e estão armazenadas no GenBank, com o número de estruturas tridimensionais com enovelamentos distintos e armazenadas no *Protein Data Bank*[1] (PDB) (BERMAN et al., 2000), podemos observar uma grande lacuna entre o número de sequências de proteínas que podemos gerar e o número de novos enovelamentos que podemos determinar através de métodos experimentais tais como: difração de raio X e Ressonância Magnética Nuclear (NMR, sigla em inglês).

O processo experimental utilizado para determinar a estrutura tridimensional

[1] www.rcsb.org/pdb

de uma proteína é caro (devido aos custos associados à cristalografia de raio X e microscopia eletrônica (NMR)) e também demorado. Uma forma barata, eficiente e eficaz para determinar de forma rápida a estrutura tridimensional de proteínas poderia beneficiar muitos campos de pesquisa como a medicina, a biotecnologia e a indústria farmacêutica. A predição da estrutura 3D de proteínas é atualmente um dos maiores problemas investigados pela Bioinformática Estrutural (TRAMONTANO, 2006; ZHANG; VERETNIK; BOURNE, 2005).

Descobrir o enovelamento de uma proteínas somente a partir de sua sequência linear de resíduos de aminoácidos é também um grande desafio para a área da otimização e para a matemática (LANDER; WATERMAN, 1999). Este problema é classificado na área de complexidade algoritmica como um problema NP-completo (CRESCENZI et al., 1998). Os principais desafios estão relacionados com a explosão no número de possíveis formas que a estrutura da proteína pode assumir. Uma longa cadeia polipeptídica pode assumir um imenso número de estados conformacionais.

Ao longo dos últimos anos diversas estratégias computacionais foram apresentadas como uma solução do problema da predição da estrutura tridimensional de proteínas (WOOLEY; YE, 2010). Estes métodos podem ser divididos em duas classes (FLOUDAS et al., 2006): (I) Métodos de primeiros princípios que não utilizam informações da base experimental (OSGUTHORPE, 2000); (II) Métodos de primeiros princípios que utilizam informações da base experimental (ROHL et al., 2004; SRINIVASAN; ROSE, 1995); (III) Métodos de reconhecimento de enovelamentos (BOWIE; LUTHY; EISENBERG, 1991; JONES; TAYLOR; THORNTON, 1992; BRYANT; ALTSCHUL, 1995); e (IV) Métodos baseados em modelagem comparativa de sequências (MARTÍ-RENOM et al., 2000; SÁNCHEZ; SALI, 1997). O primeiro grupo, que não se utiliza informações de proteínas com estruturas conhecidas, tem como objetivo predizer novas formas de enovelamento somente por meio da simulação computacional dos fenômenos fisico-químicos relacionados ao processo de enovelamento da proteínas tal qual ocorre na natureza. Esta classe de métodos utiliza o conceito de energia livre (hipótese de Anfinsen) para encontrar o estado nativo da proteína (ANFINSEN et al., 1961; ANFINSEN, 1973).

Os grupos II, III e IV representam os métodos de predição que conseguem realizar de forma rápida e eficiente a predição da estutura 3D de proteína quando modelos estruturais de proteínas com estruturas conhecidas e bibliotecas de enovelamentos são utilizados (KOLINSKI, 2004). Nos métodos de primeiros princípios que utilizam informações da base experimental, regras são extraídas de proteínas com estrutura conhecida e então utilizadas para construir novas conformações. ROBETTA (ROHL et al., 2004; SIMONS et al., 1999B), I-TASSER (ZHANG, 2007) e LINUS (SRINIVASAN; ROSE, 1995) são exemplos de métodos pertencentes a este grupo. Métodos baseados em análise comparativa por homologia podem ser utilizados sempre que for possível detectar uma relação evolucionária entre a sequência alvo e a sequência da proteína modelo, cuja estrutura 3D é conhecida (SÁNCHEZ; SALI, 1997). A estrutura destas proteínas são similares no sentido que resíduos de aminoácidos com propriedades fisico-químicas e estruturas idênticas ocupam as mesmas posições em proteínas homólogas. Os métodos de reconhecimento de enovelamentos são motivados pela noção de que estruturas são mais estáveis que sequências, isto é, proteínas com sequências diferentes podem ter enovelamentos similares. Métodos de reconhecimento de enovelamentos via alinhavamento estão limitados a biblioteca de enovelamentos derivados do PDB.

O progresso mais significativo observado na última competição de métodos de predição CASP[2] (9°edição) foi o dos métodos de primeiros princípios que utilizam informações da base experimental (KOOP et al., 2007; COZZETTO et al., 2009; ZHANG, 2008B; XU et al., 2011). Entretanto, conforme revelado pelos experimentos, os maiores desafios para o desenvolvimento de melhores métodos de predição estão focados no desenvolvimento de novas estratégias computacionais para produção, identificação e utilização de modelos estruturais na base experimental (SODING, 2005). No último CASP9 não observou-se progressos nos métodos de primeiros princípios que não utilizam informações da base experimental (JAUCH et al., 2007; BEN-DAVID et al., 2009; FLOUDAS et al., 2006; XU et al., 2011).

Apesar do significante progresso dos métodos de predição, é ainda necessário o desenvolvimento de novas estratégias para extração, representação e manipulação de informações estruturais de estruturas 3D determinadas experimentalmente, bem como, o desenvolvimento de novas estratégias computacionais que utilizem estas informações para realizar a predição, unicamente a partir da sequência de resíduos de aminoácidos, a sua estrutura 3D correspondente. O desenvolvimento de métodos computacionais que reduzem o esforço computacional e permitem a predição da estrutura 3D de proteínas é representado como um dos maiores desafios do século XXI no campo da Bioinformática Estrutural e da Biologia Molecular.

Neste tese uma nova estratégia computacional para a predição da estrutura tridimensional de proteínas foi proposta (MOIRAE), implementada e testada. Trata-se de uma estratégia baseada em primeiros princípios que utiliza informações estruturais da base experimental. A técnica proposta manipula informações estruturais do PDB com o propósito de gerar intervalos de ângulos de torção da cadeia principal da proteína alvo. Estes ângulos de torção são utilizados como entradas para uma estratégia de busca baseada em algoritmos genéticos. A estratégia de busca desenvolvida utiliza utiliza um operador de busca local como forma a acelerar a busca pela estrutura nativa da proteína alvo. Pode-se listar as seguintes principais contribuições deste trabalho:

- O desenvolvimento de uma nova estratégia computacional para coletar e representar informações estruturais de estruturas de proteínas determinadas experimentalmente;

- O desenvolvimento de uma estratégia de busca baseada em algoritmos genéticos com um operador de busca local para percorrer o espaço de busca conformacional buscando encontrar a estrutura nativa de proteínas;

- O desenvolvimento de uma estratégia computacional baseada em fragmentos e a combinação desta com conceitos de métodos *ab initio* para realizar a predição da estrutura 3D de proteínas;

- A participação em eventos e a interação com grupos de pesquisas nacionais e internacionais relacionados ao tema; e

- A predição da estrutura 3D de 20 estudos de casos (diferentes proteínas);

A seguir estão listadas as publicações originadas deste trabalho de pesquisa:

[2] Critical Assessment of Structure Prediction. predictioncenter.org

- DORN, M.; BURIOL, L.S.; LAMB, L.C. "A hybrid genetic algorithm for the 3-D protein structure prediction problem using a path-relinking strategy". IEEE Congress on Evolutionary Computation - CEC", New Orleans, 2011. p. 2691-2698.
  QUALIS CAPES: A1 (Ciência da Computação)

- DORN, M.; BURIOL, L.S.; LAMB, L.C. "Combining machine learning and optimization techniques to determine 3-D structures of polypeptides" IJCAI - Doctoral Mentoring Consortium, 2011, Barcelona.
  QUALIS CAPES: A1 (Ciência da Computação)

- GONÇALVES, W.W.; DORN, M.; BURIOL, L.S.; LAMB, L.C. "A Structured-Population Genetic Algorithm for the 3-D Protein Structure Prediction Problem". Brazilian Symposium on Bioinformatics, Brasilia, 2011. v. 1. p. 17-24.
  QUALIS CAPES: B2 (Ciência da Computação)

- ANDRADES, R.; DORN, M. ; FARENZENA, D.S.; LAMB, L.C. "Aplicação de Técnicas de Inteligência Artificial e Mineração de Dados no Design de Proteínas". XXIII Salão de Iniciação Científica UFRGS, 2011, Porto Alegre.

- TABAJARA, L.M. ; FARENZENA, D.S.; DORN, M.; LAMB, L.C. "Resolução de problemas através de computação humana utilizando redes sociais". XXIII Salão de Iniciação Científica UFRGS, 2011, Porto Alegre.

- DORN, M.; LAMB, L.C.; BURIOL, L.S. "An artificial neural network based method for the prediction of approximated 3-D structures of mini-globular proteins". 6th International Conference of Brazilian Association for Bioinformatics and Computational Biology, 2010, Ouro Preto.

Os seguintes artigos encontram-se atualmente em avaliação:

- DORN, M.; BURIOL, L.S.; LAMB, L.C. "A Molecular Dynamics and Knowledge-based Computational Strategy to Predict Native-like Structures of Polypeptides". Expert Systems with Applications, Elsevier, 2012.
  IMPACT FACTOR: 2.029
  QUALIS CAPES: A1 (Ciência da Computação) B2 (Biologia)

- DORN, M.; BURIOL, L.S.; LAMB, L.C. "Protein Tertiary Structure Prediction: Methods and Computational Strategies" Chemical Reviews, ACS, 2012.
  IMPACT FACTOR: 33.036
  QUALIS CAPES: A1 (Biologia) A1 (Química)

- DORN, M.; ANDRADES, R.; FARENZENA, D.S.; LAMB, L.C. "A Novel Cluster-DEE-based Strategy to Empower Protein Design" Artificial Intelligence in Medicine, Elsevier, 2012.
  IMPACT FACTOR: 1.568
  QUALIS CAPES: A1 (Engenharia) A2 (Ciência da Computação)

K.1.1 Conclusões

O estudo de proteínas e a predição de estrutura 3D é um dos problemas de pesquisa mais importantes da Bioinformática Estrutural. Predizer a estrutura 3D de uma proteínas que não possuem modelos armazenados no PDB é uma tarefa extremamente difícil e em alguns casos impossível. Ao longo dos últimos anos, diversos métodos computacionais, algoritmos e sistemas foram desenvolvidos com o propósito de solucionar o problema da predição de estruturas de proteínas. Entretanto, o problema continua desafiando cientistas da computação, biólogos, matemáticos, químicos, físicos e bioinformatas por motivo da complexidade e da grande dimensionalidade do espaço de busca conformacional das proteínas. Esperimentalmente, a geração de sequencias de proteínas é considerada mais fácil que a determinação de sua estrutura 3D. Entretanto, o conhecimento da estrutura 3D de proteínas permite que pesquisadores tenham importantes informações sobre a função que a mesma desempenha na célula. A dificuldade em determinar a estrutura 3D de proteínas gerou uma grande discrepância entre o volume de sequencias de proteínas geradas por projetos Genoma e o número de estruturas 3D de proteínas que tem estruturas conhecidas. Isto não somente mostra claramente a necessidade, mas também motiva a realização de pesquisas futuras no desenvolvimento de estratégias computacionais para a predição da estrutura nativa de proteínas.

Analisando o progresso dos métodos de predição no CASP ao longo dos últimos anos, nós podemos observar que é ainda necessário o desenvolvimento de novas estratégias computacionais para extração, representação e manipulação de dados estruturais de estruturas 3D de proteínas determinadas experimentalmente, bem como o desenvolvimento de estratégias computacionais para utilizar esta informação na predição, a partir da sequência linear de aminoácidos, da estrutura nativa da proteína. Neste trabalho desenvolvemos uma estratégia computacional baseada em primeiros principios que utiliza informação da base experimental para realizar a predição da estrutura nativa de proteínas. MOIRAE manipula informações estruturais do PDB para criar intervalos de ângulos de torção para a cadeia principal da proteína. Conforme pôde ser observado pelos experimentos realizados, o uso desta estratégia reduz o espaço conformacional da proteína alvo. Os intervalos de ângulos de torção calculados na primeira fase do método MOIRAE são utilizados como entrada para uma estratégia de busca baseada em um algoritmo genético. A estratégia de busca desenvolvida proporciona um mecanismo eficiente para a predição da estrutura 3D de proteínas. Isto é obtido através do uso de um operador de busca local, o qual, faz com que o algoritmo genético possa escapar de mínimos locais.

