Secondary Structure-Based Template Selection for Fragment-Assembly Protein Structure Prediction

by

Jad ABBASS

Submitted in partial fulfilment of the requirements of

Kingston University for the degree of

Doctor of Philosophy

July 2018

[Page intentionally left blank]

Supervision:

Dr Jean-Christophe Nebel (First Supervisor)[1]

Prof Nashat Mansour[2]

Prof Souheil Khaddaj[3]

[1]Bioinformatics Research Group
Department of Computer Science
School of Computer Science & Mathematics
Faculty of Science, Engineering and Computing
Kingston University
Penrhyn Road
Kingston upon Thames
Greater London
KT1 2EE
United Kingdom

[2]Department of Computer Science and Mathematics
School of Arts and Sciences
Lebanese American University
Chouran
Beirut
P.O. Box 13-5053
Lebanon

[3]Component and Distributed Systems Research Group
Department of Computer Science
School of Computer Science & Mathematics
Faculty of Science, Engineering and Computing
Kingston University
Penrhyn Road
Kingston upon Thames
Greater London
KT1 2EE
United Kingdom

[Page intentionally left blank]

Abstract

Proteins play critical biochemical roles in all living organisms; in human beings, they are the targets of 50% of all drugs. Although the first protein structure was determined 60 years ago, experimental techniques are still time and cost consuming. Consequently, *in silico* protein structure prediction, which is considered a main challenge in computational biology, is fundamental to decipher conformations of protein targets. This thesis contributes to the state of the art of fragment-assembly protein structure prediction. This category has been widely and thoroughly studied due to its application to any type of targets. While the majority of research focuses on enhancing the functions that are used to score fragments by incorporating new terms and optimising their weights, another important issue is how to pick appropriate fragments from a large pool of candidate structures. Since prediction of the main structural classes, i.e. mainly-alpha, mainly-beta and alpha-beta, has recently reached quite a high level of accuracy, we have introduced a novel approach by decreasing the size of the pool of candidate structures to comprise only proteins that share the same structural class a target is likely to adopt. Picking fragments from this customised set of known structures not only has contributed in generating decoys with higher level of accuracy but also has eliminated irrelevant parts of the search space which makes the selection of *first models* a less complicated process, addressing the inaccuracies of energy functions. In addition to the challenge of adopting a unique template structure for all targets, another one arises whenever relying on the same amount of corrections and fine tunings; such a phase may be damaging to "easy' targets, i.e. those that comprise a relatively significant percentage of alpha helices. Owing to the sequence-structure correlation based on which fragment-based protein structure prediction was born, we have also proposed a customised phase of correction based on the structural class prediction of the target in question. After using secondary structure prediction as a "global feature" of a target, i.e. structural classes, we have also investigated its usage as a "local feature" to customise the number of candidate fragments, which is currently the same at all positions. Relying on the known facts regarding diversity of short fragments of helices, sheets and loops, the fragment insertion process has been adjusted to make "changes" relative to the expected complexity of each region. We have proved in this thesis the extent to which secondary structure features can be used implicitly or explicitly to enhance fragment assembly protein structure prediction.

[Page intentionally left blank]

To my father

[Page intentionally left blank]

Acknowledgment

I would like to express my sincere thanks to my first supervisor, Dr Jean-Christophe Nebel, for his smart guidance, invaluable advice, and extreme patience during my successes and struggles. Thank you so much for the devotion you have shown to me; I had never had similar support from any of my previous teachers. Also, I would like to thank my local supervisor Prof Nashat Mansour at the Lebanese American University in Beirut for his motivation in the pre-PhD phase and his support in publishing the literature review.

I would like to thank everyone who has explicitly or implicitly, intentionally or even unintentionally, helped me in some way to complete my doctoral degree no matter how tiny that help was. Amongst those, staff, colleagues, and friends at Kingston University and the Lebanese International University, and indeed my family members and friends.

Finally, I would like to dedicate this thesis to my father who has been my role model in patience, strength and wisdom; things I strongly needed during my PhD journey.

[Page intentionally left blank]

List of Publications

Chapter 2

- **Abbass, J.**, Nebel, J.-C., & Mansour, N. (2013). Ab Initio Protein Structure Prediction: Methods and challenges. In M. Elloumi & A. Y. Zomaya (Eds.), Biological Knowledge Discovery Handbook (pp. 703–724). Hoboken, New Jersey: John Wiley & Sons, Inc.

Chapter 4

- **Abbass, J.**, & Nebel, J.-C. (2015). Customised fragments libraries for protein structure prediction based on structural class annotations. BMC Bioinformatics, 16, 136.

Chapter 5

- **Abbass, J.**, & Nebel, J.-C. (2017). Reduced Fragment Diversity for Alpha and Alpha-Beta Protein Structure Prediction using Rosetta. Protein & Peptide Letters, 24, 215–222.

Chapter 6

- **Abbass J.**, & Nebel J.-C. (2018). Weighted proteins regions for fragment cardinalities based on secondary structure prediction. BMC Bioinformatics, to be submitted in summer 2018.

[Page intentionally left blank]

Glossary of Terms

3-mers/9-mers
3-mers and 9-mers represent protein substructures, also known as fragments, of length 3 and 9 respectively. Such fragments are extracted from proteins of known structures and represent the main building block used in fragment-based protein structure prediction computational tools.

Amino acid
A protein's building block which comprises besides the side chain (R) two main groups: amine (NH2) and carboxyl (COOH). The side chain is the only part that differs amongst the twenty amino acids. The important atom that links all the three groups together is called C-alpha, denoted as Cα.

Alpha helix
The most common and regular secondary structure motif found in proteins. It has a spiral-like shape stabilised by hydrogen bonds between each couple of **amino acids** that are located on the top of the other in the spiral, i.e. 3 to 4 **amino acids** away. An alpha helix has an average length of 10 amino acids.

Best decoy
The **decoy** that corresponds to the highest **GDT TS** score with the native structure regardless of the value of the energy function.

Beta strand
It is a "zig zag" shape motif found in proteins. When two or more beta strands are pleated on each other, they form the second most abundant secondary structure, beta sheet. Beta sheets are either parallel or antiparallel based on the orientation of the strands and, as **alpha helices**, are stabilised by **hydrogen bonds** however between **amino acids** of different strands. Average length of beta sheets is approximately 6 **amino acids**.

Coil
Any secondary structure that is neither an **alpha helix** nor a beta sheet is called a coil. A coil has no well-defined shape and it serves as either a terminal or a connecting "part" between two **alpha helices**, two **beta strands**, or an **alpha helix** and a **beta strand**. Its length can reach up to 15 **amino acids** and it is consequently very hard to predict its structure due to its high degree of spatial flexibility.

Chaperone
A special kind of proteins that help other proteins to fold.

Correlation coefficient
It describes the correlation between two datasets of **GDT**. Whereas the first dataset is related to the standard predictions, the second one is related to the customised ones; for

a list of 25 targets, for instance, each dataset comprises 25 values. Correlation coefficient ranges from -1 to +1; values that are less than -0.5 and greater than +0.5 are believed to reveal strong relationships.

Decoy

In order to cope with the large search space, most protein structure prediction computational tools rely on generating thousands of candidate structures, known as decoys, where each decoy represents, in principle, a different search trajectory. Typically, the decoy(s) with the lowest energy score is(are) then considered as the best prediction(s).

Denaturation

During the denaturation process, a protein loses its tertiary structure that corresponds to the native state due to external factors such as radiation, heat or strong acid/base. In cases when the original protein structure can be recovered, the reverse process is called renaturation.

Dictionary of secondary structures of proteins (DSSP)

It is a computer program that, when fed with a protein's spatial coordinates, assigns each **amino acid** to a secondary structure, mainly either **helix** (H), beta sheet (E) or **coil** (C).

Dihedral angles

Also known as torsion angles, they are mainly two angles, phi φ and psi ψ, that describe two important rotations in each **amino acid**. Protein's dihedral angles are able to define the backbone's fold. Since Rosetta keeps all remaining local structural parameters such as bond lengths and remaining angles at their "ideal" values during the sampling process, the only degrees of freedom/change are the torsion angles.

Domain

For relatively small proteins, the whole conformation is often considered a domain. Some larger proteins may comprise more than one domain if each domain can be seen as an independent folding unit that has its own hydrophobic core and its own function(s), i.e. has the characteristics of a typical **globular protein**.

E-value

E-value (Expect value), which is associated with the results of sequence alignment using **PSI BLAST** tool, is a statistical measure that is calculated via the score, i.e. the measure of similarity of aligning two sequences, the size of the protein in question and the size of the database being hit. It is the estimated number of times one would expect to see the same score, however, by chance. In the broad sense, the lower e-value, the more significant the hit is.

Electrostatic force
The electrostatic force between two atoms could be attractive or repulsive as a result of different or similar charges respectively.

Entropy
A thermodynamic metric that measures the amount of "randomness" or disorder in a certain system. During the folding process, a protein's entropy decreases.

Enzyme
Enzymes are the largest category of proteins. Their key role is to catalyse, i.e. accelerate, chemical reactions.

Eukaryotic cells
Eukaryotic cells contain a nucleus surrounded by a membrane. Organisms who have such cells are called Eukaryotes; humans are amongst this group.

Family
A protein's family is the set of proteins that are likely to have a common evolutionary origin. This is evidenced by either a sequence identity that typically exceeds 30% or a lower sequence identity, however, associated with similar function and/or structure.

First model
The **decoy** that corresponds to the lowest energy score.

Free energy
Also known as Gibbs free energy, free energy is a thermodynamic potential. Once a protein reaches a chemical equilibrium, typically, after the folding process ends, free energy's value reaches a minimal value.

Free modelling (FM) targets
Proteins that do not have explicit template structures due to the low sequence similarity between the target on a side and the entries belonging to the database of proteins of known structures on the other side. Such targets are typically computationally predicted using either *ab initio* or fragment-based approaches. They are considered quite challenging in the field of protein structure prediction. In this study, we may refer to FM targets as "hard" targets.

Heuristic
In large optimisation problems, a heuristic is an approach to reach an approximate solution by employing, for instance, an informative function to help exploring the search space. Whilst heuristics are usually designed for each specific problem, metaheuristics can be applied to a broad range of optimisation problems.

Hydrogen bond
A hydrogen bond is a non-covalent bond that is formed as an electrostatic attraction force between a hydrogen atom and another electronegative atom.

Globular protein
Such a protein is called so because of its "globe-shape". Globular proteins are water-soluble whereas the fibrous and membrane ones are not.

Monte Carlo methods
In optimisation problems, Monte Carlo methods are based on generating a relatively high number of random samples for the sake of attaining the optimal or near-optimal solution(s). A simple analogy can be described by randomly throwing several balls in the hope of reaching the lowest point in a landscape containing hills and valleys.

Native structure
The native structure of a protein is the conformation that corresponds to the lowest free energy; without reaching that unique structure, a protein, in principle, cannot perform its function(s),

p-value
Throughout this thesis, p-value or probability value, which is the probability of obtaining corresponding results by chance, is calculated as the probability that is associated with the Student's t-test value. The latter is obtained as a paired two-tailed distribution of both datasets: **GDT** of standard predictions and their corresponding **GDT** of our customised predictions.

Peptide
The main difference between a protein and a peptide is the size as peptides are typically shorter than proteins and consequently are not involved in as many biochemical reactions as proteins do.

Peptide bond
A covalent bond that links two **amino acids** together.

Polypeptide
A protein may consist of more than one polypeptide/chain and a polypeptide may not have a well-defined compact shape. All proteins are polypeptides, but the inverse is not always true; for instance, nylon is classified as a polypeptide, but it is not a protein.

Profile
A protein **family's** profile is considered as a kind of pattern that shows how conserved each position in the **amino acid** sequence is. It is the result of a multiple sequence alignment of query protein's sequence against a certain database. Usage of profile can be further extended whenever additional search iteration(s) are employed; a sequence's profile can be used instead of a single sequence to conduct wider search in a database.

Quality assessment tools
These tools are developed to help users to select the "best" candidate structure(s) amongst a large number of generated **decoys**. Such tools are considered an alternative to selecting the **decoy**(s) with the lowest energy score. They allow dealing with the inaccuracies of energy functions.

Protein data bank (PDB)
The protein data bank is the world's largest repository of proteins of known structures. Structures of the deposited proteins are represented as the spatial coordinates of the atoms which are determined via wet laboratory techniques such as X-ray crystallography, nucleic magnetic resonance, and electron microscopy.

Ramachandran map probability
Each **amino acid** has an associated probability for each possible range of value for each **dihedral angle**, namely phi φ and psi ψ, based on the statistical data collected from the **PDB**.

Residue
Throughout this thesis, residue and **amino acid** are used interchangeably.

Root mean square deviation
A metric that evaluates the structural similarity between two proteins' conformations. It includes an optimisation superimposition of both proteins. The value of the root mean square deviation is the average of the Euclidean distances between each pair of atoms.

Rotamer
A rotamer represents the "ideal" coordinates of side-chains of proteins, i.e. low-energy conformations of side chains collected from statistical data of proteins of known structures.

Similar amino acids
In contrast to "identical amino acids", which requires two amino acids to be exactly the same, "similar amino acids" can be associated to a pair that shares some physiochemical properties such as hydrophobicity.

Structural class
A structural class is a high-level structural classification for proteins that is based on the abundances of the secondary structure elements and their organisation. A **domain** is usually classified into three main classes: mainly-alpha, mainly-beta or alpha-beta.

Template-based modelling (TBM) targets
Proteins that have at least one explicit template structure due to the relatively high sequence similarity between the target on a side and the template structure(s) on the other side. Such targets are typically computationally predicted using comparative

modelling. They are not considered not challenging in the field of **protein structure prediction**. In this study, we may refer to TBM targets as “easy” targets.

Tertiary structure
Three-dimensional **(3D)** structure and tertiary structure are used interchangeably in this thesis.

Turn
A short **coil** - no longer than 5 **amino acids** – that is stabilised by **hydrogen bonds**. The main difference between a **coil** and a turn is that the former has no well-defined shape and its length can reach up to 15 amino acids. Throughout this thesis, secondary structure predictions are based on the well-known three-state: h for **helix**, e for beta sheet and c for everything else, i.e. turns are not treated differently; everything different from helices and sheets are simply **coils**.

Index of Abbreviations

2D	Two Dimensional
3D	Three Dimensional
AA	Amino Acid
BLAST	Basic Local Alignment Search Tool
CASP	Critical Assessment of protein Structure Prediction
CATH	Class Architecture Topology Homology
CS	Chemical Shifts
DNA	Deoxyribonucleic Acid
DSSP	Dictionary of Secondary Structures of Proteins
EM	Electron Microscopy
E-value	Expected Value
FF	Force Fields
FM	Free Modelling
FSS	Few Secondary Structures
GA	Genetic Algorithms
GDT TS (or GDT)	Global Distance Test Total Score
HH	Hydrophobic Hydrophobic
HMM	Hidden Markov Model
HP	Hydrophobic Polar
IDP	Intrinsically Unordered Protein
IUP	Intrinsically Unstructured Protein
MC	Monte Carlo
MD	Molecular Dynamics
MG	Molten Globule
MODAS	MODular Approach for Structural class prediction
MSA	Multiple Sequence Alignment

NMR	Nuclear Magnetic Resonance
NN	Neural Networks
NOE	Nuclear Overhauser Effect
PDB	Protein Data Bank
PSI BLAST	Position Specific Iterative Basic Local Alignment Search Tool
PSP	Protein Structure Prediction
QA	Quality Assessment
QM	Quantum Mechanics
RMSD	Root Mean Square Deviation
SA	Simulated Annealing
SCOP	Structural Classification of Proteins
SP	Small Proteins
SS	Secondary Structure
SVM	Support Vector Machine
TBM	Template-based Modelling
TM	Transmembrane
TS	Tertiary Structure
vdW	van der Waals

List of Figures

List of Tables

Contents

1 Introduction

Everyone has heard of DNA; it is a long chain that comprises our genetic code, believed to be unique for each of us and sometimes referred to "the secret of life". Around 1.5% of human DNA codes for protein sequences; proteins have much shorter chains, but they are more structurally and functionally diverse and complicated. The origin of the word 'Protein' is Greek ("*proteios*"). It means "of primary importance" and it was suggested in the 19th century. Despite the fact that very little was known about proteins then, eventually the name proved quite appropriate; proteins are the cell's basic building blocks, constitute 20% of our human body, are involved in more than 4,000 biochemical reactions, are targets of more than 50% of drugs and when misfolded may lead to fatal diseases. Although critical biological functionalities have been found to be associated to the non-coding sequences of DNA (Sloan et al., 2016), a better understanding of the products expressed by the 20,000 human genes is certainly required to improve human health.

Protein-related fields of research include prediction of functions, structures, interactions and interfaces. However, a protein's 3D structure may be seen as the key to most protein mysteries since it determines its function and dictates possible interactions. The main topic of this thesis is to predict a protein's structure via computational means. A protein is originally a linear sequence of amino acids that folds into a generally unique conformation where free energy is believed to be minimal. This process typically takes place inside a cell, i.e. *in vivo*; since scientists cannot accurately monitor that process, they try to "replicate" it in wet laboratories, i.e. *in vitro* so they can capture the spatial coordinates of the native structure. Despite advancements in computational biology for more than two decades, *in silico* – i.e. using computer simulations - techniques for protein structure prediction haven't been classified as "trusted" by biologists and drug designers yet (Moult, Fidelis, Kryshtafovych, Schwede, & Tramontano, 2018).

From a drug design perspective, determination of a protein's native structure represents a crucial step since it allows gaining important insights of the molecular mechanisms involved in many diseases (Ramirez-Alvarado, Kelly, & Dobson, 2010). Despite the advancements achieved in both wet laboratories and computational techniques, protein structure determination still faces many challenges. Bioinformatics is usually considered as "the last chance" when neither X-ray crystallography, nor

Nuclear Magnetic Resonance (NMR), nor Electron Microscopy (EM) can be used due to time, cost or/and experimental constraints. Whereas performing protein folding simulation conforming to Newton's second law may appear as an attractive approach, it is only practical when applied on very small targets while using state-of-the-art supercomputers and/or grid computing (Baker, 2014). Consequently, many current computational methods rely on Monte Carlo simulations and heuristic search techniques besides reduction of amino acids and energy functions' representations (Kmiecik et al., 2016).

In this study, we will be exploring the different state-of-the-art methodologies used to predict a protein's structure, especially fragment-based approaches. We propose three different novel ideas applied to a state-of-the-art tool, called Rosetta, which are supported by tangible improvements.

The rest of this short chapter is organised as follows. The next section introduces the core of our research area: protein folding and structure prediction. Afterwards, we present concisely the aim and objectives and our scientific contribution in sections 1.2 and 1.3 respectively. The outline of the structure of the whole thesis is presented in the last section of the chapter.

1.1 Protein Folding and Structure Prediction

Typically, a protein is useless, sometimes harmful, unless it folds into its generally unique shape. Such a process takes place over a timescale of microseconds in nature although the number of all possible conformations is tremendous; such dilemma has been under question till now (Dill & Chan, 1997; Dill & MacCallum, 2012; Levinthal, 1968; Zwanzig, Szabo, & Bagchi, 1992). Although the final structure is the most "crucial part", the folding pathway has been under thorough study since it may reveal important clues, mainly to help computational biologists mimic the real trajectory towards the native structure (Dill, 1985; Voelz, Bowman, Beauchamp, & Pande, 2010). Probably the first finding in this regard for globular proteins notes that hydrophobic amino acids tend to be in the centre of the structure to avoid the surrounding water molecules, whilst the hydrophilic ones prefer to stay in contact with the external aqueous environment, see Figure 1.1.

Studies and proposed theories on protein folding have been published from different scientific perspectives: chemical, physical and biological (Dill, Ozkan, Shell, & Weikl, 2008; Luo, 2014; Scheraga, 2015). Computational techniques have played a key role by running simulations of that process to mimic nanosecond by nanosecond how atoms interact according to the "standard" Newton's second law. Although successful attempts have been recorded, only very powerful supercomputers and grid computing systems were able to achieve success to such experiments (Tyka et al., 2011; Voelz et al., 2010).

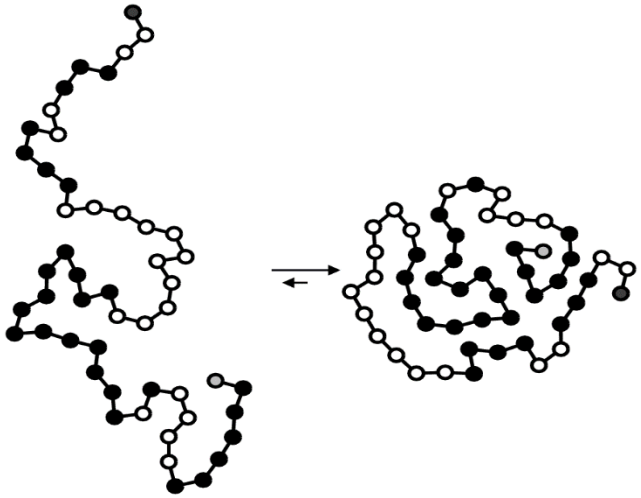

Figure 1.1: Pictorial description of globular protein folding. The left part represents the primary sequence, i.e. the linear chain of amino acid whereas the right part shows the folded structure. The black-filled, white-filled, dark grey-filled, and light grey-filled spheres represent the hydrophobic, hydrophilic amino acids, C terminal and N terminal respectively. For the sake of simplicity, the figure is shown in 2D.

Christian Anfinsen – one of the pioneers in the field of protein structures – has formulated two notable theories: the first one states that the native structure is the one that has the lowest free energy value (Anfinsen, Haber, Sela, & White, 1961), the second describes protein folding as a pure physical process, i.e. the tertiary structure can be solely determined by the sequence of amino acids (Anfinsen, 1973), see Figure 1.2. The above two principles represent the basis for the most challenging computational technique known as *ab initio* protein structure prediction. From the first theory's perspective, Protein Structure Prediction (PSP) is an optimisation problem where the energy function plays the role of heuristic as an attempt to reach the global minimum energy in the tremendous search space. Anfinsen's second theory has paved the way to computationally represent an approximate value of the interactions that take place

amongst the atoms and amino acids without taking into consideration any external effects.

Computational approaches used to determine a protein's structure can be categorised into two main groups: template-based and template free modelling; whereas the first one relies mainly on the known proteins structures deposited in the world's largest repository – the Protein Data Bank (PDB) - by trying to find either some level of sequence-sequence similarity or sequence-structure compatibility between a template conformation and the target in question, the second group, i.e. template free – also known as *ab initio* – relies solely on both of Anfinsen's theories. *Ab initio* approaches are closer to the natural case than template-based methods where most advances have been focusing on detecting more remote homologues and better modelling sequence-structure compatibility. In their turn, template-based techniques are divided further into to sub-categories: homology modelling and threading.

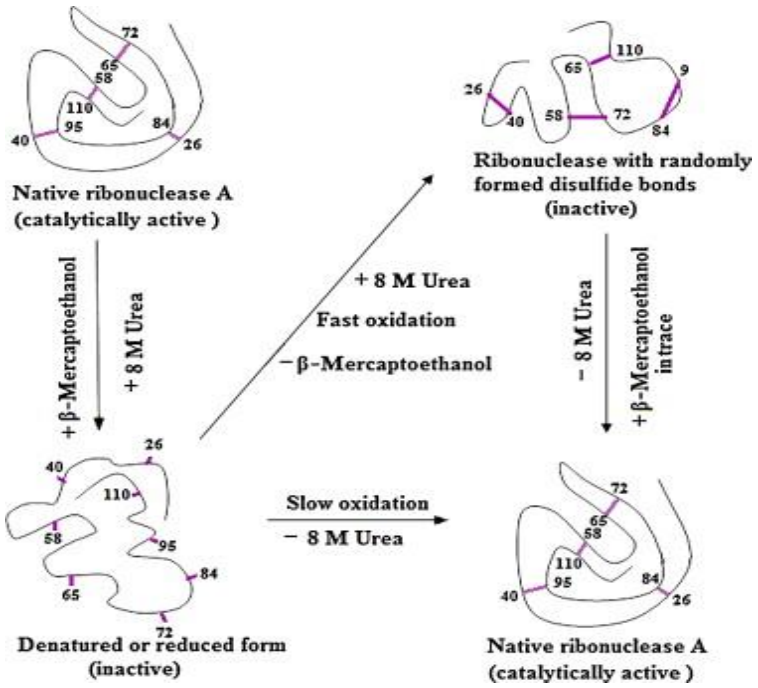

Figure 1.2: A depiction of Anfinsen's experiment; the native structure (top left) was denatured to form two inactive shapes (bottom left and top right). Both were again biologically activated (renatured) and the protein returned to its native shape. Taken from (Amani & Naeem, 2013).

Homology modelling or comparative modelling is considered the simplest way to build the target in question and it is based on a quite old hypothesis: similar sequences infer similar structure (Browne et al., 1969). Whenever sequence similarity exceeds 30%, models with good accuracy are typically expected (Lam, Das, Sillitoe, &

Orengo, 2017). Figure 1.3 shows the process of comparative modelling; the sequence alignment was performed using ClustalW (Larkin et al., 2007) and the model was built using MODELLER (B Webb & Sali, 2014).

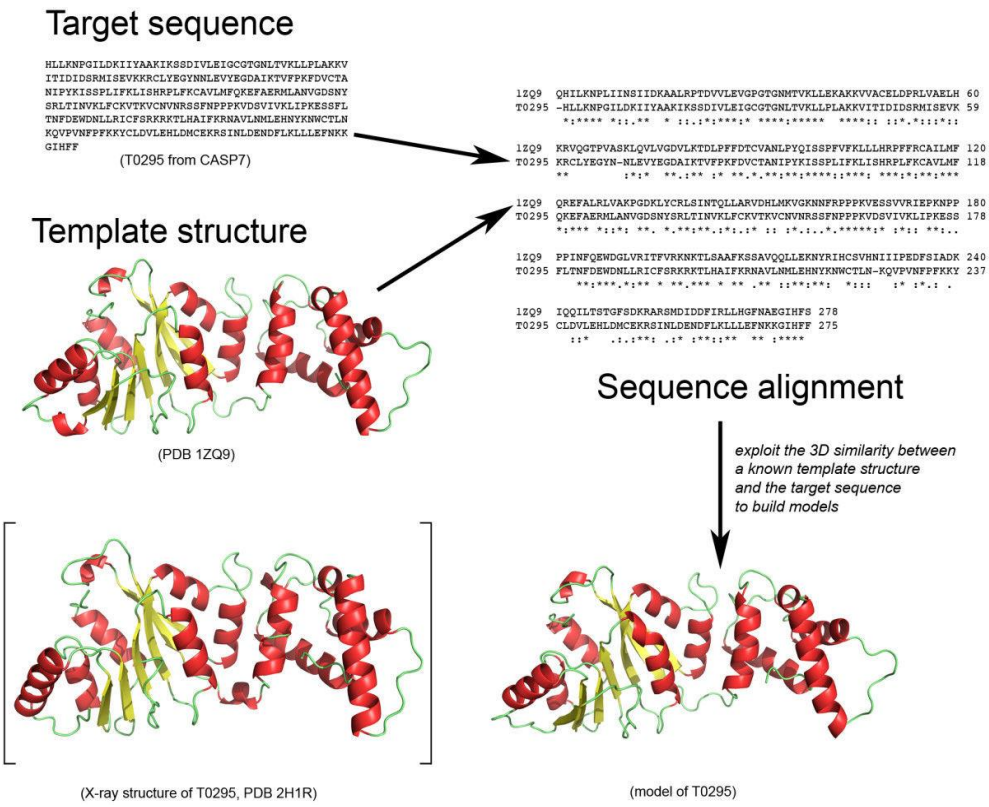

Figure 1.3: A simplified pictorial illustration of the homology modelling process. The top left part shows the target sequence as well as a template structure that was chosen due to the high sequence alignment similarity shown on the top right. The native structure of the target T0295 is shown between square brackets whereas the built one using the template is displayed next to it. Taken from (di Luccio & Koehl, 2011).

Fold recognition or threading is a more complicated and computationally expensive template-based modelling and is typically used whenever comparative modelling fails; even if no remarkable sequence similarity has been detected, a target may still fit into one of the known structures (Chothia, 1992). Threading techniques rely on fitness scores as target's amino acids are placed on known structures to evaluate how convenient and compatible those structures are. Most threading techniques do not model the whole target, rather the core regions only; see Figure 1.4.

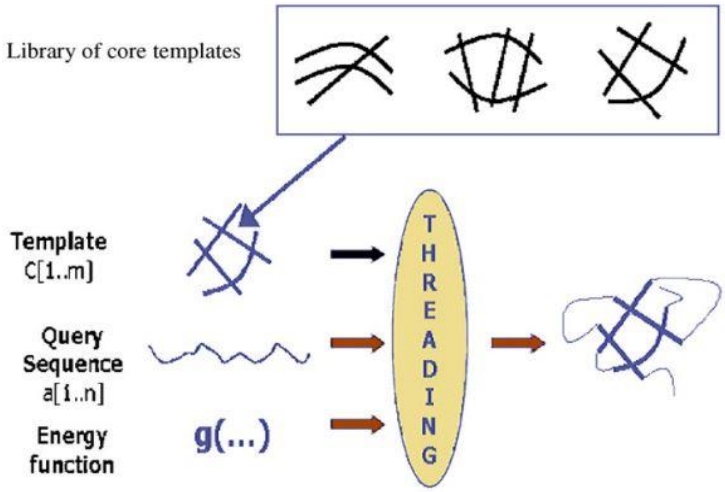

Figure 1.4: Simplified threading process. The best core template is chosen based on the score of the energy function. The size of the query sequence is n, whereas the size of the core template used for threading is m. Since n is larger than m, the remaining regions are built using different techniques such as *ab initio*. Taken from (Ngom, 2006).

Ab initio methods are by far the hardest approach for PSP. As their name implies, in these approaches, proteins are built from scratch. "Standard" *ab initio* algorithms mimic the natural folding process by using Force Fields (FF) – an approximation of the quantum mechanical representation of the interactions amongst atoms - however, due to the high computational cost their usage has been limited to small proteins (Khoury, Smadbeck, Kieslich, & Floudas, 2014).

An in-between approach that combines the strength of both template-based and *ab initio* modelling is fragment-based protein structure prediction. It is able to predict template-free targets but is not as computationally expensive as "pure" *ab initio* methods. Instead of having a single amino acid as the unit of construction, a short sequence of amino acids – treated as a rigid part – is taken into consideration. Such approaches have been the target of very active research for the sake of their enhancement and improvement as they were ranked the best in the latest blind competition: Critical Assessment of the Structure Prediction of proteins – round 12 – (CASP12) in 2016.

1.2 Aim and Objectives

The main aim of this study is to improve fragment-based protein structure prediction, taking advantage of the state-of-the-art Rosetta tool (Leaver-Fay et al., 2011), by incorporating new parameters and criteria whenever template structures and

fragments are chosen to build complete conformation models. The fundamental objectives are concisely presented as follows.

- Creating smaller but more relevant template structure sets, i.e. allowing focusing on the most promising parts of the search space:

  For the sake of sampling as many decoys as possible and therefore covering as much of the search space as possible, fragment-based approaches rely on a relatively large set of PDB's structures. For instance, Rosetta uses a group of 16,800 template structures of average size 257 amino acids; such a number has been able to let Monte Carlo simulations cover a large number of possible conformations (Gront, Kulp, Vernon, Strauss, & Baker, 2011). However, out of the decoys produced, it turned out that many of them are quite far from the native structures although they represent local energy minima (Kim, Blum, Bradley, & Baker, 2009). In this thesis, we will show that in many cases using only 20% of Rosetta's standard set of template structures allows not only producing decoys of better quality, but also reducing the number of irrelevant regions where search trajectories end.

- Incorporating proteins' structural class prediction as an additional and valuable criterion for selection of fragments:

  In all state-of-the-art fragment-based tools such as I-TASSER (Y. Zhang, 2008), FragFold (Jones, 2001) and Rosetta (Das & Baker, 2008), the procedure for the selection of fragments relies mainly on different techniques of sequence alignment and additional criteria, such as secondary structure prediction and Ramachandran map probabilities. In this thesis, we present a new factor that restricts the usage of fragments based on the structural class of their sources, that should match with the target's structural class prediction. Such a factor was able to play the role of a "preliminary filter" of all fragments before Rosetta's standard filters are applied; tangible improvements were recorded over state-of the-art prediction methods.

- Taking advantage of a proteins' sequence-structure correlation associated with the various secondary structures to create customised fragment files, where the number of candidate fragments varies based on the predicted secondary structure

so that the effort in modelling each target's region is customised according to its needs:

Rosetta uses 25 9-mers and 200 3-mers for each position in the target to be chosen randomly whilst the conformations are being built. Such a strategy – to adopt the same number of fragments at each position - is common amongst all popular fragment assembly methods. However, independent and thorough studies have shown that the sequence-structure relationship is not static for all secondary structures, specifically not for short sequences (Bystroff, Simons, Han, & Baker, 1996; de Oliveira, Shi, & Deane, 2015; Fiser et al., 2000; Sibanda & Thornton, 1985; Vanhee et al., 2011). Owing to that known fact, we have developed a novel approach to build fragment files so that, for instance, the number of candidates falls sharply whenever an alpha helix – an easier protein substructure to predict – is predicted to occur along the length of the fragment.