Quando comparado a outros métodos de predição também classificados como de primeiro princípio e que utilizam informação da base experimental, MOIRAE apresenta vantagens em termos de tempo demandado para predizer a estrutura 3D de proteínas. ROBETTA obteve os melhores resultados no CASP ao longo dos últimos anos, entretanto, este método faz uso de plataformas computacionais de alto desempenho. Claramente, estes apresentam melhores resultados, entretanto, a estratégia proposta é uma nova idéia de estratégia computacional para predizer a estrutura 3D de proteínas. As principais contribuições deste trabalho são: o uso de técnicas computacionais para desenvolver um nova e efetiva estratégia computacional para um relevante problema biológico; o desenvolvimento de uma estratégia computacional para manipular os modelos estruturais obtidos do PDB , a qual, reduz o espaço de busca conformacional; o desenvolvimento de uma estratégia de busca baseada em algortimo genético combinada com um operador de busca local, o qual, consegue percorrer o espaço de busca conformacional da proteína e escapar de mínimos locais; e a combinação de um método baseado em fragmentos com um modelo de primeiros princípios (Algoritmo genético que minimiza a energia potencial da proteína).

Concluíndo, a predição da estrutura nativa de uma proteína é um problema difícil e pesquisas futuras precisarão ser feitas. O desenvolvimento de novas estratégias, a adaptação e a investigação de novos métodos e a combinação de diferentes técnicas computacionais é necessária. Em resumo, existem diversas oportunidades de pesquisa e caminhos a serem explorados neste campo, com relevantes aplicações multidisciplinares na ciência da computação, bioinformática, química, bioquímica e ciências médicas. Este trabalho abriu diversas e importantes linhas de pesquisa, com grandes aplicações na biologia computacional e bioinformática. Por exemplo, o método proposto poderia ser testado com outras classes de proteínas; outras estretégias de busca também poderiam ser estudadas tais como: PSO; Recozimento simulado; GRASP e busca TABU, as quais, poderiam apresentar bons resultados na predição da estrutura 3D de proteínas. As estruturas preditas pela estratégia MOIRAE poderiam ser usadas como entrada em métodos de refinamento baseados em mecânica molecular (MM), como por exemplo, Dinâmica Molecular (DM). Espera-se com isto que o espaço de busca conformacional seja drasticamente reduzido e os métodos puramente *ab initio* podem demandar um tempo computacional muito menor para encontrar estruturas nativas de proteínas. Isto poderia reduzir o tempo total destes métodos que usualmente iniciam com a estrutura da proteína totalmente extendida.

REFERENCES

ABAGYAN, R.; TOTROV, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. **J. Mol. Biol.**, [S.l.], v.235, n.3, p.983–1002, 1994.

ALEXANDROV, N.; NUSSINOV, R.; ZIMMER, R. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING, Big Island of Hawaii, USA. **Anais. . .** [S.l.: s.n.], 1996. p.53–72.

ALTSCHUL, S. et al. Basic local alignment search tool. **J. Mol. Biol.**, [S.l.], v.215, n.3, p.403–410, 1990.

ALTSCHUL, S. et al. Issues in searching molecular sequence databases. **Nat. Genet.**, [S.l.], v.6, p.119–129, 1994.

ALTSCHUL, S. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res.**, [S.l.], v.25, n.17, p.3389–3402, 1997.

ANDERSON, D. BIONC: a system for public-resource computing and storage. In: IEEE/ACM INTERNATIONAL WORKSHOP ON GRID COMPUTING, 5., Pittsburgh, USA. **Anais. . .** [S.l.: s.n.], 2004. p.4–10.

ANDERSON, J.; TRAVESSET, A. Molecular dynamics on graphic processing units: hoomd to the rescue. **Comput. Sci. Eng.**, [S.l.], v.10, n.6, p.6–10, 2008.

ANDREONI, W.; CURIONI, A. New advances in chemistry and materials science with CPMD and parallel computing. **Parallel Comput.**, [S.l.], v.26, p.819–842, 2000.

ANFINSEN, C. Principles that govern the folding of protein chains. **Science**, [S.l.], v.181, n.96, p.223–230, 1973.

ANFINSEN, C. et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.47, p.1309–1314, 1961.

APOSTOLICO, A.; GIANCARLO, R. Sequence alignment in molecular biology. **J. Comput. Biol.**, [S.l.], v.5, n.2, p.173–196, 1998.

ARNOLD, K. et al. The SWISS-MODEL workspace: a web-based environment for protein structure homology modeling. **Bioinformatics**, [S.l.], v.22, n.2, p.195–201, 2006.

ARORA, N.; JAYARAM, B. Energetics of base pairs in B-DNA in solution: an appraisal of potential functions and dielectric treatments. **J. Phys. Chem. B**, [S.l.], v.102, p.6139–6144, 1998.

ASZÓDI, A.; TAYLOR, W. R. Homology modeling by distance geometry. **Folding Des.**, [S.l.], v.1, n.5, p.325–334, 1996.

BAHAMISH, H.; ABDULLAH, R.; SALAM, R. Protein tertiary structure prediction using artificial bee colony algorithm. In: THIRD ASIA INTERNATIONAL CONFERENCE ON MODELLING AND SIMULATION, 2009., Bandung, Bali. **Proceedings. . .** [S.l.: s.n.], 2009. p.258–263.

BAJORATH, J.; STENKAMP, R.; ARUFFO, A. Knowledge-based model building of proteins: concepts and examples. **Protein Sci.**, [S.l.], v.2, n.11, p.1797–1810, 1994.

BANNER, D.; KOKKINIDIS, M. Structure of the ColE1 rop protein at 1.7 A resolution. **J. Mol. Biol.**, [S.l.], v.196, p.657–675, 1987.

BARTHEL, D. et al. ProCKSI: a decision support system for protein (structure) comparison, knowledge, similarity and information. **BMC Bioinf.**, [S.l.], v.8, p.1–22, 2007.

BATES, P. A. et al. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. **Proteins**, [S.l.], v.5, p.39–46, 2001.

BAXEVANIS, A. Practical aspects of multiple sequence alignment. **Methods Biochem. Anal.**, [S.l.], v.39, p.172–188, 1998.

BAXEVANIS, A.; QUELLETTE, B. **Bioinformatics**: a practical guide to the analysis of genes and proteins. 2.ed. New York, USA: John Wiley and Sons, Inc., 1990. 488p.

BEN-DAVID, M. et al. Assessments of CASP8 structure predictions for template free targets. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.77, n.9, p.50–65, 2009.

BERG, B. A.; NEUHAUS, T. Multi-canonical algorithms for first order phase transitions. **Phys. Lett. B**, [S.l.], v.267, n.2, p.249–253, 1991.

BERMAN, H. et al. The protein data bank. **Nucleic Acids Res.**, [S.l.], v.28, n.1, p.235–242, 2000.

BLANC, E. et al. Solution structure of P01, a natural scorpion peptide structurally analogous to scorpion toxins specific for apamin-sensitive potassium channel. **Proteins**, [S.l.], v.24, p.359–369, 1996.

BLUNDELL, T. et al. Knowledge-based prediction of protein structures and the design of novel molecules. **Nature**, [S.l.], v.326, p.347–352, 1987.

BOAS, F. E.; HARBURY, P. B. Potential energy functions for protein design. **Curr. Opin. Struct. Biol.**, [S.l.], v.17, n.2, p.199–204, 2007.

BONNEAU, R.; BAKER, D. Ab initio protein structure prediction: progress and prospects. **Annu. Rev. Biophys. Biomol. Struct.**, [S.l.], v.30, p.173–189, 2001.

BOWIE, J. U.; EISENBERG, D. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and empirical guiding fitness function. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.91, n.10, p.4436–4440, 1994.

BOWIE, J. U.; LUTHY, R.; EISENBERG, D. A method to identify protein sequences that fold into a known three-dimensional structure. **Science**, [S.l.], v.253, n.5016, p.164–170, 1991.

BRADLEY, P.; MISURA, K.; BAKER, D. Toward high-resolution de Novo structure prediction for small proteins. **Science**, [S.l.], v.309, n.5742, p.1868–1871, 2005.

BRANDEN, C.; TOOZE, J. **Introduction to protein structure**. 2.ed. New York, USA: Garlang Publishing Inc., 1998. 410p.

BRIFFEUIL, P. et al. Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance reliability of predictions. **Bioinformatics**, [S.l.], v.14, n.4, p.357–366, 1998.

BROOKS, R. et al. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. **J. Comput. Chem.**, [S.l.], v.4, n.2, p.187–217, 1983.

BRUDNO, M. et al. Fast and sensitive multiple alignment of large genomic sequences. **BMC Bioinf.**, [S.l.], v.4, n.66, p.1–11, 2003.

BRYANT, S. H.; ALTSCHUL, S. Statistics of sequence-structure threading. **Curr. Opin. Struct. Biol.**, [S.l.], v.5, n.2, p.236–244, 1995.

BRYANT, S.; LAWRENCE, C. An empirical energy function for threading protein sequence through the folding motif. **Proteins: Struc., Func. Gen.**, [S.l.], v.16, n.1, p.92–112, 1993.

BUJNICKI, J. Protein structure prediction by recombination of fragments. **ChemBioChem**, [S.l.], v.7, n.1, p.19–27, 2006.

BUJNICKI, J. et al. Structure prediction meta server. **Bioinformatics**, [S.l.], v.17, p.750–751, 2001.

BURIOL, L. et al. A hybrid genetic algorithm for the weight setting problem in OSPF/IS-IS routing. **Networks**, [S.l.], v.46, n.1, p.36–56, 2005.

CAI, Z. et al. Solution structure of BmBKTx1, a new BKCa1 channel blocker from the Chinese scorpion Buthus martensi Karsch. **Biochemistry**, [S.l.], v.43, p.3764–3771, 2004.

CANUTESCU, A.; SHELENKOV, A.; DUNBRACK JR., R. A graph-theory algorithm for rapid protein side chain prediction. **Proteins**, [S.l.], v.12, n.9, p.2001–2014, 2001.

CASARI, G.; SIPPL, M. J. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. **J. Mol. Biol.**, [S.l.], v.224, n.3, p.725–732, 1992.

CASE, D. et al. The AMBER biomolecular simulation program. **J. Comput. Chem.**, [S.l.], v.26, n.16, p.1668–1688, 2005.

CHAGOT, B. et al. An unusual fold for potassium channel blockers: nmr structure of three toxins from the scorpion opisthacanthus madagascariensis. **Biochem. J.**, [S.l.], v.388, p.263–271, 2005.

CHEN, H.; ZHOU, H. Prediction of solvent accessibility and sites of deleterious mutation from protein sequence. **Nucleic Acids Res.**, [S.l.], v.33, n.10, p.3193–3199, 2005.

CHENG, J. A multi-template combination algorithm for protein comparative modeling. **BMC Struct. Biol.**, [S.l.], v.8, n.18, p.1–13, 2008.

CHIKENJIA, G.; FUJITSUKAB, Y.; TAKADAC, S. A reversible fragment assembly method for de novo protein structure prediction. **J. Chem. Phys.**, [S.l.], v.119, n.13, p.6895–6903, 2003.

CHIVIAN, D. et al. Ab initio methods. **Methods Biochem. Anal.**, [S.l.], v.44, p.547–557, 2003.

CHRISTEN, M. et al. The GROMOS software for biomolecular simulation: gromos05. **J. Comput. Chem.**, [S.l.], v.26, n.16, p.1719–1751, 2005.

CLAESSENS, M. et al. Modelling the polypeptide backbone with 'sparse parts' from known protein structures. **Protein Eng.**, [S.l.], v.2, n.5, p.335–345, 1989.

CLARKE, N. et al. Structural studies of the engrailed homeodomain. **Protein Sci.**, [S.l.], v.3, p.1779–1787, 1994.

CORNELL, W. et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. **J. Am. Chem. Soc.**, [S.l.], v.117, n.19, p.5179–5197, 1995.

CORPET, F. Multiple sequence alignment with hierarchical clustering. **Nucleic Acids Res.**, [S.l.], v.16, n.22, p.10881–10890, 1988.

COZZETTO, D. et al. Evaluation of template-based models in CASP8 with standard measures. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.77, n.9, p.18–28, 2009.

CREIGHTON, T. E. Protein folding. **Biochem. J.**, [S.l.], v.270, p.1–16, 1990.

CRESCENZI, P. et al. On the complexity of protein folding. **J. Comput. Biol.**, [S.l.], v.5, n.3, p.423–466, 1998.

CUTELLO, V.; NARZISI, G.; NICOSIA, G. A multi-objective evolutionary approach to the protein structure prediction problem. **J. R. Soc., Interface**, [S.l.], v.3, n.6, p.139–151, 2006.

CZAPLEWSKI, C. et al. Application of Multiplexed Replica Exchange Molecular Dynamics to the UNRES Force Field: tests with alpha and alpha+beta proteins. **J. Chem. Theory Comput.**, [S.l.], v.5, n.3, p.627–640, 2009.

DANDEKAR, T.; ARGOS, P. Potential of genetic algorithms in protein folding and Protein Eng. simulations. **Protein Eng.**, [S.l.], v.5, n.7, p.637–645, 1992.

DANDEKAR, T.; ARGOS, P. Folding the main chain of small proteins with the genetic algorithm. **J. Mol. Biol.**, [S.l.], v.236, n.3, p.844–861, 1994.

DARDEN, T.; YORK, D.; PEDERSEN, L. Particle mesh Ewald: an n.log n method for ewald sums in large systems. **The J. Chem. Phys.**, [S.l.], v.98, n.12, p.10089–10091, 2009.

DAS, R. et al. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. **Proteins**, [S.l.], v.68, n.S8, p.118–128, 2007.

DAUPLAIS, M. et al. On the convergent evolution of animal toxins. Conservation of a diad of functional residues in potassium channel-blocking toxins with unrelated structures. **J. Biol. Chem**, [S.l.], v.272, p.4302–4309, 1997.

DAYHOFF, M.; SCHWARTZ, R.; ORCUTT, B. A model of evolutionary change in proteins. **Atlas of Protein Sequence and Structure**, [S.l.], v.5, n.3, p.345–352, 1978.

DEBE, D. et al. STRUCTFAST: protein sequence remote homology detection and alignment using novel dynamic programming and profile-profile scoring. **Proteins**, [S.l.], v.64, n.4, p.960–967, 2006.

DEMBO, R.; STEIHAUG, T. Truncated-Newton algorithms for large-scale unconstrained optimization. **Math. Program.**, [S.l.], v.26, p.190–212, 1983.

DERREUMAUX, P. From polypeptide sequences to structures using Monte Carlo simulations and an optimized potential. **J. Chem. Phys.**, [S.l.], v.111, n.5, p.2301–2310, 1999.

DERREUMAUX, P. et al. A Truncated Newton minimizer adapted for CHARMM and biomolecular applications. **J. Comput. Chem.**, [S.l.], v.15, n.5, p.532–552, 1994.

DESJARLAIS, J.; CLARKEB, N. Computer search algorithms in protein modification and design. **Curr. Opin. Struct. Biol.**, [S.l.], v.8, n.4, p.471–475, 1998.

DEWAR, M. Development and status of MINDO/3 and MNDO. **J. Mol. Struct.**, [S.l.], v.100, p.41–50, 1983.

DILL, K. Theory for the folding and stability of globular proteins. **Biochemistry**, [S.l.], v.24, n.6, p.1501–1509, 1985.

DONALDSON, L.; WOJTYRA, U. Solution structure of the dimeric zinc binding domain of the chaperone ClpX. **J. Biol. Chem.**, [S.l.], v.278, p.48991–48996, 2003.

DORN, M.; BREDA, A.; SOUZA, O. Norberto de. A hybrid method for the protein structure prediction problem. **Lect. Notes Bioinf.**, [S.l.], v.5167, p.47–56, 2008.

DORN, M.; SOUZA, O. Norberto de. CReF: a central-residue-fragment-based method for predicting approximate 3-d polypeptides structures. In: ACM SYMPOSIUM ON APPLIED COMPUTING, 2008., Vila Gale in Fortaleza, Ceara, Brazil. **Proceedings. . .** [S.l.: s.n.], 2008. p.1261–1267.

DORN, M.; SOUZA, O. Norberto de. Mining the Protein Data Bank with CReF to predict approximate 3-D structures of polypeptides. **Int. J. Data Min. and Bioin.**, [S.l.], v.4, n.3, p.281–299, 2010.

DORN, M.; SOUZA, O. Norberto de. A3N: an artificial neural network n-gram-based method to approximate 3-d polypeptides structure prediction. **Expert Syst. Appl.**, [S.l.], v.37, n.12, p.7497–7508, 2010B.

DUNBRACK JR., R.; COHEN, F. Bayesian statistical analysis of protein side-chain rotamer preferences. **Protein Sci.**, [S.l.], v.6, n.8, p.1661–1681, 1997.

DUNBRACK JR., R.; KARPLUS, M. Backbone-dependent rotamer library for proteins: application to side-chain prediction. **J. Mol. Biol.**, [S.l.], v.230, n.2, p.543–574, 2003.

DUNKER, A. et al. Intrinsically disordered protein. **J. Mol. Graph. Model.**, [S.l.], v.19, n.1, p.26–59, 2001.

DUNKER, A. et al. Function and structure of inherently disordered proteins. **Curr. Opin. Struct. Biol.**, [S.l.], v.18, n.6, p.756–764, 2008.

EDGAR, R. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Res.**, [S.l.], v.32, n.5, p.1792–1797, 2004.

EISENBERG, D.; MCLACHLAN, A. Solvation energy in protein folding and binding. **Nature**, [S.l.], v.319, p.199–203, 1986.

EISENMENGER, F. et al. SMMP a modern package for simulation of proteins. **Comput. Phys. Commun.**, [S.l.], v.138, p.192–212, 2001.

EISENMENGER, F. et al. An enhanced version of SMMP - open-source software package for simulation of proteins. **Comput. Phys. Commun.**, [S.l.], v.174, p.422–429, 2006.

EISENSTAT, S.; WALKER, H. Choosing the forcing terms in an inexact Newton method. **SIAM J. Sci. Comput.**, [S.l.], v.17, n.1, p.16–32, 1996.

ELBER, R. Computer simulations of protein folding: classical trajectories by optimization of action. **Comput. Phys. Commun.**, [S.l.], v.169, n.1-3, p.277–283, 2005.

ELBER, R. et al. MOIL- A Program for Simulation of Macromolecules. **Comput. Phys. Commun.**, [S.l.], v.91, n.1-2, p.159–189, 1995.

ELBER, R.; KARPLUS, M. A method for determining reaction paths in large molecules: application to myoglobin. **Chem. Phys. Lett.**, [S.l.], v.139, n.5, p.375–380, 1987.

ESWAR, N. et al. Tools for comparative protein structure modeling and analysis. **Nucleic Acids Res.**, [S.l.], v.31, n.13, p.3375–3380, 2003.

ESWAR, N. et al. Comparative protein structure modeling with MODELLER. **Curr. Protoc. Bioinf.**, [S.l.], v.15, p.5.6.1–5.6.30, 2006.

FEIG, M. et al. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. **Proteins**, [S.l.], v.41, n.1, p.86–97, 2000.

FENG, D.; DOOLITTLE, R. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. **J. Mol. Evol.**, [S.l.], v.25, n.4, p.351–360, 1987.

FIELD, M. J. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. **J. Phys. Chem.**, [S.l.], v.102, n.18, p.3586–3616, 1998.

FINKELSTEIN, A.; PTITSYN, O. Why do globular proteins fit the limited set of folding patterns? **Prog. Biophys. Mol. Biol.**, [S.l.], v.50, n.3, p.171–190, 1987.

FISCHER, D. Servers for protein structure prediction. **Curr. Opin. Struct. Biol.**, [S.l.], v.16, p.178–182, 2006.

FISER, A.; DO, R.; SALI, A. Modeling of loops in protein structure. **Protein Sci.**, [S.l.], v.9, n.9, p.1753–1773, 2000.

FITCH, W.; MARGOLIASH, E. Construction of phylogenetic trees. **Science**, [S.l.], v.155, n.760, p.279–284, 1967.

FLOREANO, D.; MATTIUSSI, C. **Bio-Inspired artificial intelligence**. 1.ed. Boston, USA: MIT Press, 2008. 659p.

FLOUDAS, C. A. **Deterministic Global Optimization**: theory, methods and application. 2.ed. Netherlands: Springer, 2004. 549p.

FLOUDAS, C. Computational methods in protein structure prediction. **Biotechnol. Bioeng.**, [S.l.], v.97, n.2, p.207–213, 2007.

FLOUDAS, C. et al. Advances in protein structure prediction and de novo protein design: a review. **Chem. Eng. Sci.**, [S.l.], v.61, n.3, p.966–988, 2006.

FOGOLARI, F.; BRIGO, A.; MOLINARI, H. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. **J. Mol. Recognit.**, [S.l.], v.15, n.6, p.377–392, 2002.

FONSECA, R.; PALUSZEWSKI, M.; WINTER, P. Protein structure prediction using bee colony optimization metaheuristic. **J. Math. Model. Alg.**, [S.l.], v.9, n.2, p.181–194, 2010.

FRAENKEL, A. S. Complexity of protein folding. **Bull. Math. Biol.**, [S.l.], v.55, n.6, p.1199–1210, 1993.

FUJITSUKA, Y.; CHIKENJI, G.; TAKADA, S. SimFold energy function for de novo protein structure prediction: consensus with rosetta. **Proteins**, [S.l.], v.62, n.2, p.381–398, 2006.

FUJITSUKA, Y. et al. Optimizing physical energy functions for protein folding. **Proteins**, [S.l.], v.54, n.1, p.88–103, 2004.

GARCEZ, A. d'Avila; LAMB, L.; GABBAY, D. **Neural-Symbolic Cognitive Reasoning**. 1.ed. New York, USA: Springer, New York, 2009.

GARCEZ, A.; LAMB, L. A connectionist computational Mmdel for epistemic and temporal reasoning. **Neural Computation**, [S.l.], v.18, n.7, p.1711–1738, 2006.

GARCEZ, A.; LAMB, L.; GABBAY, D. Connectionist modal logic: representing modalities in neural networks. **Theoretical Computer Science**, [S.l.], v.371, n.1-2, p.34–53, 2007.

GIBAS, C.; JAMBECK, P. **Developing bioinformatics computer skills**. 1.ed. USA: O'Reilly, 2001. 448p.

GIBBS, N.; CLARKE, A.; SESSIONS, R. Ab initio protein structure prediction using physicochemical potentials and a simplified off-lattice model. **Proteins**, [S.l.], v.43, n.2, p.186–202, 2001.

GINALSKI, K. et al. ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. **Nucleic Acids Res.**, [S.l.], v.31, n.13, p.3804–3807, 2003.

GINALSKI, K. et al. 3D-Jury: a simple approach to improve protein structure predictions. **Bioinformatics**, [S.l.], v.19, n.8, p.1015–1018, 2003.

GLYKOS, N.; CESARENI, G. Protein plasticity to the extreme: changing the topology of a 4-alpha-helical bundle with a single amino acid substitution. **Structure Fold. Des.**, [S.l.], v.7, p.597–603, 1999.

GOHLKEA, H.; HENDLICHA, M.; KLEBE, G. Knowledge-based scoring function to predict protein-ligant interactions. **J. Mol. Biol.**, [S.l.], v.295, p.337–356, 2000.

GONNET, G.; COHEN, M.; BENNER, S. Exhaustive matching of the entire protein sequence database. **Science**, [S.l.], v.256, n.5062, p.1443–1445, 1992.

GORDON, D.; MARSHALL, S.; MAYO, S. Energy functions for protein design. **Curr. Opin. Struct. Biol.**, [S.l.], v.9, n.4, p.509–513, 1999.

GOUDA, H. et al. Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures. **Biochemistry**, [S.l.], v.31, p.9665–9672, 1992.

GRASSO, C.; LEE, C. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. **Bioinformatics**, [S.l.], v.20, n.10, p.1546–1556, 2004.

GREER, J. Comparative modeling methods: application to the family of the mammalian serine protease. **Proteins**, [S.l.], v.7, n.4, p.317–334, 1990.

GRIBSKOV, M. Methods in Molecular Biology. In: GRIFFIN, A.; GRIFFIN, H. (Ed.). **Profile analysis**: methods in molecular biology. [S.l.: s.n.], 1994. v.25, p.247–266.

GRIBSKOV, M.; MCLACHLAN, A.; EISENBERG, D. Profile analysis: detection of distantly related proteins. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.84, n.13, p.4355–4358, 1987.

GRIPPEN, G. The tree structural organization of proteins. **J. Mol. Biol.**, [S.l.], v.126, n.3, p.315–332, 1978.

GUARNIERI, F.; SITILL, W. A rapidly convergent simulation method: mixed monte carlo/stochastic dynamics. **J. Comput. Chem.**, [S.l.], v.15, n.11, p.1302–1310, 1994.

GUEST, M. et al. The GAMESS-UK electronic structure package: algorithms, developments and applications. **Mol. Phys.**, [S.l.], v.103, n.6, p.719–747, 2005.

GUNASEKARAN, K. et al. Extended disordered proteins: targeting function with less scaffold. **Trends Biochem. Sci.**, [S.l.], v.28, n.2, p.81–85, 2003.

GUNSTEREN, W. van; BERENDSEN, H. Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. **Angew. Chem., Int. Ed. Engl.**, [S.l.], v.29, n.9, p.992–1023, 1990.

HAGLER, A. T. et al. Urey-Bradley force field, valence force field, and ab initio study of intramolecular forces in tri-tert-butylmethane and isobutane. **J. Am. Chem. Soc.**, [S.l.], v.101, n.4, p.813–819, 1979.

HALGREN, T. A. Potential energy functions. **Curr. Opin. Struct. Biol.**, [S.l.], v.5, n.2, p.205–210, 1995.