- Applying an appropriate "amount" of corrections and tuning to an initial model to prevent excessive changes which may have a damaging effect:

  Rosetta uses 200 3-mers for each position in the target of interest as an attempt to explore neighbouring regions. However, we will demonstrate in this thesis that for a category of targets such a large number of fragments of size 3 cause "damage" to some parts of the conformation that had already reached a decent accuracy during the coarser structure prediction phase.

- Tackling energy function inaccuracies by narrowing the size of the explored area:

  Exploration-exploitation trade-off is a common issue in all optimisation problems, especially in protein structure prediction using fragment assembly techniques (Simoncini, Schiex, & Zhang, 2017). However, reaching a fair compromise between these does not only raise the probability of reaching "good" regions in the search space but also narrows the gap between the decoys with low energy scores and decoys with high accuracy. In this research, we have decreased the level of exploration in our three contributions. However, each time this was done in a different way, which makes the selection of the *first models* – models that are associated with the lowest energy score – a more accurate process.

1.3 Scientific Contribution

In this thesis, novel ideas are presented leading to improvements over the standard Rosetta protein structure predictions. Our novel ideas result in the following scientific contributions:

- A novel fragment selection process where usage of template structures is restricted to those who share the same structural class prediction as the target (chapter 4). This novel idea was the basis of our contribution to CASP12 – under the name of "Rosetta_at_Kingston" group – as we had the opportunity to compete against the formal Rosetta research group and were able to show better results for 40% of the targets despite the huge gap between their computational and human resources and ours.

- A structure refinement process depending on a target's structural class prediction (chapter 5). We have shown that the standard number of 3-mers, i.e. 200, which is used in the structure refinement phase, is not only unnecessary but also destructive for alpha and alpha-beta proteins. Indeed, for those classes, the main 9-mers insertion phase is sufficient to "deliver" conformations close to the native-like structure. As a consequence, refinement only requires being light touch.

- A protein structure prediction process which takes advantage of the sequence-structure correlation which is present amongst the three different secondary structures to select the number and diversity of possible fragment alternatives (chapter 6). The sequence-structure correlation amongst the three different secondary structures has been established for a long time (Sibanda & Thornton, 1985); alpha helices are very "conservative" whilst loops tend to have a large range of structural variety and beta strands are somewhere in between. As a consequence, it is proposed that fragments that are predicted to be either pure helices, strands or loops should have an increasing number of available candidates. Adopting this novel approach to create fragment files allows the structure prediction process to focus on complex regions by conducting extensive fragment insertions, while limiting the number of the insertions in the simpler regions.

1.4 Thesis Outline

This thesis is divided into 7 chapters as follows:

In this chapter, we have presented a concise description of the novelty of our findings after a short definition of the problem.

Chapter 2 is dedicated to a thorough literature review of protein folding and protein structure prediction. An earlier version of this chapter was published as a book chapter by John Wiley and Sons (Abbass, Nebel, & Mansour, 2013).

Chapter 3 describes a popular and challenging fragment-based protein structure prediction method called Rosetta (Lyskov et al., 2013) that has been ranked the best amongst its competitors.

Chapter 4 proposes the adoption of a small but customised template structures set for each group of targets that share some "global" properties. We have demonstrated that based on the target's structural class prediction, an ad hoc fragments library should be built to produce better decoys by exploring less but exploiting deeper regions that are likely to be near the native-like structure. This work has been published as an article in BMC Bioinformatics (Abbass & Nebel, 2015).

Chapter 5 relies on the same principle as chapter 4, i.e. the structural class of the target is predicted first. However, the standard fragment library is kept but the number of 3-mers has been adjusted accordingly. This contribution has been published in a journal paper in Protein Peptides and Letters (Abbass & Nebel, 2017).

Chapter 6 illustrates how secondary structure prediction can play an additional role besides its original one as a factor whenever it comes to choosing a fragment; the number of available fragments varies between the target's positions based on the secondary structure that it is likely to adopt starting at those positions. This work will be the core of a future publication in a bioinformatics journal.

Chapter 7 presents the conclusions, discussion and future works.

2 Literature Review

2.1 Introduction

This chapter presents an exhaustive review of protein structure prediction starting from an introduction to the structure of proteins (section 2.2), an overview of the main theories of protein folding (section 2.3) and ending in a thorough investigation of the different means of protein 3D structure determination (section 2.4). Section 2.4 comprises three sub-sections; experimental techniques (section 2.4.1), protein 3D structure prediction (section 2.5) and CASP competition (section 2.4.3). Protein 3D structure prediction (section 2.5) is the largest one, which discusses the most common computational means used in this regard, such as comparative modelling (section 2.5.4.2), fold recognition (section 2.5.4.3), *ab initio* modelling (section 2.5.4.4) and fragment-based techniques (section 2.5.4.5). Sections related to different *ab initio* and fragment-based techniques are described in greater detail.

2.2 Overview on Protein Structures

Proteins represent approximately 20% of a eukaryotic cell's weight, that is, the largest percentage after water. They are involved in the most critical functions: structural proteins are an organism's basic building blocks; enzymes, the largest class of these, are known to be involved in 4,000 biochemical reactions (Bairoch, 2000), and transmembrane proteins are essential in maintenance of the cellular environment.

Proteins are sequences of amino acids which fold through a high-speed spontaneous process into a unique conformation; this conformation typically represents a global energy minimum (Anfinsen et al., 1961). Proteins are so small that they cannot be seen via optical microscopes. Typically, the sizes of proteins range from about 3 to 10 nanometres, and finding their structure experimentally is relatively difficult and expensive. Such experiments are usually conducted by using either X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy or Electron Microscopy (EM). Proteins are initially built from a sequence of amino acids; however, the length of this chain varies from tens to many thousands of amino acids. For instance, insulin is a protein of 51 amino acids only, while another protein called titin has an approximate length of 28,000.

Besides amino acids and primary structures that are discussed in the next subsections, secondary structure and tertiary structures are elaborately described in sections 2.2.3 and 2.2.4 respectively.

2.2.1 Amino Acids

All amino acids have the same molecular structure; a central carbon atom called Cα, an amine group NH2, a carboxyl group COOH, a hydrogen atom H, and a side chain denoted as R (See Figure 2.1). Amino acids differ in structures of their side chains.

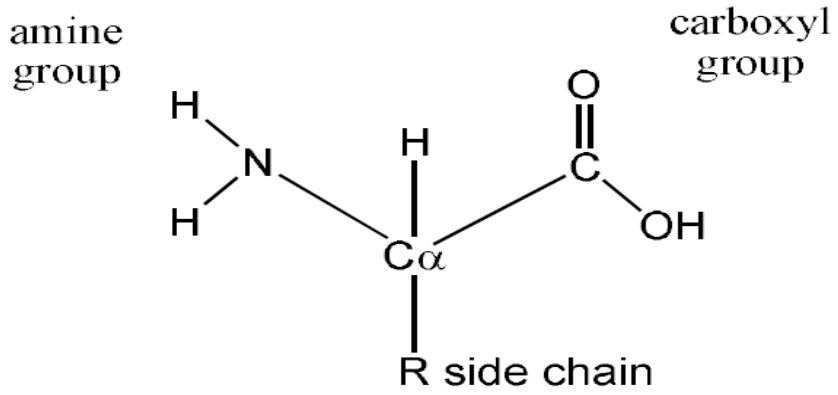

Figure 2.1: A template structure of an amino acid. All components, except for the side chain, are common amongst all amino acids. Taken from ("Proteomics," 2007).

There are 20 different amino acids; the only differences are the chemical structures and characteristics of their side chains (See Table 2.1 for a list of all amino acids). Amongst those characteristics, polarities and charges are the most important, since they determine the way they interact with other chemical compounds and the surrounding water molecules. According to their polarity and their charges, any amino acid can be classified into one of two main categories. (1) Nonpolar amino acids: have no or little connection with water molecules and for this reason they are also called "hydrophobic". Accordingly, hydrophobic amino acids try to occupy the space inside the protein molecule, in an attempt to avoid any contact with any surrounding H_2O molecules. In many proteins, such amino acids are dominant. (2) Polar amino acids, also known as hydrophilic, are, as the name implies, more soluble in water and tend to be placed on the exterior of proteins to be in contact with water; Table 2.1 shows all 20 amino acids along with their chemical properties.

Table 2.1: List of the 20 amino acids' names, abbreviations, symbols and chemical characteristics.

Amino acid	3-letter abbreviation	1-letter symbol	Chemical characteristics
Alanine	ALA	A	Nonpolar, hydrophobic
Arginine	ARG	R	Polar, hydrophilic
Asparagine	ASN	N	Polar, hydrophilic
Aspartic acid	ASP	D	Polar, hydrophilic
Cysteine	CYS	C	Polar, hydrophilic
Glutamine	GLN	Q	Polar, hydrophilic
Glutamic acid	GLU	E	Polar, hydrophilic
Glycine	GLY	G	Polar, hydrophilic
Histidine	HIS	H	Polar, hydrophilic
Isoleucine	ILE	I	Nonpolar, hydrophobic
Leucine	LEU	L	Nonpolar, hydrophobic
Lysine	LYS	K	Polar, hydrophilic
Methionine	MET	M	Nonpolar, hydrophobic
Phenylalanine	PHE	F	Nonpolar, hydrophobic
Proline	PRO	P	Nonpolar, hydrophobic
Serine	SER	S	Polar, hydrophilic
Threonine	THR	T	Polar, hydrophilic
Tryptophan	TRP	W	Nonpolar, hydrophobic
Tyrosine	TYR	Y	Polar, hydrophilic
Valine	VAL	V	Nonpolar, hydrophobic

Two amino acids can be joined together to form one molecule called a dipeptide, through a peptide bond between the carbon atom of the carboxyl group COOH of the first amino acid and the nitrogen atom in the amine group NH_2 of the second one. Figure 2.2 shows how a dipeptide can be formed by a peptide bond between two amino acids.

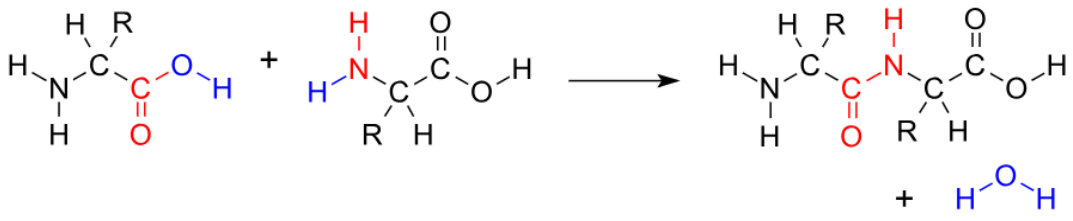

Figure 2.2: Dipeptide formation and release of a water molecule. The peptide bond takes place between the carbon atom in the carboxyl group of the first amino acid and the nitrogen atom in the amine group of the second amino acid.

2.2.2 Primary Structure

Many peptide bonds lead to the formation of a chain of amino acids, formally known as polypeptide. Except for some cases where a polypeptide doesn't satisfy a protein's characteristics, a polypeptide is often used as another name of what it is called linear, or primary, structure of a protein. By convention, primary structures are represented as sequences of one-letter symbols where each symbol stands for an amino acid in the chain. Also by convention, the left end is the N atom and the right end is the C atom. The chain of carbon and nitrogen atoms connected by peptide bonds is called the backbone of the protein. The upper part of Figure 2.5 shows an extended primary structure as well as the freedom of movement of the chain which are discussed in section 2.2.4.

2.2.3 Secondary Structure

A protein can be seen as a sequence of secondary structures, as they constitute the main components of the tertiary structures. Most often, secondary structures are formed due to the occurrence of hydrogen bonds between oxygen and hydrogen atoms that maintain their spatial stabilities. There are two common substructures that can be found in folded chains; alpha-helices and beta-strands, and they are joined by structures called loops. Alpha-helices, beta-strands, and loops constitute the secondary structure elements. Discovery of alpha helices and beta sheets took place in 1950 by Linus Pauling and co-workers (Pauling, Corey, & Branson, 1951). Automated/computerised classification of secondary structures is back to 1983 when the Dictionary Secondary Structures of Proteins (DSSP) was published (Kabsch & Sander, 1983); this was slightly improved in 2002 (Andersen, Palmer, Brunak, & Rost, 2002). However, DSSP is still the standard tool used in the Protein Data Bank (PDB) (P. W. Rose et al., 2017) to classify secondary structures of proteins based on the coordinates of the amino acids' atoms.

2.2.3.1 Alpha Helices

An alpha helix is a spiral-like structure where the side chains of the amino acids occupy the exterior part. The number of amino acids per turn is approximately 3.6 and the rotation of each amino acid is estimated as 100°. Since an alpha helix has a structure similar to a screw, amino acid number i + 4 is located approximately above amino acid number i. Therefore, the hydrogen bonds are between oxygen atoms in the CO group on amino acid i and hydrogen atoms of the NH group in amino acid i+ 4. Accordingly,

such a type of alpha helix is often known as 4-alpha helix and its structure is presented in Figure 2.3.

There are also two rare types of alpha helices called 3_{10} and Π alpha helices, having the 3 or 5 amino acids per turn respectively. The direction of the spiral determines whether an alpha helix is right-handed or left-handed, abbreviated respectively as α R-helix and α L-helix. However, the latter is very rare.

2.2.3.2 Beta Sheets

A beta sheet is formed by two expanded sequences of amino acids and is also maintained by hydrogen bonds between oxygen atoms in the CO group in the first sequence and hydrogen atoms of the NH groups in the other sequence. A beta sheet can be either parallel or antiparallel based on the directions of the amino acid sequences (See Figure 2.4). Beta sheets can contain more than two amino acid sequences, increasing the width of the structure. Antiparallel beta sheets are slightly more stable due the shorter distance between the hydrogen and oxygen atoms which makes the hydrogen bond stronger.

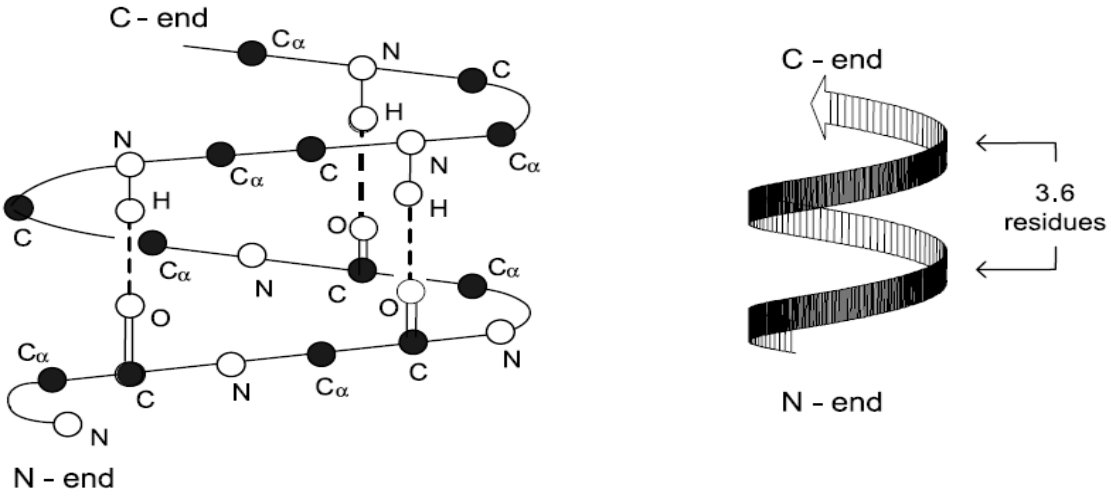

Figure 2.3: Illustrates a right-handed alpha helix in detail (left side) and symbolic representation (right side). Note: Side chains are not shown for clarity purposes. Taken from ("Proteomics," 2007).

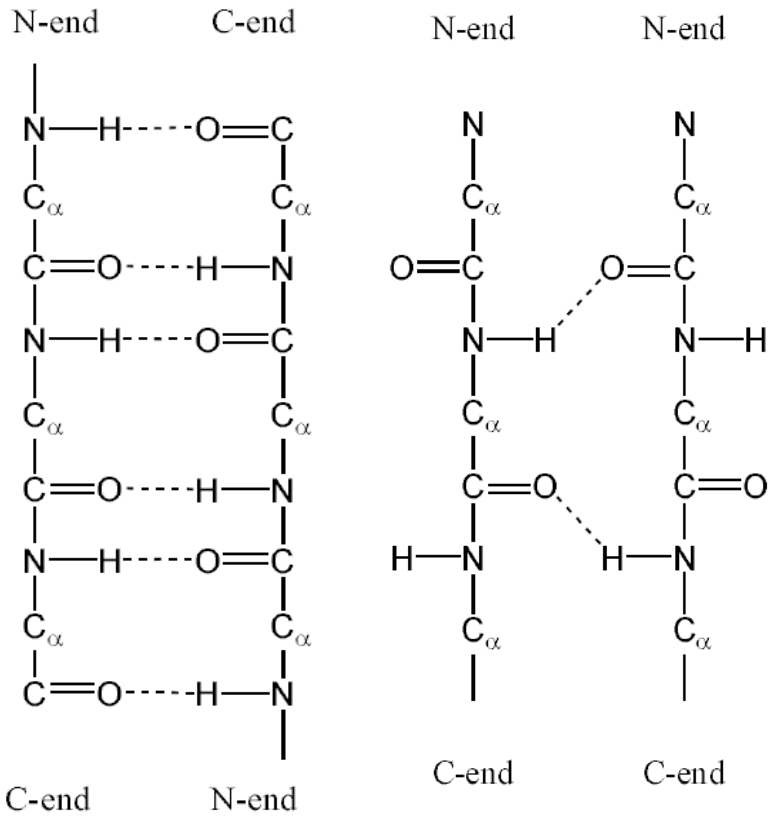

Figure 2.4: Antiparallel beta sheet structure (left side) and parallel beta sheet structure (right side). Note: Side chains are not shown for clarity purposes. Taken from ("Proteomics," 2007).

2.2.3.3 Turns

Turns are also known as loops and are considered as the third dominant secondary structure type, however, in contrast to alpha helices and beta sheets, turns are non-repetitive motif elements. They are in charge of the globular shapes of proteins and of reversing the direction of the amino acid chain. Turns contain polar and charged residues and are found mainly on the surface of a protein. According to the number of residues that constitute a turn, it can be classified into 5 types: delta-turn, gamma-turn, beta-turn, alpha-turn, and pi-turn. Gamma-turns, can be divided into two subcategories: inverse and classic (Bystrov, Portnova, Tsetlin, Ivanov, & Ovchinnikov, 1969), and are believed to be the most common in proteins and play a critical role in the folding process. Furthermore, they are responsible for stabilizing beta-strands before they transfer into beta-sheets (Milner-White, 1990).

2.2.3.4 Coils

Coils are simply unordered secondary structures; they have no defined arrangement in the DSSP. In the DSSP legend on the PDB website, such regions are defined as "empty: no secondary structure assigned"

2.2.4 Tertiary and Quaternary Structures

Tertiary structure is essential since it is the final three-dimensional conformation of a protein. It has been shown that the tertiary structure is unique for a protein and typically corresponds to a global minimum energy (Anfinsen et al., 1961). The heart of the protein folding problem lies in predicting the appropriate tertiary structure. The peptide bond that connects the carbonyl C of the i^{th} amino acid to the alpha amine N of the $(i+1)^{th}$ amino acid forms with the four neighbouring atoms a planar arrangement. Accordingly, these 6 atoms reside and move in a plane. Cα-C, as seen in Figure 2.5, serves as an axis of rotation for the first plane and Cα-N serves as an axis of rotation for the second plane. These two angles, which are independent of each other, are called ***phi*** (φ) and ***psi*** (ψ) respectively and they can vary from -180^o to $+180^o$. They are also known as dihedral angles or torsion angles. The side chain can also rotate around the Cα with an angle called *chi* (χ). Other rotations may also take place within the side chain itself and are referred as *chi1* (χ_1), *chi2* (χ_2), and *chi3* (χ_3). (Further information regarding restrictions of the degree of freedom of the main angles is explained in section 2.3.3 – Ramachandran plot). Figure 2.6 shows the tertiary structure of conjugative transfer protein (PDBID: 4EW7).

Not all proteins are composed of one polypeptide chain; actually, many of them are made up of two, three, four, five, six or even more chains and are known as dimers, trimers, tetramers, pentamers, hexamers and so on. The structural assembly of more than one chain is known as the quaternary structure of a protein. The subunits that make up such complex structures, also known as monomers, are not necessary identical, as shown in the bottom three template structures in the left part of Figure 2.7. On the other hand, when the subunits are the same, the quaternary structure is likely to be symmetric, as shown in the right part of Figure 2.7.

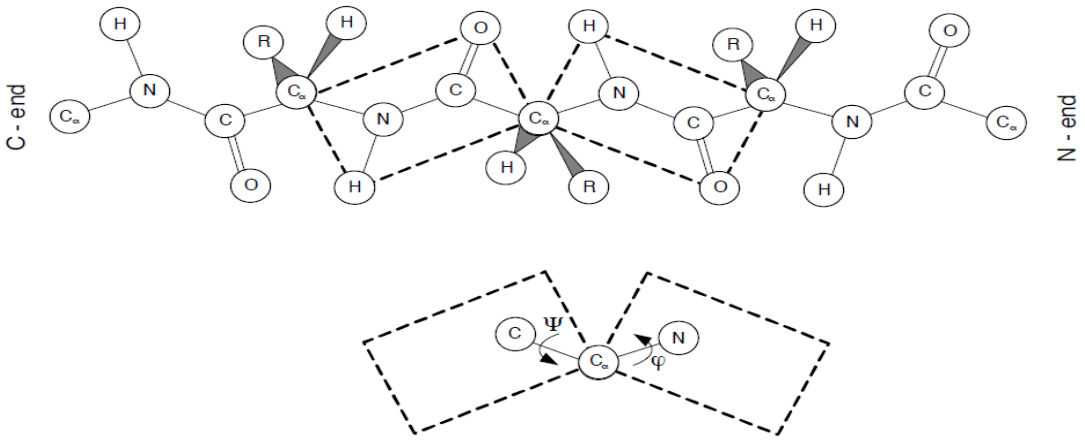

Figure 2.5: Extended chain of five residues showing the six atoms that stay in a planar conformation as well as the rotations that correspond to *phi* (φ) and *psi* (ψ). Taken from ("Proteomics," 2007).

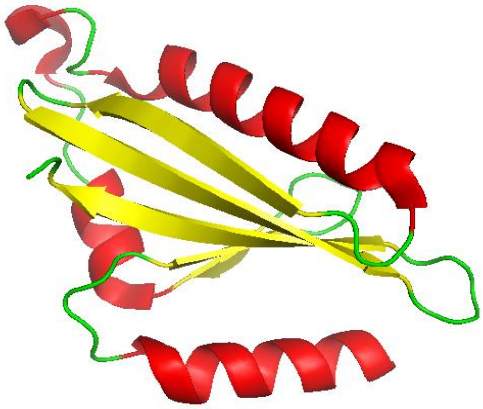

Figure 2.6: The crystal structure of conjugative transfer PAS_like domain from the Salmonella enterica organism. Colours are shown based on the secondary structures: red for helices, yellow for beta sheets and green for loops. Image produced using PyMol (Schrödinger, LLC, 2015).

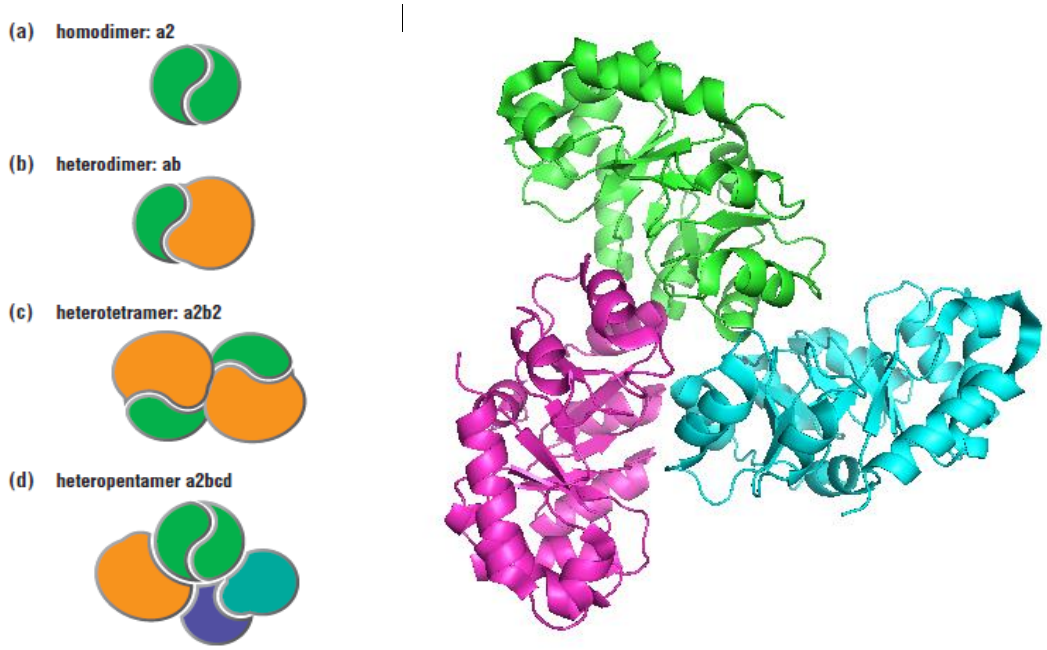

Figure 2.7: Left: Schematic representation of some possible assembly of monomers. Taken from (Petsko & Ringe, 2004). Right: Crystal structure of a homotrimer (PDBID: 1FQ0) where the three subunits are identical and thus a symmetric architecture is formed. Image produced using PyMol.

2.2.4.1 Disulphide Bonds

A disulphide bond is a covalent bond, sometimes referred to Sulphur Sulphur bond (SS-bond) or Disulphide Bridge, between two thiol groups. In proteins, thiol is an active side chain of an amino acid called cysteine. A disulphide bond takes place between two cysteines. Two residues of that kind and the corresponding disulphide bond are often called cystines. Such bonds are created during the process of folding and play a critical role in the stability of the tertiary structure since they cause the nucleus of the hydrophobic core to be constituted; a crucial step towards the compact shape, i.e. a dramatic decrease in terms of entropy (Hatahet, Nguyen, Salo, & Ruddock, 2010; Ruoppolo, Vinci, Klink, Raines, & Marino, 2000).

However, formation of such bond depends on the final conformation itself; it is the final structure that determines how close two cysteines residues are to each other and, therefore, whether they are able to establish a disulphide bond between them. Studies have proved that mis-pairing of cysteine residues may stop proteins from reaching their native conformation, and therefore lead to a misfolded structure (Tu & Weissman, 2004).

2.3 Protein Folding Milestones at a Glance

The notions of polypeptide bond, dipeptide, tripeptide, and polypeptide were first introduced by Emil Fischer, a German Chemist at the University of Berlin, in 1894. He looked for new methods to identify individual amino acids and discovered a new type called cyclic amino acid. Later, he proposed the "Lock and Key Model" regarding interaction between enzymes. He said that enzymes would completely ignore any non-rigid molecules. He is considered as the first scientist to mention that a rigid 3D structure determines a protein's function (Kunz, 2002; Lichtenthaler, 2002).

The Chinese biochemist Hsien Wu is believed to be the first who introduced the concept of protein denaturation in 1931. He showed that denaturation was purely due to an unfolding process and not to any chemical alteration of the protein, and that a wrong fold could lead to a loss of protein's function (Wahid, Ahmad, Nor, & Rashid, 2017; H. Wu, 1995).

Myoglobin, an Oxygen-binding protein found mainly in muscles (Ordway, 2004), was the first protein whose structure was revealed using X-ray crystallography in 1958. John Kendrew, an English biochemist, and co-workers in Cavendish laboratory in Cambridge described that milestone in details in three publications in Nature (Kendrew et al., 1958, 1960; Perutz et al., 1960). Kendrew and Perutz won the Nobel Prize in Chemistry in 1962 for their notable work.

Besides the above three findings, the next three subsections introduce four main milestones regarding protein folding, namely, Anfinsen's theory, Levinthal's paradox, the Ramachandran plot and Anfinsen's dogma.

2.3.1 Anfinsen's Theory

In 1961, Christian Anfinsen et al. proposed a theory concerning the native structure of proteins (Anfinsen et al., 1961). They stated that the correct conformation has the lowest potential energy among all possible structures. Although this theory has not been proved and seems to be contradicted by a few experimentally determined structures, it has been widely accepted. This is the basis of *ab initio* protein structure prediction which searches for the optimal solution using heuristics.

2.3.2 Levinthal's Paradox

In 1969, Cyrus Levinthal raised the question as to why and how a sequence of amino acids can fold into its functional native structure despite that the number of

geometrically possible structures is extremely large (Levinthal, 1968). This 3-page article has been known as Levinthal's Paradox (Zwanzig et al., 1992). He compares the quite large number of possible conformations with the folding time in nature which is measured using millisecond or even microsecond time scales. Even if a protein had the ability to fold into 100 billion different structures per second, hundreds of billion years would be required by a small protein to explore all possible conformations.

2.3.3 Ramachandran Plot

Another milestone was the Ramachandran plot that was first developed in 1963 (Ramachandran, Ramakrishnan, & Sasisekharan, 1963) and further elaborated in 1968 (Ramachandran & Sasisekharan, 1968). It reveals the possible local conformations in protein structures which lead to their secondary structure, i.e. the presence of α-helices and β-sheets. This is illustrated in Figure 2.8. The main importance of such a plot is narrowing the search space of the angles phi (φ) and psi (ψ) since some values are not possible due to steric collisions. Allowed ranges of values and their corresponding secondary structures are shown in Figure 2.9.

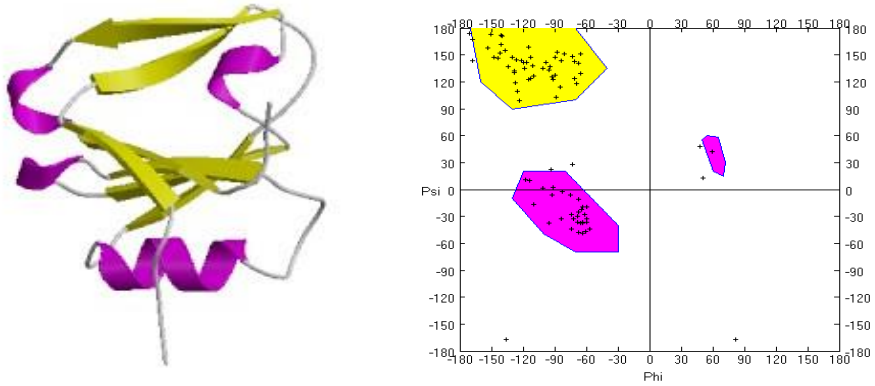

Figure 2.8: Structure of a fragment of the human hepatocyte growth factor (pdb:3hms) and positions of each amino acids on the Ramachandran plot according to their main rotation angles, i.e. phi and psi in degrees. The yellow and pink colours represent beta sheet and alpha helix configurations, respectively.

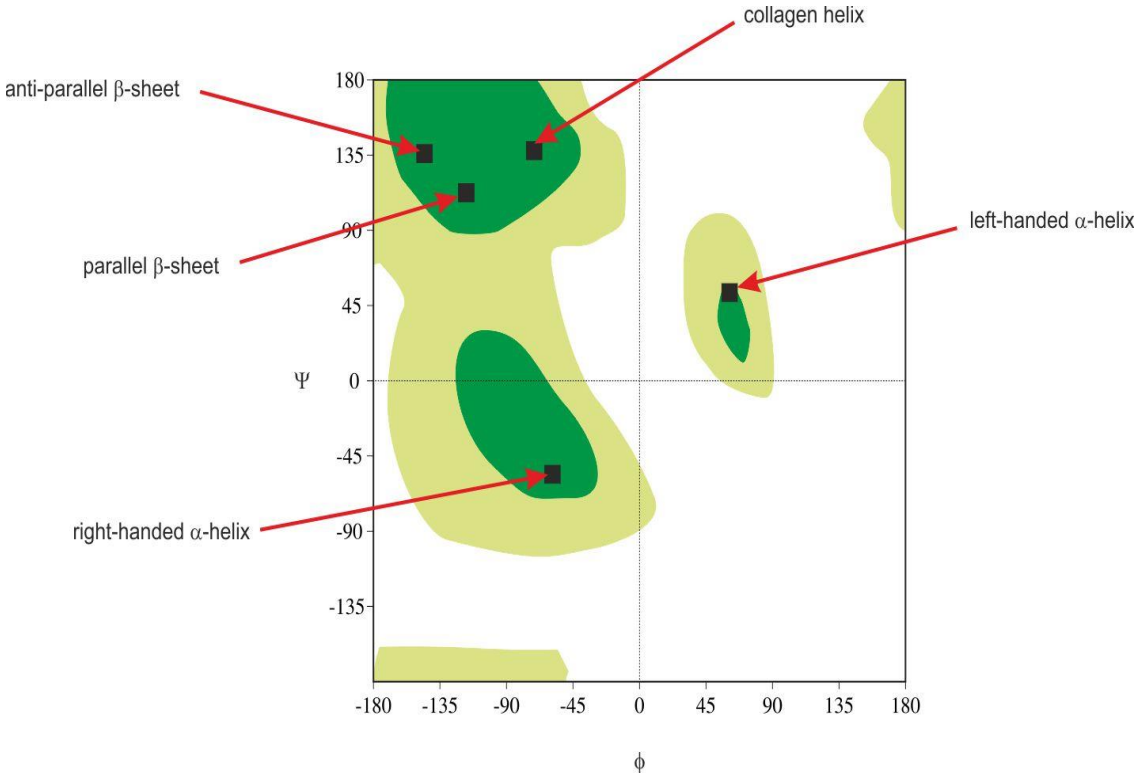

Figure 2.9: Ramachandran plot showing the most favoured regions (dark green) and allowed regions but rare (light green) for the torsion angles in alpha helices and beta sheets. Regions in white are not possible due to stric collision. Taken from http://laborant.pl/index.php/mapa-ramachandrana-narzedzie-do-okreslania-jakosci-struktur-peptydow-i-bialek) with permission.