HAO, M.; SCHERAGA, H. Designing potential energy functions for protein folding. **Curr. Opin. Struct. Biol.**, [S.l.], v.9, n.2, p.184–188, 1999.

HART, W.; ISTRAIL, S. Robust proofs of NP-hardness for protein folding: general lattices and energy potentials. **J. Comput. Biol.**, [S.l.], v.4, n.1, p.1–22, 1997.

HARVEY, M.; GIUPPONI, G.; FABRITIIS, G. D. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. **J. Chem. Theory Comput.**, [S.l.], v.5, n.6, p.1632–1639, 2009.

HAVEL, T.; SNOW, M. A new method for building protein conformations from sequence alignments with homologues of knowledge structure. **J. Mol. Biol.**, [S.l.], v.217, n.1, p.1–7, 1991.

HAYKIN, S. **Neural Networks**: a comprehensive foundation. 2.ed. New York, USA: Prentice Hall Inc., 1998.

HE, Y. et al. Exploring the parameter space of the coarse-grained UNRES force field by random search: selecting a transferable medium-resolution force field. **J. Comput. Chem.**, [S.l.], v.30, n.13, p.2127–2135, 2009.

HENDLICH, M. et al. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. **J. Mol. Biol.**, [S.l.], v.216, n.1, p.167–180, 1990.

HENIKOFF, S.; HENIKOFF, J. Amino acid substitution matrices from protein blocks. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.89, p.10915–10919, 1992.

HENIKOFF, S.; HENIKOFF, J. Performance evaluation of amino acid substitution matrices. **Proteins**, [S.l.], v.17, n.1, p.49–61, 1993.

HENIKOFF, S.; HENIKOFF, J. Protein family classification based on searching a database of blocks. **Genomics**, [S.l.], v.19, p.97–107, 1994.

HERGES, T. et al. Stochastic optimization methods for structure prediction of biomolecular nanoscale systems. **Nanotechnology**, [S.l.], v.14, p.1161–1167, 2003.

HESS, B. et al. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. **J. Chem. Theory Comput.**, [S.l.], v.4, n.3, p.435–447, 2008.

HIGGINS, D.; SHARP, P. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. **Gene**, [S.l.], v.73, n.1, p.237–244, 1988.

HIGGINS, D.; THOMPSON, J.; GIBSON, T. Using CLUSTAL for multiple sequence alignments. **Methods Enzymol.**, [S.l.], v.266, p.383–401, 1996.

HILDEBRAND, A. et al. Fast and accurate automatic structure prediction with HHpred. **Proteins**, [S.l.], v.77, n.S9, p.128–132, 2009.

HILL, C. et al. Crystal structure of defensin HNP-3, an amphiphilic dimer: mechanisms of membrane permeabilization. **Science**, [S.l.], v.251, p.1481–1485, 1991.

HIROSAWA, M. et al. Comprehensive study on iterative algorithms of multiple sequence alignment. **CABIOS, Comput. Appl. Biosci.**, [S.l.], v.11, n.1, p.13–18, 1995.

HOLLAND, J. **Adaptation in natural and artificial systems**. 1.ed. Boston, USA: The MIT Press, 1975.

HOLM, L. et al. A database of protein structure families with common folding motifs. **Protein Sci.**, [S.l.], v.1, n.12, p.1691–1698, 1992.

HOLM, L.; SANDER, C. Mapping the protein universe. **Science**, [S.l.], v.273, n.5275, p.595–602, 1996.

HOQUE, M.; CHETTY, M.; DOOLEY, L. A new guided genetic algorithm for 2D hydrophobic-hydrophilic model to predict protein folding. In: IEEE CONGRESS ON EVOLUTIONARY COMPUTATION, Edinburgh, UK. **Anais...** [S.l.: s.n.], 2005. p.259–266.

HOQUE, M.; CHETTY, M.; DOOLEY, L. A guided genetic algorithm for protein folding prediction using 3D hydrophobic-hydrophilic model. In: IEEE CONGRESS ON EVOLUTIONARY COMPUTATION, Vancouver, Canada. **Anais...** [S.l.: s.n.], 2006. p.2339–2346.

HOQUE, M.; CHETTY, M.; SATTAR, A. Genetic algorithm in *ab initio* protein structure prediction using low resolution model: a review. In: SIDHU, A. S.; DILLON, T. (Ed.). **Biomedical Data and Applications**. [S.l.: s.n.], 2009. v.224, p.317–342.

HORNAK, V. et al. Comparison of multiple Amber force fields and development of improved protein backbone parameters. **Proteins**, [S.l.], v.65, p.712–725, 2006.

HORST, R.; TUY, H. **Global Optimization**: deterministic approaches. 3.ed. Berlin, Germany: Springer, 2010. 728p.

HOVMOLLER, T.; OHLSON, T. Conformation of amino acids in protein. **Acta Crystallogr.**, [S.l.], v.58, n.5, p.768–776, 2002.

HUANG, E. et al. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. **J. Mol. Biol.**, [S.l.], v.257, n.3, p.716–725, 1996.

HUANG, E.; SAMUDRALA, R.; PONDER, J. Distance geometry generates native-like folds for small helical proteins using the consensus distances of predicted protein structures. **Protein Sci.**, [S.l.], v.7, n.9, p.1998–2003, 1998.

HUGHEY, R.; KROGH, A. Hidden Markov models for sequence analysis extension and analysis of the basic method. **CABIOS, Comput. Appl. Biosci.**, [S.l.], v.12, n.2, p.95–107, 1996.

HUNG, L. et al. PROTINFO: new algorithms for enhaced protein structure predictions. **Nucleic Acids Res.**, [S.l.], v.33, p.77–80, 2005.

HUTCHINSON, E.; THORNTON, J. PROMOTIF: a program to identify and analyze structural motifs in proteins. **Protein Sci.**, [S.l.], v.5, n.2, p.212–220, 1996.

HUTTER, J.; CURIONI, A. Dual-level parallelism for ab initio molecular dynamics: reaching teraflop performance with the cpmd code. **Parallel Comput.**, [S.l.], v.31, n.1, p.1–17, 2005.

ISHIDA, T. et al. Development of an ab initio protein structure prediction system ABLE. **Genome Inf.**, [S.l.], v.14, p.228–237, 2003.

JACOBSON, M. et al. Force field validation using protein side-chain prediction. **J. Phys. Chem. B**, [S.l.], v.106, n.44, p.11673–11680, 1968B.

JACOBSON, M. et al. A hierarchical approach to all-atom loop prediction. **Proteins**, [S.l.], v.55, p.351–367, 2004.

JACOBSON, M.; FRIESNER, R. X.; HONIG, B. On the role of the crystal environment in determining protein side-chain conformations. **J. Mol. Biol.**, [S.l.], v.320, n.3, p.597–608, 2002.

JASKOWSKI, W. et al. **Found. Comput. Decis. Sci.**, [S.l.], v.31, n.1, p.3–15, 2007.

JAUCH, R. et al. Assessment of CASP7 structure predictions for template free targets. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.69, n.8, p.57–67, 2007.

JAYARAM, B. et al. Bhageerath: an energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins. **Nucleic Acids Res.**, [S.l.], v.34, n.21, p.6195–6204, 2006.

JI, H. et al. Inhibition of human immunodeficiency virus type 1 infectivity by the gp41 core: role of a conserved hydrophobic cavity in membrane fusion. **J. Virol.**, [S.l.], v.73, p.8578–8586, 1999.

JIANG, T.; XU, Y.; ZHANG, M. Protein structure prediction by protein threading and partial experimental data. In: **Current Topics in Computational Molecular Biology**. London, England: The MIT Press, 2002. p.468–502.

JOHNSON, M. et al. Knowledge-based protein modeling. **Crit. Rev. Biochem.**, [S.l.], v.29, n.1, p.1–68, 1994.

JOHNSTON, M.; FERNÁNDEZ-GALVÁN, I.; VILLÀ-FREIRA, J. Framework-based design of a new all-purpose molecular simulation application: the adun simulator. **J. Comput. Chem.**, [S.l.], v.26, n.15, p.1647–1659, 2005.

JONES, D. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. **Proteins**, [S.l.], v.S1, p.185–191, 1997.

JONES, D. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. **J. Mol. Biol.**, [S.l.], v.287, n.4, p.797–815, 1999.

JONES, D. Protein secondary structure prediction based on position-specific scoring matrices. **J. Mol. Biol.**, [S.l.], v.292, n.2, p.195–202, 1999B.

JONES, D. Predicting novel protein folds by using Fragfold. **Proteins**, [S.l.], v.45, n.S5, p.127–132, 2001.

JONES, D.; MILLER, R.; THORNTON, J. Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. **Proteins**, [S.l.], v.23, n.3, p.387–397, 1995.

JONES, D.; TAYLOR, W.; THORNTON, J. A new approach to protein fold recognition. **Nature**, [S.l.], v.358, n.6381, p.86–89, 1992.

JONES, S.; THORNTON, J. Prediction of protein-protein interaction sites using patch analysis. **J. Mol. Biol.**, [S.l.], v.272, n.1, p.133–14, 1997B.

JOO, K. et al. Profile-based nearest neighbor method for pattern recognition. **J. Korean Phys. Soc.**, [S.l.], v.44, n.3, p.599–604, 2004.

JORGENSEN, W. et al. Comparison of simple potential functions for simulating liquid water. **J. Chem. Phys.**, [S.l.], v.79, p.926–936, 1983.

JORGENSEN, W.; MAXWELL, D.; TIRADO-RIVES, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. **J. Am. Chem. Soc.**, [S.l.], v.118, n.45, p.11225–11236, 1996.

JORGENSEN, W.; TIRADO-RIVES, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.102, n.19, p.6665–6670, 2005.

JORGENSEN, W.; TIRADO-RIVES, J. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. **J. Comput. Chem.**, [S.l.], v.26, n.16, p.1689–1700, 2005B.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers**, [S.l.], v.22, n.12, p.2577–637, 1983.

KABSCH, W.; SANDER, C. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.81, n.10, p.1075–1078, 1984.

KACZANOWSKI, S.; ZIELENKIEWICZ, P. Why similar protein sequences enconde similar three-dimensional structures? **Theor. Chem. Acc.**, [S.l.], v.125, p.643–650, 2010.

KARCI, A.; DEMIR, M. Estimation of protein structures by classification of angles between α-carbons of amino acids based on artificial neural networks. **Expert Syst. Appl.**, Tarrytown, v.36, n.3, p.5541–5548, 2009.

KARPLUS, K.; BARRETT, C.; HUGHEY, R. Hidden Markov models for detecting remote protein homologies. **Bioinformatics**, [S.l.], v.14, n.10, p.846–856, 1992.

KARPLUS, K. et al. What is the value added by human intervention in protein structure prediction? **Proteins**, [S.l.], v.5, p.86–91, 2001.

KARPLUS, K. et al. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. **Proteins**, [S.l.], v.56, n.S6, p.491–496, 2003.

KARPLUS, M. The Levinthal paradox: yesterday and today. **Folding Des.**, [S.l.], v.2, n.1, p.S69–S75, 1997.

KELLEY, L.; GARDNER, S. P.; STUTCLIFFE, M. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally-related subfamilies. **Protein Eng.**, [S.l.], v.9, p.1063–1065, 1996.

KELLEY, L.; MACCALLUM, R.; STERNBERG, M. Enhanced genome annotation using structural profiles in the program 3D-PSSM. **J. Mol. Biol.**, [S.l.], v.299, p.501–522, 2000.

KEPLEIS, J.; FLOUDAS, C. Prediction of β-sheet topology and disulfide bridges in polypeptides. **J. Comput. Chem.**, [S.l.], v.24, n.2, p.191–208, 2002B.

KHALILI, M. et al. Molecular dynamics with the united-residue model of polypeptide chains. II. Langevin and Berendsen-bath dynamics and tests on model alpha-helical systems. **J. Phys. Chem. B**, [S.l.], v.109, n.28, p.13798–13810, 2005.

KIEFER, F. et al. The SWISS-MODEL Repository and associated resources. **Nucleic Acids Res.**, [S.l.], v.37, p.D387–D392, 2009.

KIM, J.; PRAMANIK, S.; CHUNG, M. Multiple sequence alignment using simulated annealing. **CABIOS, Comput. Appl. Biosci.**, [S.l.], v.10, n.4, p.419–426, 1994.

KITCHEN, D. B. et al. Docking and scoring in virtual screening for drug discovery: methods and applications. **Nat. Rev. Drug Discovery**, [S.l.], v.3, n.11, p.935–949, 2004.

KLEPEIS, J.; ANDROULAKIS, M.; FLOUDAS, C. Predicting solvated peptide conformations via global minimization of energetic atom-to-atom interactions. **Comput. Chem. Eng.**, [S.l.], v.22, p.765–788, 1998B.

KLEPEIS, J.; FLOUDAS, C. Free energy calculations for peptides via deterministic global optimization. **J. Chem. Phys.**, [S.l.], v.110, p.7491–7512, 1999.

KLEPEIS, J.; FLOUDAS, C. Ab initio prediction of helical segments in polypeptides. **J. Comput. Chem.**, [S.l.], v.23, p.245–266, 2002.

KLEPEIS, J.; FLOUDAS, C. ASTRO-FOLD: a combinatorial and global optimization framework for ab initio prediction of the three-dimensional structures of proteins from the amino acid sequence. **Biophys. J.**, [S.l.], v.85, p.2119–2146, 2003.

KLEPEIS, J.; FLOUDAS, C. Ab initio tertiary structure prediction of proteins. **J. Global Optim.**, [S.l.], v.25, p.113–140, 2003B.

KLEPEIS, J.; IERAPETRITOU, M. G.; FLOUDAS, C. Protein folding and peptide docking: a molecular modeling and global optimization approach. **Comput. Chem. Eng.**, [S.l.], v.22, p.3–10, 1998.

KOEHL, P.; LEVITT, M. A brighter future for proteins structure prediction. **Nat. Struct. Mol. Biol.**, [S.l.], v.6, p.108–111, 1999.

KOLINSKI, A. Protein modeling and structure prediction with a reduced representation. **Acta Biochim. Pol.**, [S.l.], v.51, p.349–371, 2004.

KOLINSKI, A.; BUJINICKI, J. Generalized protein structure prediction based on combination of fold-recognition with de Novo folding and evaluation of models. **Proteins**, [S.l.], v.7, p.84–90, 2005.

KOONIN, E.; GALPERIN, M. . Norwell, USA: Kluwer, 2002. 463p.

KOOP, S. et al. Assessment of CASP7 predictions for template-based modleing targets. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.69, n.8, p.38–56, 2007.

KOPPENSTEINER, W. A.; SIPPL, M. J. Knowledge-based potentials-back to the roots. **Biochemistry**, [S.l.], v.63, p.247–252, 1995.

KOZA, J. R. **Molecular dynamics simulations**: elementary methods. 1.ed. USA: John Wiley and Sons, Inc., 1992. 512p.

KRESSE, G.; MARSMAN, M.; FURTHMULLER. **VASP the Guide**. Computational Physics, Faculty of Physics, Wien University, Wien, Austria, 2009. 163p.

KRIEGER, E. et al. Improving physical realism, stereo-chemistry, and side-chain accuracy in homology modeling: four approaches that performed well in casp8. **Proteins**, [S.l.], v.77, n.S9, p.114–122, 2009.

KROGH, A. et al. Hidden Markov models in computational biology: application to protein modeling. **J. Mol. Biol.**, [S.l.], v.235, n.5, p.1501–1531, 1994.

KRYSHTAFOVYCH, A. et al. Evaluation of residue-residue contact predictions in CASP9. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.79, n.S10, p.119–125, 2011B.

KRYSHTAFOVYCH, A.; FIDELIS K. ANDMOULT, J. CASP9 results compared to those of previous casp experiments. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.79, n.S10, p.196–207, 2011B.

KRYSHTAFOVYCH, A.; FIDELIS, K.; TRAMONTANO, A. Evaluation of model quality predictions in CASP9. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.79, n.S10, p.91–106, 2011.

KUANG, S. et al. **TINKER Summary Sheet**. OPENMD Manual.

KUNDROT, C.; PONDER, J.; RICHARDS, F. Algorithms for calculating excluded volume and its derivatives as a function of molecular conformation and their use in energy minimization. **J. Comput. Chem.**, [S.l.], v.12, n.3, p.402–409, 1991.

LAMBERT, C. et al. ESyPred3D: prediction of proteins 3d structures. **Bioinformatics**, [S.l.], v.18, n.9, p.1250–1256, 2002.

LANDER, E.; WATERMAN, M. **The secrets of life**: a mathematician's introduction to molecular biology. Washington D. C., USA: National Academy Press, 1999. 300p.

LANGDON, W.; POLI, R. **Foundations of genetic programming**. 1.ed. Berlin, Germany: Springer-Verlag, 2010. 260p.

LARKIN, M. et al. Clustal W and Clustal X version 2.0. **Bioinformatics**, [S.l.], v.23, n.21, p.2947–2948, 2007.

LASKOWISKI, R.; WATSON, J.; THORNTON, J. ProFunc: a server for predicting protein functions from 3d structure. **Nucleic Acids Res.**, [S.l.], v.33, p.89–93, 2005.

LASKOWISKI, R.; WATSON, J.; THORNTON, J. Protein function prediction using local 3D templates. **J. Mol.Biol.**, [S.l.], v.351, p.614–626, 2005B.

LASKOWSKI, R. et al. PROCHECK: a program to check the stereochemical quality of protein structures. **J. Appl. Crystallogr.**, [S.l.], v.26, n.2, p.283–291, 1993.

LASKOWSKI, R. et al. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by nmr. **J. Biomol. NMR**, [S.l.], v.8, p.477–486, 1996.

LAVALLE, S. M. **Planning Algorithms**. 1.ed. New York, USA: Cambridge University Press, 2006.

LAZARIDIS, T.; KARPLUS, M. Effective energy functions for protein structure prediction. **Curr. Opin. Struct. Biol.**, [S.l.], v.10, n.2, p.139–145, 2000.

LE GRAND, S.; MERZ JR., K. The application of the genetic algorithm to the minimization of potential energy functions. **J. Global Optim.**, [S.l.], v.3, n.1, p.49–66, 1993.

LEE, J. et al. Efficient parallel algorithms in global optimization of potential energy functions for peptides, proteins, and crystals. **Comput. Phys. Commun.**, [S.l.], v.128, n.1-2, p.399–411, 2000.

LEE, J. et al. Optimization of parameters in macromolecular potential energy functions by conformational space annealing. **J. Phys. Chem. B**, [S.l.], v.105, n.30, p.7291–7298, 2001.

LEE, J. et al. Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. **Proteins**, [S.l.], v.56, n.4, p.704–714, 2004.

LEE, J.; SCHERAGA, H. Conformational space annealing by parallel computations: extensive conformational search of met-enkephalin and of the 20-residue membrane-bound portion of melittin. **Int. J. Quantum Chem.**, [S.l.], v.75, p.255–265, 1999.

LEE, J.; SCHERAGA, H.; RACKOVSKY, S. New optimization Method for conformational energy calculations on polypeptides: conformational space annealing. **J. Comput. Chem.**, [S.l.], v.18, n.9, p.1222–1232, 1997.

LEE, J.; SCHERAGA, H.; RACKOVSKY, S. Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. **Biopolymers**, [S.l.], v.46, n.2, p.103–116, 1998.

LEHNINGER, A.; NELSON, D.; COX, M. **Principles of Biochemistry**. 4.ed. New York, USA: W.H. Freeman, 2005. 1100p.

LENGAUER, T.; RAREY, M. Computational methods for biomolecular docking. **Curr. Opin. Struct. Biol.**, [S.l.], v.6, n.3, p.402–406, 1996.

LESK, A. M. **Introduction to Bioinformatics**. 1.ed. New York, USA: Oxford University Press Inc., 2002. 308p.

LESK, A. M. **Introduction to Protein Science**. 2.ed. New York: Oxford University Press, 2010. 455p.

LEVINTHAL, C. Are there pathways for protein folding? **J. Chim. Phys. Phys.-Chim. Biol.**, [S.l.], v.65, n.1, p.44–45, 1968.

LEVITT, M. Molecular Dynamics of native protein: computer simulation of trajectories. **J. Mol. Biol.**, [S.l.], v.168, n.3, p.595–620, 1983.

LEVITT, M. Accurate modeling of protein conformation by automatic segment matching. **J. Mol. Biol.**, [S.l.], v.226, p.507–533, 1992.

LEVITT, M.; CHOTHIA, C. Structural patterns in globular proteins. **Nature**, [S.l.], v.261, n.5561, p.552–558, 1976.

LEWIS, P.; MOMANY, F.; SCHERAGA, H. Chain reversals in proteins. **Biochim. Biophys. Act.**, [S.l.], v.303, n.2, p.211–229, 1973.

LI, H. et al. Emergence of preferred structures in a simple model of protein folding. **Science**, [S.l.], v.273, n.5275, p.666–669, 1996.

LI, S. C. et al. Fragment-HMM: a new approach to protein structure prediction. **Proteins**, [S.l.], v.17, n.11, p.1925–1934, 2008.

LI, Y.; ZHANG, Y. REMO: a new protocol to refine full atomic protein models from c-α traces by optimizing hydrogen-bonding networks. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.76, p.665–676, 2009.

LI, Y.; ZHANG, Y. Atomic-level protein structure refinement using fragment guided molecular dynamics conformation sampling. **Structure**, [S.l.], v.19, n.12, p.1784–1795, 2011.

LIFSON, S.; WARSHEL, A. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and nAlkane Molecules. **J. Chem. Phys.**, [S.l.], v.49, n.5116, p.14, 1968.

LILJAS, A. et al. . Singapore: World Scientific Printers, 2001. 572p.

LIMBACH, H. et al. ESPResSo: an extensible simulation package for research on soft matter systems. **Comput. Phys. Commun.**, [S.l.], v.174, n.9, p.704–727, 2006.

LIPMAN, D.; ALTSCHUL, S.; KECECIOGLU, J. A tool for multiple sequence alignment. **Proc. Natl. Acad. Sci. U.S.A.**, [S.l.], v.86, n.12, p.4412–4415, 1989.

LIPMAN, D.; PEARSON, W. Rapid and sensitive protein similarity searches. **Science**, [S.l.], v.227, n.4693, p.1435–1441, 1985.

LIU, J. et al. Crystal structure of the unique RNA-binding domain of the influenza virus NS1 protein. **Nat. Struct. Biol.**, [S.l.], v.4, p.896–899, 1997.

LIU, J. et al. A seven-helix coiled coil. **Proc. Natl. Acad. Sci. U.S.A.**, [S.l.], v.103, n.42, p.15457–15462, 2006.

LIWO, A. et al. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. **J. Comput. Chem.**, [S.l.], v.18, p.849–873, 1997.

LIWO, A. et al. United-residue force field for off-lattice protein-structure simulations; III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials. **J. Comput. Chem.**, [S.l.], v.19, p.259–276, 1998.

LIWO, A. et al. Protein structure prediction by global optimization of a potential energy function. **Proc. Natl. Acad. Sci. U.S.A.**, [S.l.], v.96, p.5482–5485, 1999.

LIWO, A. et al. Implementation of molecular dynamics and its extensions with the coarse-grained UNRES force field on massively parallel systems; towards millisecond-scale simulations of protein structure, dynamics, and thermodynamics. **J. Chem. Theory Comput.**, [S.l.], v.6, n.3, p.890–909, 2010.

LLOYD, S. Least squares quantization in PCM. **IEEE Trans. Inf. Theory**, [S.l.], v.28, n.2, p.129â–137, 1982.

LO CONTE, L. et al. SCOP: a structural classification of protein database. **Nucleic Acids Res.**, [S.l.], v.28, n.1, p.257–259, 1999.

LODISH, H. et al. **Molecular Cell Biology**. 5.ed. New York, USA: Scientific American Books, W.H. Freeman, 1990. 970p.

LU, H.; SKOLNICK, J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.44, p.223–232, 2001.

LUKE, S. **Essentials of metaheuristics**. 1.ed. [S.l.]: Lulu, 2009. 227p.

LUNDSTROM, J. et al. Pcons: a neural-network based consensus predictor that improves fold recognition. **Protein Sci.**, [S.l.], v.10, p.2354–2362, 2001.

LUTHY, R.; BOWIE, J.; EISENBERG, D. Assessment of protein models with three-dimensional profiles. **Nature**, [S.l.], v.356, p.83–85, 1992.

LYUBARTSEV, A.; LAAKSONEN, A. M.DynaMix - a scalable portable parallel MD simulation package for arbitrary molecular mixtures. **Comput. Phys. Commun.**, [S.l.], v.128, p.565–589, 2000.

MACKE, T.; CASE, D. Modeling unusual nucleic acid structures. In: **Molecular Modeling of Nucleic Acids**. [S.l.: s.n.], 1998. v.682, p.379–393.

MACKERELL, A. J.; BANAVALI, N.; FOLOPPE, N. Development and current status of the CHARMM force field for nucleic acids. **Biopolymers**, [S.l.], v.56, n.4, p.257–265, 2001.

MACKERELL, A. J.; FEIG, M.; BROOKS, C. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. **J. Comput. Chem.**, [S.l.], v.25, n.11, p.1400–1415, 2004.

MACKERELL JR., A. Empirical force fields for biological macromolecules: overview and issues. **J. Comput. Chem.**, [S.l.], v.25, n.13, p.1584–1604, 2004.

MACKERELL JR., A. et al. CHARMM: the energy function and its parameterization with an overview of the program. In: AL., P. v. R. Schleyer et (Ed.). **The Encyclopedia of Computational Chemistry**. [S.l.]: John Wiley and Sons, 1998. v.1, p.271–277.