2.3.4 Anfinsen's Dogma

In 1973, Anfinsen demonstrated that the conformation of a protein can be inferred only from its sequence of amino acids (Anfinsen, 1973). He introduced his thermodynamic hypothesis, later known as Anfinsen's dogma, which says that protein folding is a pure physical, not biological process, that depends only on the specific amino acid sequence and the surrounding solvent. This theory has been considered the main support and motive for advocates of *ab initio* protein structure prediction.

2.4 Protein 3D Structure Determination

A protein's lack of structure, or mis-fold, may be harmful, since a protein's biological function is highly related to its three-dimensional structure. It has been shown that many serious diseases like Alzheimer's, Parkinson's, Creutzfeldt-Jakob disease, cystic fibrosis, and many cancers are linked to protein mis-folding (Dobson, 2001; Prusiner, 1998; Thomas, Qu, & Pedersen, 1995), see Figure 2.10. Thus, determining or predicting the native structure of proteins may not only contribute to a better understanding of the biochemistry of diseases, but also have an invaluable

contribution to drug design. Consequently, predicting a final structure from a sequence of amino acids has been described as deciphering “the second half of the genetic code” (Kolata, 1986) and referred to the “holy grail of computational biochemistry”(Hansmann & Okamoto, 1999). This quest has attracted researchers from many disciplines including biology, biochemistry, biophysics and computer science.

This section contains three subsections: Experimental techniques (section 2.4.1), protein 3D structure prediction (section 2.4.2) and evaluation metrics (section 2.4.3). The section on protein 3D structure prediction is particularly detailed since it is at the core of the research presented in this thesis.

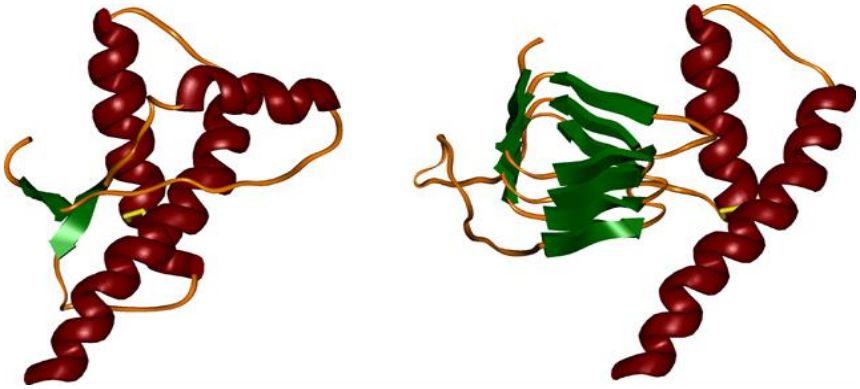

Figure 2.10: Structures of a normal prion protein (PrPc) and the corresponding disease-causing prion (PrPSc). The misfolded molecule is believed to be responsible for Creutzfeldt–Jakob disease. (Retrieved from:http://www.cmpharm.ucsf.edu/cohen/media/pages/gallery.html).

2.4.1 Experimental Techniques

So far, the only trusted and formal way to determine a protein’s 3D structures is via experimental techniques, namely, X-Ray crystallography, Nuclear Magnetic Resonance (NMR) and Electron Microscopy (EM). Only structures resolved by those means can be deposited in the PDB. As shown in Table 2.2, it is clear that advancements in experimental techniques have led to a dramatic increase of the number of available structures during the last two decades when compared to the previous two. Moreover, those data extracted from the PDB show that X-ray crystallography is the most widely used method as it represents more than 89% of the protein database’s entries.

Electron microscopy, which is the least used experimental technique, has recently been gaining importance. Since 2017 till mid 2018, 862 strucutres were deposited using EM, which represent ~40% of the number deposited in the past two decades (from 1998 till 2018). Articles and reviews describing this approach to structure determination can be found in the following publications (Jonic & Vénien-Bryan, 2009; Murata & Wolf, 2018; Nannenga & Gonen, 2014; Saibil, 2000; Unwin & Henderson, 1975). The next two sections describe concisely the two main techniques and their implications on the frequency of proteins deposited in the PDB.

Table 2.2: Number of structures deposited in the PDB over the last 4 decades.

	Year	1978	1998	2018[a]
Number of structures	Total	42	8,607	141,415
	X-Ray Crystallography	42	7,190	126,588
	NMR	0	1,376	12,254
	Electron Microscopy	0	1	2,186

[a] as of June 21, 2018

2.4.1.1 X-Ray Crystallography

In 1958, X-ray crystallography succeeded in resolving the first 3D structure of a protein called Myoglobin (Kendrew et al., 1958). Before 1958, there had been many interesting studies, attempts and findings concerning the birth of X-ray crytallography starting from 1927 (Ott, 1927) up to 1954 (Woolfson, 1954); the complete history of that period can be found in a review (Hauptman, 2015). As mentioned in the previous section, this technique has benefited from a lot of research, advancements and improvements; more than 15 Nobel prizes have been awarded in related topics (Galli, 2014). Accordingly, it has been the dominant experimental way to determine the spatial structure of proteins in the PDB.

This experimental technique can be divided into two main phases: (1) the first one is called crystallisation, which involves converting a solution of a given protein into a well-organised solid, known as crystal. It can be seen as a solid-liquid separation process, where proteins (or any other molecules) can cause an X-ray to diffract, (2) in the second phase X-rays are launched onto the molecule from different angles. The diffraction technique whose theory was pioneered by Bragg and Bragg (Bragg & Bragg,

1913), allows to eventually determine the coordinates of all atoms based on the scattering angles resulting from diffracted rays (Pechkova & Nicolini, 2003b; "Proteomics," 2007). See Figure 2.11 for a simplified diagram of the process. Advanced technical details about X-ray crystallography are beyond the scope of this thesis and can found in many key publications (Ladd & Palmer, 2013; Pechkova & Nicolini, 2003a; Rondeau & Schreuder, 2015; Sherwood & Cooper, 2011; Y. Shi, 2014; Smyth & Martin, 2000a; Su et al., 2015).

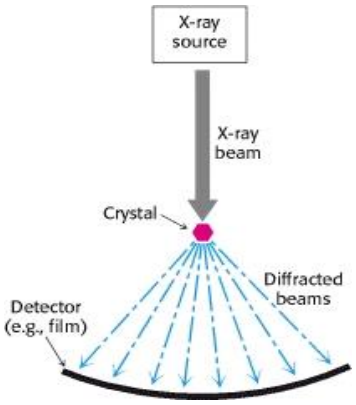

Figure 2.11: Simplified picturial description of the X-ray crystallography process to determine the spatial coordinates of proteins' atoms. Taken from (J. M. Berg, Tymoczko, & Stryer, 2002).

Despite the success this tecnique has gained, many challenges still hinder some experiments and therefore, their corresponding results' accuracy. The first problem is related to the time, effort and even errors whenever it is applied to membrane proteins and protein complexes. The second one is the inability to crystallise the protein in the first place, which might be caused by several factors such as presence of some contaminants or the concentration of the protein (Smyth & Martin, 2000b). The third one is the likelihood of radiation damage due to the x-rays whenever the crystal are either not large enough or not well ordered (Bill et al., 2011; Nannenga & Gonen, 2014).

2.4.1.2 Nuclear Magnetic Resonance

Nuclear Magnetic Resonance (NMR) spectroscopy has the ability to determine the atomic structure of macromolecules and, consequently, the spatial coordinates of proteins, by relying on the magnetic properties of proteins' nuclei. The sample protein should first be purified and dissolved in a watery environment. Afterwards, multidimensional nuclear magnetic resonance is performed thousands of times. Owing to the fact that each nucleus reacts differently in the varying magnetic field, the

structure of the protein can be inferred. Further technical details of this technology can be found in the following publications (Cavanagh, Fairbrother, Palmer III, Rance, & Skelton, 1996; Downing, 2004; Jahnke & Widmer, 2004; Montelione, Zheng, Huang, Gunsalus, & Szyperski, 2000; Rule & Hitchens, 2006).

This technique has some limitations regarding determining alone the 3D structure of proteins (Rehm, Huber, & Holak, 2002); one of them is related to the weight of the molecules as this should not exceed a certain threshold for the sake of an experiment's success (Pechkova & Nicolini, 2003b). Such issues explain the small percentage of the total contributions in the deposited structures in the PDB produced by NMR (around 9%). Nevertheless, NMR has made very important contributions to structural biochemistry, including secondary structure determination, the dynamics of proteins, the structure-function relationship, and most importantly, screening of protein samples to find out which ones are suitable for structure determination either by X-ray crystallography or NMR spectroscopy (Rehm et al., 2002; Renner & Holak, 2001; Rossi et al., 2010; Shuker, Hajduk, Meadows, & Fesik, 1996; Staunton, Owen, & Campbell, 2003).

2.5 Protein 3D Structure Prediction

It is estimated that about 100,000 different proteins can be found in the human body (Dunker & Kriwacki, 2011). However, fewer than 4% of them have been deposited so far in the PDB. From a drug design perspective, determination of a protein's native structure represents a crucial step, since this allows to gain important insights into molecular mechanisms involved in many diseases (Knowles, Vendruscolo, & Dobson, 2014; Ramirez-Alvarado et al., 2010). Since experimental techniques for protein structure determination, as mentioned in the previous section, are very expensive and time consuming and sometimes not possible, there is a great incentive in generating such knowledge via computational means. Therefore, Bioinformatics is usually considered as "the last chance" and probably "the only promising hope" to overcome such a dilemma.

Whereas performing protein folding simulations conforming to Newton's second law may appear as an attractive approach, it is only practical when applied to very small targets while using state-of-the-art supercomputers and grid computing (Baker, 2014), since even for short protein sequences, the search space is enormous and is computationally intractable (an NP-hard problem) (Zwanzig et al., 1992). The field of

protein structure prediction (PSP) aims at predicting computationally the three-dimensional (3D) structure of proteins from their sequences of Amino Acids (AAs). This has been claimed to be one of the most complicated optimisation problems computer scientists have ever faced (Dill & MacCallum, 2012).

2.5.1 Introduction

The approaches used to predict the final 3D structure of a protein are usually classified into three categories: comparative, threading, and *ab initio* modelling. Whereas *ab initio* modelling is based solely on amino acid sequences, the first two classes rely on the PDB as they infer new structures from previous known structures. Consequently, these two approaches are also referred to as template-based modelling (H. Zhou & Zhou, 2005). The main difference between comparative modelling and threading is that the former requires the existence of the structure of a homologous sequence in the PDB. In other words, comparative modelling relies on sequence–sequence alignment, while threading is based on a sequence–structure alignment. Both comparative and threading modelling methods are further discussed in sections 2.54.2 and 2.54.3 respectively.

In principle, *ab initio* approaches do not rely on previous known structures. They are based on thermodynamic rules expressing interactions amongst atoms and energy functions and, thus, the most stable structure is found by determining the minimum energy configuration through one of the Force Field (FF) energy models. For this reason, *ab initio* approaches can also be referred as physics-based methods (Dill et al., 2008), Free Modelling (FM) (Hagler, Huler, & Lifson, 1974), or *de novo* (Song et al., 2005). An FF model aims at evaluating structures using an energy-scoring function. This function usually quantifies chemical interactions and physical forces that occur within the conformation. Initially, when the PDB was relatively small and thus the chance of finding the sought structure was low, *ab initio* methods were seen as a "fall-back position" when both comparative modelling and threading failed (Jones, 1997). However, this view is changing in light of the progress achieved by *ab initio* methods. Different *ab initio* approaches and related topics are described in greater detail in the "Computational Means" section.

In order to evaluate and stimulate the development of computational methods that attempt to predict the native structure of a protein, a biannual community-wide experiment called Critical Assessment of protein Structure Prediction (CASP) was

created in 1994 by John Moult (Moult, Pedersen, Judson, & Fidelis, 1995). This event is now the benchmark for research groups that work in the field of PSP: Prediction methods are evaluated through blind tests of proteins structures. In the latest edition – CASP12 (2016) - more than 37,000 3D models were submitted in the category of Tertiary Structure (TS) predictions. In principle, targets are classified as either Template-Based Modelling (TBM) or Free-Modelling (FM). The former – known as "easy" targets – have homologs in the PDB whereas, the latter - known as "hard" targets - lack such templates, therefore, computational techniques should predict their structures "from scratch", i.e. using *ab initio* techniques. The results of the competition are publicly available on the website of the community (predictioncenter.org) and published as a special issue of a journal (Moult, Fidelis, Kryshtafovych, Schwede, & Tramontano, 2014). At the end of this chapter, a whole section (Section 2.4.4), is dedicated to CASP's general rules, results, analyses and discussions about the latest competition.

2.5.2 Computational and Biological Challenges

Although more and more protein 3D structures have been resolved experimentally - in April 2018, the PDB contained 129,211 entries - and this number increases at a roughly linear pace, the gap between the number of available sequences and known structures continues to widen dramatically; see Table 2.3. Consequently, computational techniques remain essential to protein structure prediction.

Table 2.3: Comparison of Growth of the size of PDB and UniProtKB/TrEMBL over the past 10 years.

Year	Number of Nonredundant Sequence Entries in UniProtKB/TrEMBL[b]	Number of Known Structures Found in PDB[a]
2018 [c]	111,425,245	129,211
2017	87,291,332	126,659
2016[d]	71,002,161	116,320
2015	89,451,166	106,306
2014	88,589,455	97,732
2013	48,701,576	88,993
2012	28,395,832	80,216
2011	18,510,272	72,049
2010	12,347,303	64,603
2009	8,926,016	57,340
2008	6,964,485	50,441
2007	5,072,048	43,919

[a] http://www.rcsb.org/stats/growth/protein
[b] http://www.ebi.ac.uk/uniprot/TrEMBLstats/.
[c] March 28, 2018: date of UniProtKB/TrEMBL latest release.
[d] In 2016, a thorough review was made to remove "similar" entries.

Despite the enormous advances that have taken place in the three types of computational approaches during the past two decades, they all suffer from inconsistency: although they may be successful at predicting some particular targets, they fail for others (Kmiecik et al., 2016; Moult et al., 2014). All the latest blind assessments - mainly CASP - of state-of-the-art PSP computational packages have shown that relying on such computerised tools instead of experimental means for depositing structures of proteins in the PDB is a somewhat "hopeless goal" (Dill & MacCallum, 2012; Kryshtafovych, Monastyrskyy, & Fidelis, 2016). For instance, reliability of predictions for large- and medium-size proteins (i.e. with 300+ residues) tend to be far from acceptable levels. Moreover, models whose accuracy can be considered as equivalent to experimental techniques are still limited to template-based modelling associated with very high sequence identity (Kryshtafovych et al., 2016).

Furthermore, even in the latter "easy" cases where extremely high scores were generated by trusted sequence similarity tools such as PSI-BLAST (Altschul et al., 1997) and HHsearch (Söding, Biegert, & Lupas, 2005), exceptions have been recorded (Alexander, He, Chen, Orban, & Bryan, 2009; Yanan He, Chen, Alexander, Bryan, & Orban, 2008; Jones et al., 1996). Some of those exceptions were found as a consequence of a challenge known as Paracelsus challenge (G. D. Rose & Creamer, 1994). Authors had launched a competition for scientists to find two different proteins with high sequence similarity, but having different architecture/fold/structure. Contributors' findings were probably beyond Rose and Creamer's expectations especially when Bryan and co-workers designed a protein resulting from a single amino acid mutation which displayed a totally different fold from the initial protein (Alexander et al., 2009).

Ab initio approaches have been always associated with a wide range of searching and sampling techniques such as molecular dynamics (McCammon, Gelin, & Karplus, 1977), Monte Carlo (Hansmann & Okamoto, 1999; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983) and genetic algorithms (Holland & H., 1992). However, since all such similar ways are classified under the category of Explore-and-Exploit (Christen & Van Gunsteren, 2008; Perez, Morrone, & Dill, 2017), they all suffer from the associated trade-off (Berger-Tal, Nathan, Meron, Saltz, & Houston, 2014; Zimmerman & Bowman, 2015). Such a trade-off, as well as the tremendous size of the search space has caused such algorithms to get stuck at local minima. A pictorial folding funnel by Ken Dill revealed two decades ago the difference between the ideal and the real funnels (Dill & Chan, 1997); see Figure 2.12. It is worth noting that Dill himself published a recent paper describing a Bayesian method called MELD that aims to accelerate molecular dynamics simulations by using external experimental data as attempt to solve the exploitation/exploration dilemma (Perez et al., 2017).

Most current PSP methods, regardless of their category, comprise a final phase called "refinement", where structures are subject to fine tuning. Such a process is supposed to increase the accuracy of the predicted conformations; structure refinement has unexpectedly been found to be a task as complicated as structure prediction since the results of refinement stage have often been disappointing (Khoury et al., 2014; Maccallum et al., 2011; Modi & Dunbrack, 2016; Nugent, Cozzetto, & Jones, 2014).

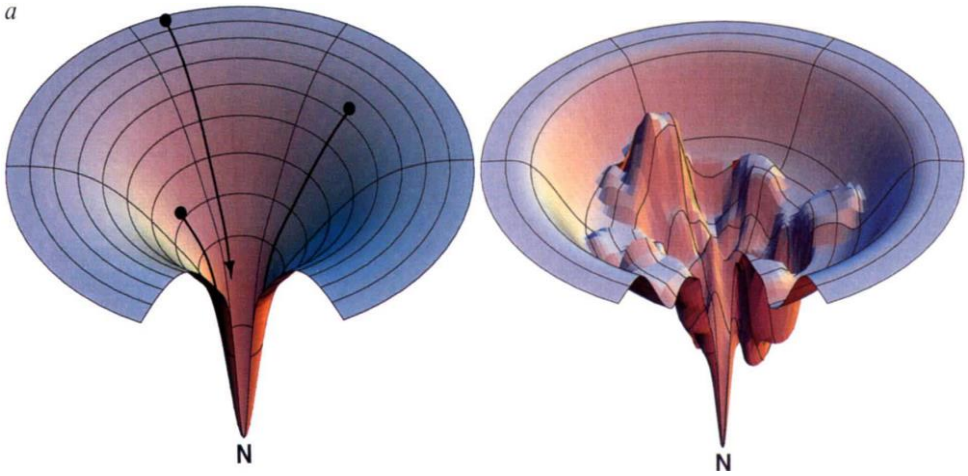

Figure 2.12: Ideal versus real funnel showing the energy landscape where path(s) to reach the native state face too many local minima. Taken from (Dill & Chan, 1997)**.**

Since current computational tools often generate thousands or even tens of thousands of candidate structures, an attractive compromise would be to, at least, be able to assess a priori the quality of generated models, i.e. detect the conformation which is the closest to the native structure among those candidate conformations. As a consequence, the development of Quality Assessment (QA) programs (Elofsson et al., 2017; Kalman & Ben-Tal, 2010; Konopka, Nebel, & Kotulska, 2012) has become an important field of research. Although they are quite accurate at ranking a set of alternative models according to their accuracy, their ability to evaluate the quality of a single model is still limited (R. Cao & Cheng, 2016; Kryshtafovych, Fidelis, & Tramontano, 2011).

Besides the computational issues stated above, it is important to realize that even if they were solved, PSP would remain a challenge for many classes of proteins. Indeed, nature has evolved proteins whose folding includes additional biological complexities which are rarely considered by any existing computational method. Here we will discuss four of these classes: membrane proteins, proteins whose folding is chaperone-assisted, proteins with more than one stable structure, and intrinsically unstructured proteins.

While PSP methodologies generally target proteins located within the cell, membrane proteins are located in the membrane of the cell, which presents a very different chemical environment. Therefore, membrane proteins, except for some peripheral enzymes, are not water soluble and their environment is heterogeneous and complicated, see Figure 2.13b (Cross, Sharma, Yi, & Zhou, 2011). These proteins are

particularly important because not only do they represent around 30% of proteins found in eukaryotic cells (Hopkins & Groom, 2002) but also they are the target of more than 50% of current drugs (Ahmad et al., 2010; Venko, Choudhury, & Novič, 2017; White, 2009). Consequently, a deep understanding of their structure and function has been assessed as invaluable for the world of drug design. However, the fact they are not water soluble represents a serious obstacle to determine their structures experimentally (Cross et al., 2011; C. Zhou, Zheng, & Zhou, 2004). As a consequence, up to April 2018, only 3000 structures have been recorded by MemProtMD (Thomas D. Newport, Mark S.P. Sansom and Phillip J. Stansfeld, found on http://memprotmd.bioch.ox.ac.uk.), which represents only 2.27% of the total number of proteins found in the PDB. The small number of high-resolution templates available so far has limited advances in computational approaches, especially comparative modelling techniques which rely on the availability of templates (Forrest, Tang, & Honig, 2006). As a consequence, prediction of membrane proteins is currently excluded from CASP. Attempts have been made to predict these proteins using *ab initio* methods. However, this means that, besides the AA sequence, a detailed structural and thermodynamic knowledge of the membrane environment is also required (Pollack, Scheiber, Pfaller, & Schramek, 1997). *Ab initio* methods with membrane database information, such as ROSETTA (Barth, Schonbrun, & Baker, 2007; Barth, Wallner, & Baker, 2009; Yarov-Yarovoy, Schonbrun, & Baker, 2006) have been proposed but with limited success. The most recent update was published under the name of RosettaMP (Alford et al., 2015). Recently, a research group in Germany combined CS-Rosetta – the special version of Rosetta that is fed with chemical shifts – to predict alpha-helical membrane proteins aided by external information such as chemical shifts and NOE distance restraints and RosettaMP. Although their dataset comprised only 5 targets, their approached showed quite promising results. Their results have demonstrated, not only how complicated it is to apply *de novo* protein structure prediction to membrane proteins, but also that additional data are needed to derive decent levels of accuracy (Reichel et al., 2017). One would conclude that the main limitations of the current state-of-the-art protein structure prediction tools lies in handling such category of targets.

Examples of notable attempts to develop ad hoc computational tools to predict the topology of membrane proteins are (1) TopPred (von Heijne, 1992) followed by an improved version called TopPred II (Claros & Heijne, 1994). (2) TopCons (Bernsel, Viklund, Hennerdal, & Elofsson, 2009; Tsirigos, Peters, Shu, Käll, & Elofsson, 2015). (3) TMHMM (Krogh, Larsson, Von Heijne, & Sonnhammer, 2001). (4) SCAMPI

(Bernsel et al., 2008). (5) MEMSAT developed by David Jones and co-workers at UCL (Jones, 2007; Jones, Taylor, & Thornton, 1994). (6) HMMTOP (Tusnády & Simon, 2001). (7) Phobius (Käll, Krogh, & Sonnhammer, 2007).

Some researchers developed algorithms that try to predict the type and/or location of such proteins (K.-C. Chou & Cai, 2005; K.-C. Chou & Elrod, 1999).

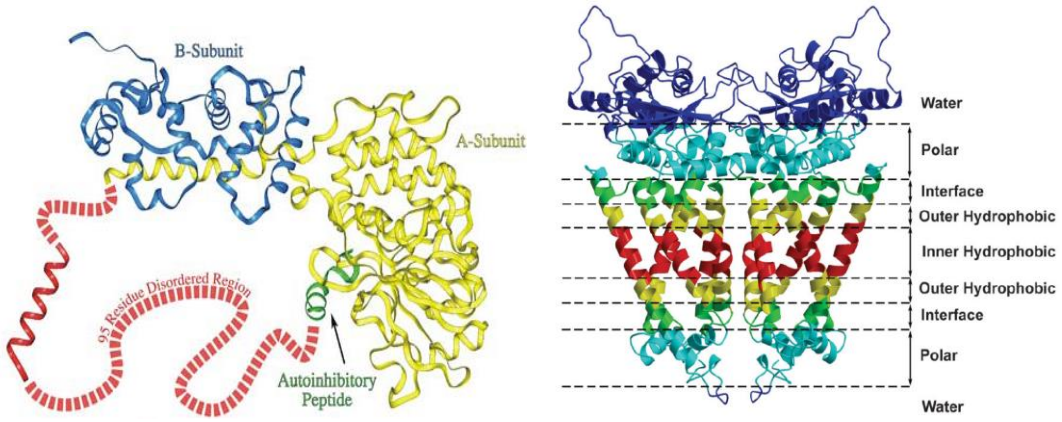

Figure 2.13: (a) The 3D structure of a protein called calcineurin. The discrete part shows a 95-residue disordered region. Taken from (Dunker et al., 2001). (b) The bilayer and surrounding solvent region of membrane proteins is divided into four layers: Water-exposed, interface, outer hydrophobic, and inner hydrophobic. Taken from (Yarov-Yarovoy et al., 2006).

In addition to membrane proteins which fold in a non-aqueous environment, some proteins are not able to fold correctly on their own. They require the assistance of a specific type of proteins, called chaperones, to conform to their proper 3D structures (Bukau, Weissman, & Horwich, 2006) or to prevent them from aggregating (Hartl & Hayer-Hartl, 2002). Those molecular chaperones can distinguish between folded and unfolded proteins by their ability to recognize hydrophobic AAs in the unfolded forms (Engin & Hotamisligil, 2010). Since the role of chaperones is not well understood biologically and they create folding environments which are several orders of magnitude more complex than those currently modelled (see Figure 2.14), existing PSP methods do not consider chaperone assisted folding and are unlikely to do it for the foreseeable future.

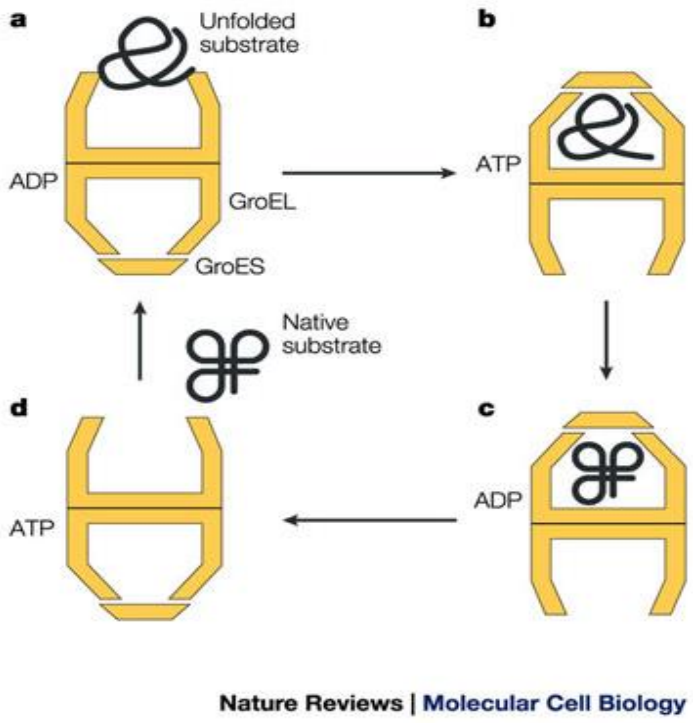

Figure 2.14: Model of chaperone-assisted folding. Taken from (Young, Agashe, Siegers, & Hartl, 2004).

Whereas PSP methods attempt to find the structure of a protein, it has been shown that some proteins can have more than one partially folded intermediate state that may play a critical biological role (Laidig & Daggett, 1996; Pande & Rokhsar, 1998). The *molten globule* (MG) is a distinct intermediate but not very thermodynamically stable conformational state, that takes place between the unfolded state (U) and the native one (N) for most proteins (Bhattacharyya & Varadarajan, 2013); see Figure 2.15 (Ptitsyn, 1995). An MG is characterized by its compact size with the presence of significant amounts of most secondary structures, however with the absence of a specific tertiary structure due to the tight packing of side chains and high mobility of the loops (Regan, 2003). A popular study suggests that MG can be seen as a third phase (Pande & Rokhsar, 1998) where the N/U/MG diagram of protein phases is similar to the solid/vapor/liquid diagram of fluid phases. In other words, each state has its own thermodynamic phase which corresponds to a local minimum of the free energy. MG along with the jagged surface of the energy landscape represents a real obstacle, especially for heuristic algorithms in *ab initio* methods. MGs are currently not addressed by PSP.

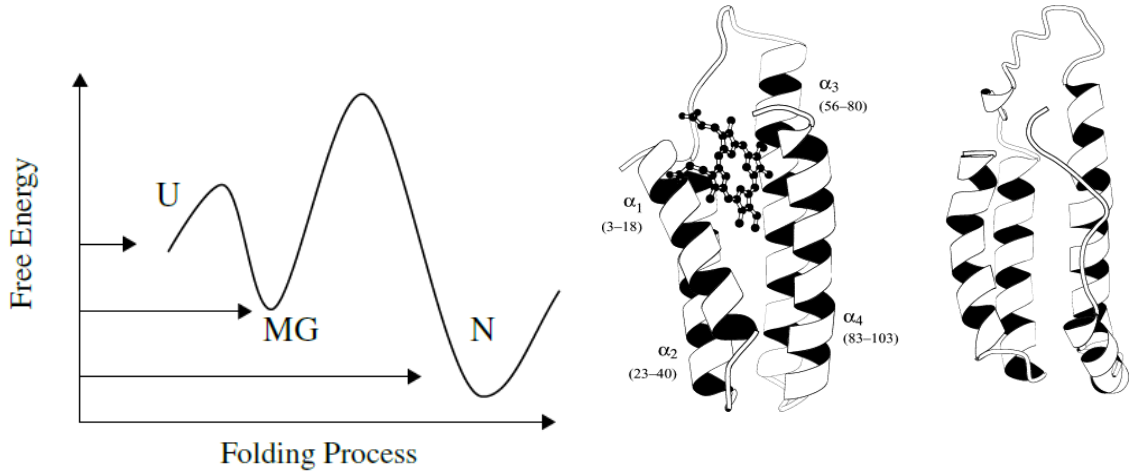

Figure 2.15: (Left) The free energy of the MG state is lower than that of the unfolded state but higher than that of the native one. (Middle and right) Native and molten globule structures of cytochrome *b*562. Taken from (Laidig & Daggett, 1996).

Finally, the intrinsically unstructured proteins (IUPs), also known as natively unfolded or intrinsically disordered proteins (IDPs), may be the most interesting type of proteins in terms of breaching the standard wisdom about proteins (Dunker et al., 2013; Dyson & Wright, 2005; Seshadri, Salma, & Chhatbar, 2009). They simply lack, partially or completely, stable tertiary structure but are able to perform their functions in the cell (see Figure 2.13a). Such a finding was first seen in 1996 (Kriwacki, Hengst, Tennant, Reed, & Wright, 1996) and has been considered so far the first experimental evidence that the lack of structure does not necessarily make some kind of protein harmful or even useless, and thus these are exceptions to the conventional rules of proteins (Dunker & Kriwacki, 2011). It has been suggested that 32% of all human proteins could have some unstructured regions in which the lock-and-key concept cannot be applied (Y. Cheng, LeGall, Oldfield, Dunker, & Uversky, 2006). To date, around 803 partially or totally unstructured proteins have been recorded in the Database of Protein Disorder (http://www.disprot.org) (Piovesan et al., 2017; Sickmeier et al., 2007). Unstructured parts of proteins are indeed a problem with respect to predicting the structure of such molecules; however, they play a critical role in many biological functions. For this reason, the CASP community has created a special category since CASP5 to predict the disordered regions in proteins (Melamud & Moult, 2003). Since then, this category has been gaining importance; more than 94 targets were released and 28 groups submitted their predictions in CASP10 (Monastyrskyy, Kryshtafovych, Moult, Tramontano, & Fidelis, 2014).

2.5.3 Evaluation Metrics

The standard score used to evaluate the similarity of two protein structures is the Root Mean Square Deviation (RMSD) (Coutsias, Seok, & Dill, 2004; Kabsch, 1976, 1978). Although typically, RMSD calculates the minimum average distance of each pair of atoms between the two superimposed conformations, it can be calculated by taking into account Cα atoms only, denoted as Cα-RMSD. The weakness of RMSD, in addition to its correlation with the length of proteins, is that any deviation of a fragment would dramatically change the score even when the remaining regions show a perfect superimposition. More extensions have appeared, such as weighted RMSD (wRMSD) to focus on specific sets of atoms (Kufareva & Abagyan, 2012) and distance variant of RMSD, $RMSD_d$ which can be considered a more "global" metric where each protein is represented by its internal distance matrix (Liebert, 2000).

The global distance test-total score (GDT_TS) was introduced as a part of the LGA (Local Global Alignment) method and since then it has been widely accepted in the community mainly due the fact it is less sensitive to outliers than the RMSD (Zemla, 2003). GDT_TS is the formal criterion CASP uses in order to qualify and assess Tertiary Structure (TS) prediction. It is defined as the average of the percentage of residues that are less than 1, 2, 4, and 8 angstroms. Besides GDT_TS, GDT_HA (High Accuracy) is sometimes used for homology modelling results where high quality superimpositions are expected. It is defined as the average of the percentage of residues that superimpose to within 0.5, 1, 2, and 4 angstroms.

Whilst the GDT_TS overcomes the problem of RMSD in terms of the effects of outliers, the Template Modelling Score (TM-score) (Y. Zhang & Skolnick, 2004b) overcomes the dependence of RMSD on the length of the proteins being aligned. TM-score has become as popular as GDT_TS since it focuses on fold similarities rather than local structural alignments by taking into account all atom pairs rather than those that are below a certain distance cut-off.

MaxSub was introduced by Siew and co-workers by maximising the regions that are close within a standard threshold of 3.5 Å; that is, to detect the MAXimum SUBstructure (Siew, Elofsson, Rychlewski, & Fischer, 2000). Other structural alignment metrics include: FlexE (Perez, Yang, Bahar, Dill, & MacCallum, 2012), Contact Area Distance Score (CAD-score) (Olechnovič, Kulberkyte, & Venclovas, 2013) and Local Distance Difference Test (LDDT) (Mariani, Biasini, Barbato, &

Schwede, 2013). Some structural similarity metrics are dedicated to fragments rather than whole conformations such as the Amplitude Spectrum Distance (ASD) (Galiez & Coste, 2015), and Binet-Cauchy (BC) score (Guyon & Tufféry, 2014). The main metric used to assess structure prediction during the course of this thesis is the global distance test-total score (GDT_TS). Metrics were calculated using MaxCluster, a tool for protein structure comparison and clustering (Siew et al., 2000).

2.5.4 Computational Methods

As discussed earlier, experimental techniques are still time and cost consuming; consequently, computational techniques are essential to produce proposed conformations of protein targets. While excellent results can be produced *in silico* when homologous structures are available, despite advancements in the field of Bioinformatics, structure predictions remain far from being accurate and reliable when attempting to identify a protein's native conformation from its sequence alone (Dill & MacCallum, 2012); See Table 2.4.

Ab initio methods (also known as *de novo*, template-free, or physics-based modelling) mimic Anfinsen's thermodynamic principle by seeking the lowest possible energy conformation that a sequence can adopt. Initially, physics-based methods were proposed, sampling the conformation space until reaching that minimal energy. Although successful predictions have been achieved using Monte Carlo methods and molecular dynamics simulations (Jooyoung Lee et al., 2000; Lindorff-Larsen, Piana, Dror, & Shaw, 2011; Shaw et al., 2010), their extensive computational requirements have limited their application to small proteins. Usage of approximations and heuristics has been a strategy to reduce computational costs; however, this has led to the production of less accurate models.

The classification of PSP approaches into three categories, that is, comparative modelling, threading, and *ab initio* methods, is becoming more and more blurred especially for those techniques that are classified under the *ab initio* category. Some scientists believe that any usage of information about known structures may breach the "classical" rule of *ab initio*: the sequence of amino acids should be the sole source of data. For example, although some methods don't explicitly employ any template structures or even substructures from the PDB, they use knowledge extracted from known secondary, super-secondary and tertiary structures, for instance to predict the secondary structures or derive some terms in the energy functions. However, the big and

endless debate has focused on hybrid and fragment-based protein structure prediction pipelines. This arose for two main historical reasons: (1) when a fragment-based method was first introduced in CASP in 1996, it was assessed to be *ab initio* since it was applied in the discovery of new folds (Jones, 1997). For this reason, the "New Fold" term replaced "*ab initio*" in CASP4 as an attempt to avoid any confusion and include techniques that, to some extent, use the PDB (Klepeis & Floudas, 2003a; Moult, Fidelis, Zemla, & Hubbard, 2001); (2) Some early attempts in this category used relatively short fragments (Bowie & Eisenberg, 1994; Rohl, Strauss, Misura, & Baker, 2004). Consequently, the term "coarse grained *ab initio* protein structure predictors" was accepted based on the perspective that the unit of construction was a set of amino acids rather than a single amino acid (Abbass & Nebel, 2015, 2017). However, the launch of similar pipelines that involve longer fragments such as ROBETTA (Chivian & Baker, 2004), I-TASSER (Pandit, Zhang, & Skolnick, 2006) and QUARK (Xu & Zhang, 2012) has widened the gap between fragment-based and *ab initio* modelling. Although global template information is not used, these cannot be considered as a direct implementation of Anfinsen's hypothesis since they do not use the protein sequence as sole input. In the literature, there have been different categorisations to differentiate between fragment-based and biophysics-based methods. They have been called, respectively, *ab initio* approaches relying on sequence and structural databases and true *ab initio* approaches (Klepeis & Floudas, 2003a), or first-principles methods that employ database information and first principles methods without database information (Floudas, 2007), or physics-based & knowledge-based or simply *de novo* or *ab initio* (Punta et al., 2007), and *ab initio* with database information and pure *ab initio* methods (Abbass et al., 2013). Since fragment-based approaches, such as Rosetta, are at the core of this thesis, to prevent any confusion, the term "Fragment-based Approaches" will be used in this chapter to describe all such modelling techniques regardless of the size of the fragments. They will be reviewed in a separate sub-section (Section 2.5.4.10) rather than within the "*Ab initio* Modelling" section.

The rest of this section will be as follows: Section 2.5.4.1 provides an overview of the early and popular computational techniques like the "Lattice Model" and "LINUS". "Comparative Modelling" and "Fold Recognition" will be covered subsequently in Sections 2.5.4.2 and 2.5.4.3. "*Ab initio* Modelling" in Section 2.5.4.4 covers all important fundamentals of this category ranging from Molecular Dynamics, Force Fields, Monte Carlo Simulation and so on. The last Section (2.5.4.5) is dedicated to Fragment-based approaches along with their successful pipelines.

2.5.4.1 Early Techniques

2.5.4.1.1 Lattice Model

The lattice model is considered as one of the earliest techniques used for protein modelling and protein structure prediction; it was first introduced by Ken Dill in 1985 (Dill, 1985; Lau & Dill, 1989) and has been under study since then (Himu, Jahangir, Ridoy, Dhar, & Shatabda, 2015; Jana, Sil, & Das, 2017; Maher, Albrecht, Loomes, Yang, & Steinhöfel, 2014). Such a model is another way to simplify proteins' representation: atoms are represented by dots/points either in two (2D) or three dimensions (3D) and their motions are simulated using optimization algorithms. Although such a model offers low accuracy, it has succeeded in revealing some modest properties and features contributing to a better understanding of protein folding problems (Pande & Rokhsar, 1999) and the computation of minimum energy conformations (Hart & Newman, 2006).

A popular lattice model is the hydrophobic-hydrophilic lattice model, often abbreviated as the HP model. As the name implies, such model primarily focuses on hydrophobic and hydrophilic interactions which constitute the main source for computing minimum energy. In this model, each amino acid is considered to be either hydrophobic, i.e. non-polar (H), or hydrophilic, i.e. polar (P). The energy of a specific conformation can be evaluated as the number of HH contacts, however, excluding contacts between adjacent amino acids, if any. A degree of hydrophobicity was introduced in the HP model to improve accuracy (Agarwala et al., 1997). The most dominant lattice models for 2D and 3D are respectively the square and simple cubic lattices. However, other lattices have been studied as well such as the triangular lattice (Agarwala et al., 1997), the face centred cubic (FCC) (Agarwala et al., 1997; Hart & Newman, 1997), the cubic lattice with diagonal edges on each face (Heun, 1999), and other crystallographic lattices (Hart & Istrail, 2000). Popular lattice models are shown in Figure 2.16.

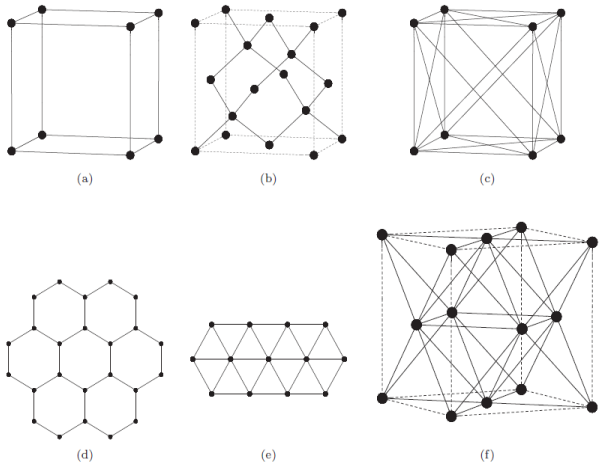

Figure 2.16: Popular 2D and 3D lattice models. (a) Simple cubic, (b) diamond, (c) cubic with planar diagonals, (d) hexagonal, (e) triangular, and (f) face-centred-cubic. Taken from (Hart & Newman, 2006).

In two-dimension HP model, each amino acid is labelled either as hydrophobic or hydrophilic on a 2D grid. The *Best model* is defined as having the highest hydrophobic pair interactions and thus the lowest energy, see Figure 2.17.

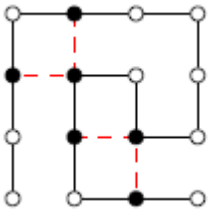

Figure 2.17: An example of a 2D square lattice. Circles represent hydrophilic residues: filled circles represent hydrophobic residues, and red dashed lines represent an HH contact. This lattice is the optimal conformation (4 contacts) of the sequence PPHPHPPPPHPHHPHP. Taken from (Hart & Newman, 2006).

Although HP lattice models are simple and, as a consequence, save computations associated with finding the minimum energy conformation, PSP using various lattice models were shown to be NP-hard more than two decades ago: this includes the cubic (Fraenkel, 1993; Paterson & Przytycka, 1996), diamond (Ngo & Marks, 1992), and cubic with planar diagonals (Unger & Moult, 1993a) lattices. Accordingly, many approximation and optimisation algorithms were used such as Monte Carlo simulation (Kolinski & Skolnick, 2004; Thachuk, Shmygelska, & Hoos, 2007), simulated annealing (Teso, Di Risio, Passerini, & Battiti, 2010), tabu search (T. Jiang, Cui, Shi, & Ma, 2003; Lesh, Mitzenmacher, & Whitesides, 2003), genetic algorithms (GA) (König & Dandekar, 2001; Unger & Moult, 1993b), and ant colony optimization (ACO) (Shmygelska & Hoos, 2003, 2005).

In HP, when side chains are considered, they are modelled using three different symbols: backbone, hydrophobic side chain, and hydrophilic side chain. Only contacts between hydrophobic side chains are taken into consideration when calculating the energy of a specific conformation. (Bromberg & Dill, 1994).

Similar to the HP models, off-lattice models represent proteins' conformations as a set of tangent spheres (in case of 3D) or tangent circles (in case of 2D) of equal radius. Consequently, an HH contact is counted when two circles/spheres representing hydrophobic amino acid/side chain are tangent. A picture showing a representation of a 2D off-lattice HP model with side chains can be seen in Figure 2.18. As in 2D a circle may have at most 6 different tangent circles, a circle that corresponds to a hydrophobic amino acid may be tangent to at most 4 other hydrophobic amino acids (Hart & Newman, 2006). As shown in Figure 2.18, adjacent amino acids are taken into consideration only when there is a contact interface amongst them. In 3D, each sphere may have at most 12 different tangent spheres; therefore, a sphere representing a hydrophobic amino acid may be tangent to at most 10 other hydrophobic amino acids and a sphere representing a hydrophobic side chain may be tangent to 11.

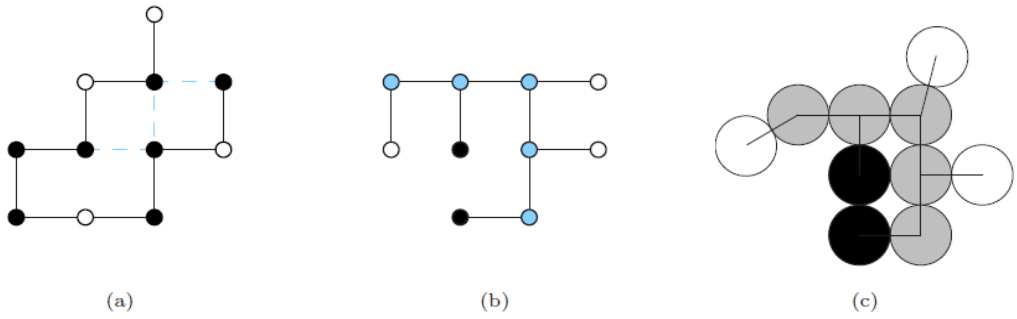

Figure 2.18: (a) The standard 2D HP square lattice model. (b) HP side chains lattice model. (c) Off-Lattice HP model (with side chains). White circles represent hydrophilic amino acid/side chain, circles filled with black represent hydrophobic amino acid/side chain, and circles filled with grey represent backbone element. Taken from (Hart & Newman, 2006).

In an alternative model, the accessible surface area lattice model, the energy for a specific conformation is calculated by evaluating hydrophobic burial instead of HH contacts. Accessible surface area (ASA) aims to quantify the area of the surface of protein that is in contact with the solvent. As a concept, quantifying hydrophobic burial via solvent accessible area was first studied by Lee and Richards (B. T. Lee & Richards, 1971). Similar to most HP lattice models, amino acids/side chains are represented by either an H or P. Many potential criteria were introduced to calculate the ASA for a specific protein's conformation and improve it. They include (1) minimizing the

number of empty lattice points that are in contact with all hydrophobic amino acids and, (2) maximizing the number of covered hydrophobic amino acids (Hart & Istrail, 2000).

2.5.4.1.2 Folding by Hierarchic Condensation

In 1995, Rose and Srinivasan introduced an *ab initio* Monte Carlo–based method called LINUS which stands for Local Independently Nucleated Units of Structure (Srinivasan, Fleming, & Rose, 2004; Srinivasan & Rose, 1995). LINUS relies on a hierarchical procedure that simulates the folding process as discrete hierarchical phases. The rationale behind this approach comes from the decomposition of globular protein structures into secondary structures, super secondary structures and domains (Crippen, 1978; G. D. Rose, 1979). The term "folding by hierarchic condensation" was first introduced by Rose himself in 1979. He had proposed that close chain sites interact to form small structures, which in turn interact iteratively to form larger structures.

The essence of the hierarchical approach lies in constraining some favourable conformations found in previous stages, so that good structures are accumulated. The algorithm works as follows: it starts from an initial extended conformation where both the ϕ and ψ angles are set to 120° with a small interval of allowed interactions. After each iteration, this interval increases and a new conformation is chosen using the Metropolis criterion (Metropolis et al., 1953). Each iteration involves the random selection of three residues whose torsion angle values are amended, whereas bond angles and length are kept constant to "ideal" values (Engh & Huber, 1991). At each stage, a simplified energy function is used to assess the most favourable conformation amongst a set of candidates. The function only considers three main types of interactions: steric overlap, hydrogen bonds, and polarity, that is, hydrophobic burial. A protein's geometric representation in LINUS can be seen as "medium grain", where all non-hydrogen atoms are taken into account.

In 2000, LINUS participated in CASP4 when "new fold" approaches were considered to be in their infancy (Srinivasan & Rose, 2002). Although their overall RMSDs ranged from 8.7 to 16.2 Å, some fragments of around 50 AAs displayed RMSDs of around 4Å, which highlighted the strength of the method. Furthermore, from a secondary-structure prediction perspective, α-helix predictions were evaluated as one of the best among all competitors. Unfortunately, LINUS's latest contribution to the competition was in 2002 and no further publications/improvements have been released since 2004.

2.5.4.2 Comparative Modelling

The idea of comparative modelling, also known as homology or sequence alignment, is quite simple and based on the old principle saying that sequence similarity leads to structure similarity (Browne et al., 1969; Chothia & Lesk, 1986; Evers & Klebe, 2004). In other words, homology methods compare the amino acid sequence of the protein of interest with the sequences of known proteins stored in the PDB. The first successful structure revealed using comparative modelling was lysozyme in 1969 (Browne et al., 1969). Despite all advances in the remaining approaches, homology is still the best one in terms of accuracy when appropriate conditions are met. This success is due to two facts: first, the database of proteins of known structure is expanding and thus the probability to find homologues increases, and second, small changes in the sequence often also yield just small changes in the structure (Martí-Renom et al., 2000). For instance, if identity exceeds 50%, the results of comparative modelling are usually expected to be of high quality. Moreover, for a 30-50% sequence identity, predictions are shown to have more than 80% of the Cα- atoms within 3.5 Å of their true positions. It is worth noting that even for sequence similarity around 30%, some excellent results have been recorded (Rost, 1999). For less than 30% sequence identity, results will probably display major errors (Kopp & Schwede, 2004; Vitkup, Melamud, Moult, & Sander, 2001).

Comparative modelling approaches, at a high level, are in general composed of two phases (Floudas, Fung, McAllister, Mönnigmann, & Rajgaria, 2006): the first one involves selection of the potential templates from the database based on their alignment to the target sequence, and the second one refines side chain geometry and regions of low sequence identity. The first step, often called template identification, is achieved using many algorithms and packages, namely, pair-wise sequence alignment methods like basic local alignment search tool (BLAST) (Altschul, 1990), multiple-sequence alignment methods such as position specific iterative basic local alignment search tool (PSI-BLAST) (Altschul et al., 1997), profile based approaches (Gribskov, McLachlan, & Eisenberg, 1987; Marti-Renom, Madhusudhan, & Sali, 2004) and Hidden Markov Models (HMM) (K Karplus, Barrett, & Hughey, 1998; Kevin Karplus, 2009). Use of HMM for profile-based search is considered one of the most important advances that have taken place in sequence comparison techniques (Deng & Cheng, 2014; Johnson, Eddy, & Portugaly, 2010; Remmert, Biegert, Hauser, & Soding, 2012). For easy targets,

the refinement phase is less important and the difference between average and best predictions is not critical (Tramontano & Morea, 2003).

At a lower level, homology can be seen as either a four-phase (Martí-Renom et al., 2000) or a five-phase procedure (Floudas, 2007). The latter, described by Floudas, can be summarized by the following five points: (a) selection of potential sequences in PDB, (b) an alignment procedure, (c) modelling regions with high accuracy, (d) modelling regions with low accuracy including side chains and loops and (e) refining and assessing the accuracy of the generated model. Examples of popular computer software that implement different homology modelling methods are 3D-JIGSAW (Bates, Kelley, MacCallum, & Sternberg, 2001), SWISS-MODEL (Biasini et al., 2014; Schwede, Kopp, Guex, & Peitsch, 2003), MODELLER (Sali & Blundell, 1993; Benjamin Webb & Sali, 2016), NEST (Petrey et al., 2003), OPLS (Jacobson et al., 2004), SABERTOOTH (Teichert, Minning, Bastolla, & Porto, 2010), and FUGUE (J. Shi, Blundell, & Mizuguchi, 2001). A comprehensive comparative study of several sequence alignment tools is available (Yan, Xu, Yang, Walker, & Zhang, 2013).

2.5.4.3 Fold Recognition

Fold recognition or threading can be seen as a more general but more complicated approach than comparative modelling in protein structure prediction. In practice, fold recognition is applied when comparative modelling fails. In other words, no homologous template is found in the database. The motive behind using fold recognition techniques comes from the observation that the number of structures is much smaller than the number of sequences, that is, the space of different folds is much smaller than the space of different sequences (Chothia, 1992; Govindarajan, Recabarren, & Goldstein, 1999; Z X Wang, 1998; Zhi Xin Wang, 1996; C. T. Zhang, 1997; Chao Zhang & DeLisi, 1998). That observation is based also on the fact that different sequences may share the same structure (Sippl & Flöckner, 1996). The set of different folds is built applying clustering techniques to the proteins of known structure and then labelled into families (Andreeva, Howorth, Chothia, Kulesha, & Murzin, 2014; Berman et al., 2000; Sillitoe et al., 2015). The last point is very important since it is based on an investigation over the new protein structures being deposited in the PDB: they could all be classified in one of the current set of folds determined by structural classification databases such as SCOP (Andreeva et al., 2008) and CATH (Sillitoe et al., 2015) (SCOP and CATH will be revisited in detail in Chapter 4).

Threading methods, which are computationally expensive, are often referred to sequence-structure alignment since they try to fit the target sequence to a known structure using statistical knowledge. Simply, they work by placing an amino acid from the target sequence into a known structure and try to evaluate its "fitness". In this regard, advanced sequence comparison approaches have been used such as Hidden Markov Models (Kevin Karplus et al., 1999) and PSI-BLAST searches (Altschul et al., 1997) – the same methods used in homology techniques – however, to detect remote homologues. Although there is no specific "threshold" to consider to use threading instead of homology modelling which makes the decision to move from the former to the latter methods somewhat blurred (Floudas et al., 2006), a 30% sequence similarity is typically seen as an approximate boundary between homologues and remote homologues. Under that threshold, proteins are said to be "in the twilight-zones" and building an accurate structure is quite challenging (Khor, Tye, Lim, & Choong, 2015; Mihăşan, 2010). The main challenge of threading is gaining the ability to find suitable protein templates; profile-profile alignment techniques have been proved to be the best choice due to their strength to detect very weak homologues (Khor et al., 2015). A recent review shows that a combination of several matches is needed to build a relatively successful scoring function, though accurate results are not guaranteed (Lam et al., 2017). An additional challenge comes from the fact that, even when threading succeeds in finding a suitable template, it may not cover the whole target. Regions that are missing from the template are typically built using another threading "process" or even *ab initio* modelling. Consequently, an assembly methodology is needed.

Another set of techniques that relies on secondary structure prediction has also been used in threading. Klepeis and Floudas showed that there are cases where secondary structure similarity might reach 80% whilst their corresponding sequence similarity does not exceed 10% (Klepeis & Floudas, 2003b). Those results have given more motivation to adopting secondary structure prediction results besides the score of the fitness of a particular known structure relative to the sequence in question (Przybylski & Rost, 2004). However, global optimal protein threading was shown to be an NP-complete problem (Lathrop, 1994). For this reason, many threading algorithms discard residue pair-wise interaction whenever the fitness score is calculated in order to decrease the amount of computation (Bates et al., 2001).

2.5.4.4 Ab initio Modelling

Ab initio approaches are motivated by three important points. First, they are a direct implementation of Anfinsen's thermodynamics hypothesis since they consider the protein sequence the sole source and they search for the minimum free energy of the protein in its environment. Second, whenever comparative and threading modelling fail, that is, homologues and remote homologues were not detected, only an *ab initio* method can, in principle, derive the native structure. Third, these methods, if the natural folding trajectory is followed, give researchers some insights into the folding mechanisms and pathways that are essential to biochemists.

Since some *ab initio* methods attempt to replicate *in silico* the folding process, quantum mechanics should be used to model and estimate the interactions that take place among atoms. Currently, despite the availability of high-performance computing facilities, the computational complexity is such that no comprehensive protein structure prediction systems which are based on quantum mechanics have been recorded. Instead, *ab initio* methods rely on force fields (FF) or energy functions which attempt to express a variety of atomic interactions such as van der Waals, torsion angles, electrostatics, and bond length. Energy functions are usually associated with a search procedure in order to locate the conformation with the minimum-energy function value. The most popular optimization methods are molecular dynamics (Alder & Wainwright, 1957, 1959) and Monte Carlo simulations (Metropolis, 1987; Metropolis et al., 1953).

Despite the usage of FF, *ab initio* techniques remain computationally expensive, which has limited their scope to the prediction of the structure of small protein chains (Jooyoung Lee, Wu, & Zhang, 2009). To address this limitation, researchers have proposed ways of simplifying the PSP task. First, they suggested simplifying the atomic representation of a protein by considering only some atoms (Pillardy et al., 2001; Sun, 1993) or using lattice models (Agarwala et al., 1997; Hart & Newman, 2006). Second, they investigated the narrowing of FF terms by considering few dominant forces (Srinivasan et al., 2004). Finally, protein conformational space was reduced using dihedral angle restrictions to limit their motions (Klepeis, Pieja, & Floudas, 2003; Rohl, Strauss, Chivian, & Baker, 2004).

Despite the variety of the proposed *ab initio* methods, they all rely on minimization of an energy function over the conformation parameters. A general approach for *ab initio* methods is based on a four-step procedure aiming at finding the conformation which has the lowest energy: (1) Start with an unfolded/arbitrary folded conformation;

(2) generate alternative conformations using some heuristics; (3) estimate their corresponding energy; and (4) go to step 2 and repeat until the ending criterion is reached. In general, three parameters play critical roles in any *ab initio* method: energy function accuracy, search algorithm efficiency, and selection of the *Best model*s among several structures.

2.5.4.4.1 Force Fields

The force field models are empirical and attempt to provide an atomic description of quantum mechanics; they quantify bonded and non-bonded interactions between atoms so that the inner energy of a whole molecular system can be estimated by adding the values associated to each interaction between pairs of atoms. Whereas bonded interactions, also known as intramolecular or internal terms, are expressed by terms related to bonds, angles, and torsion angles, non-bonded ones deal with van der Waals and electrostatic interactions and are known as intermolecular forces. Energy functions can be classified into two major groups: physics-based and knowledge-based. While the latter is based on knowledge and statistics extracted from the known protein structures by observing folded protein properties (Arab, Sadeghi, Eslahchi, Pezeshk, & Sheari, 2010; Skolnick, 2006), the former relies on basic physical theories such as molecular mechanics (Boas & Harbury, 2007; Cheatham & Young, 2000; MacKerell, Wiórkiewicz-Kuczera, Karplus, & MacKerell, 1995). Eventually, knowledge-based energy functions have become the most popular as researchers in the field consider them as a "shortcut" due to their availability and low computational calculations (Tian et al., 2011).

ECEPP (Arnautova, Jagielska, & Scheraga, 2006), CHARMM (Brooks et al., 1983), AMBER (Pearlman et al., 1995), UNRES (Oldziej et al., 2005), GROMOS (Schmid et al., 2011), MARTINI (De Jong et al., 2013) and the one used in ASTRO-FOLD (Klepeis & Floudas, 2003a) are physics-based. On the other hand, CABS (Kolinski, 2004), DOPE (Shen & Sali, 2006), dDFIRE (Y. Yang & Zhou, 2008), GOAP (H. Zhou & Skolnick, 2011), ROTAS (J. Park & Saitou, 2014) and the energy functions used in TASSER (Y. Zhang & Skolnick, 2004a), Chunk-TASSER (H. Zhou & Skolnick, 2007), and I-TASSER (S. Wu, Skolnick, & Zhang, 2007) are knowledge based. ROSETTA (Rohl, Strauss, Chivian, et al., 2004) is a popular example of a PSP methodology where the energy function is a combination of terms of both types.

Besides the classification of force fields terms as physics-based and knowledge-based, they can be classified also as either coarse-grained or fine-grained (also known as all-atom). Obviously, a coarse-grained energy function doesn't produce accurate scores, but it makes the energy landscape easier to navigate, therefore, easier to avoid local minima; See Figure 2.19.

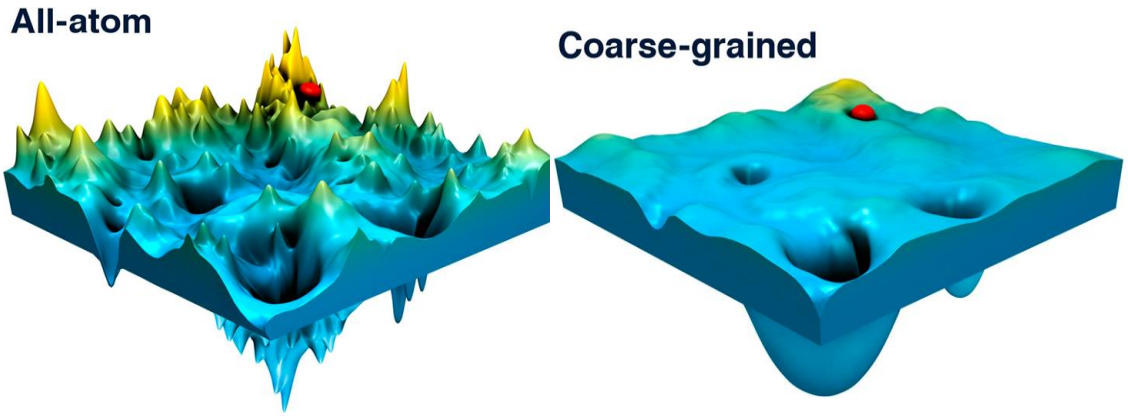

Figure 2.19: All-atom versus coarse-grained energy landscape. The figure illustrates the effect of the smoothening of the energy landscape in a coarse-grained model as compared to an all-atom model. The flattening enables efficient exploration of the energy landscape in search for the global minima, while avoiding traps in the local minima. Taken from (Kmiecik et al., 2016).

2.5.4.4.1.1 Physics based energy functions

A typical physics-based energy function comprises 3 bonded and 2 non-bonded terms: bond lengths, bond angle geometry, dihedral angles, electrostatic forces (Coulombic term) and van der Waals (vdW) (Lennard-Jones) terms respectively. Most of those terms are calculated via quantum mechanics (QM). Ranging from all-atom to coarse-grained physics-based force fields, computational cost may dramatically vary (from high level coarse-grain atomic representation such as two beads per amino acid: Cα and side chain centroid to full-atom representation) and some terms might be added or removed/simplified respectively (Kmiecik et al., 2016).

Bond lengths term, simply, increases whenever the bond length increases or decreases beyond the equilibrium bond length. Bond angle geometry term, also known as valence angles term, works in the same manner as the first term; it tries to keep angles close to the preferred value. Torsion angles terms are counted for four atoms and its degree of freedom is more "flexible" than the first two; there is more than one equilibrium value that yield minimum value. The non-bonded terms that calculate intermolecular interactions need much more computational time, since theoretically,

such interactions take place between any pair or more of atoms that are not bonded, no matter if they are within the same molecule or not. Electrostatic forces, which are typically calculated using Coulomb's law, result from at least two electric charges; therefore, they can be either repulsive or attractive. Taking into account a group of 3, 4 and more atoms to calculate that force is computationally extremely costly as each term for each group will require a nested loop of N^3, N^4… time complexity. As a result, an atom-pair formula is taken into consideration instead. Furthermore, since their values vanish quickly over distances, a threshold is usually set, beyond which no calculation is carried out. Van der Waals forces are mostly attractive unless atoms are very close then the interaction turns out to be suddenly repulsive. The Dutch scientist has shown that the electron clouds surrounding each atom create such a force. The Lennard-Jones term presents an estimated and simplified value of van der Waals interactions, both repulsive and interactive, as a pairwise atom interactions form.

Regardless of the classifications of physics-based force fields as coarse-grained or fine-grained, a huge amount of research has been deployed for the sake of an optimised weighted-term force field (Arnautova, Vorobjev, Vila, & Scheraga, 2009; Krupa et al., 2015; Leaver-Fay et al., 2013; O'Meara et al., 2015; Wroblewska, Jagielska, & Skolnick, 2008). In the next two paragraphs, CHARMM and UNRES have been selected to be concisely introduced; further details describing each term in each force field are beyond the scope of this thesis and can be found in the literature.

CHARMM (Chemistry at HARvard Macromolecular Mechanics) is considered one of the oldest and pioneered physics-based force fields that started in 1983 (Brooks et al., 1983) and its latest update, CHARMM36, was in 2014 when two additional bonded terms were added (S. Lee et al., 2014). CAHRM36m is a special version of the latest release launched in 2016 that is designed specifically for intrinsically disordered proteins (Huang et al., 2016). Concisely, the latest version comprises 6 bonded terms and 2 non-bonded ones and the threshold used for electrostatic and van der Waals interactions is 12 Å. Changes in the latest release can be found in the following publication (Vanommeslaeghe & Mackerell, 2015).

UNRES (UNited RESidue) is a protein model representation that is considered a very reduced one (two beads only per amino acid) associated with its corresponding coarse-grained force fields (Liwo et al., 2014; Liwo, Czaplewski, Pillardy, & Scheraga, 2001). UNRES force fields have been employed in several successful attempts in protein folding and structure prediction (Kachlishvili et al., 2014; Liwo, Khalili, &

Scheraga, 2005; Maisuradze, Senet, Czaplewski, Liwo, & Scheraga, 2010; R. Zhou et al., 2014) and achieved excellent results in CASP (Yi He et al., 2013). Moreover, it has shown competing results whenever used in loop prediction and protein-protein interaction.

2.5.4.4.1.2 Knowledge-based energy functions

Owing to the thorough observation and statistical analysis of the experimentally determined structures' features such as distances amongst atoms and molecules, potentials of knowledge-based force fields are determined. First work in this regard is believed to be credited to Tanaka and Scheraga in 1976; they came up with a term that describes the interaction between two amino acids (Tanaka & Scheraga, 1976). In the same context, Moult and Samudrala created a function which, when fed with pairwise atomic distance data, has the ability to provide a probabilistic value that describes the likelihood of a given sequence and structure to be correlated (Samudrala & Moult, 1998). Although such a function can be considered a simple quality assessment score, an energy score can be derived from it. Rosetta's statistical potentials were derived in almost the same way (Simons, Ruczinski, et al., 1999; Simons, Kooperberg, Huang, & Baker, 1997). Similar to physics-based, knowledge-based force fields can be decomposed into more than one term; each one is related to a specific spatial feature such as ideal torsion angles (Amir, Kalisman, & Keasar, 2008; Gront & Kolinski, 2005; Levy-Moonshine, Amir, & Keasar, 2009).