MACKERREL, A. **Empirical force fields**. [S.l.]: Springer, 2010. 45-69p.

MADHUSUDHAN, M. et al. Variable gap penalty for protein sequence-structure alignment. **Protein Eng. Des. Sel.**, [S.l.], v.19, n.3, p.129–133, 2006.

MAISURADZE, G. et al. Investigation of protein folding by coarse-grained molecular dynamics with the UNRES force field. **J. Phys. Chem. A**, [S.l.], v.114, n.13, p.4471–4485, 2010.

MARELIUS, J. et al. Q: an md program for free energy calculations and empirical valence bond simulations in biomolecular systems. **J. Mol. Graphics Modell.**, [S.l.], v.16, n.4-6, p.213–225, 1999.

MARIANI, V. et al. Assessment of template based protein structure predictions in CASP9. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.79, n.S10, p.37–58, 2011.

MARQUES, J. . 1.ed. Lisboa, Portugal: IST Press, 1999. 320p.

MARSILI, S. et al. Orac: a molecular dynamics simulation program to explore free energy surfaces in biomolecular systems at the atomistic level. **J. Comput. Chem.**, [S.l.], v.31, n.5, p.1106–1116, 2010.

MARTÍ-RENOM, M. et al. Comparative protein structure modeling of genes and genomes. **Annu. Rev. Biophys. Biomol. Struct.**, [S.l.], v.29, n.16, p.291–325, 2000.

MARTÍ-RENOM, M.; MADHUSUDHAN, M.; SALI, A. Alignment of protein sequences by their profiles. **Protein Sci.**, [S.l.], v.13, n.4, p.1071–1087, 2004.

MARTIN, M.; SIEPMANN, J. Novel configurational-bias Monte Carlo method for branched molecules. Transferable potentials for phase equilibria. 2. united-atom description of branched alkanes. **J. Phys. Chem. B**, [S.l.], v.103, n.21, p.4508–4517, 1999.

MARTÍNEZ, L. et al. PACKMOL: a package for building initial configurations for molecular dynamics simulations. **J. Comput. Chem.**, [S.l.], v.30, n.13, p.2157–2164, 2009.

MCLACHLAN, A. Rapid comparison of protein structures. **Acta Crystallogr.**, [S.l.], v.A38, p.871–873, 1992.

MEINKE, J. et al. SMMP v. 3.0 - Simulating proteins and protein interactions in Python and Fortran. **Comput. Phys. Commun.**, [S.l.], v.178, n.6, p.459–470, 2008.

MILNER-WHITE, E. et al. One type of gamma-turn, rather than the other gives rise to chain-reversal in proteins. **J. Mol. Biol.**, [S.l.], v.204, n.3, p.777–782, 1988.

MITRA, S.; ACHARYA, T. **Data Mining**: pratical machine learning tools and techniques. 2.ed. San Francisco, USA: Elsevier, 2005. 525p.

MOHANTY, D. et al. Correlation between knowledge-based and detailed atomic potentials: application to the unfolding of the gcn4 leucine zipper. **Proteins**, [S.l.], v.35, n.4, p.447–452, 1999.

MOMANY, F. et al. Energy parameters in polypeptides VII, geometric parameters, partial charges, non-bonded interactions, hydrogen bond interactions and intrinsic torsional potentials for naturally occurring amino acids. **J. Phys. Chem.**, [S.l.], v.79, n.22, p.2361–2381, 1975.

MONASTYRSKYY, B. et al. Evaluation of disorder predictions in CASP9. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.79, n.S10, p.107–118, 2011.

MORRIS, A. et al. Stereochemical quality of protein structure coordinates. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.12, p.345–364, 1992.

MOULT, J. A. Decade of CASP: progress, bottlenecks an prognosis in protein structure prediction. **Curr. Opin. Struct. Biol.**, [S.l.], v.15, n.3, p.285–289, 2005.

MOULT, J. et al. A large-scale experiment to assess protein structure prediction methods. **Proteins: Struc., Func. Gen.**, [S.l.], v.23, p.2–5, 1995.

MOULT, J. et al. Critical assessment of methods of protein structure prediction (CASP): round ii. **Proteins: Struc., Func. Gen.**, [S.l.], v.29, p.2–6, 1997.

MOULT, J. et al. Critical assessment of methods of protein structure prediction (CASP): round iii. **Proteins: Struc., Func. Gen.**, [S.l.], v.37, p.2–6, 1999.

MOULT, J. et al. Critical assessment of methods of protein structure prediction (CASP): round iv. **Proteins: Struc., Func. Gen.**, [S.l.], v.45, p.2–7, 2001.

MOULT, J. et al. Critical assessment of methods of protein structure prediction (CASP): round vi. **Proteins: Struc., Func. Gen.**, [S.l.], v.61, p.3–7, 2005.

MOULT, J. et al. Critical assessment of methods of protein structure prediction: round vii. **Proteins: Struc., Func. Gen.**, [S.l.], v.69, p.3–9, 2007.

MOULT, J. et al. Critical assessment of methods of protein structure prediction: round viii. **Proteins: Struc., Func. Gen.**, [S.l.], v.77, p.3–9, 2009.

MOULT, J. et al. Critical Assessment of methods of protein structure prediction (CASP) - Round IX. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.79, n.S10, p.1–5, 2011.

MOULT, J.; FIDELIS, K.; HUBBARD, T. Critical assessment of methods of protein structure prediction (CASP): round v. **Proteins: Struc., Func. Gen.**, [S.l.], v.53, p.334–339, 2003.

MOUNT, D. W. **Bioinformatics**: sequence and genome analysis. 1.ed. New York, USA: Cold Spring Harbor Laboratory Press, 2001. 564p.

MURZIN, A. G. et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. **J. Mol. Biol.**, [S.l.], v.247, n.4, p.536–540, 1995.

NAGADOI, A. et al. Solution structure of the transactivation domain of ATF-2 comprising a zinc finger-like subdomain and a flexible subdomain. **J. Mol. Biol.**, [S.l.], v.287, p.593–607, 1999.

NANIAS, M.; CZAPLEWSKI, C.; SCHERAGA, H. Replica Exchange and Multicanonical Algorithms with the coarse-grained UNRES force field. **J. Chem. Theory Comput.**, [S.l.], v.2, n.3, p.513–528, 2009.

NARANG, P. et al. A computational pathway for bracketing native-like structures for small alpha helical globular proteins. **Phys. Chem. Chem. Phys.**, [S.l.], v.7, n.11, p.2364–2375, 2005.

NARANG, P. et al. Protein structure evaluation using an all-atom energy based empirical scoring function. **J. Biomol. Struct. Dyn.**, [S.l.], v.23, n.4, p.385–406, 2006.

NÉMETHY, G.; PRINTZ, M. P. The γ-Turn, a Possible Folded Conformation of the Polypeptide Chain. Comparison with the β-Turn. **Macromolecules**, [S.l.], v.5, n.6, p.755, 1972.

NEUMAIER, A. Molecular modeling of proteins and mathematical prediction of protein structure. **SIAM Rev.**, [S.l.], v.39, p.407–460, 1997.

NGO, J.; MARKS, J.; KARPLUS, M. The Protein Folding Problem and Tertiary Structure Prediction. In: MERZ JR, K.; GRAND, S. (Ed.). **Computational complexity, protein structure prediction and the Levinthal Paradox**. Boston, USA: Birkhauser, 1997. p.435–508.

NOTREDAME, C. Recent progresses in multiple sequence alignment: a survey. **Pharmacogenomics**, [S.l.], v.31, n.1, p.131–144, 2002.

NOTREDAME, C. Recent evolutions of multiple sequence alignment algorithms. **PLoS Comput. Biol.**, [S.l.], v.8, n.3, p.1405–1408, 2007.

NOTREDAME, C.; HIGGINS, D.; HERINGAL, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. **J. Mol. Biol.**, [S.l.], v.302, n.1, p.205–217, 2000.

NOTREDAME, C.; HOLM, L.; HIGGINS, D. COFFEE: an objective function for multiple sequence alignments. **Bioinformatics**, [S.l.], v.14, n.5, p.407–422, 1998.

OHSEN, N. von; SOMMER, I.; ZIMMER, R. Profile-Profile alignment: a powerful tool for protein structure prediction. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING, Lihue, USA. **Proceedings...** [S.l.: s.n.], 2003. v.8, p.252–263.

OLDZIEJ, S. et al. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. **Proc. Natl. Acad. Sci. U.S.A.**, [S.l.], v.102, n.21, p.7547–7552, 2005.

ONUFRIEV, A.; BASHFORD, D.; CASE, D. Modification of the Generalized Born Model Suitable for Macromolecules. **J. Phys. Chem. B**, [S.l.], v.104, p.3712–3720, 2000.

ONUFRIEV, A.; BASHFORD, D.; CASE, D. Effective Born radii in the generalized Born approximation: the importance of being perfect. **J. Comput. Chem.**, [S.l.], v.23, n.14, p.1297–1304, 2002.

ONUFRIEV, A.; BASHFORD, D.; CASE, D. Exploring protein native states and large-scale conformational changes with a modified generalized born model. **Proteins**, [S.l.], v.55, p.383–394, 2004.

OOI, T. et al. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.84, n.10, p.3086–3090, 1987.

ORENGO, C. et al. CATH - a hierarchic classification of protein domain structures. **Structure**, [S.l.], v.5, n.8, p.1093–1108, 1997.

ORENGO, C. et al. The CATH database provides insights into protein structure function relationship. **Nucleic Acids Res.**, [S.l.], v.27, n.1, p.275–279, 1999.

OSGUTHORPE, D. Ab initio protein folding. **Curr. Opin. Struct. Biol.**, [S.l.], v.10, n.2, p.146–152, 2000.

OTA, M.; NISHIKAWA, K. Assessment of pseudo-energy potentials by the best-five test: a new use of the three-dimensional profiles of proteins. **Protein Eng.**, [S.l.], v.10, n.4, p.339–351, 1997.

O'TOOLE, J.; DAHLER, J. Boltzmann equation and inverse collisions. **J. Chem. Phys.**, [S.l.], v.33, n.5, p.1487–1495, 1960.

OUZOUNIS, C. et al. Prediction of protein structure by evaluation of sequence structure fitness aligning sequences to contact profiles derived from three-dimensional structures. **J. Mol. Biol.**, [S.l.], v.232, n.3, p.805–825, 1993.

PAPADIMITRIOU, C. H.; STEIGLITZ, K. **Combinatorial optimization**: algorithms and complexity. 1.ed. New Jersey, USA: Dover Publications, 1998. 512p.

PAPPU, R.; HART, R.; PONDER, J. Analysis and application of potential energy smoothing and search methods for global optimization. **J. Phys. Chem. B**, [S.l.], v.102, n.48, p.9725–9742, 1998.

PARK, B.; HUANG, E.; LEVITT, M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. **J. Mol. Biol.**, [S.l.], v.266, n.4, p.831–846, 1997.

PARK, S. A study of fragment-based protein structure prediction: biased fragment replacement for searching low-energy conformation. **Genome Inf.**, [S.l.], v.16, n.2, p.104–113, 2005.

PASCHEK, D.; GEIGER, A. **Performing Molecular Dynamics Simulations - User's Guide and Manual for the MOSCITO Simulation Package**. Physikalische Chemie IIa, Dortmund University, Dortmund, Germany, 2003. 80p.

PASTOR, M. et al. Combinatorial approaches: a new tool to search for highly structured beta-hairpin peptides. **Proc. Natl. Acad. Sci. USA**, [S.l.], v.99, p.614–619, 2002.

PAULING, L.; COREY, R. The pleated sheet, a new layer configuration of polypeptide chains. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.37, n.5, p.251–256, 1951.

PAULING, L.; COREY, R.; BRANSON, H. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.37, n.4, p.205–211, 1951.

PEARLMAN, D. et al. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. **Comput. Phys. Commun.**, [S.l.], v.91, n.1-3, p.1–41, 1995.

PEARSON, W.; LIPMAN, D. Improved tools for biological sequence comparison. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.85, n.8, p.2444–2448, 1988.

PEDERSEN, J.; MOULT, J. Protein folding simulations with genetic algorithms and a detailed molecular description. **J. Mol. Biol.**, [S.l.], v.269, n.2, p.240–259, 1997.

PEITSCH, M. ProdMod and Swiss-Model: internet-based tools for automated comparative protein modeling. **Biochem. Soc. Trans.**, [S.l.], v.24, n.1, p.274–279, 1996.

PEITSCH, M.; JONGENEEL, C. A 3-D model for the CD40 ligand predicts that it is a compact trimer similar to the tumor necrosis factors. **Int. Immunol.**, [S.l.], v.5, n.2, p.233–238, 1993.

PHILLIPS, J. et al. Scalable molecular dynamics with NAMD. **J. Comput. Chem.**, [S.l.], v.26, n.16, p.1781–1802, 2005.