2.5.4.4.2 Pure Physics-based Approaches

Physics-based approaches can be considered the only computational mean that, in principle, is guaranteed to reach a native-like conformation, since they mimic nanosecond by nanosecond the natural folding process that occurs *in vivo* using molecular dynamics. All-atom (protein and solvent) physics-based protein folding simulations need humongous computational resources to be performed even for very small proteins. Accordingly, this kind of *ab initio* techniques is the least used by research groups, especially when contributing in CASP as the time needed to complete a comprehensive folding simulation exceeds by far the time between the release of the sequence and the deadline to submit spatial coordinates. The next section describes molecular dynamics' principles, applications to physics-based protein modelling and successful attempts using supercomputers and grid computing systems.

2.5.4.4.2.1 Molecular Dynamics

Molecular dynamics (MD) simulation is a computational method that calculates the time-dependent behaviour of a molecular system. It attempts to mimic a protein's motion based on Newton's equation of motion, $F = ma$, where F is the force applied on the particle, m and a are respectively the particle's mass and acceleration. Motions occurring in a protein can be classified into four categories according to scale and time: Local motions involve atomic fluctuation and side-chain motions, medium-scale motions include loop and helices motion, large-scale motions describe motions between domains, and finally global motions include helix–coil transition and the folding/unfolding process.

Originally, MD was introduced to study hard-sphere interactions (Alder & Wainwright, 1957, 1959), simulation of liquid argon (Rahman, 1964), and phases of liquid water (Stillinger & Rahman, 1974). Protein MD simulation was conducted for the first time in 1977 on the bovine pancreatic trypsin inhibitor (BPTI) (McCammon et al., 1977). Since then, MD has given valuable information regarding protein fluctuations, stability, conformational changes, folding pathways and contribution in experimental methods like X-ray crystallography and NMR. The main limitation of MD is its tremendous computational time. Typically, a single CPU requires around a day to simulate a nanosecond, whereas a protein folds generally on the tens of microsecond time scale (Voelz et al., 2010). Consequently, whenever massively parallel processing capabilities are absent, MD is often used in structure refinement rather than simulating the whole folding process starting from a random coil.

The advancement of supercomputers as well as the grid computing systems technologies has allowed a fully MD-based folding simulation using the natural folding path to take place; a breakthrough, scientists had waited for. Although the number of those successful simulations is still quite limited and the size of targets is very small, they have been considered as milestones. The first notable attempt using a supercomputer is back to 1998 and credited to Duan and Kollman when a one microsecond-simulation was performed using MD yielding a 150-nanosecond stable conformation. The protein's length was 36 and the final conformation was 4.5Å close to the native structure (Duan & Kollman, 1998). The same protein – Villin – was also predicted with much higher accuracy several years later (Lei & Duan, 2007; Lei, Wu, Liu, & Duan, 2007). Another remarkable achievement was the prediction of a 20-residue protein – Trpcage - that led to approximately 1Å Cα RMSD stable structure by

three different research groups (Chowdhury, Lee, Xiong, & Duan, 2003; Pitera & Swope, 2003; Simmerling, Strockbine, & Roitberg, 2002). Shaw and co-workers have had a relatively high number of successful predictions, all using all-atoms proteins and solvent MD simulations, (Koldsø, Jensen, Jogini, & Shaw, 2017; Kruse et al., 2012; Lindorff-Larsen, Maragakis, Piana, & Shaw, 2016; Lindorff-Larsen et al., 2011; Piana, Klepeis, & Shaw, 2014; Sborgi et al., 2015; Shaw et al., 2010; Cheng Zhang et al., 2012)

Folding@home, a large distributed grid computing system developed and managed by Pande Laboratory at Stanford University, was introduced in 2000 for the sake of protein folding simulation using MD (Shirts & Pande, 2000). Volunteers all around the world can install dedicated software connected to the server of Folding@home that employs unused and idle processors on their PCs. As long as the CPU is idle, processes are being downloaded, executed and once finished, uploaded to the server. One of the earliest successes was a 300-millisecond simulation (equivalent to 1000 CPU years) of a 36-mer that led to a 1.7 Å native-like conformation (Zagrovic, Snow, Shirts, & Pande, 2002). Since its launch, more than 200 papers have been published; notable and interesting achievements can be found in the following papers (Kohn et al., 2004; Ponder et al., 2010; Snow, Nguyen, Pande, & Gruebele, 2002; Sorin & Pande, 2005; L. P. Wang et al., 2017; Zagrovic et al., 2002). Folding@home was ranked as the most powerful distributed computing network in 2007 by Guinness.

2.5.4.4.3 Approximation and Randomisation Techniques

Instead of aiming at simulating the folding mechanism as physics-based approaches do, this category of *ab initio* methods focuses only on predicting as accurately as possible a protein's final configuration. Such techniques still conform to the "standard rules" of *ab initio*. However, instead of applying the physics-based way that relies on MD and thus the real trajectory, randomised and heuristic sampling and search paradigms are used instead. Monte Carlo (MC) simulations and heuristics are the dominant techniques in this regard. However the trajectories followed by such methods are random, therefore, they reveal no information about the folding pathway even if they succeed to reach a native-like structure.

2.5.4.4.3.1 Monte Carlo Simulations

The Monte Carlo (MC) method was established in the 1940s to approximately solve intractable problems (Metropolis, 1987; Metropolis et al., 1953; Metropolis &

Ulam, 1949). It is based on generating several random samples of the problem and aggregating their results to constitute the final one. In general, Monte Carlo methods can be summarized in four steps: (1) definition of the domain of the problem, (2) generation of many random samples that cover the domain, (3) calculation of the result for each sample, and (4) estimation of the final solution based on the sample results.

Starting from randomly constructed conformation, random minor changes, such as some rotations, within the degree of freedom of the angles are applied. Although, such moves involve “minor changes” as just stated, search pathways may achieve a relatively long step and probably jump over potential barriers, but, probably miss a “good region”. Whenever, fragment-based approaches are carried out, “minor changes” are simply random fragment substitutions.

Monte Carlo simulations are very popular to discover the conformational space of proteins. However, since they may converge toward local minima due to the jagged surface of the energy landscape, many extensions of MC have been proposed. They include multi-canonical ensemble (B. A. Berg & Neuhaus, 1992), entropic ensemble (Jooyoung Lee, 1993), replica exchange MC method (REM) (Kihara, Lu, Kolinski, & Skolnick, 2001), parallel hyperbolic sampling (PHS) (Y. Zhang, Kolinski, & Skolnick, 2003), Electrostatically driven Monte Carlo (EDMC) (Ripoll & Scheraga, 1988, 1989), conformational family Monte Carlo (CFMC) (Pillardy et al., 2001; Pillardy, Czaplewski, Wedemeyer, & Scheraga, 2000) and Monte Carlo with minimization (MCM) (Z. Li & Scheraga, 1987). Monte Carlo with Simulated Annealing (SA) (Kirkpatrick et al., 1983) has been very popular in protein structure prediction; Rosetta has been employing such an optimization algorithm since its birth.

2.5.4.5 Fragment-based Approaches

Fragment-based techniques were in principle meant to be “coarse-grained” *ab initio* method despite the fact that they take advantage of the known 3D structures which are held in the PDB. However, unlike comparative modelling and fold recognition methods which take advantage of full structures, they only extract peptide fragments, secondary structures, and statistical information. The initiative behind these methods followed the observation that the ratio of sequence universe/fold universe decreases whenever the sequence shortens.

Typically, all fragment-based methods start by a fully extended chain or simply a random conformation followed by random substitutions of fragments from a fragment library. The value of the energy function of any changes is the key to acceptance or

rejection of the newly generated conformation. Heuristic algorithms, such as simulated annealing, may accept a worse conformation using Metropolis criterion for the sake of retrieving a better low-energy basin.

Most fragment assembly paradigms involve several search trajectories that run independently to explore low-energy regions and choose a candidate conformation. Afterwards, this pool of candidates, also called decoys, which can reach several dozens of thousands in some pipelines, is subject to clustering and quality assessment techniques to choose the "*Best model*".

The next section – 2.5.4.5.1 - introduces the main concepts and motivation behind the success of this kind of protein structure prediction. Section 2.5.4.5.2 investigates the general principles that govern the main steps from building the fragment library to the refinement phase. The five subsequent sections describe notable methods.

2.5.4.5.1 Introduction and Motivation

Motivated by the fact there is a strong correlation between sequence and structure at the local level (Lu & Liu, 2007), fragment based protein structure prediction methods were first proposed in 1994 by Bowie and Eisenberg (Bowie & Eisenberg, 1994). They rely on the concatenation of short rigid fragments excised from actual protein structures to construct putative protein models. Still, unlike homology and threading modelling, fragment-based predictors are able to handle template-free modelling (FM) targets; sometimes with very high accuracy, especially for small proteins (Song et al., 2005).

As a "compromise" between *ab initio* and fold recognition modelling, fragment-based protein structure prediction packages have been developed (Subramani, Wei, & Floudas, 2012). Methods such as FRAGFOLD (Kosciolek & Jones, 2014), Rosetta (Leaver-Fay et al., 2011), I-TASSER (J. Yang et al., 2015), and QUARK (Xu & Zhang, 2012) have demonstrated the strength of such approaches. Regardless of the fragments' length used by those methods, their popularity is supported by five main points: (1) since the "smallest element" considered in computation is a set of amino acids instead of a single one, entropy in conformational search space is decreased in a dramatic way, (2) short sub-sequences converge towards a relatively limited number of sub-structures, (3) usage of Monte Carlo simulations instead of Molecular Dynamics has allowed making those methods much faster than pure physics-based ones, (4) the fragments used are already of low-energy, therefore, local interactions need not to be calculated within

the fragments after each substitution; a feature that makes such approaches much less expensive than their competitors and (5) from a short fragment perspective, a structure can be built from fragments of other structures that belong to totally different architectures/folds/structures; see Figure 2.20. It is worth noting that a thorough study by Zhang and Skolnick entitled "The protein structure prediction problem could be solved using the current PDB library" has supported such a hypothesis (Y. Zhang & Skolnick, 2005). Accordingly, FM targets, which indeed lack any template structures in the PDB, do not represent a conformation that has a totally new arrangement and shape of secondary structures. Consequently, an ideal short fragment library should be able to predict any FM target. That observation is probably the reason that led CASP organisers to replace the name of the category of "New Fold" to "Free Modelling" starting at CASP6 (Klepeis & Floudas, 2003a). ("Fold" is the second level in the structural classification hierarchy of SCOP, however, since 2008 no new folds have been identified in SCOP1: https://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=fold-scop. The corresponding level in CATH is called "topology", and no new topologies have been added since 2010: https://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=fold-cath).

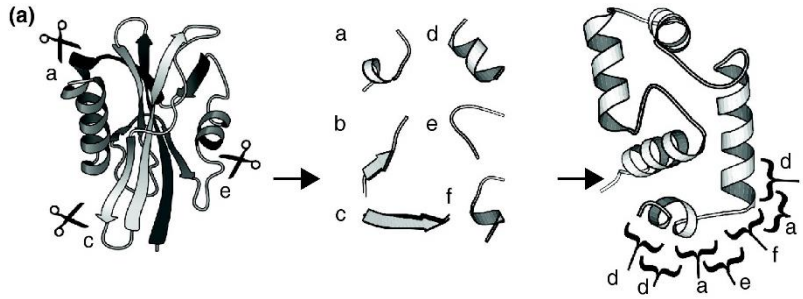

Figure 2.20: Six fragments were taken from a structure (left) to form a small set of fragments (centre) and five of them – a fragment may be used more than once – were able to construct a part of another structure (right). Taken from (Verschueren et al., 2011).

Besides FragFold, I-TASSER, Quark and Rosetta, there have been a few other predictors built following the fragment-based paradigm. Undertaker uses fragments of very different length excised from three different libraries: (1) generic fragments whose length is 2-4, (2) one that contains 9-12 length fragments and (3) a one that relies on fold recognition techniques to extract fragments of larger size. Sampling is conducting using genetic algorithm (Kevin Karplus et al., 2003). PROFESY (PROFile Enumerating SYstem) on the other hand, adopts a library of 15-residue fragments and the assembly phase is conducted using Conformational Space Annealing (CSA) (Julian Lee, Kim,

Joo, Kim, & Lee, 2004). Fragment-HMM is a consensus method; it includes threading, homology modelling, fragment packing, refinement and quality assessment to generate a final candidate model. Position-specific hidden Markov Model is used to sample the target sequence (S. C. Li, Bu, Gao, Xu, & Li, 2008). EdaFold (EDA stands for Estimation for Distribution Algorithm) relies on iterative and "dependent" sampling: fragments that are frequently found in previous low-energy packed conformations are reconsidered in subsequent predictions by raising their usage's probability (Simoncini, Berenger, Shrestha, & Zhang, 2012; Simoncini & Zhang, 2013).

Nevertheless, fragment-based methods continue to fail reaching reasonable accuracy for many CASP's targets, which has paved the way for further investigations, comparative studies, improvements, amendments and tuning (Abbass & Nebel, 2015, 2017; Baeten et al., 2008; Bhattacharya, Adhikari, Li, & Cheng, 2016; J. Cheng, Eickholt, Wang, & Deng, 2012; Guyon & Tufféry, 2010; Helles, 2008; Kandathil, Handl, & Lovell, 2016; S. C. Li, Bu, Xu, & Li, 2008; Olson, Molloy, Hendi, & Shehu, 2012; S.-J. Park, 2005; Trevizani, Custódio, Dos Santos, & Dardenne, 2017; Uziela & Wallner, 2016; Vanhee et al., 2009; T. Wang, Yang, Zhou, & Gong, 2016).

2.5.4.5.2 Principles

These methods, first, search in the PDB for known structure fragments which match sub-sequences of the protein of interest. Once candidate fragments have been selected, compact structures can be formed by randomly assembling fragments using stochastic techniques such as simulated annealing. Then, with the aid of scoring functions the fitness of each conformation is evaluated and the most promising ones are optimized. Scoring functions are loosely related to energy functions and fragment assembly along with optimization algorithms which are conceptually similar to free-energy optimization. Besides, it has been shown that including a similarity measure between the secondary structure of a candidate fragment and the corresponding predicted one in the target improved scoring functions. This justifies the usage of external resources to predict the secondary structures of the target sequence in Rosetta (Simons, Ruczinski, et al., 1999).

In order to eliminate the 'discrete' nature of the process of associating the best sub-structures to given sub-sequences, first, continuous overlapping fragments along the sequence are used, second, weighted knowledge-based energy functions are applied to measure the fitness of fragments using non-local interactions, and third, all-atom refinement is conducted. Such a procedure aims at emulating the actual protein folding

mechanism which is believed to follow a ‘local-to-global/divide-and-conquer’ process. This would explain the high speed of the folding process observed in nature (Dill & MacCallum, 2012; Hockenmaier, Joshi, & Dill, 2007; Voelz & Dill, 2007). Regarding the choice of fragment length, several studies concluded that their optimal size should be around 10 amino acids (Bystroff et al., 1996; Xu & Zhang, 2013). Moreover, it was shown that at least a set of 100 fragments should be explored for each position to produce native-like conformations (Xu & Zhang, 2013).

Success of a protein prediction process using fragment assembly relies on three fundamentals: energy function accuracy, search method efficiency and quality of the fragment library (Kandathil et al., 2016). Weakness in any of them yields wrong search trajectories, thus, inadequate quality of decoys. The majority of fragment packing methods use a coarse-grained atomic representation during the sampling phase so that the smoothness of the search space may help avoid local minima, see Figure 2.19. Full-atom representation is then obtained gradually during optimisation and refinement phases, mostly using knowledge-based ideal values.

2.5.4.5.3 FRAGFOLD

Although some researchers refer to Bowie and Eisenberg as being the small-fragment assembly pioneers (Bowie & Eisenberg, 1994), FRAGFOLD is considered the first fragment-based method developed by Jones in 1996 (Jones, 1997). Its results in CASP2 (1996) seemed promising for a totally innovative approach and paved the way for the development of similar methods. Beside pairwise and solvation potentials, Jones took into consideration compactness of low-energy folds, hydrogen bonds, and steric overlaps to constitute a weight-based energy function. Its minimization was carried out using a simulated annealing approach. FRAGFOLD’s main contribution has been the usage of two types of fragments: super-secondary structural motifs (variable length of 9 to 31 residues) which have been shown to be parts of the polypeptides that form early but remain stable during the folding process, and miscellaneous fragments extracted from high-resolution proteins (fixed length of 9-mers). The first fragments library contains four types of super-secondary structural fragments, that is, α-hairpin, α-corner, β-hairpin, and β–α–β unit, which are defined as motifs containing two or three sequential secondary structures extracted from a library of protein structures.

Since its first launch, FRAGFOLD has been continuously improved, including an extended library of super-secondary structures (Jones, 2001), several enhancements of secondary-structure prediction algorithms, and the removal of the compactness-

related term in their energy function (Jones & McGuffin, 2003). FRAGFOLD3, the version that contributed in CASP6, has included an all-atom representation feature instead of the old and simplified one that considered two points only for each amino acid. Moreover, in the third version, the number of super-secondary structures has been raised to 6 (Jones et al., 2005). In CASP6, in the Free-modelling sections (called New Fold (NF) category then), 4 out of 8 targets FRAGFOLD3 achieved reasonable accuracy. Jones achieved excellent results in CASP9, where his system was overall ranked in the 24th place out of 174 in terms of the *first model*. In CASP12, since no further publications/improvements has taken place since 2005, no tangible improvement was recorded and its overall rank was 23rd amongst 128. FRAGFOLD's main components, including THREADER (Jones, 1998) and PSIPRED (Jones, 1999), can be downloaded at http://bioinf.cs.ucl.ac.uk.

2.5.4.5.4 I-TASSER

TASSER is another successful fragment-based method and was initially created in 2004 by Zhang and Skolnick (Y. Zhang & Skolnick, 2004a). Later, it led to the development of two significantly improved versions: Chunk-TASSER (H. Zhou & Skolnick, 2007) and I-TASSER (S. Wu et al., 2007), the latter has been the most successful version since then and replaced all previous versions (more details on I-TASSER will be covered in the subsequent paragraphs). TASSER is a hierarchical approach that encompasses three phases which gave it its name: threading/assembly/refinement. The first step is based on a threading program called PROSPECTOR 3 (Skolnick, Kihara, & Zhang, 2004). It is based on an iterative sequence–structure alignment algorithm that results in three category targets—easy, medium, and hard—that depend on score value and alignment consistency. Note that using threading techniques in I-TASSER does not prevent its classification as a fragment-based approach since it is only applied on protein sequence subsets to choose the appropriate fragments. Then, the assembly step uses parallel hydrophobic Monte Carlo sampling by rearranging the fragments (Y. Zhang, Arakaki, & Skolnick, 2005). In order to decrease computation, a preliminary model is built using only *Cα* and side-chain coordinates. Finally, a refinement stage is performed using a clustering program called SPICKER (Y. Zhang & Skolnick, 2004c). The full-atom optimization is conducted using the CHARMM22 force field. TASSER achieved an average RMSD of 5.4 Å on all CASP6's 90 targets. Further improvements were achieved at CASP7 with an average RMSD of 4.9 Å on 124 models (H. Zhou et al., 2007) by using better

templates from 3D-jury and applying two additional threading software, that is, SP3 (H. Zhou & Zhou, 2004) and SPARKS (H. Zhou & Zhou, 2005).

Departing from Bowie and Eisenberg's principles, but still considered as belonging to the fragment-assembly category, I-TASSER (Iterative Threading ASSEmbly Refinement) combines *ab initio* modelling and threading (Abbass et al., 2013). Since the length of the fragments chosen from threading has no upper limit (greater than or equal to 5), this method is suitable for both FM and template-based modelling (TBM) targets. As Rosetta, I-TASSER initially generates low resolution conformations, which are then refined. More specifically, structure prediction relies on three main stages (Roy, Kucukural, & Zhang, 2010). First, sequence profile and predicted secondary structure are used for threading through a representative set of the PDB. The highly-ranked template hits are selected for the next step. Second, structural assemblies are built using a coarse representation involving only C-alphas and centres of mass of the side chains. While fragments are extracted from the best aligned regions of the selected templates, pure *ab initio* modelling is used to create sections without templates. Fragment assemblies are performed by a modified version of the replica-exchange Monte Carlo simulation technique (REMC) (Y. Zhang, Kihara, & Skolnick, 2002) constrained by a knowledge-based force field including PDB-derived and threading constraints, and contact predictions. Generated conformations are then structurally clustered to produce a set of representatives, i.e. cluster centroids. Third, those structures are refined during another simulation stage to produce all atom models. This mixed strategy has proved extremely successful since the "Zhang-Server" (Y. Zhang, 2014), which is a combined pipeline of I-TASSER and QUARK (see next section for details about QUARK), has been ranked as the best server for protein structure prediction in five successive CASP experiments (CASP7-11) (Roy et al., 2010; Y. Zhang, 2014; Y. Zhang et al., 2002), when all target categories are considered. However, when only FM targets associated with *ab initio* approaches are taken into account, Rosetta tends to provide more accurate models than I-TASSER (Ben-David et al., 2009; Jauch, Yeo, Kolatkar, & Clarke, 2007; Kinch, Yong Shi, et al., 2011; Tai, Bai, Taylor, & Lee, 2014). See section 2.4.3 for details regarding the results of the latest version: CASP12.

2.5.4.5.5 QUARK

Xu and Yang identified force fields and search strategies as the main limitations to accurate structure prediction (Xu & Zhang, 2012). They proposed a new approach

dedicated for *ab initio* structure modelling, QUARK, which attempts to address them, while taking advantage of I-TASSER and Rosetta's strengths. In addition to sequence profile and secondary structure, QUARK also uses predicted solvent accessibility and torsion angles to select, like Rosetta and unlike I-TASSER, small fragments (size up to 20 residues) using a threading method for each sequence fragment. Then, using a semi-reduced model, i.e. the full backbone atoms and the side-chain centre of mass, and a variety of predicted structural features, an I-TASSER like pipeline is followed: assembly generation using Replica Exchange Monte Carlo (REMC) simulations, conformation clustering and production of a few all-atom models. In this phase, not only does QUARK allow more conformational movements than I-TASSER, but also utilises a more advanced force field comprised of 11 terms including hydrogen bonding, SA and fragment based distance profile, see (Xu & Zhang, 2012) for details. When QUARK started contributing to CASP in its 9th experiment, it was outperformed by Rosetta; however, positions were inverted in the subsequent versions (Kinch, Yong Shi, et al., 2011; Tai et al., 2014).

2.5.4.5.6 Rosetta

Rosetta, developed at the University of Washington at Seattle, is arguably the most popular fragment assembly approach, and it was assessed as the most accurate *de novo* PSP by CASP6 when it contributed for the fourth time (Jauch et al., 2007); see Figure 2.21. Its results in FM targets in CASP have been truly remarkable since then; in CASP12, Rosetta server achieved the first place in the free modelling category (See Section 2.4.3 for more details about the results of CASP12).

Studies highlighting local sequence-structure relationships (Han & Baker, 1996) suggested that methods built on Bowie and Eisenberg's principles should only consider short fragments. As a result, Rosetta, a fully fragment-based protein structure prediction suite, offered to generate conformations from assemblies of short fragments (3-mers and 9-mers) excised from high resolution protein structures. Using the target's sequence, for each position, the best 9-mers and 3-mers are selected. This is performed not only using sequence similarity and the sequence profile, but also by considering secondary structure prediction information generated from several sources as well as Ramachandran map probabilities. The secondary structure predictions are taken from three resources, PSIPRED (Z.-C. Li, Zhou, Lin, & Zou, 2008), SAM-T99 (Tian Liu & Jia, 2010), and JUFO (Kurgan & Chen, 2007) which represents a crucial factor (Gront

et al., 2011). Then, the process of building conformations is conducted using two levels of search and refinement: coarse and fine grained associated with their respective energy functions. In the first level, low-resolution conformations are generated by representing the chain by heavy atoms of the backbone besides a single centroid for the side chains, whereas in the second one, all atoms are modelled. In addition to keeping the fragments rigid during the simulation as most methods do, Rosetta maintains bond angles and length at some ideal values to reduce the search space. Accordingly, the sole degrees of freedom in the coarse-grained search are the backbone torsion angles, whereas, side chains' are only taken into account in the fine-grained stage (Song et al., 2005). A noteworthy observation concerning the force fields type used in both scoring functions is the usage of both physics and knowledge-based terms (Rohl, Strauss, Misura, et al., 2004).

During the coarse-grained search and refinement and in order to generate a conformation's backbone along with its side chain centroids, Rosetta operates in two main steps: first, 9-mer fragments are inserted within the initial fully extended conformation; second, insertions of 3-mer fragments are used to refine the structure previously generated. 9-mers and 3-mers are protein fragments extracted for each amino acid - except for the protein C terminus - of the protein of interest from a template database according to some similarity criteria. Eventually, Rosetta converts the coarse-grained conformation into an all-atom representation by adding all missing atoms using knowledge-based information extracted from known structures (Song et al., 2005). All services related to Rosetta, including downloads, can be found at http://www.rosettacommons.org/ and more details about Rosetta are covered in the next chapter.

Figure 2.21: CASP6 Target T0281 - 70 residues - (PDB code: 1WHZ). The blue structure represents the native one, whereas the magenta represents Rosetta server's predicted structure. With an RMSD of 1.5 Å, Rosetta's model is believed to be the first notable successful *ab initio* prediction in the history of CASP.

2.5.4.5.7 Robetta

From a hierarchical perspective, Robetta protein structure prediction server (Kim, Chivian, & Baker, 2004a) is considered the "mother" of Rosetta whenever it participates in CASP. In the abstracts book (found at http://www.predictioncenter.org/casp12/index.cgi) of all participant groups that describes the methods used by each group, targets' sequences are firstly passed through Robetta. Robetta is a fully automated server, when fed with a sequence in question, follows two different routes: (1) a comparative modelling and (2) *de novo* approach using Rosetta. Concisely, domains and/or regions with high sequence similarity score are then forwarded to Rosetta-based tool for homology modelling called RosettaCM (Song et al., 2013), whereas regions with low homologs detection (mainly long loops), are simply modelled using the "traditional" Rosetta. Once all regions are built, the assembly phase is performed by an iterative domain assembly method (Wollacott, Zanghellini, Murphy, & Baker, 2007) by inserting fragments into the "connecting" parts using the same scoring function as in *de novo* Rosetta. Side chains are then added following the same standard as in *de novo* modelling. Accordingly, whenever an FM target is submitted, Robetta will simply act as Rosetta.

The first phase comprises determining domains' start and ending and candidate templates for easy regions. Such a procedure is achieved using highly ranked sequence alignment tools for homology and threading modelling: HHSearch (Söding, 2005), Sparks (Y. Yang, Faraggi, Zhao, & Zhou, 2011), RaptorX (Peng & Xu, 2011), BLAST (Altschul, 1990), PSI-BLAST (Altschul et al., 1997), FFAS03 (Jaroszewski, Godzik, & Rychlewski, 2000; Rychlewski, Li, Jaroszewski, & Godzik, 2008) or 3D-Jury (Ginalski, Elofsson, Fischer, & Rychlewski, 2003; Ginalski & Rychlewski, 2003). Furthermore, if GREMLIN (Kamisetty, Ovchinnikov, & Baker, 2013), a state-of-the-art contact prediction method that employs meta-genome sequences (Ovchinnikov et al., 2017) provides accurate results, those results are fed to Rosetta as restraints for both sampling and refinement phases. For the sake of quality assessment phase – where top model(s) should be chosen amongst a large set of decoys (up to 300,000) – ProQ2 is carried out (Uziela & Wallner, 2016). The steps described in this section can be considered the pipeline followed by "BAKER-ROSETTASERVER" group mentioned in the next section.

2.5.5 CASP Competition

The state of the field of PSP has been monitored and quantitatively evaluated since 1994 by the biennial CASP event. This community-wide experiment has grown significantly from a set of 33 targets which attracted around 100 models (CASP1, 1994) to a set of 82 targets which led to the submission of more than 37,000 models (CASP12, 2016). Analysis of the outcome of the latest CASP shows that *ab initio* methods are still considered for many successful groups as a backup plan when template-based techniques fail. Consequently, the majority of algorithms/servers whose category is "hybrid" uses *ab initio* as the last approach to be applied. This is mainly due to the facts that, homology modelling is very accurate, and, *ab initio* methods have a very high computational cost even when parallel processing is available. In the free modelling category, it is clear that fragment-based techniques like I-TASSER and Rosetta perform much better than pure *ab initio* ones like ASTRO-FOLD.

2.5.5.1 Introduction

Although CASP was mainly created for the sake of assessment of computational means for protein structure prediction, more categories have been added/removed since the second round in 1996 when docking category was introduced (Dixon, 1997). In round 12 (2016), CASP comprised 6 various categories besides the Tertiary Structure (TS) competition. Other categories include "contact prediction" that involves prediction of contact maps of all residues and "Data Assisted", which is similar to TS competition; however, groups are provided in advance additional information such as NMR sparse data. Nevertheless, PSP (TS category) remains the most challenging and interesting category. For this purpose, every two years, a set of protein sequences are released gradually across a couple of months during which research groups from around the world attempt to predict their 3D structures by submitting putative models (up to 5 per target). Once a target's submission deadline has passed, determination of its native structure is conducted *in vitro*. If successful, a thorough evaluation is performed on the submitted models. In the first 6 rounds – that is, till CASP6 – targets were classified into three categories: "Comparative Modelling", "fold recognition" and "*ab initio*" or "new fold". Afterwards, targets released by CASP have usually been classified into two main categories: template-based modelling (TBM) and Free modelling (FM). Whereas the TBM category comprises "easy targets" for which structures of homologous proteins have already been deposited in the PDB, FM targets represent the greatest challenge in the competition since only groups that rely on *ab initio* methods can

contribute. Due to the complexity of the task, any minor improvement regarding accuracy of FM targets amongst competing groups is considered worthwhile. Recently, in CASP12 experiments, FM and TBM terms have been replaced by “high accuracy models” and “topology” respectively.

2.5.5.2 Classification of Targets

Whenever a target (sequence of amino acids) is released, contributors submit their candidate models for the whole sequence, but, the unit of assessment adopted by CASP is the domain rather than the whole target. The approaches used to divide the targets into domains and classify them as either FM or TBM targets is quite a crucial step achieved by CASP organisers and assessors, so that, a dedicated paper is published for each round since CASP5 (2003) to explain the followed procedures (Clarke et al., 2007a, 2007b; Kinch et al., 2016; Kinch, Qi, Hubbard, & Grishin, 2003; Kinch, Shi, et al., 2011; Taylor et al., 2014; Tress, Ezkurdia, & Richardson, 2009); note that the corresponding paper for CASP12 has not been published yet and that of CASP ROLL and CASP11 were merged into one. Structures are usually released as either “all groups” or “server only”. Concisely, the criterion used is the sequence to known-structure similarity scores; a high score leads a target to be classifies as “server only”, whereas a low score leads a target to be under the “all group” category. However, interestingly, both the domain determination and classification (FM/TBM) operations are carried out after the completion of submission phase as the quality of models submitted is an important feed to both processes (Kinch et al., 2016).

Taking into consideration domains rather than the whole conformation for evaluation purposes is due to several reasons: some targets are relatively long and therefore contain several domains; evaluating the whole prediction may seem unfair especially that a “global” structural superimposition is likely to yield bad scores and therefore may not reveal the accuracy of some prediction in some “independent” regions, namely domains. Moreover, some targets contain more than one domain; whereas one (or more) of them is (are) classified as FM, the other(s) might be classified as TBM. Not to add that even in single-domain targets, structures of tails which are most of the time classified as coils (i.e. they have not defined secondary structure by DSSP) are not of interest and, as a consequence, including them and their predictions are most likely to lead to very poor quality. The domain determination/organisation process, that is, splitting targets into domain(s), is carried out during the phase of post-submission. It is a procedure that relies mainly on two sources: (1) the availability of

templates for specific units and (2) comparison between the scores of servers of the whole targets versus those of the split domains.

Over the years, classification of FM and TBM targets has become a very difficult task; for instance, some participating groups have the ability to detect remote evolutionary relationships to known structures. The standard way adopted by CASP is to use PSI-BLAST and HHPRED to detect any homologs and based on the results targets are classified, however, for some targets/domains, sequence-folds scores are "on the edge", thus, they require further analysis (Kinch et al., 2016). In addition, some insertions and deletions that exist in the core of a certain target's unit may make it appear dissimilar to a specific Pfam (Punta et al., 2012) family for instance. As of CASP11, a new database was introduced to help the CASP community to detect remote homologues of the targets: the Evolutionary Classification of Protein Domains (ECOD) (H. Cheng et al., 2014). It is worth noting that 19 domains in CASP12 have FM and TBM regions in the same time and are formally annotated as "FM/TBM" as we will see in the next section.

2.5.5.3 Analysis and Results of Predictions

For the sake of evaluating the state-of-the-art of PSP approaches and the CASP competition, the results of the latest round, that is, CASP12, which comprises 96 domains/targets, are presented herein. The classification of the targets/domains is as follows: 39 FM, 38 TBM and 19 FM/TBM. Besides the whole group of targets, there will be a separate table for each of the three categories.

Tables 1, 2, 3 and 4 shown in the appendix are taken from the data released by CASP12 community on their website (http://predictioncenter.org/casp12/zscores_final.cgi). Typically, contributors, when publishing their own papers to describe their results, follow their own way of evaluation (such as the *Best model* amongst the five submitted or even the average of the score of the five models) and scores (such as GDT_TS, RMSD, TM). In this thesis, I follow exactly the criteria adopted by the CASP organisers: the cumulative Z-score of the GDT_TS of the *first model* after removing the outliers where Z-score is below the threshold (- 2.0). Z-score's possible range is typically from -3 to +3. Furthermore, the number of groups shown in the three tables is 43 not 128 (the total of groups participated in CASP12), since only server groups were selected; the main aim of this thesis and specifically this section is to assess the computational techniques of protein structure prediction without taking into account any human intervention.

As shown in Tables 1 and 2 in the appendix - the overall ranking of all targets and the overall ranking of FM targets only respectively - Zhang lab at the University of Michigan led by Yang Zhang (Zhang-Server and QUARK) and the Baker lab at the University of Washington led by David Baker (BAKER-ROSETTASERVER) top all remaining research groups all around the world. Whilst in the overall results of all targets, both Zhang-Server (I-TASSER) and QUARK performed better than Rosetta, the latter captured the first place in the hardest category, FM targets. The concise and obvious conclusion, as mentioned before during the chapter: when it comes to a domain where no reliable templates can be found, Rosetta's fragments perform a relatively "good job". Interestingly, Rosetta did the same in the TBM category as shown in Table 3.

Further data have been collected from CASP12 webpages (http://www.predictioncenter.org/casp12/results.cgi) regarding the top three contributors, namely "Zhang-Server", "Quark" and "BAKER-ROSETTASERVER" by extracting the GDT_TS of the submitted models in all categories. The average of the GDT_TS of each group in each category is calculated and summarised in Table 2.4. Such a table will help to (1) assess the performance of the top participated groups and thus to assess the "current status" of the computational techniques in PSP; (2) evaluate each category aside All, FM, TBM and FM/TBM targets; (3) to analyse the performance of each methodology in each category of targets. It is worth noting that we have collected all data without taking into consideration the value of Z-scores; therefore Rosetta appears in the third place for FM whereas it appears in the first in the official ranking in Table 2.

Table 2.4: CASP12's Top Three Servers' Detailed GDT_TS Scores

		All - 96	**FM – 39**	**TBM - 38**	**FM/TBM - 19**
Average (Std. dev.)	Zhang-Server	**52.0** (22.3)	**30.9** (12.6)	**72.4** (12.5)	**54.4** (11.3)
	QUARK	**51.3** (22.5)	**30.7** (12.1)	**71.8** (12.9)	**52.6** (14.5)
	BAKER-ROSETTASERVER	**50.4** (23.9)	**30.5** (14.8)	**72.3** (14.2)	**47.3** (16.1)
Min - Max	Zhang-Server	9.2 – 96.6	9.2 -54.7	50.9 – 96.6	39.2 – 80.1
	QUARK	9.4 – 96.3	9.4 - 57.0	49.5 – 96.3	23.3 - 79.3
	BAKER-ROSETTASERVER	12.1 – 97.6	12.1 – 77.1	34.1 – 97.6	20.3 - 76.5

Table 2.4 shows that the best group reached 52/100 as an average of GDT_TS score of 96 targets which is considered as on "moderate" quality [40 – 59] on the scale. However, the status is much worse in FM domains, where on average, the best three pipelines are still in the "poor" (sometimes referred to "random") scale [0 - 39] of GDT_TS. The reason why applying Newton's second law on grid computing and/or supercomputers is still the only trusted way whenever templates are not found. Regarding the TBM category, results are quite good; oddly, "pure" comparative and threading modelling contributing groups were not able to beat those three pipelines (Note that in such category credits should be awarded to Robetta rather than Rosetta). One would refer this to the following reason: relying on several homologs to select fragments from leads probably to better conformations rather than relying on one or very limited number of homologs.

Regarding the quality of the near-native structure, a very recent study suggests that short fragments are likely to produce higher quality; on the other hand, multiple fragment lengths are able to generate overall better decoys as long fragments have a positive effect in the early stages of simulations (Trevizani et al., 2017). A reasonable explanation of Rosetta's largest standard deviation amongst the three competitors; whenever Rosetta succeeds to reach a "good region" on the conformational space, ability to explore that region is higher than the others due to the short fragments policy (maximum value of GTD_TS in FM category is by far better than the two remaining competitors), however, whenever simulations tend to end up in a "bad region", final structures' accuracy is relatively low due to the lack of usage of long fragments.

After more than 3 decades of continuous work on the diverse types of computational techniques to predict a protein's tertiary structure, the best server has just passed the mid of the "moderate" region on the scale. Therefore, it can be concluded that, despite a lot of progress, there is still a lot of scope of improvement for protein structure prediction. Obviously, hybrid approaches involving mainly fragment-based techniques are the most promising methods amongst all other competitors. Amongst these techniques, one should notice remarkable differences such as the size of fragments, the alignment approaches used to excise the fragments and the subset of the PDB used as a template structures library.

3 Rosetta

3.1 Introduction

Currently, Rosetta is a very large package comprising programs, scripts, tools, for different types of macromolecular modelling such as ligand docking (Lemmon & Meiler, 2012), protein-protein docking (Sircar, Chaudhury, Kilambi, Berrondo, & Gray, 2010), protein design (Guntas, Purbeck, & Kuhlman, 2010), and loop modelling (Mandell, Coutsias, & Kortemme, 2009). However, its first version was initially an implementation of a sole algorithm for *ab initio* protein structure prediction written in FORTRAN (Simons et al., 1997). In 2005, Rosetta 2 was launched using C++. The automatic translation process adopted then to move from FORTRAN to C++ made it inconsistent for further development; there was no other choice than rewriting Rosetta from scratch as a fully object-oriented C++ suite known as Rosetta 3 that was launched in 2009 (Leaver-Fay et al., 2011). Different programs have been gradually added and improvements have taken place during the course of development which has led to the current Rosetta software suite. The latest version (v3.9) was launched in 2018. We started using Rosetta version 3.4 in the first contribution – Chapter 4; for the consistency of the thesis, we continued adopting the same version for all experiments.

The rest of this chapter is organised as follows. The next section tells a brief history of the phase that preceded the launch of Rosetta. Section 3.3 describes the birth of Rosetta and its associated tools. The four energy functions - especially *Score12* - that have been adopted since 2004 are then introduced. Sections 3.4 and 3.5 explore technical details of the fragment picking and fragment assembly processes respectively. The final section is dedicated to the conclusion for this chapter.

3.2 Pre-Rosetta Phase

Before the launch of an *ab initio* protein structure prediction that is based on fragment-assembly, David Baker and his co-workers had been investigating some ideas related the conformational conservation of the short sequences found in different proteins (Han & Baker, 1996), the recurring short sequence motifs used to identify protein family borders (Han & Baker, 1995) and the strong correlations between local sequences and structures (Bystroff et al., 1996). Regardless of a specific secondary structure, their thorough study illustrated that the sequence-structure correlation shows dramatically increasing relative entropy as the length goes from 3 to 8 amino acids, then a slow increase till 10 amino acids (the peak of relative entropy value) followed by a

slow decrease till 15 amino acids. More specifically, the authors investigated some inconsistencies researchers had faced to reach an accurate local sequence-structure mapping. Taking into account fragments of length range of 3 to 15 amino acids, another study was conducted to investigate all types of fragments including those that lie in a transition region between different secondary structures; the latter was considered new as previous papers had excluded such parts. Some "ideal" sequence lengths were found as follows: 13 and 15 amino acids for helix caps (helix-turn-helix motifs), 7 to 11 for helices, 3 and 5 for β-strands, loops and turns. As for the transition fragments results were not as accurate as pure secondary structure regions, however sequences of length 7 and 9 residues showed some level of successful mapping for turn-to-sheet. Furthermore, another study reveals that local sequence motifs that are likely to recur within a protein family. They concluded that local interactions favour a limited number of substructures which in turn dramatically decreases the search space or, in other words, decreases the entropy. The University of Washington's group suggested that most probably this is what happens *in vivo* giving a reasonable explanation of the fast folding process. All those findings had paved the way for the Baker research group to adopt a simple hypothesis: a protein structure can be constructed from a set of short substructures.

In 1997, an "informal Rosetta" paper was published by the Baker lab describing a relatively simple algorithm for predicting the tertiary structure of a protein using a fragment assembly approach aided by simulated annealing and a Bayesian scoring function (Simons et al., 1997). Although some concepts have been changed in the subsequent releases, one can consider this paper as the first step towards Rosetta. Once fragments are determined based on sequence similarity a purely knowledge-based scoring function is employed for measuring non-local interactions to build the final conformation. An important particularity of this scoring function is the exploitation of Bayes' statistical theorem using a large database where the sequences have known structures:

$$P(\text{Structure} \mid \text{Sequence}) = P(\text{Structure}) \times \frac{P(\text{Sequence} \mid \text{Structure})}{P(\text{Sequence})}$$

The meaning of the above formula is quite simple: looking for the most likely structure that could be associated with a given sequence. In short, some fragment-assembled conformations, even lacking any steric collisions, are not likely to be found in nature. *P(structure)*, which is a sequence-independent term, is proportional to the

value of *exp(-radius of gyration^2)* since an assembled conformation that doesn't comply with some structural constraints often leads to a more expanded shape than the native one. On the other side, *P (Sequence |Structure)* captures some sequence-dependent criteria such as amino acid propensities and the positions of hydrophobic residues. Simulated annealing along with the Metropolis criterion was used as a search paradigm where each move is simply a substitute of a 9-residue fragment. Each amino acid is represented by the main chain's heavy atoms and the Cβ of the side chain (a virtual one is created for glycine). Less than two years later, a follow-up paper was published suggesting an improvement to the scoring function (Simons, Ruczinski, et al., 1999). They introduced some sequence-independent terms, such as the properties of packing β-strands to form β-sheets, besides the old sequence-dependent ones; yet using the same Bayes' statistical theorem.

David Baker and Christopher Bystroff contributed to CASP2 in the "*ab initio*" category (8 targets) under the group "BAKER" relying on the publications mentioned in the previous paragraph; the fragment library was called "I-Sites" and comprised fragments of 3 to 15 residues length (Bystroff & Baker, 1997). Although, overall, the results were not good, they succeeded to reach a reasonable accuracy for one target.

3.3 Rosetta's History and Development; At a Glance

The first paper in which "Rosetta" appeared in the literature was in 1999, entitled "*Ab initio* Protein Structure Prediction of CASP III Targets Using ROSETTA" (Simons, Bonneau, Ruczinski, & Baker, 1999). Whilst keeping the same knowledge-based energy function mentioned in the previous section, they finally decided to adopt fragments of fixed size; 9-mers represented the core of the building process, whereas 3-mers played a refinement role. Since then, those fragments' lengths have been continuously adopted in Rosetta. Results were truly encouraging; The group was ranked the best in the *ab initio* PSP category (Orengo, Bray, Hubbard, LoConte, & Sillitoe, 1999) in CASP3.

In 2003, Robetta went online – Robetta Server - allowing users to submit their sequences for prediction using either comparative modelling or *ab initio*. Moreover, users could build Rosetta's fragments online for local execution of Rosetta (Kim et al., 2004a). In 2005, Rosetta@Home, a distributed grid computing systems, that uses volunteers' idle processors all around the world to execute processes related to different Rosetta applications and studies including CASP predictions, was introduced (Baker,

2014; Das et al., 2007; Tyka et al., 2011). Besides Rosetta@Home, another initiative was taken to involve the public mainly in protein structure prediction; a game called "Foldit" was launched in 2008 (Cooper, Khatib, et al., 2010). Outcomes have been beyond the developers' expectations as valuable findings were found out and published based on this work (Cooper, Baker, et al., 2010; Eiben et al., 2012; Gilski et al., 2011; Khatib et al., 2011). A highly improved version called "Foldit StandAlone" was recently made available (Kleffner et al., 2017). In 2012, a much bigger server than Robetta called ROSIE (the Rosetta Online Server that Includes Everyone) was launched (Lyskov et al., 2013). Currently it includes 18 Rosetta applications that can be executed without usage of local computers. It is worth noting that Rosetta@Home is used to run a large amount of processes being submitted every day to ROSIE.

3.4 Energy Functions

Rosetta's energy function, which combines knowledge-based and physics-based terms, has passed through four main phases: *Score12* (Rohl, Strauss, Misura, et al., 2004), *Talaris13* (Leaver-Fay et al., 2013), *Talaris14* (O'Meara et al., 2015) and *REF2015* (Alford et al., 2017). Relying mainly on the new Dunbrack rotamer library (Shapovalov & Dunbrack, 2011), a wider range of experimentally high resolution conformations in the PDB, and a thorough optimisation process to adjust weights, *Talaris13* was introduced instead of *Score12*. *Talaris14* was simply an error-corrected version of *Talaris13* as only a hydrogen bond's weight was changed and the remaining weights were adjusted accordingly. Although the paper giving details of *Talaris14* paper was published in 2014, its widespread usage commenced in 2016. *REF2015* has been the official energy function since July 2017; it includes some updates such as optimised electrostatic parameters and additional terms using the Lennard-Jones potential for hydrogen atoms. It is worth mentioning that the four versions of the energy functions comprise weighted-based terms and the changes that took place cannot be considered as "major". Moreover, the unit of all energy functions is Rosetta energy unit (REU); a Rosetta-specific metric that cannot be converted into standard physical units such as kilo calories per mole (kcal/mol). Chapters 4, 5 and 6's experiments were all carried out using *Score12*.

3.4.1 Score12

The *Score12* force field lasted around 10 years as the default energy function of Rosetta3 (Rohl, Strauss, Misura, et al., 2004). It has been considered the "gold

standard" as during those years Rosetta achieved many milestones such as reaching native-like conformations for small proteins (Song et al., 2005) as well as some of CASP's targets' for high accurate predictions (Bradley et al., 2005; Chivian et al., 2005; Das et al., 2007; Raman et al., 2009). The *Score12* energy function comprises two versions: coarse-grained for low resolution, where a residue is represented by the backbone's heavy atoms besides the centroid of the side chain, and fine-grained for high resolution, that is, for all-atom representation. Low resolution terms include (1) secondary structure pairing terms – a knowledge-based score to evaluate the favourable hydrogen bonding value between any couple of strands and helix-strand packing, (2) radius of gyration, also known as packing density, which is used to favour compact folds using van der Waals attraction forces, (3) van der Waals repulsion term, (4) solvation term (Lazaridis & Karplus, 1999) that includes both a bonus and penalty value and (5) pair-interaction electrostatic forces for up to 12 Å distance of separation. High resolution terms include, in addition to the last three terms above: (1) a Hydrogen bond score (Kortemme, Morozov, & Baker, 2003), (2) Ramachandran and torsion angles (phi and psi) preferences, (3) Dunbrak rotamer energy – a knowledge-based term to assess the likelihood of a certain rotamer to exist (Dunbrack, 2002; Dunbrack & Cohen, 1997) and (4) the reference energy for each residue type in its unfolded state.

3.5 Fragment Picking

In 2011, a new fragment picking tool called "picker" was introduced to replace the old one - "nnmake" - and has been in use since then (Gront et al., 2011). The latest protein database file, where fragments are excised from, comprises 16,801 high resolution template structures of average size of 257 amino acids. The "picker" tool comprises three protocols: best fragments, quota and flexible loop design protocols. Whereas the last one is used for protein design, the first two are used for fragment picking. The Quota protocol is the one that is dedicated primarily for *ab initio* protein structure prediction; consequently, the one adopted in our experiments. As its name implies, it applies the principle of "quota" for secondary structure prediction taken from three different resources (explained further below).

The scoring function, on which the selection of candidate fragments is based, is evaluated at each position in the sequence in question (except the last 8 and 2 positions in case of 9-mer and 3-mers respectively) typically to generate 25 and 200 9-mers and 3-mers respectively. The overall scoring function is the sum of 7 weighted terms: secondary structure predictions from three resources, PsiPred (McGuffin, Bryson, &

Jones, 2000), Jufo (Leman, Mueller, Karakas, Woetzel, & Meiler, 2013) and SAM (Kevin Karplus, 2009), their corresponding scores in the Ramachandran map, and the sequence profile by PSI-BLAST (Altschul et al., 1997). It is worth noting that due to the overall higher accuracy achieved by PsiPred, the default factors of PsiPred, Jufo and SAM are 0.6, 0.2 and 0.2 respectively. . In this regard, the "quota" protocol works as follows. Since none of these three predictors is optimal, using a unique total score based on which all 9-mers and 3-mers will be selected will be a biased approach. For instance, if a fragment's middle residue is predicted to be 40% a helix, 30% a strand and 30% loop, all fragments will be chosen as helix. Instead, "Quota" protocol will guarantee that the corresponding percentages of fragments will be generated from each pool. For example, if the three predictors provide a total probability of having a 3-mer at a certain position as helix is 50%, therefore, amongst the 200 3-mers, 100 fragments will be helical taken as follows: 60 from PsiPred, 20 from Jufo and 20 from SAM pools. See Figure 3.1. Moreover, in quota protocol, the overall scoring function, mentioned earlier, is simply useless, as the Quota scoring function for each pool takes over. In other words, the following three terms: profile, secondary structure prediction of the middle residue and the Ramachandran map probability value of the middle residue constitutes the scoring function based on which the fragments from the corresponding pool are picked. The default weights of each term are as follows: 1, 1 and 2 respectively. Note that the secondary structure prediction of the middle residue is the one that determines the "overall" secondary structure of the fragment.

3.6 Fragment Assembly

Starting from a fully extended chain, the fragment assembly process takes place via a Monte Carlo search; a sequence window of length 9 is randomly selected and one of the available 25 candidate fragments in its turn is randomly selected. Once the torsion angles of that window are replaced by those of the chosen fragment's, the coarse-grained energy score is calculated; the minimisation process is performed using Simulated Annealing (SA) (Kirkpatrick et al., 1983). Therefore, if the energy score after an insertion is smaller than that of the previous conformation, it will be accepted, otherwise, the Metropolis criterion (Metropolis et al., 1953), for the sake of avoiding getting trapped local minima, may also accept it with lower probability for larger energy increases. In short, probability of accepting "bad" moves, whose formula is shown below, decreases exponentially with the ΔE, which describes how worse the energy increases.

$$P = e^{-\frac{\Delta E}{kT}}$$

Whilst k is constant known as “Boltzmann constant”, T represents the “Temperature”, a parameter that plays a key role in the Metropolis criterion. In natural annealing, temperature is first set to a high value, then, gradually, decreases until the material reaches the shape needed. In simulated annealing, “temperature” changes in the same context; it is first set to a certain high value, then decreased, consequently decreases the probability of accepting a “bad” move (in Rosetta, it is a fragment insertion that results in increasing the energy of the conformation). In both "natural" and "simulated" annealing, the heating and gradual cooling cycle can be repeated several times. Indeed, once a fragment replacement is accepted, the “temperature” is set again to its initial value, i.e. the acceptance probability is back to its default value.

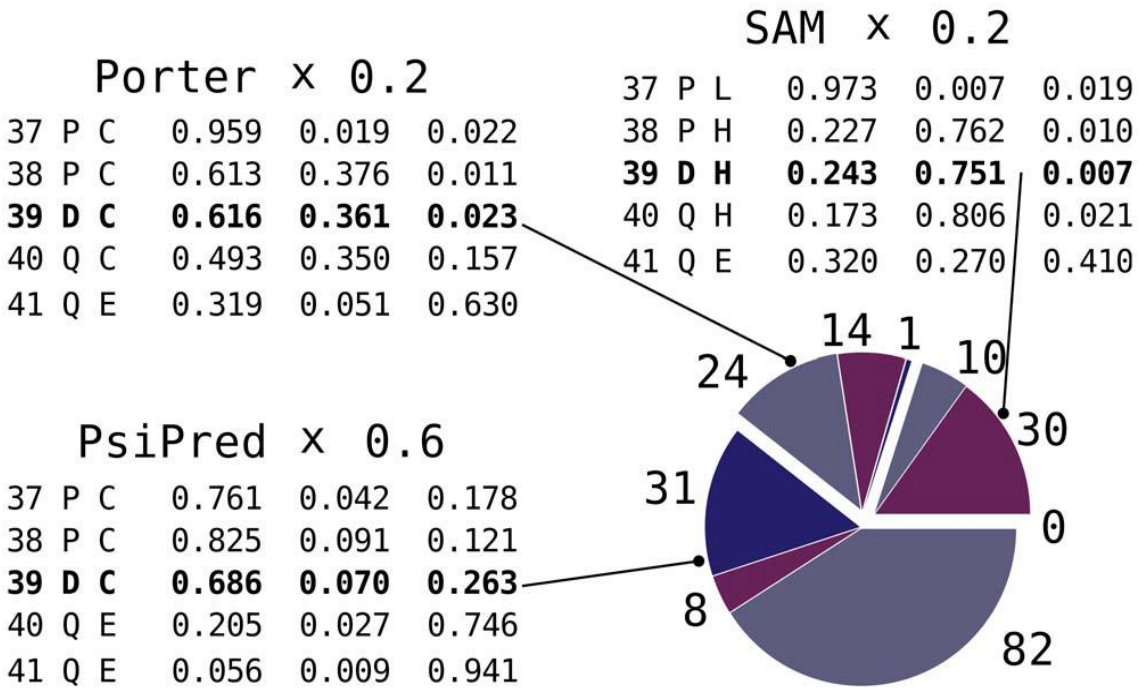

Figure 3.1: Percentages of helical (purple), coil (light blue) strand (dark blue) fragments, whose middle residue’s number is 39 in the sequence of Ubiquitin, taken from each predictor’s pool. The right-most three columns in each predictor show the probability of being coil, strand or helix respectively (Porter corresponds to Jufo). Taken from (Gront et al., 2011).

The 9-mer insertion phase involves 28,000 insertion attempts, however terms of the coarse-grained energy score are added gradually. For instance, in the first 2,000 attempts, only steric overlaps, i.e. van der Waals terms, are considered, whereas in the last 4,000 insertion attempts, the complete energy function is estimated. Once the 9-mer insertion phase is finished, 8,000 insertion attempts using fragments of size 3 are

performed, taking into account the whole coarse-grained energy function. After the overall 36,000 insertion attempts, a simulation will end up with a conformation with heavy backbone atoms only. Optionally, all additional atoms are then added using "ideal" values, and fine-tuned using an all-atom energy score, also known as a fine-grained energy function. The technical term of this phase in Rosetta is called "relax". In this thesis, all generated decoys are full-atom ones.

3.7 Conclusion

Rosetta, using fragments of sizes 3 and 9 amino acids, has reached a "compromised" fragment length for all secondary structures and secondary structure motifs. Furthermore, usage of profile-profile, which has been demonstrated to be the state of the art for detecting homologues, in addition to secondary structure similarity and Ramachandran maps propensities for picking fragments has made the fragment picker combine both relevant sequence-dependent and sequence-independent features. Furthermore, the "quota" protocol has allowed Rosetta to diversify "hard" regions by allowing any secondary structure prediction, no matter how low its associated probability is, to have a "chance" in the assembly phase. Unsurprisingly, Rosetta has been on the top of free modelling (FM) targets in CASP12, since all the above features are remarkable from an FM targets' perspective. It is worth noting that besides the two formal Rosetta groups, more than 12 participating groups have explicitly or implicitly relied on Rosetta in CASP12.

4 Protein structure prediction based on structural class annotations

4.1 Introduction

As mentioned in the previous chapters, when homologous structures are not available in the PDB, fragment-based protein structure prediction has become the approach of choice. However, it still has many issues including poor performance when targets' lengths are above 150 residues, excessive running times, sub-optimal energy functions and specifically, quite general-purpose fragments' libraries comprised of a large number of protein templates to extract fragments from. Those libraries contain a large variety of fragments at each position which makes the probability of picking an appropriate one extremely low. Such an issue has been a topic of significant importance as it was clearly shown that a set "good" fragments in terms of quality and quantity is able, to some extent, to overcome the remaining problems, especially energy functions' inaccuracies (Bhattacharya et al., 2016; Handl, Knowles, Vernon, Baker, & Lovell, 2012; S. C. Li, Bu, Xu, et al., 2008; Simoncini et al., 2017; Trevizani et al., 2017). Taking advantage of the reliable performance of structural class prediction software, we propose in this chapter to address some of the limitations of fragment-based methods by integrating structural constraints in their fragment selection process. Using Rosetta, we evaluated our proposed pipeline on 70 former CASP targets containing up to 150 amino acids and show how structural class predictions can be used, for the first time, as a valuable input for a fragments generation tool.

This chapter is organised as follows. The next section visits the main and noteworthy previous studies, improvements and findings that are related to the fragments selection process. Afterwards, a review of proteins' structural class classifications and predictions is presented followed by a detailed description of the proposed methodology. The last two sections present the results and discussion. Most importantly, a whole section is dedicated to present and discuss the results of our contribution to CASP11 using this methodology.

4.2 Related Work and Motivation

All fragment-based protein structure prediction methods described in the literature review are sequence-dependent since fragments are extracted from templates

selected using sequence-based information. However, it has also been proposed to create databases of fragment models, which are chosen independently of their amino acid compositions to constitute conformation assemblies (Baeten et al., 2008; Kolodny, Koehl, Guibas, & Levitt, 2002; Vanhee et al., 2011). Fragments are only defined by their 'shape' and substituted into the query sequence at positions where amino acids can conform to those shapes. Although such techniques have not been competitive against sequence-dependent predictors, they have shown interesting results in modelling loops (Kolodny et al., 2002; Vanhee et al., 2011).

A promising approach has been the integration of spatial constraints within standard fragment-based systems. So far, this has been performed using predicted contact maps – a matrix that represents the approximate value of the distance between each pair of amino acids (Kosciolek & Jones, 2014; Mao, Tejero, Baker, & Montelione, 2014; M. Michel et al., 2014; Mirco Michel, Menéndez Hurtado, Uziela, & Elofsson, 2017; Ovchinnikov et al., 2017; Ovchinnikov, Kim, et al., 2016; Ovchinnikov, Park, et al., 2016; Ramelot et al., 2009; S. Wu, Szilagyi, & Zhang, 2011). However, since accurate prediction of a contact map currently relies on the availability of a relatively large protein family (ideally more than 1000 homologous protein sequences) (Skwark, Raimondi, Michel, & Elofsson, 2014), their usage is not suitable for all protein targets. Moreover, low quality contact maps lead invariably to poor models, since incorrect constraints prevent appropriate exploration of the native structure conformation space. As a conclusion, there is a need for the design of alternative constraints to fragment-based protein structure prediction.

Although fragment assembly methods have been ranked as the most successful techniques for free-modelling predictions, yet many issues remain and need to be addressed (Dill & MacCallum, 2012). First, successful attempts to produce accurate conformations have been mainly restricted to targets whose lengths are less than 100 residues (Xu & Zhang, 2012) due to the enormous search space, even though protein fragments are used instead of individual amino acids. Second, even for small proteins, processing times are prohibitive for the typical user; Rosetta, for instance, needs on average 150 CPU days per target (S. Wu et al., 2007). Third, despite effective use of Monte Carlo simulations along with fragment replacements, a structure's global energy minimum is likely to be missed. In addition, the design of the most appropriate force field is still open research question as current ones often fail to recognise native structures (Jooyoung Lee et al., 2009; Xu & Zhang, 2012). Finally, the large number of

decoys produced by most of those methods constitutes an additional barrier to identification of native-like conformations since there is no straightforward correspondence between free energy values and similarity to a native structure. As a consequence, design of model quality assessment programs has become an active research area of its own (R. Cao, Wang, Wang, & Cheng, 2014; Konopka et al., 2012).

4.3 Protein Structural Class Classifications

Categorisation of protein structural classes was first introduced by Levitt and Chothia in 1976 (Levitt & Chothia, 1976) when proteins were found to belong to one of four classes: (1) all-alpha proteins; (2) all-beta proteins; (3) alpha + beta proteins where beta strands tend to be segregated and are likely to form antiparallel beta sheets; (4) alpha / beta proteins where alpha helices and beta strands are rather mixed and therefore polypeptide chains are expected to contain parallel beta sheets. Two decades later, Chothia and co-workers established a manually curated online database the Structural Classification Of Proteins (SCOP) (Murzin, Brenner, Hubbard, & Chothia, 1995). The first level of its hierarchy was initially divided into five classes: the original four and a 'multi-domain' class. Later on two further classes were added, namely 'Membrane and cell surface proteins and peptides' and 'Small proteins' (SP) (Lo Conte, Brenner, Hubbard, Chothia, & Murzin, 2002). However, currently only the "small proteins" class exists in the database besides the original four (Andreeva et al., 2014).

Two years after the initial release of SCOP, an alternative database, CATH – named after the first four levels of its hierarchy: Class, Architecture, Topology and Homology – was established (Orengo et al., 1997). Since this showed that there was no clear separation between alpha + beta and alpha/beta proteins (Berman et al., 2000; Michie, Orengo, & Thornton, 1996), CATH has been based on only 4 classes: (1) mostly alpha; (2) mostly beta; (3) alpha beta and (4) Few secondary structures (FSS) (Sillitoe et al., 2015). Despite differences between SCOP and CATH, a comparative study (Csaba, Birzele, & Zimmer, 2009) has shown the top level of both hierarchies, i.e. 'Class', is relatively consistent in comparison to the remaining levels since it is defined according to high level structural features.

Assigning a protein structure to a specific class is not trivial. Whereas CATH uses an automated and explicit method (Michie et al., 1996), SCOP relies on manual inspection. Except for discrimination between 'alpha/beta' and 'alpha + beta', the critical criterion is the percentage of helix and strand content of the protein structure.

Many studies have been conducted to establish the best thresholds for classification, which led to a variety of values (K.-C. Chou, 1995; K.-C. Chou, Liu, Maggiora, & Zhang, 1998; P. Chou, 1989; Eisenhaber, Frömmel, & Argos, 1996; Klein & Delisi, 1986; Kneller, Cohen, & Langridge, 1990; Kurgan, Zhang, Zhang, Shen, & Ruan, 2008; Nakashima, Nishikawa, & Ooi, 1986). Eventually, a thorough comparative study established that the 15% helix and 10% strand thresholds are optimal – those are used by CATH - see Figure 4.1- even if overlapping regions exist between adjacent classes (L. A. Kurgan et al., 2008). Some instances of disagreement between CATH and SCOP structural class classification are mainly a result of the disagreement of domain classification in the first place, especially between 'alpha+beta' and 'mainly beta'. This is due to two causes: the similarity of beta sheets in both classes and whether an alpha helix can be considered a part of the domain or simply a peripheral. Example of such a disagreement between SCOP and CATH is the haemagglutinin (PDBID: 1HGG). Whilst SCOP considers it as mainly-beta ignoring a helical part, CATH treats the whole conformation as one domain and classifies it under alpha-beta (Hadley & Jones, 1999). It is worth noting that CATH employs the distance between various secondary structure as a secondary criterion for classification to cope with this problem. Based on certain thresholds (H-H: 8Å, H-E: 10Å and E-E: 21Å), a secondary structure element can be considered then whether it is a part of the folding unit or not.

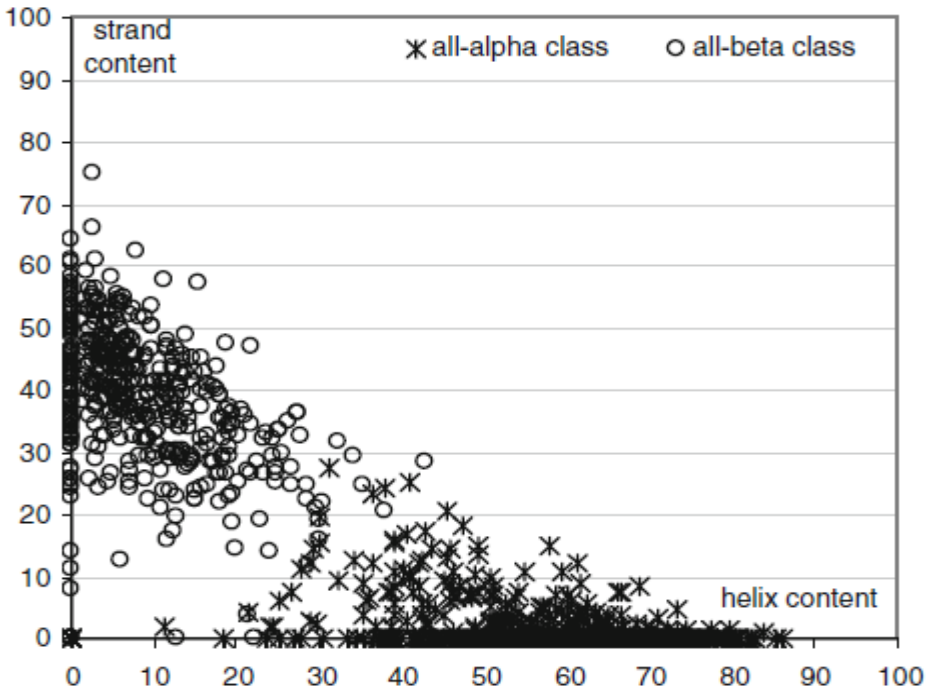

Figure 4.1: Scatter plot of helix and strand content percentages (X-axis and Y-axis respectively) for a large set of proteins classified as either all-alpha or all-beta classes. Taken from (Kurgan, Zhang, et al., 2008).