PLIMPTON, S. Fast parallel algorithms for short-range molecular dynamics. **J. Comput. Phys.**, [S.l.], v.117, p.1–19, 1995.

POKALA, N.; HANDEL, T. Review: protein design - where we were, where we are, where we're going. **J. Struct. Biol.**, [S.l.], v.134, n.2-3, p.269–281, 2000.

POKAROWSKI, P.; KOLINSKI, A.; SKOLNICKZ, J. A minimal physically realistic protein-like lattice model: designing an energy landscape that ensures all-or-none folding to a unique native state. **Biophys. J.**, [S.l.], v.84, n.3, p.1518–1526, 2003.

PONDER, J. **TINKER Summary Sheet**. Jay Ponder Lab, Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine.

PONDER, J.; RICHARDS, F. An efficient Newton-like method for molecular mechanics energy minimization of large molecules. **J. Comput. Chem.**, [S.l.], v.8, n.7, p.1016–1024, 1987.

PROCACCI, P. et al. ORAC: a molecular dynamics program to simulate complex molecular systems with realistic electrostatic interactions. **J. Comput. Chem.**, [S.l.], v.18, n.15, p.1848–1862, 1997.

QIU, D. et al. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. **J. Phys. Chem. A**, [S.l.], v.101, p.3005–3014, 1997.

RAMACHANDRAN, G.; SASISEKHARAN, V. Conformation of polypeptides and proteins. **Adv. Protein Chem.**, [S.l.], v.23, p.238–438, 1968.

RAPAPORT, D. C. **The art of molecular dynamics simulation**. 2.ed. Cambridge, UK: Cambridge University Press, 2004. 549p.

REFSON, K. Moldy: a portable molecular dynamics simulation program for serial and parallel computers. **Comput. Phys. Commun.**, [S.l.], v.126, n.3, p.310–329, 2000.

RICHARDS, F. Areas, volumes, packing and protein structure. **Annu. Rev. Biophys. Bioeng.**, [S.l.], v.6, p.151–176, 1977.

RICHARDS, F.; KUNDROT, C. Identification of structural motifs from protein coordinate data: secondary structure and first level super-secondary structure. **Proteins**, [S.l.], v.3, n.2, p.71–84, 1988.

RICHARDSON, J. The anatomy and taxonomy of protein structure. **Biopolymers**, [S.l.], v.34, p.167–339, 1981.

ROHL, C. et al. Protein structure prediction using Rosetta. **Methods Enzymol.**, [S.l.], v.383, n.2, p.66–93, 2004.

ROSE, G. Hierarchic organization of domains in globular proteins. **J. Mol. Biol.**, [S.l.], v.134, n.3, p.447–470, 1979.

ROSE, G.; GIERASCH, L.; SMITH, J. Turns in peptides and proteins. **Adv. Protein Chem.**, [S.l.], v.37, p.1–109, 1985.

ROSE, G.; WOLFENDEN, R. Hydrogen bonding, hydrophobicity, packing and protein folding. **Annu. Rev. Biophys. Biomol. Struct.**, [S.l.], v.22, p.381–415, 1993.

BOHR, H.; BRUNAK, S. (Ed.). **Fitting 1-D predictions into 3-D structures**. Boca Raton: CRC Press, 1995. 132p.

ROST, B. TOPITS: threading one-dimensional predictions into three-dimensional structures. **Proc. Int. Conf. Intell. Syst. Mol. Biol.**, [S.l.], v.3, p.314–321, 1995b.

RUMELHART, D.; HINTON, G.; WILLIAMS, R. Learning representations by back-propagating errors. **Nature**, [S.l.], v.323, p.533–536, 1986.

RUSSELL, R.; BARTON, G. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. **J. Mol. Biol.**, [S.l.], v.244, n.3, p.332–350, 1994.

RYCHLEWSKI, L. et al. Comparison of sequence profiles. Strategies for structural predictions using sequence information. **Protein Sci.**, [S.l.], v.9, n.2, p.232–241, 2000.

SA, J. de. . 1.ed. Berlin: Springer, 2001. 318p.

SADREYEV, R.; GRISHIN, N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. **J. Mol. Biol.**, [S.l.], v.326, n.1, p.317–336, 2003.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Mol. Biol. Evol.**, [S.l.], v.4, n.4, p.406–425, 1987.

SALI, A. Modelling mutations and homologous proteins. **Curr. Opin. Biotechnol.**, [S.l.], v.6, n.4, p.437–451, 1995.

SALI, A.; BLUNDELL, T. Comparative protein modeling by satisfaction of spatial restraints. **J. Mol. Biol.**, [S.l.], v.234, n.3, p.779–815, 1993.

SÁNCHEZ, R.; SALI, A. Advances in comparative protein-structure modeling. **Curr. Opin. Struct. Biol.**, [S.l.], v.7, n.2, p.206–214, 1997.

SARISKY, C.; MAYO, S. The beta-beta-alpha fold: explorations in sequence space. **J. Mol. Biol.**, [S.l.], v.307, p.1411–1418, 2001.

SASIN, J.; KUROWSKI, M.; BUJNICKI, J. STRUCLA: a www meta-server for protein structure comparison and evolutionary classification. **Bioinformatics**, [S.l.], v.19, p.252–254, 2003.

BOURNE, P.; WEISSIG, H. (Ed.). **Fundamentals of protein structure**: structural bioinformatics. [S.l.: s.n.], 2003. 15p.

SCHMIDT, T. et al. Assessment of ligand-binding residue predictions in CASP9. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.79, n.S10, p.126–136, 2011.

SCHUELER-FURMAN, O. et al. Progress in modeling of protein structures and interactions. **Science**, [S.l.], v.310, n.5748, p.638–642, 2005.

SCHUG, A. et al. Investigation of the parallel tempering method for protein folding. **J. Phys.: Condens. Matter**, [S.l.], v.17, p.1641–1650, 2005.

SCHWEDE, T. et al. SWISS-MODEL: an automated protein homology-modeling server. **Nucleic Acids Res.**, [S.l.], v.31, n.13, p.3381–3385, 2003.

SCOTT, W. et al. The GROMOS biomolecular simulation program package. **J. Phys. Chem. A**, [S.l.], v.103, n.19, p.3596–3607, 1999.

SELEZENEV, A. et al. SAGE MD: molecular-dynamic software package to study properties of materials with different models for interatomic interactions. **Comput. Mater. Sci.**, [S.l.], v.28, n.2, p.107–124, 2003.

SETUBAL, J.; MEIDANIS, J. **Introduction to Computational Molecular Biology**. 1.ed. Boston, USA: PWS Publishing Company, 1997. 300p.

SHEN, H. et al. Implementation of a Serial Replica Exchange Method in a Physics-Based United-Residue (UNRES) Force Field. **J. Chem. Theory Comput.**, [S.l.], v.4, n.8, p.1386–1400, 2008.

SHEN, H.; LIWO, A.; SCHERAGA, H. An improved functional form for the temperature scaling factors of the components of the mesoscopic UNRES force field for simulations of protein structure and dynamics. **J. Phys. Chem. B**, [S.l.], v.113, n.25, p.8738–8744, 2009.

SHI, J.; BLUNDELL, T.; MIZUGUCHI, K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. **J. Mol. Biol.**, [S.l.], v.310, n.1, p.243–257, 2001.

SHINDYALOV, I.; BOURNE, P. Protein structure alignment by incremental combinatorial extension CE of the optimal path. **Protein Eng.**, [S.l.], v.11, n.9, p.739–747, 1998.

SIEW1, N. et al. MaxSub: an automated measure for the assessment of protein structure prediction quality. **Bioinformatics**, [S.l.], v.16, n.9, p.776–785, 2000.

SIMONS, K. et al. Assembly of protein tertiary structures from fragments with similar local sequences using simulated anneling and Bayesian score functions. **J. Mol. Biol.**, [S.l.], v.268, n.1, p.209–225, 1997.

SIMONS, K. et al. Improved recognition of native-like structures using a combination of sequence-dependent and sequence-independent features of proteins. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.34, n.1, p.82–95, 1999.

SIMONS, K. et al. Ab initio protein structure prediction of CASP III targets using ROSETTA. **Proteins**, [S.l.], v.3, n.3, p.171–176, 1999B.

SIPPL, M. Knowledge-based potentials for proteins. **Curr. Opin. Struct. Biol.**, [S.l.], v.5, n.2, p.229–235, 1995.

SIPPL, M.; HENDLICH, M.; LACKNER, P. Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments: development of strategies and construction of models for myoglobin, lysozyme, and thymosin beta 4. **Protein Sci.**, [S.l.], v.1, p.625–640, 1992.

SMITH, G. et al. Ab initio structure determination and refinement of a scorpion protein toxin. **Acta Crystallogr.**, [S.l.], v.53, p.551–557, 1997.

SMITH, J. The co-evolution of memetic algorithms for protein structure prediction. **Stud. Fuzziness Soft Comput.**, [S.l.], v.166, p.105–128, 2005.

SMITH, T. The art of matchmaking: sequence alignment methods and their structural implications. **Structure**, [S.l.], v.7, n.1, p.R7–R12, 1999.

SMITH, W.; FORESTER, T. DL POLY 2.0: a general-purpose parallel molecular dynamics simulation package. **J. Mol. Graphics**, [S.l.], v.14, n.3, p.136–141, 1996.

SMITH, W.; YONG, C.; RODGER, P. DL POLY: application to molecular simulation. **Mol. Simul.**, [S.l.], v.28, n.5, p.385–471, 2002.

SODING, J. Protein homology detection by HMM-HMM comparison. **Bioinformatics**, [S.l.], v.21, n.7, p.951–960, 2005.

SODING, J.; BIEGERT, A.; LUPAS, A. The HHpred interactive server for protein homology detection and structure prediction. **Nucleic Acids Res.**, [S.l.], v.33, n.Web Server Issue, p.244–248, 2005.

SOLER, J. et al. Levinthal's paradox. **J. Phys.: Condens. Matter**, [S.l.], v.14, p.2745–2779, 2001.

SPOEL, D. van der. The solution conformation of amino acids from molecular dynamics simulations of Gly-X-Gly peptides: comparison with nmr parameters. **Biochem. Cell Biol**, [S.l.], v.76, n.2-3, p.164–170, 1998.

SPOEL, D. van der; BERENDSEN, H. Molecular dynamics simulations of Leu-Enkephalin in water and DMSO. **Biophys. J.**, [S.l.], v.72, n.5, p.2032–2041, 1997.

SPOEL, D. van der et al. Molecular dynamics simulations of peptides from BPTI: a closer look at amide-aromatic interactions. **J. Biomol. NMR**, [S.l.], v.8, n.3, p.229–238, 1996.

SPOEL, D. van der et al. GROMACS: fast, flexible, and free. **J. Comput. Chem.**, [S.l.], v.26, n.16, p.1701–1718, 2005.

SPOEL, D. van der; VOGEL, H.; BERENDSEN, H. Molecular dynamics simulations of N-terminal peptides from a nucleotide binding protein. **Proteins**, [S.l.], v.24, n.4, p.450–466, 1996.

SRINIVASAN, R.; ROSE, G. LINUS - A hierarchic procedure to predict the fold of a protein. **Proteins**, [S.l.], v.22, n.2, p.81–99, 1995.

SRINIVASAN, R.; ROSE, G. Ab initio prediction of protein structure using LINUS. **Proteins**, [S.l.], v.47, n.4, p.489–495, 2002.

SRINIVASAN, S.; MARCH, C.; SUDARSANAM, S. An automated method for modeling proteins on known templates using distance geometry. **Protein Sci.**, [S.l.], v.2, n.2, p.227–289, 1993.

STADLER, J.; MIKULLA, R.; TREBIN, H. IMD: a software package for molecular dynamics studies on parallel computers. **Int. J. Mod. Phys. C**, [S.l.], v.8, n.5, p.1131–1140, 1997.

STERNBERG, M. **Protein Structure Prediction**: a practical approach. 1.ed. New York, USA: Oxford University Press, 1997. 320p.

STILL, W.; YEO H.C. AMD KOLATKAR, P.; CLARKE, N. Semianalytical treatment of solvation for molecular mechanics and dynamics. **J. Am. Chem. Soc.**, [S.l.], v.112, p.6127–6129, 1990.

SUN, S. A genetic algorithm that seeks native states of peptides and proteins. **Biophys. J.**, [S.l.], v.69, n.2, p.340–355, 1995.

SWENDSEN, R.; WANG, J. Replica Monte Carlo simulation of spin glasses. **Phys. Rev. Lett.**, [S.l.], v.57, n.21, p.2607–2609, 1986.

TAYLOR, W. A flexible method to align large numbers of biological sequences. **J. Mol. Evol.**, [S.l.], v.28, n.1-2, p.161–169, 1988.

TAYLOR, W. Multiple protein sequence alignment: algorithms and gap insertion. **Methods Enzymol.**, [S.l.], v.266, p.343–367, 1996.