Since knowledge of a protein's structural class based on its sequence may reveal crucial information concerning folding types and functions (K.-C. Chou, 2005b; K.-C. Chou & Zhang, 1995) and can be considered as a first step towards solving the structure prediction problem, sequence based class prediction has become an active research area (K.-C. Chou, 2011). Proposed approaches take advantage of either 1) machine learning techniques such as Support Vector Machines (SVM) (Anand, Pugalenthi, & Suganthan, 2008; Dehzangi, Paliwal, Lyons, Sharma, & Sattar, 2014; Hayat & Khan, 2012a), Artificial Neural Networks (Jahandideh, Abdolmaleki, Jahandideh, & Asadabadi, 2007), rough sets (Y. Cao et al., 2006), bagging (Dong, Yuan, & Cai, 2006), ensembles (Chen, Kurgan, & Ruan, 2008; Dehzangi, Paliwal, Sharma, Dehzangi, & Sattar, 2013; Hayat, Khan, & Yeasin, 2012; J.-Y. Yang, Peng, & Chen, 2010) and Meta-Classifiers (Cai, Feng, Lu, & Chou, 2006; Feng, Cai, & Chou, 2005); or 2) features that reveal class-related information like physiochemical-based information (Dehzangi et al., 2013; Z.-C. Li et al., 2008), pseudo amino acid composition (K.-C. Chou, 2000; Y.-S. Ding, Zhang, & Chou, 2007), amino acid sequence reverse encoding (Deschavanne & Tufféry, 2008; Mizianty & Kurgan, 2009), Position Specific Scoring Matrix (PPSM) profile (Hayat & Khan, 2012b) and structural based information including secondary structure prediction (Jones, 1999; Kurgan & Chen, 2007; Kurgan, Zhang, et al., 2008; Tian Liu & Jia, 2010). Detailed reviews can be found in (K.-C. Chou, 2005a; Kurgan & Homaeian, 2006). Although state-of-the-art tools, including SCPRED (Kurgan, Cios, & Chen, 2008), MODAS (Mizianty & Kurgan, 2009), RKS-PPSC (J.-Y. Yang et al., 2010), PSSS-PSSM (S. Ding, Li, Shi, & Yan, 2014), AADP-PSSM (Taigang Liu, Zheng, & Wang, 2010), SCEC (Chen et al., 2008), AATP (S. Zhang, Ye, & Yuan, 2012), AAC-PSSM-AC (Taigang Liu, Geng, Zheng, Li, & Wang, 2012) and PSSP-RFE (L. Li et al., 2014) report overall accuracy up to 90%, challenges remain, in particular with proteins with low sequence similarity and discrimination between alpha/beta versus alpha + beta classes (S. Ding et al., 2014). It is worth noting that most tools only deal with the four original SCOP classes which comprise around 90% of annotated domains (K.-C. Chou, 2005a).

4.4 Proposed Methodology

4.4.1 Overview

As highlighted in the literature review chapter, the main limitation of fragment-based protein structure prediction approaches, as with all ab-initio methods, is their inability to sample efficiently the enormous protein configuration space which increases

exponentially with protein sequence length. However, production of accurate predictions is improved if, for each given position, there is high proportion of available fragments fitting closely to the native one (Handl et al., 2012): the higher the quality of the fragment libraries, the more focus the conformation search is on the sub-space containing the native structure. We propose to exploit this property by further customising fragment libraries according to the nature of the protein target. More specifically, we suggest tailoring the set of template proteins which are the source of those libraries so that their quality is increased by removing unnecessary and misleading conformations. We formulate the hypothesis that protein structures which share structural information with a protein target are more likely to provide better fitting fragments than structurally unrelated proteins. Since sequence based structural class prediction has become relatively mature, we have decided to use such information to select the relevant template structures.

From those principles, we have designed this new fragment-based protein structure prediction methodology, see Figure 4.2. First, the structural class is predicted from the amino acid sequence of the protein target. Second, a target specific list of template structures is generated by taking the intersection set between the high-resolution templates sharing the same structural class from the default template protein set (a subset of the PDB) and Rosetta's default database of templates – Vall - which contains 16,801 template structures of average length of 257. Finally, the target sequence and its associated template list are submitted to a fragment-based protein structure prediction, which produces customised fragment libraries and generates a set of putative structures of the protein target (Abbass & Nebel, 2015).

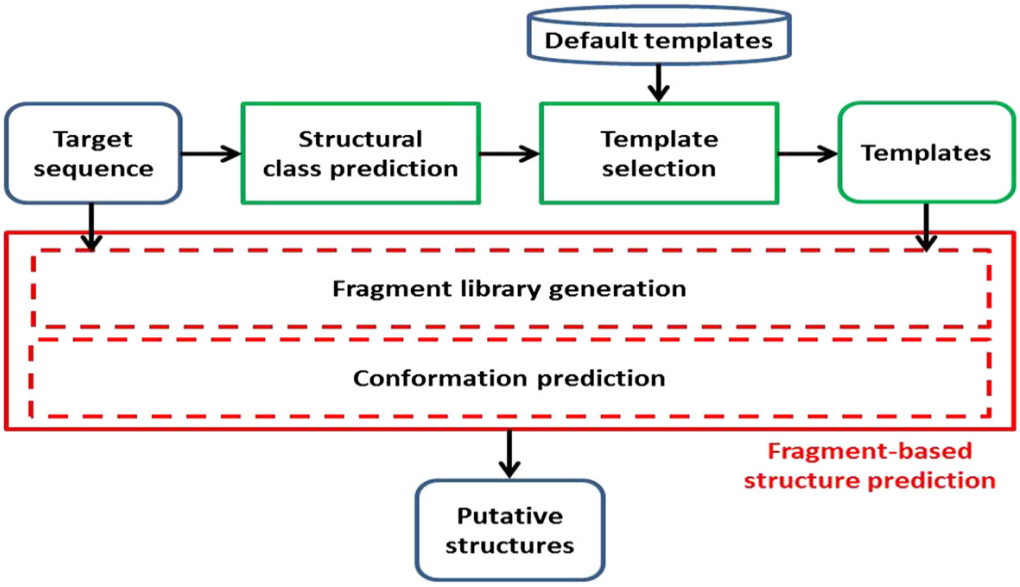

Figure 4.2: Proposed fragment-based protein structure prediction methodology.

In this chapter, we conduct an exhaustive evaluation of our methodology on a set of recent CASP targets. First, we provide a detailed presentation of the proposed methodology for fragment-based protein structure prediction. Second, we compare the quality of decoys with and without class annotations, including the case when structural classes are predicted from a sequence. Third, we analyse the influence of the class type on structure prediction performance. Fourth, we study the impact of using class annotations in terms of convergence towards the best conformation. Fifth, a blind assessment of the models is conducted. Finally, we discuss the validity of the proposed methodology and its potential application.

4.4.2 Procedure

4.4.2.1 Fragment-based Protein Structure Prediction Software

In Rosetta, fragment-based protein structure prediction relies on high resolution template proteins from which to excise fragments. When using the standard Rosetta framework, the database of template proteins on Rosetta's web server can be used (Kim, Chivian, & Baker, 2004b). However, the Rosetta package also offers the facility – a local fragment builder called 'Fragment_Picker' (Gront et al., 2011) and a local copy of the database of template proteins called "vall" – to build user-specific fragment libraries by using a user defined set of templates. Indeed, the standard Rosetta set of fragments can be built either using Rosetta's fragment picker's web server or using the Rosetta suite's built-in fragment picker.

Here, our approach takes advantage of that capacity under the 'Quota' protocol which is specifically designed for *ab initio* predictions, so that the high-resolution template proteins selected by structural class annotation of the target become the source of the fragment libraries. We have used the latest version of the "vall" supported by Rosetta3, which comprises 16,801 high resolved proteins of different classes and folds and of average length of 257 amino acids. A list of a class's PDB code is provided to the "Fragment_Picker", so that the intersection of that set and "vall" is used as the fragment libraries' source. Details about the size of each subset are found in Table 4.1. As shown, the new set of templates has dramatically decreased in size; for instance, for CATH's mainly alpha-based predictions, only 1,905 out of 16,801 are used. Furthermore, CATH's template libraries are larger than SCOP's, except for the few secondary structures. The versions of CATH and SCOP used in this study are 4.0 and 1.75 respectively; the ones adopted by the PDB.

Table 4.1: A comparison showing the difference in terms of the number of template structures amongst the standard, CATH and SCOP-based experiments. The first part– light grey shaded – is dedicated to CATH whereas the second one – darker grey shaded – is related to SCOP. The second and fourth rows show the exact number of templates used to build each of the 9 customised fragment libraries (4 for CATH and 5 for SCOP). In both rows, one would notice that the total number of templates is not equal to the size of the VALL, this due to the fact that some proteins in VALL were not annotated by CATH or SCOP, consequently they were excluded.

CATH Structural Classes (Number of Templates)	Mainly Alpha (10,194)	Mainly Beta (10,532)	Alpha Beta (22,685)		FSS (531)
Templates Common with VALL (16,801)	1,905	1,826	5,119		84
SCOP Structural Classes (Number of Templates)	All Alpha (4,807)	All Beta (7,534)	Alpha + Beta (7,824)	Alpha/ Beta (9,186)	SP (853)
Templates Common with VALL (16,801)	1,149	1,270	1,672	2,128	152

4.4.2.2 Structural Class Annotations

Our novel approach relies on structural class annotations of target sequences. Both SCOP and CATH are widely used databases, attracting diverse publics according to appreciation of their different degrees of automation. Since SCOP-based annotations rely largely on a manual process, they are preferred by many biologists as it is seen to be "more natural" (Kurgan, Zhang, et al., 2008). On the other hand, CATH's higher degree of automation makes annotations more systematic and allows processing a larger share of the PDB. Here both classification schemes are considered in our evaluation. Since we wish to both validate the concept of using class-specific fragment libraries for protein structure predictions and demonstrate its practicality, all protein targets were annotated twice based on either their known structure – classifications seen as the gold standard - or their sequence.

First, structural class annotations, according to both SCOP and CATH classifications, were conducted on all protein targets using their structure. Note that all selected targets only contained a single domain. Initially, when available, annotations were extracted from the SCOP and CATH databases. If a target was present only in one

of the two, the second annotation could generally be deduced directly. However, in the case of a protein belonging to CATH's class 'alpha beta', manual inspection was used to allocate it to either the alpha/beta or alpha + beta class in the SCOP classification. Alternatively, when targets did not have any annotation in either database, we classified them manually based on the secondary structure content of their PDB entry as provided by the Dictionary of Secondary Structures of Protein (DSSP) (Kabsch & Sander, 1983) and the thresholds adopted by CATH (Michie et al., 1996).

Second, class annotations were predicted from the sequence alone. As seen in the "Protein structural class classifications" section, structural class prediction is a very mature field where accuracy reaches up to 90%. Among the most competitive methods, MODAS (K.-C. Chou, 2000) - MODular Approach to Structural class prediction – is particularly suitable for our application since it is freely available online and it provides predictions for the main classes of SCOP, from which CATH-like annotations can automatically inferred. MODAS classifiers are based on a SVM which operates on combined features from both predicted secondary structure and multiple sequence alignment profiles.

4.4.2.3 Evaluation Framework

In order to evaluate the proposed framework, predictions have to be performed using protein sequences the structures of which are known. Since we intend to simulate *ab initio* protein structure prediction, it is important to make sure that information about the actual native and potential homologous structures is not exploited. As a consequence, we have excluded all homologues in all experiments (we haven't included this step in Figure 4.2 for the sake of simplicity). There are several criteria to run a template-free protein structure prediction in the literature; in all our experiments – in this chapter and next two – we have stick to Rosetta's "default" *ab initio* predictions' policy. This is achieved during the fragment building process where all proteins that are classified as homologous are excluded from the database of protein templates known as "vall". Such a classification is mentioned in Baker lab's *de novo* key experiments (Song et al., 2005) and defined by the "exclude homologous" flag on the "fragment-picker" online server: to remove all proteins that appear in the result of running PSI BLAST of the protein target against the database of templates where the E-value's threshold is 0.05. Position specific iterative BLAST (PSI-BLAST) is arguably one of the most popular and successful multiple sequence alignment tools due to its ability to detect remote homologous. The PSI-BLAST's E-value, for a particular database size and

sequence size, is the expected number of hits in order to receive a sequence comparison score by chance as good as the one observed. For instance, a hit of E-value of 10 – the PSI-BALST's default value – simply means that using the same database and size of sequence in question, one might expect to see 10 hits having the same sequence similarity score. Although E-value is not a probability, very low values (typically less than 0.01) are interpreted to be closely identical to p-values, (see https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html).

First, structural class annotation is conducted according to the aim of the experiment, i.e. concept validation or practicality demonstration using either CATH or SCOP. Second, all structures of the "vall" belonging to same structural class are extracted. Note that "vall" already contains high-quality templates since a 2.5 Angstrom resolution cut-off was already applied to produce high quality fragments. Third, the target and all its homologues (based on PSI-BLAST with an E-value < 0.05) were removed from the set of collected structures. Fourth, the fragment libraries were constructed by providing Rosetta's fragment-picker with this set of protein templates. All the default options – see chapter 3 - were kept including parameter weights and the number of fragments at each position, i.e. 25 for 9-mers and 200 for 3-mers. Finally, since picking and assembling fragments to construct a whole conformation is a stochastic process that relies on Monte Carlo simulation, it needs to be performed a large number of times. As it is appropriate to produce as many as possible structures for each target in an attempt to cover the highest number of permutations amongst the total number of fragments, the recommended value of 20,000 models was chosen for all experiments (Barth et al., 2007).

4.4.2.4 Dataset, Databases and Software Tools

The target dataset comprises 70 proteins selected from the latest CASP contests. First, only proteins containing fewer than 150 amino acids were considered, since larger targets would show a complexity which is generally believed to be beyond the capabilities of Rosetta. Second, the selection process was aimed at producing a set of FM targets showing diversity in terms of structural class. However, in order to be able to produce statistically significant results, the initial set was extended using TBM targets. In any case, the experimental protocol was designed so that predictions would be made independently of the presence of homologous structures in the template set.

In terms of structural class prediction, the two main classification schemes, i.e. CATH and SCOP, were considered. Class annotations used in experiments were collected from two sources: annotations based on actual protein structures – which are treated as the gold standard - and sequence-based predictions performed by MODAS. Finally, structure prediction was performed using the fragment based *de novo* protein structure prediction software offered by the Rosetta suite. Evaluation of the structures was performed using the GDT metric.

4.5 Results

The results are divided into two main parts. First, the quality of the decoys generated by each methodology is extensively presented and compared, without taking into consideration the energy scores. Second, the *first model* – the conformation that is associated with lowest energy score – is also compared amongst the three different experiments.

Since we have used different fragment libraries, this led to new areas in the search space being explored. Decoys' quality - regardless of the corresponding energy scores - is thoroughly discussed in the next three sub-sections. Since a very specific set of structures is used from which to extract fragments, each of the customised libraries is "narrower" than the standard one. Accordingly, the exploitation process has limited the number of funnels to be explored, ideally revealing more local minima in relatively small areas. The second part is dedicated to the blind assessment of our new methodology by comparing the *first model* – the conformation that corresponds to the lowest energy score - of each set of decoys; the same way CASP evaluates competitors' models.

4.5.1 General Performance

We have investigated two methods to compare the quality of decoys: the *best decoy* and the average of the *best 10 decoys*, both out of 20,000. All data are found in Table 4.2. This shows our approach gives an overall tangible improvement in terms of GDT, although CATH tends to be better for the *best decoy*. For the sake of evaluation and comparison (the current section and subsequent two sections), we will adopt the "average of *best 10 decoys*" since it is a much fairer assessment of the experiments by avoiding any real chance of a "lucky hit". Moreover, given inaccuracies of the energy function and the challenges in the area of quality assessment, the set of the *best 10*

decoys is more likely to be hit by the conformations associated with lowest energy scores.

Table 4.2: Overall improvements in terms of GDT of CATH and SCOP-based experiments.

	GDT overall improvement of CATH-based experiments	GDT overall improvements of SCOP-based experiments
Best decoy	6.3%	5.8%
AVG of *Best 10 decoys*	6.8%	6.9%

First, the quality of the models generated by the standard Rosetta framework, i.e. without using any structural class annotation, is compared to those produced using the "gold standard", i.e. structure based, class annotations. As Tables 4.2 and 4.3 show, the average performance for the 70 targets in terms of GDT demonstrates that class annotation allows better structure prediction (~7% improvement). This difference is statistically highly significant since the p-value < 0.0005. On the other hand, there is no significant difference between the SCOP and CATH based approaches (p-values > 0.05). The p-value is calculated as the probability associated with the Student's t-test. Since we are using the same set of protein targets, however, using two different ways of predictions – the standard and our customised ones – a paired two-tailed t-test is obtained as a paired two-tailed distribution of both datasets: 70 GDT of standard predictions and their corresponding GDT of, for instance, CATH-based predictions.

Table 4.3: Average performance (and standard deviation) in terms of GDT and associated p-values. Sequence based annotations are the one taken from MODAS predictions. GDT is the average of the GDT_TS of the 70 targets, which in turn, is the average of the highest 10 scores. The p-value is calculated from the large dataset of GDT – 70 values for each experiment- and not using the averages.

	No class annotation	CATH class annotation		SCOP class annotation	
		Structure based	Sequence based (MODAS predictions)	Structure based	Sequence based (MODAS predictions)
GDT Mean (Std. Dev.)	46.04 (13.89)	48.62 (14.22) p = 0.0002	47.64 (14.10)	48.92 (14.97) p = 0.0004	48.31 (15.14)

In addition, Table 4.3 reveals that predictions based on MODAS automatic annotations are only marginally worse than those based on structure-based class annotations, especially for SCOP. This can be explained, firstly, by the very good accuracy of MODAS predictions and, secondly, by the fact that misclassifications only appear between classes with blurred borders (Michie et al., 1996). Comparison between structure and sequence-based annotations shows that 78.5% and 81.4% of classes have been correctly predicted by MODAS for SCOP and CATH respectively. As expected, there is higher accuracy for CATH since there is no differentiation between alpha/beta and alpha + beta classes. Indeed, the confusion matrix shown in Table 4.4 highlights that confusion only occurs between alpha and alpha-beta, or beta and alpha-beta, or FSS and alpha-beta classes (misclassifications happen since targets lie on the border between those classes), but never between alpha and beta classes. Those results demonstrate that usage of a structural class predictor makes our pipeline practical and allows the generation of better models than those produced by the standard Rosetta framework. Since structural class prediction is an active research area, there is no doubt that performances obtained with predicted classes will get even closer to those attained with actual classes in the near future. Given that the aim of this contribution is to demonstrate and analyse the value of fragment libraries generated from class specific

templates, the remaining analysis in this chapter concentrates on results generated from structure-based class annotations.

Table 4.4: Confusion matrix showing CATH classes versus MODAS predicted ones

Predicted/ Gold standard	Alpha	Alpha_Beta	Beta	FSS
Alpha	15	1	1	0
Alpha_Beta	2	25	3	3
Beta	0	4	14	0
FSS	0	0	0	3

As Figure 4.3 shows, predictions based on structural class annotations outperform standard ones for a majority of targets. Actually, a higher GDT value is obtained for 70.0% and 78.6% of the targets using CATH and SCOP respectively (Figures 4.3 and 4.4). Correlation coefficient between Standard and CATH-based data sets (70 pairs of corresponding values) is approximately 0.92 whilst for SCOP-based is 0.90; both show very strong correlations. More detailed information regarding the amount of improvements and declines is shown in Table 4.5.

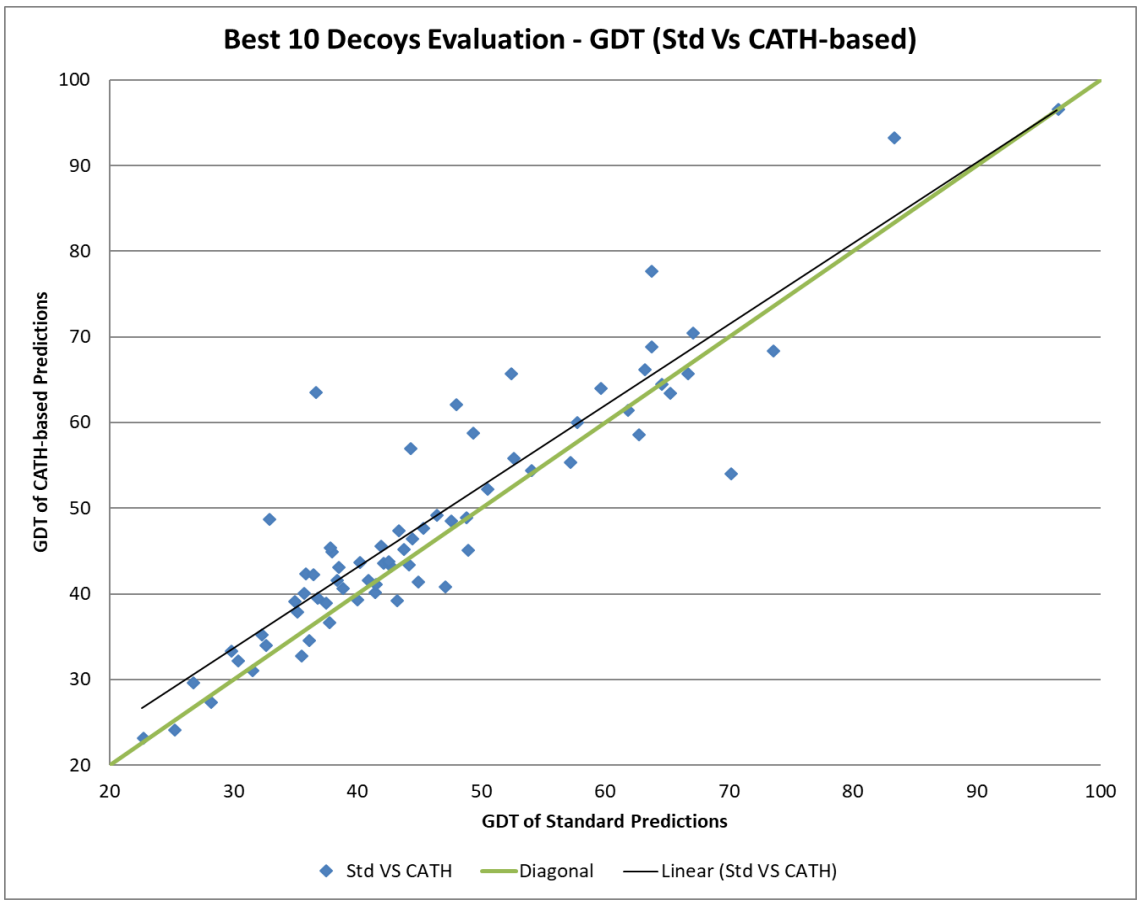

Figure 4.3: GDT of standard predictions versus CATH -based predictions for the 70 targets; 49 targets out of 70 show higher accuracy; linear regression line is shown in navy blue. Overall, our customised predictions show an improvement of 6.8%. Correlation coefficient between the two data sets (each of 70 values) is 0.92

4.5.2 Performance According to Structural Class

Since SCOP and CATH-based produces similar results, we can conclude that those classifications are equally informative in terms of protein template selection; however, that may not be case for all classes. Hence, we have conducted a more in-depth analysis by focusing on performance enhancement according to the structural class of the target (see Table 4.6). First, whatever the classification, targets from all main classes benefit significantly from template selection: the number of targets with models displaying a better GDT is between 61.1% and 100.0%. Interestingly, targets combining Alpha and Beta structures seem to gain more from the proposed methodology. One may suggest that, since structural discontinuities between secondary structure elements are key to a protein conformation, using libraries with a higher content of alpha to/from beta transition fragments leads to better conformation predictions.

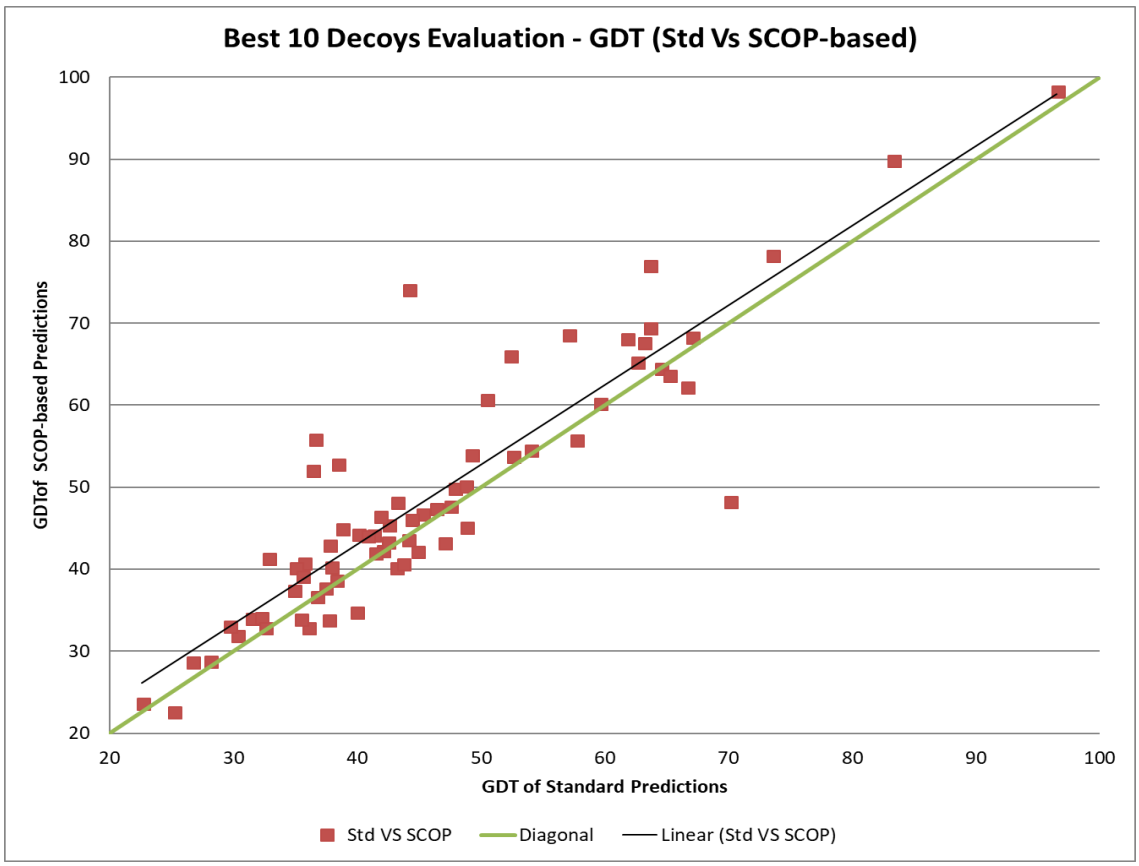

Figure 4.4: GDT of standard predictions versus SCOP-based predictions for the 70 targets; 55 targets out of 70 show higher accuracy; linear regression line is shown in navy blue. Overall, our customised predictions show an improvement of 6.9%. Correlation coefficient between the two data sets (each of 70 values) is 0.9

Table 4.5: Performance comparison for the 70 targets. In both customised predictions, approximately a 5-point increase on the GDT score is recorded on average, which corresponds to 11% of the standard predictions' GDT value.

	Percentage of improved targets (average GDT change)	Percentage of unaffected targets	Percentage of worsened targets (average GDT change)
CATH	70.00% (+4.77, i.e. +11.19%)	0.00%	30.00% (−2.53, i.e. -4.83%)
SCOP	78.57% (+4.77, i.e. +10.98%)	0.00%	21.43% (−4.01, i.e. -8.07%)

Secondly, as expected, association to less common classes that are not specific in terms of structural content, i.e. Few Secondary Structures (FSS) and Small Proteins

(SP), seem to be less beneficial with (SP) or even detrimental (FSS) to structure prediction. Although one should be cautious when discussing results for such a small number of targets, the fact that the number of templates associated with those classes is an order of magnitude lower than those of the main classes may also lead to the generation of fragment libraries which do not cover sufficiently the conformation space. Thirdly, SCOP-based predictions lead to a marginally higher number of targets with improved models (see Table 4.6 for details). One can also note that, except in the case of SP and FSS classes, the number of templates does not seem to impact on structure prediction.

Table 4.6: Performance comparison according to structural class

	CATH-based predictions		SCOP-based predictions	
Targets	Class	Proportion of Targets with better GDT	Class	Proportion of Targets with better GDT
16	Mainly Alpha	75.0%	All Alpha	75.0%
18	Mainly Beta	61.1%	All Beta	77.8%
33 (29+ 4)	Alpha Beta	75.8%	Alpha + Beta	86.2%
			Alpha / Beta	100.0%
3	Few Secondary Structures	33.3%	Small Proteins	66.6%
70	All	68.6%	All	81.4%

4.5.3 Convergence towards native-like conformations

Although we have shown that methods relying on structural class-based libraries generally generate better conformations than the standard Rosetta framework, it is important to know if this leads to a notable change in terms of model quality. To address this question, we performed classification of the average of the best 10 model for each target according to thresholds adopted in the literature. Production of models the GDT of which are above 40 is particularly important since their conformation is believed to have the same ‘shape’ as the target, which may reveal crucial information

about potential proteins' functions (Abbasi, Ghatee, & Shiri, 2013; Kavousi, Moshiri, Sadeghi, Araabi, & Moosavi-Movahedi, 2011). Models whose GDT value is greater than or equal to 85 are judged convenient to solve the phase problem in crystallography (Giorgetti, Raimondo, Miele, & Tramontano, 2005). Conformations with GDT higher than 59 are believed to be 'good enough' (S. Shi et al., 2009), whilst structures with GDT lower than 40 are considered of poor quality or even random (Kalman & Ben-Tal, 2010; J. Zhang et al., 2010). Consequently, we will adopt the following thresholds and associated classes: "Poor" for GDT < 40, "Moderate" for GDT between 40 and 59, "Good" between 60 and 84, and "High Quality" for GDT > 84. As Figure 4.5 shows, whereas the standard Rosetta framework is able to produce informative models for 61.4% of the targets, both SCOP and CATH-based schemes deliver a much larger proportion of them, 74.8% for both.

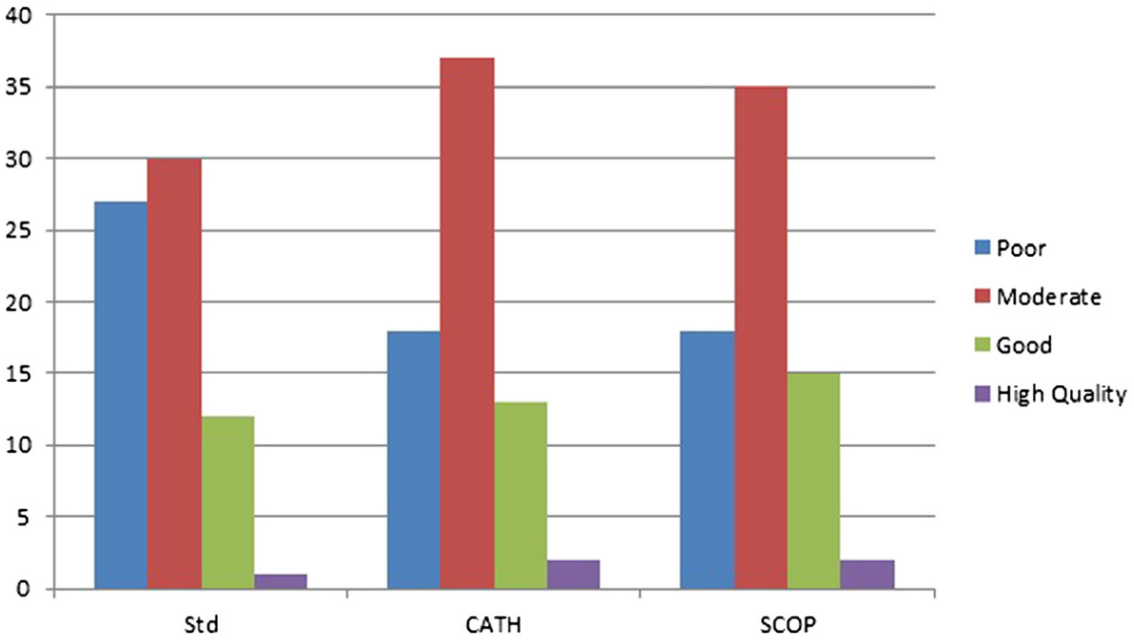

Figure 4.5: Qualitative distribution of the average GDT of the best 10 models.

Since part of the rationale of the proposed methodology is a reduction of the size of the conformation space, we calculated for each target the number of conformations which were generated in order to produce the structure with highest GDT out of the 20,000. SCOP and CATH-based experiments produce their best GDT structures after generating a smaller number of conformations than the standard Rosetta framework, converging towards those conformations, respectively, 2.8% and 6.9% faster (see Table 4.7).

Table 4.7: Average number of conformations for convergence towards the structure with highest GDT (and associated standard deviations).

Standard predictions	SCOP-based predictions	CATH-based predictions
10848 (5469)	9743 (5753)	9452 (5968)

4.5.4 Blind Assessment

Since, based on the target's structural class predictions, different fragment libraries have been generated (3 for CATH and 4 for SCOP), we have showed and discussed the effects of our novel idea on the quality of the decoys generated by Rosetta. It is worth now assessing the *first model* from each experiment to examine whether the new method contributes to overcoming the energy functions inaccuracies by tightening the bound between the *best decoy*s and *first model*.