THACHUK, C.; SHMYGELSKA, A.; HOOS, H. A replica exchange Monte Carlo algorithm for protein folding in the HP model. **BMC Bioinf.**, [S.l.], v.8, p.20–22, 2007.

THOMPSON, J.; HIGGINS, D.; GIBSON, T. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. **Nucleic Acids Res.**, [S.l.], v.22, n.22, p.4673–4680, 1994.

THOMPSON, J.; PLEWNIAK, F.; POCH, O. A comprehensive comparison of multiple sequence alignment programs. **Nucleic Acids Res.**, [S.l.], v.27, n.13, p.2682–2690, 1999.

THUKRAL, L. et al. ProRegIn: a regularity index for the selection of native-like tertiary structures of proteins. **J. Biosci.**, [S.l.], v.32, n.1, p.71–81, 2007.

TOMPA, P. Intrinsically unstructured proteins. **Trends Biochem Sci.**, [S.l.], v.27, n.10, p.527–533, 2002.

TOMPA, P.; CSERMELY, P. The role of structural disorder in the function of RNA and protein chaperones. **FASEB J.**, [S.l.], v.18, n.11, p.1169–1175, 2004.

TOUKMAJI, A.; BOARD, J. J. Ewald summation techniques in perspective: a survey. **Comput. Phys. Commun.**, [S.l.], v.95, n.2, p.73–92, 1996.

TRAMONTANO, A. **Protein structure prediction**. 1.ed. Weinheim, Germany: John Wiley and Sons, Inc., 2006. 208p.

TSUI, V.; CASE, D. Theory and applications of the generalized bron solvation model in macromolecular simulations. **Biopolymers**, [S.l.], v.56, p.275–291, 2001.

TUCKERMAN, M. et al. Exploiting multiple levels of parallelism in Molecular Dynamics based calculations via modern techniques and software paradigms on distributed memory computers. **Comput. Phys. Commun.**, [S.l.], v.128, n.1-2, p.333–376, 2000.

TUDOR, J. et al. Solution structure of ShK toxin, a novel potassium channel inhibitor from a sea anemone. **Nat. Struct. Biol.**, [S.l.], v.3, p.317–320, 1996.

TUFFERY, P. et al. A new approach to the rapid determination of protein sidechain conformations. **J. Biomol. Struct. Dyn.**, [S.l.], v.8, n.6, p.1267–1289, 1991.

TUGARINOV, V.; ZVI, A.; LEVY, R. A cis proline turn linking two beta-hairpin strands in the solution structure of an antibody-bound HIV-1IIIB V3 peptide. **Nat. Struct. Biol.**, [S.l.], v.6, p.331–335, 1999.

TURCOTTE, M.; MUGGLETON, S.; STERNBERG, M. Application of inductive logic programming to discover rules governing the three-dimensional topology of protein structure. In: INTERNATIONAL WORKSHOP ON INDUCTIVE LOGIC PROGRAMMING. **Proceedings. . .** [S.l.: s.n.], 1998. p.53–64.

TURCOTTE, M.; MUGGLETON, S.; STERNBERG, M. Automated discovery of structural signatures of protein fold and function. **J. Mol. Biol.**, [S.l.], v.306, p.591–605, 2001.

TURCOTTE, M.; MUGGLETON, S.; STERNBERG, M. The effect of relational background knowledge on learning of protein three-dimensional fold signatures. **Machine Learning**, [S.l.], v.43, n.1-2, p.81–96, 2001B.

TURCOTTE, M.; MUGGLETON, S.; STERNBERG, M. Generating protein three-dimensional fold signatures using inductive logic programming. **Comput. Chem.**, [S.l.], v.26, p.57–64, 2001C.

UNGER, R.; MOULT, J. On the applicability of genetic algorithms to protein folding. In: THE TWENTY-SIXTH HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES. **Anais. . .** [S.l.: s.n.], 1993. p.715–725.

UNGER, R.; MOULT, J. Genetic algorithms for protein folding simulations. **J. Mol. Biol.**, [S.l.], v.231, n.1, p.75–81, 1993a.

UVERSKY, V. What does it mean to be natively unfolded? **Eur. J. Biochem.**, [S.l.], v.269, n.1, p.2–12, 2001.

VÁSQUEZ, M. Modeling side-chain conformation. **Curr. Opin. Struct. Biol.**, [S.l.], v.6, n.2, p.217–221, 1996.

VAZIRANI, V. V. **Approximation Algorithms**. 1.ed. New York, USA: Springer, 2001. 256p.

VENKATACHALAM, C. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. **Biopolymers**, [S.l.], v.6, n.10, p.1425–1436, 1968.

VOIGT, C.; GORDON, D.; MAYO, S. Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. **J. Mol. Biol.**, [S.l.], v.299, n.3, p.789–803, 2000.

WALLACE, I.; BLACKSHIELDS, G.; HIGGINS, D. Multiple sequence alignments. **Curr. Opin. Struct. Biol.**, [S.l.], v.15, n.3, p.261–266, 2005.

WALLNER, B.; LARSSON, P.; ELOFSSON, A. Pcons.net: protein structure prediction meta server. **Nucleic Acids Res.**, [S.l.], v.35, p.369–374, 2007.

WANG, G.; DUNBRACK, R. PISCES: a protein sequence culling server. **Bioinformatics**, [S.l.], v.19, n.12, p.1589–1591, 2003.

WANG, Z. A re-estimation for the total numbers of protein folds and super-families. **Protein Eng.**, [S.l.], v.11, n.8, p.621–626, 1998.

WEINER, S. et al. A new force field for the molecular mechanical simulation of nucleic acids and proteins. **J. Am. Chem. Soc.**, [S.l.], v.106, p.765–784, 1984.

WESSON, L.; EISENBERG, D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. **Protein Sci.**, [S.l.], v.1, n.2, p.227–235, 1992.

WHEELER, D. et al. Database resources of the national center for biotechnology information. **Nucleic Acids Res.**, [S.l.], v.33, n.Database issue, p.39–45, 2005.

WHITE, J.; MUCHNIK, I.; SMITH, T. Modeling protein cores with Markov random fields. **Math. Biosci.**, [S.l.], v.124, n.2, p.149–179, 1994.

WILLIAMS, D. Representation of the molecular electrostatic potential by atomic multi-pole and bond dipole models. **J. Comput. Chem.**, [S.l.], v.9, n.7, p.745–763, 1998.

WITHERS-WARD, E. et al. Biochemical and structural analysis of the interaction between the UBA(2) domain of the DNA repair protein HHR23A and HIV-1 Vpr. **Biochemistry**, [S.l.], v.39, p.14103–14112, 2000.

XU, Y.; XU, D.; LIANG, J. (Ed.). **A historical perspective and overview of protein structure prediction**. [S.l.]: Springer, 2010. 1-43p.

WRIGHT, P.; DYSON, H. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. **J. Mol. Biol.**, [S.l.], v.293, n.2, p.321–331, 1999.

WU, S.; SKOLNICK, J.; ZHANG, Y. Ab initio modeling of small proteins by iterative TASSER simulations. **BMC Biol.**, [S.l.], v.5, n.17, p.1–10, 2007.

WU, S.; ZHANG, Y. LOMETS: a local meta-threading-server for protein structure prediction. **Nucleic Acids Res.**, [S.l.], v.35, n.10, p.3375–3382, 2007.

WU, S.; ZHANG, Y. ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction. **Plos One**, [S.l.], v.2, p.3400–3408, 2008.

WU, S.; ZHANG, Y. MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. **Proteins: Struc., Func. Gen.**, [S.l.], v.72, p.547–556, 2008.

WU, S.; ZHANG, Y. SEGMER:identifying protein sub-structural similarity by segmental threading. **Structure**, [S.l.], v.18, p.858–867, 2010.

XU, D. et al. Automated protein structure modeling in CASP9 by I-Tasser pipeline combined with Quark-based ab initio folding and FG-MD-based strcuture refinement. **Proteins: Struct., Funct., Bioinf.**, [S.l.], v.79, n.10, p.147–160, 2011.

XU, J. et al. Protein structure prediction by linear programming. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING, Big Island of Hawaii, USA. **Anais...** [S.l.: s.n.], 2003. p.264–275.

XU, J. et al. RAPTOR: optimal protein threading by linear programming. **J. Bioinf. Comput. Biol.**, [S.l.], v.1, n.1, p.95–117, 2003B.

XU, J.; PENG, J.; ZHAO, F. Template-based and free modeling by RAPTOR11 in CASP8. **Proteins**, [S.l.], v.77, n.S9, p.133–137, 2009.

XU, Y.; XU, D. Protein threading using PROSPECT: design and evaluation. **Proteins**, [S.l.], v.40, n.3, p.343–354, 2000.

XU, Y.; XU, D.; UBERBACHER, E. An efficient computational method for globally optimal threading. **J. Comput. Biol.**, [S.l.], v.5, n.3, p.597–614, 1998.

YAMANO, A.; HEO, N.; TEETER, M. Crystal structure of Ser-22/Ile-25 form crambin confirms solvent, side chain substate correlations. **J. Biol. Chem.**, [S.l.], v.272, p.9597–9600, 1997.

ZERELLA, R. et al. Structural characterization of a mutant peptide derived from ubiquitin: implications for protein folding. **Protein Sci.**, [S.l.], v.9, p.2142–2150, 2000.

CHEN, Y.-P. (Ed.). **Overview of structural bioinformatics**. Heidelberg: Springer, 2005.

ZHANG, Y. Template-based modeling and free modeling by I-TASSER in CASP7. **Proteins**, [S.l.], v.69, n.8, p.108–117, 2007.

ZHANG, Y. I-TASSER server for protein 3D structure prediction. **BMC Bioinf.**, [S.l.], v.9, n.40, p.1–8, 2008.

ZHANG, Y. Progress and challenges in protein structure prediction. **Curr. Opin. Struct. Biol.**, [S.l.], v.18, p.342–348, 2008B.

ZHANG, Y. I-TASSER: fully automated protein structure prediction in casp8. **Proteins**, [S.l.], v.77, n.S9, p.100–113, 2009.

ZHANG, Y. Protein structure prediction: when is it useful? **Curr. Opin. Struct. Biol.**, [S.l.], v.19, n.2, p.145–155, 2009-B.

ZHANG, Y. et al. On the origin and completeness of highly likely single domain protein structures. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.103, n.8, p.2605–2610, 2006.

ZHANG, Y.; KIHARA, D.; SKOLNICK, J. Local energy landscape flattering: parallel hyperbolic monte carlo sampling of protein folding. **Proteins**, [S.l.], v.48, p.192–201, 2002.

ZHANG, Y.; SKOLNICK, J. Scoring function for automated assessment of protein structure template quality. **Proteins**, [S.l.], v.57, n.4, p.702–710, 2004.

ZHANG, Y.; SKOLNICK, J. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. **Biophys. J.**, [S.l.], v.87, n.4, p.2647–2655, 2004B.

ZHANG, Y.; SKOLNICK, J. Automated structure prediction of weakly homologous proteins on a genomic scale. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.101, n.20, p.7594–7599, 2004C.

ZHANG, Y.; SKOLNICK, J. SPICKER: a clustering approach to identify near-native protein folds. **J. Comput. Chem.**, [S.l.], v.25, n.6, p.20–22, 2004D.

ZHANG, Y.; SKOLNICK, J. TM-align: a protein structure alignment algorithm based on tm-score. **Nucleic Acids Res.**, [S.l.], v.33, p.2302–2309, 2005.

ZHOU, H.; PANDIT, S.; SKOLNICK, J. Performance of the Pro-sp3-TASSER Server in CASP8. **Proteins: Struc., Func. Gen.**, [S.l.], v.77, n.S9, p.123–127, 2009.

ZHOU, H.; SKOLNICK, J. Ab initio protein structure prediction using chunk-TASSER. **Biophys. J.**, [S.l.], v.93, p.1510–1518, 2007.

ZHOU, H.; SKOLNICK, J. Protein structure prediction by Pro-Sp3-TASSER. **Biophys. J.**, [S.l.], v.96, n.6, p.2119–2127, 2009.

ZWANZIG, R.; SZABO, A.; BAGCHI, B. Levinthal's paradox. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.89, p.20–22, 1991.

INSTITUTO DE INFORMÁTICA UFRGS

PPGC

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

"MOIRAE: A Computational Strategy to Predict 3-D Structures of Polypeptides"

por

Márcio Dorn

Tese apresentada aos Senhores:

Prof. Dr. Hugo Verli
(UFRGS)

Prof. Dr. Leandro Krug Wives
(UFRGS)

Prof. Dr. Mario Inostroza Ponta
(USACH)

Vista e permitida a impressão.
Porto Alegre, 02/8/2012

Prof. Dr. Luis da Cunha Lamb
Orientador

Profa. Dra. Luciana Salete Buriol
Coorientadora