As shown in Figure 4.6, improvements in terms of the *first model* is remarkable; CATH and SCOP-based experiments' *first model* shows an overall improvement of 25.5% and 15.0% respectively. Correlation coefficient of Standard vs CATH-based data sets (70 pairs of corresponding values) is approximately 0.54 whilst for SCOP-based is 0.62. Although SCOP-based correlation is higher, both are classified as "strong". CATH's improvements in terms of each structural class are further elaborated in Table 5.1. Besides the tangible enhancements in terms of the decoys' quality as shown earlier, the much lower number of template structures available for building fragments has forced Rosetta, in most cases, to "concentrate" more on relatively small, but good areas. Having less structurally diverse, but "good enough" local minima conformations makes the process of choosing the *first model* more "successful", i.e. less "random". Nevertheless, more "real" results considering the *first model* – the formal criterion in CASP – will be shown in the next subsection where we present our results compared to the Rosetta's formal group's contributions to CASP11.

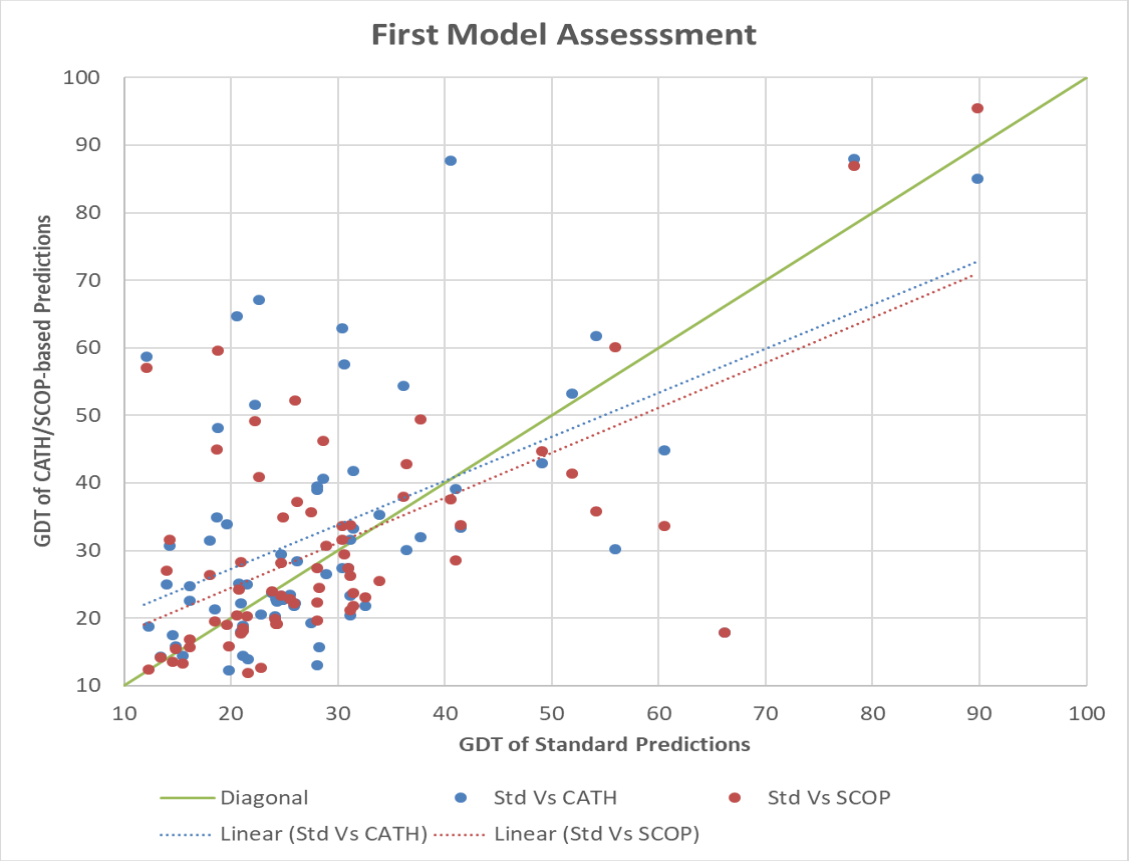

Figure 4.6: *First model*'s GDT comparison; from each experiment, the conformation (out of 20,000) that corresponds to the lowest energy score has been chosen. Correlation coefficient between Standard vs CATH-based and SCOP-based is 0.54 and 0.62 respectively.

4.6 CASP11 Results

For CASP11, we contributed 18 targets by applying the above methodology, but by generating 50,000 decoys instead of 20,000. The relatively small number of targets we were engaged in was due to the low computational resources we had then; the above experiments (70 previous CASP targets) - this chapter is based on - were conducted in parallel with our CASP11 contributions using a limited number of processors on Kingston University High Performance Computer (KUHPC). As mentioned in Chapter 2, each group can submit up to 5 candidate models; one of them should be designated as the *first model* – the main criterion CASP uses to rank the participating groups for each target. Furthermore, CASP relies on domains rather than the whole target. We followed the "default" way to choose the 5 models, i.e. the 5 conformations that correspond to the 5 lowest energy scores. Out of the 18 targets, 6 were cancelled due to advance public release of their corresponding spatial coordinates. Accordingly, the total number of domains is 14, taking into consideration the fact that one target – T0804 – has been

further divided into two separate domains (See Table 4.8 for all details including our results and our main competitors').

Table 4.8: List of targets/domains in CASP11. GDT Scores, that are shown in red and green are respectively worse and better than ours.

Length	Domain Classification	Target ID	PDB ID	Structural Class	Rosetta at Kingston	Baker	Baker-Rosetta Server	Jones - UCL	Zhang Server	Zhang	Quark	Tasser
130	FM	T0763	4Q0Y	B	15.6	14.6	17.7	20.2	16.5	18.6	16.4	19.8
97	TBM	T0769	2MQ8	AB	75.0	75.3	75.3	68.8	80.2	81.2	76.0	78.9
67	TBM	T0773	2N2U	AB	93.3	78.7	78.7	80.2	90.7	87.7	88.4	75
112	FM	T0785	4D0V	B	20.3	25.7	21.2	29.5	23.7	26.8	21.9	22.6
136	TBM	T0795	5FJL	B	13.6	66.4	64.5	59.2	71.9	71.9	71.7	66.2
212	TBM -Hard	T0800	4QRK	B	12.0	43.7	43.3	42.8	42.1	42.2	40.8	43.6
202	FM	T0804			14.2	12.1	12.1	13.7	13.5	13.3	11.6	11.7
37	FM				44.6	30.4	30.4	44.6	44.6	42.6	46.0	34.5
152	FM				18.1	14.8	14.8	17.4	17.3	16.6	14.8	15.1
68	TBM	T0816	5A1Q	A	46.7	35.6	34.6	64.0	35.3	66.2	66.2	57.4
134	TBM	T0818	4R1K	AB	20.5	34.9	34.9	41.6	41.2	41.8	41.4	57.4
114	TBM	T0822	5FU5	B	18.6	50.0	50.0	46.3	39.7	43.2	44.5	50.5
108	FM	T0824	5OMT	AB	25.9	52.6	28.5	28.9	29.4	28.7	29.2	28.7
204	FM	T0836			20.2	29.9	17.9	44.1	20.1	19.6	20.6	20.7

Since our methodology is an adapted version of Rosetta, the main competitor is "Baker-Rosetta Server": the fully automated server which represents Robetta (already introduced in Chapter 2); recall that for FM targets, Robetta is simply Rosetta. Another Rosetta-related group, called "BAKER", has been added which belong to the human assisted approaches that rely on contact maps predictions as well as human intervention.

Furthermore, we have shown values from other research groups whose tools were introduced in the literature review chapter and lie within the same category – fragment-based assembly - namely “Jones UCL” (FragFold), “Zhang Server”, “Zhang”, “Quark” and “Tasser”. Both “Zhang Server” and “Zhang” represent I-TASSER, with minor difference related to the human intervention in the latter.

Since CASP7, Rosetta has been taking advantage of their distributed computing project Rosetta@Home to tackle conformational sampling (Das et al., 2007; Ovchinnikov, Kim, et al., 2016; Raman et al., 2009). In CASP7, around 500,000 CPU hours were dedicated for each domain (Das et al., 2007). Typically, whenever the grid computing service is engaged, the number of decoys may be very large; for instance a study conducted in 2011 by the Rosetta team, they generated up to 600,000 decoys per target (Tyka et al., 2011), not to add the team of biologists, biochemists and biophysicists that are involved in the human intervention part. Nevertheless, our results show higher accuracy in terms of GDT for 6 out of 14 domains compared with both Rosetta groups as shown in Figure 4.6.

4.7 Discussion

Following an exhaustive evaluation of our methodology, we have demonstrated that usage of class annotations leads to highly statistically significantly enhanced structure prediction performance (p-values < 0.0005), even though they have been predicted from the amino acid sequence alone. Although experiments were conducted using two different types of structural classifications, i.e. CATH and SCOP, there is no convincing evidence suggesting that one is more appropriate than the other in terms of the top 10 decoys, whereas CATH outperforms SCOP whenever the *first model*, i.e. the model with lowest energy, is taken into consideration. Performance analysis according to structural type class shows that targets from all main and well-defined classes benefit from the proposed methodology.

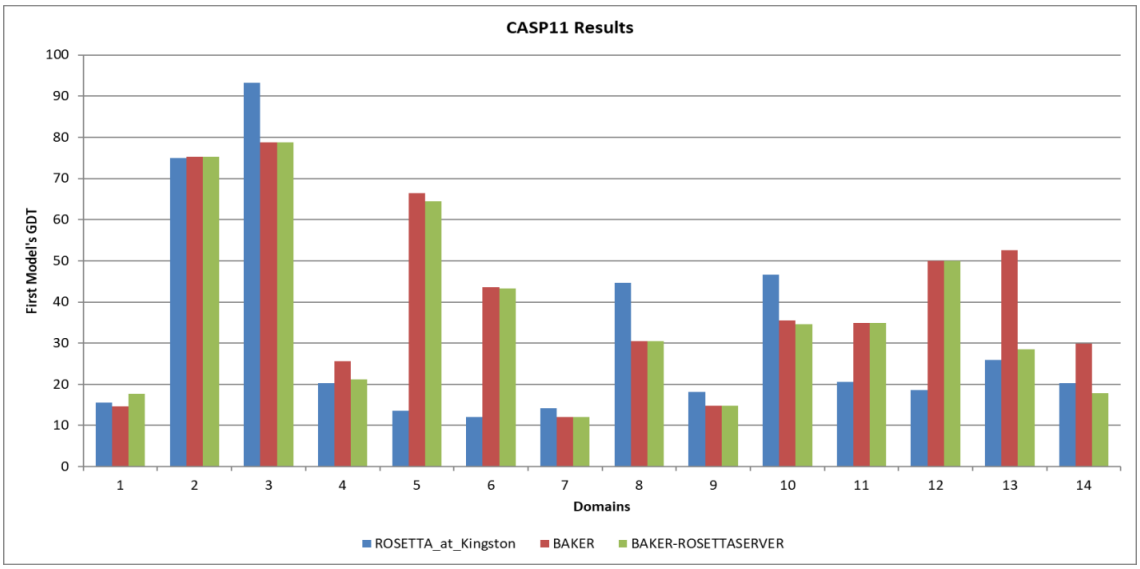

Figure 4.6: Results of 14 domains amongst three groups: Rosetta_at_Kingston, BAKER, and BAKER-ROSETTASERVER.

Moreover, the quality of the structure prediction does not appear to be influenced by the number of selected templates, if this is over one thousand. All these results support our hypothesis that, in terms of structural relevance, template quality is more important than quantity and diversity. In addition, experiments conducted using structural class prediction demonstrates the proposed methodology is practical. Further analysis of the results also shows that methods relying on class-based libraries produce conformations which are more relevant, i.e. more 'good' and 'accurate' models are generated. In addition, since structure-based predictor models converge quicker towards the *best decoy*, this substantiates our claim that usage of structurally relevant templates contribute to reducing the size of the conformation space to be explored.

4.8 Conclusion

In this chapter, we have proposed the use of structural class constraints for *ab initio* fragment-based protein structure prediction, to decrease the size of the conformation search space. Then, using Rosetta, a comprehensive evaluation of our methodology has been conducted on a set of recent CASP targets. We have demonstrated that exploitation of class annotations leads to enhanced structure prediction performance. Results also support our hypothesis that focusing on a better focused structure space contributes to quicker identification of better models.

Since our methodology produces models the quality of which, in terms of GDT, is up to 7% higher in average than those generated by a standard fragment-based predictor, we believe our approach should be considered before conducting any fragment-based protein structure prediction. Despite such progress, *ab initio* prediction

remains a challenging task, especially for proteins of average and large sizes. Apart from improving search strategies and energy functions, integration of additional constraints seems a promising route, especially if such constraints can be accurately predicted from the amino acid sequence alone.

5 Reduced Fragment Diversity for Alpha and Alpha Beta Protein Structure Prediction

5.1 Introduction

In order to generate a conformation's backbone along with its side chain centroids, Rosetta operates in two main steps: first, 9-mer fragments are inserted into the initial fully extended conformation; second, insertions of 3-mer fragments are used to refine the structure previously generated. 9-mers and 3-mers are protein fragments extracted for each amino acid - except for the protein C-terminus - of the protein of interest from a template database according to some similarity criteria. Eventually, Rosetta converts the coarse-grained conformation into an all-atom representation by adding all missing atoms using knowledge-based information extracted from known structures. The proposed approach in this chapter relies on a limitation of 3-mer fragment diversity so that conformations generated during the 9-mer phase can be refined in more depth than with the standard Rosetta settings. Similarly to the previous chapter where the methodology relied on structural class predictions, we present a new methodology to improve the *first models* of two main classes, however by keeping the same default library of fragments.

The rest of this chapter is organised as follows. First, a thorough study of the effects of the selection of 3-mers on the quality of conformations generated by Rosetta is presented. Second, we introduce a new pipeline for Rosetta protein structure prediction dedicated to alpha and alpha beta proteins. Following a description of the evaluation framework, we justify both theoretically and experimentally the principles of the proposed method. A variety of experiments are then conducted to validate them, and results are discussed in light of other relevant studies.

5.2 Overview, Motivation and Preliminary Experiments.

Rosetta's first phase with its 28,000 9-mer insertion attempts is considered the essential part of the process since it builds the general shape or fold of the structure guided by secondary structure predictions. Those insertions are divided into several sub-phases where more terms of energy functions are successively added to tighten the acceptance criterion of a fragment replacement. 9-mer insertions can be seen as relatively coarse scale operation as each insertion may change the structure being built

dramatically. Although this allows escaping from local minima, this coarse modelling phase is unlikely to reach a near-native conformation.

As a consequence, Rosetta includes a 3-mer insertion phase to improve that initial conformation by performing 8,000 additional insertion attempts. Although the 3-mer insertion phase is generally seen as structure refinement, the fact that by default Rosetta uses 200 fragments means that they can be quite diverse and insertions may on occasion lead to dramatic structural corrections. Whereas those corrections are certainly beneficial to conformations which failed to adopt the correct fold during the first phase, they may be detrimental to those which only needed some fine tuning. In this work, it is proposed to investigate and exploit that hypothesis by adapting the number of 3-mer fragments according to the perceived structural complexity of the protein target.

First, the 'correction' abilities of the 200 3-mer fragments are demonstrated by generating 20,000 decoys using Rosetta without the initial 9-mer insertion phase. Although, as expected, performance is generally below that of the standard two-phase Rosetta (-21.8% in terms of GDT of the model with the lowest energy, or *first model*), the 3-mer only version of Rosetta was still able to generate a better *first model* in 9 out of the 33 tested targets, see Figure 5.1. This experiment clearly demonstrates that usage of 200 3-mer fragments could go well beyond refinement, and also has some abilities of conformation generation. Figure 5.2 illustrates this for an example where 3-mer only insertions are able to generate a good quality *first model* (74.7 GDT) for a target of length 94. Structures are visualised using PyMol (Schrödinger, LLC, 2015). Second, it has been shown in the previous chapter that the performance of the standard version of Rosetta depends on the structural class of a protein target. Recall that the three main classes in CATH are mainly alpha, mainly beta and alpha beta. Taking into account the 67 targets of the previous study, i.e. excluding the three targets that belong to the class of few secondary structures, a detailed study on the *first models*, is shown in Table 5.1. Alpha and alpha-beta protein conformations are better predicted than mainly beta proteins.

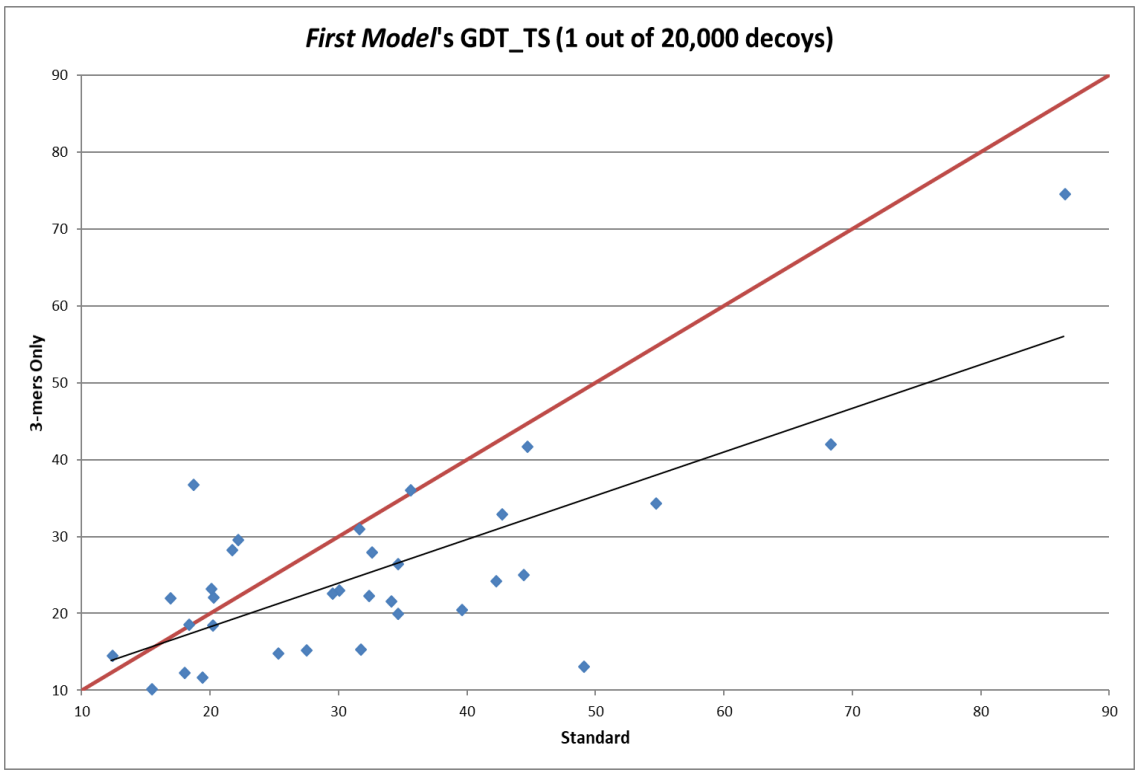

Figure 5.1: *First model*'s GDT out of 20,000 decoys of standard predictions versus predictions using 3-mers only; 9 out of 33 targets achieve better results in the "9-mers"-free Rosetta experiments. The navy blue line represents the linear regression; correlation coefficient between the two data sets (33 pairs of GDT values) is 0.73.

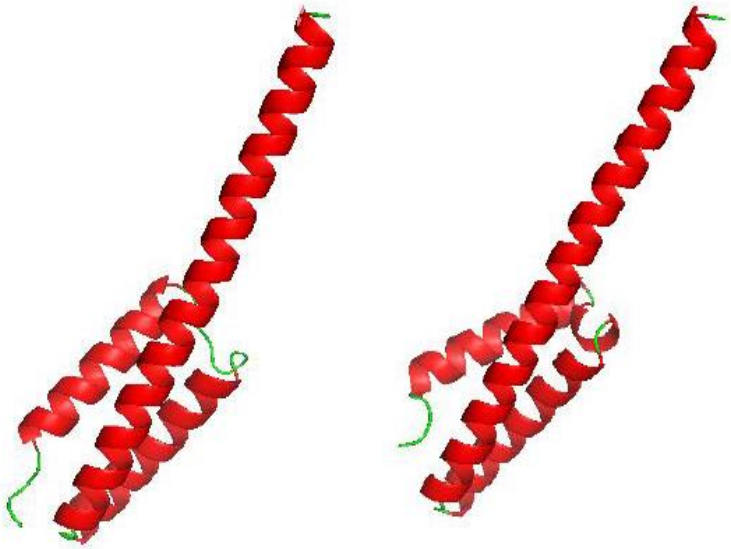

Figure 5.2: Structures of the native model (PDB ID: 4FM3) and *first model*'s conformation using the version of Rosetta using only 3-mers.

Table 5.1: Results of a thorough study on 67 targets performed in the previous chapter.

	Number of targets	Average of the *first model*'s GDT for standard Rosetta	Average of the *first model*'s GDT for CATH-based Rosetta	Improvements
Mainly Alpha	16	39.5	46.5	17.7%
Mainly Beta	18	23.4	25.7	9.6%
Alpha Beta	33	27.4	31.5	15.0%

5.3 Proposed Methodology

In view of those experimental results, it is proposed to adapt Rosetta's 3-mer phase according to a target's structural class. Since structure prediction of proteins belonging to alpha and alpha-beta structural classes is more accurate, one may infer that the 'correction' behaviour of the 3-mer phase is less needed whereas additional refinement could lead to the generation of better models. Here, it is demonstrated this behaviour can be achieved by reducing 3-mer diversity. Figure 5.3 shows a new processing pipeline describing optimisation of Rosetta's 3-mer phase according to a target's predicted structural class.

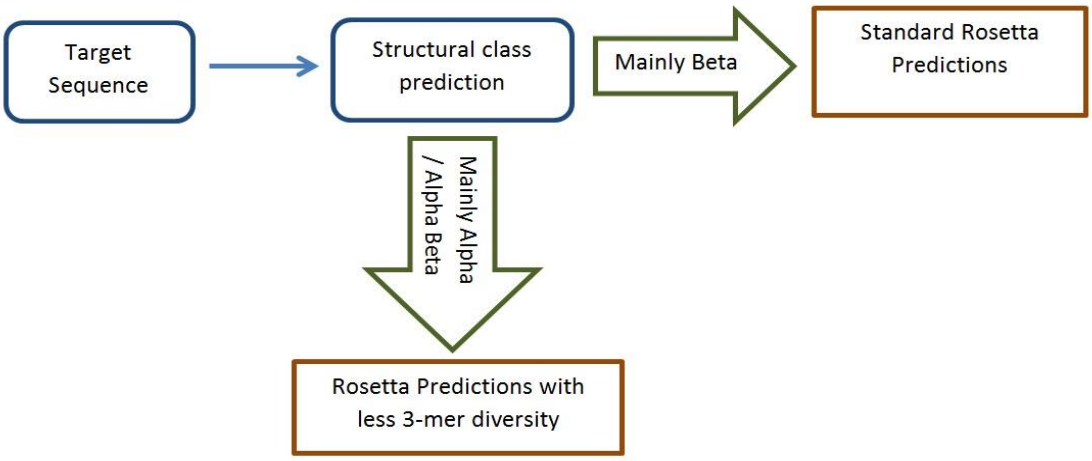

Figure 5.3: New pipeline for optimisation of Rosetta for mainly-alpha and alpha-beta protein structure predictions.

The evaluation dataset covers the three main protein structural classes, i.e. mainly alpha, mainly beta and alpha beta. It comprises 33 targets the length of which ranges from 33 to 141 amino acids. As in the previous chapter, they were selected from models of the Free Modelling and Template-Based Modelling categories assessed during CASP8, 9, 10, 11 and CASP ROLL. All homologues, defined by an E-value lower than 0.05 on PSI-BLAST (Altschul et al., 1997), were removed from Rosetta's fragment libraries. This threshold is typically used to evaluate Rosetta's ability to infer new folds, which is its raison d'être. The reason behind decreasing the size of the dataset (70 targets in the previous chapter) has been a reduced access availability to our high-performance computing (HPC) facility because, since we finished our previous study, it has become more popular for being used by other researchers from within the university. Moreover, a set of 33 has been assessed as being sufficiently large to infer conclusions, as key studies and improvements that took place on Rosetta involved even a smaller number of targets (Barth et al., 2007; Blum, Jordan, & Baker, 2010; Simoncini et al., 2012, 2017; Simoncini & Zhang, 2013). Similar to the work described in the previous chapter, the formal evaluation metric used is again the Global Distance Test – Total Score (GDT-TS) (GDT in the text).

5.4 Evaluation and Results

As Figure 5.4 shows, 14 out of the 23 targets belonging to the alpha and alpha-beta structural classes achieved a higher *first model* GDT when the number of 3-mer fragments was reduced to 100; an overall improvement of 11.5% is recorded on average. However, when all 33 targets are considered, this benefit from the reduction of the number of 3-mers decreases to +6.5%, since the GDT scores of mainly beta targets fell on average by 10.8%. Those results confirm that removal of some 'correction'

fragments improves predictions of alpha and alpha-beta structures, while it degrades the generation of beta structures, see Table 5.2. In an additional experiment, where the number of 3-mers was further decreased to 25, reveals that such a dramatic reduction of 3-mer diversity leads to similar performance when all targets are considered. However, when only alpha and alpha-beta targets are considered, usage of 25 3-mers still delivers slightly better performance (+2.5%) than the standard approach (note that Rosetta uses an additional 25 9-mers in the first phase of the prediction process). Table 5.2 displays a summary of the results of this *first model* study.

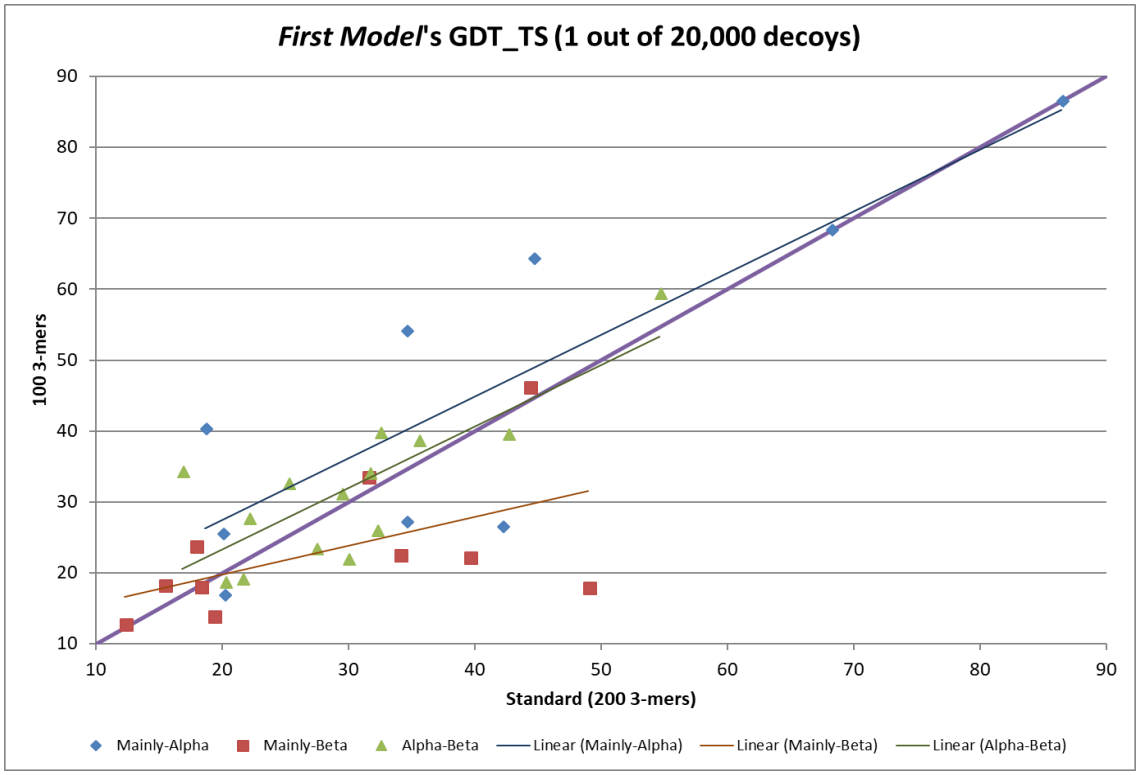

Figure 5.4: *First model*'s GDT of standard predictions versus predictions using 100 3-mer fragments only for the three structural classes along with their corresponding regression lines. The overall correlation coefficient between the two datasets (33 pairs of GDT values) is 0.79. However, dividing the data into three subsets (Mainly Alpha, Mainly Beta and Alpha-Beta) and producing a regression line for each subset gave correlation coefficients of 0.84, 0.53 and 0.79 respectively.

Table 5.2: Comparison of *first model*'s quality according to 3-mer reduction strategy relative to the standard approach.

Average GDT change of *First models* compared to standard approach					
	All three classes	Mainly alpha	Mainly beta	Alpha beta	Mainly alpha and alpha beta classes only
100 3-mers	+6.5%	19.7%	-10.8%	9.7%	+11.5%
25 3-mers	+1.0%	11.1%	-13.6%	4.0%	+2.5%

An example evidence of the accuracy of the 100 3-mers approach for alpha targets over the standard approach is displayed in Figure 5.5: except for the N and C-terminus coil regions, the conformations of which are predicted incorrectly, the structure of the *first model* generated using 100 3-mers is very close to the native one in terms of fold and alpha helix topology; on the other hand, the *first model* from the standard approach is much less accurate due to the incorrect orientation of the third alpha helix.

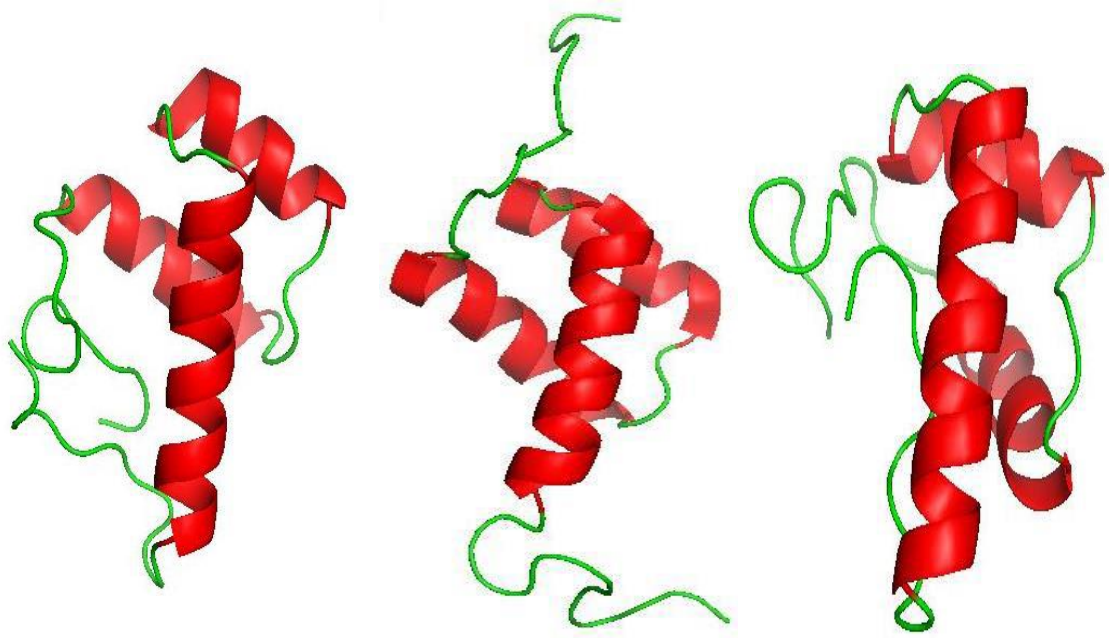

Figure 5.5: From left to right: Structures of 100 3-mer approach's *first model* (GDT = 64.5), native (PDB ID: 2LY9) and standard approach's *first model* (GDT = 44.5) of that 74-amino acid protein, respectively.

The evaluation of the structure-energy correlation amongst the three experiments is performed by calculating the percentage of the *Best model*'s GDT achieved by the *first model*'s. As shown in Table 5.3, for the mainly alpha and alpha beta classes, the

100 3-mer approach delivers *first models* which are closer +8.0% to the *Best model*, , than those of the standard approach.

Table 5.3: Comparison of structure-energy correlation in terms of GDT.

Average of percentage of the *Best model*'s GDT achieved by the *first model*'s		
	All three classes	Mainly alpha and alpha beta classes only
Standard	62.2%	60.7%
100 3-mers	62.3%	65.7%
25 3-mers	59.0%	61.9%

It is worth noting that although there were no tangible improvements in terms of the *best decoy*s – regardless of the energy scores – standard predictions were not able to perform better than our approach. In other words, even if an optimal quality assessment tool existed, our customised predictions would show the same quality for the *best decoy*s. Figure 5.6 demonstrates that 100 3-mers and 25 3-mers based predictions' *best decoy*s are as good as standard predictions' (+1.8% and +0.7% respectively).

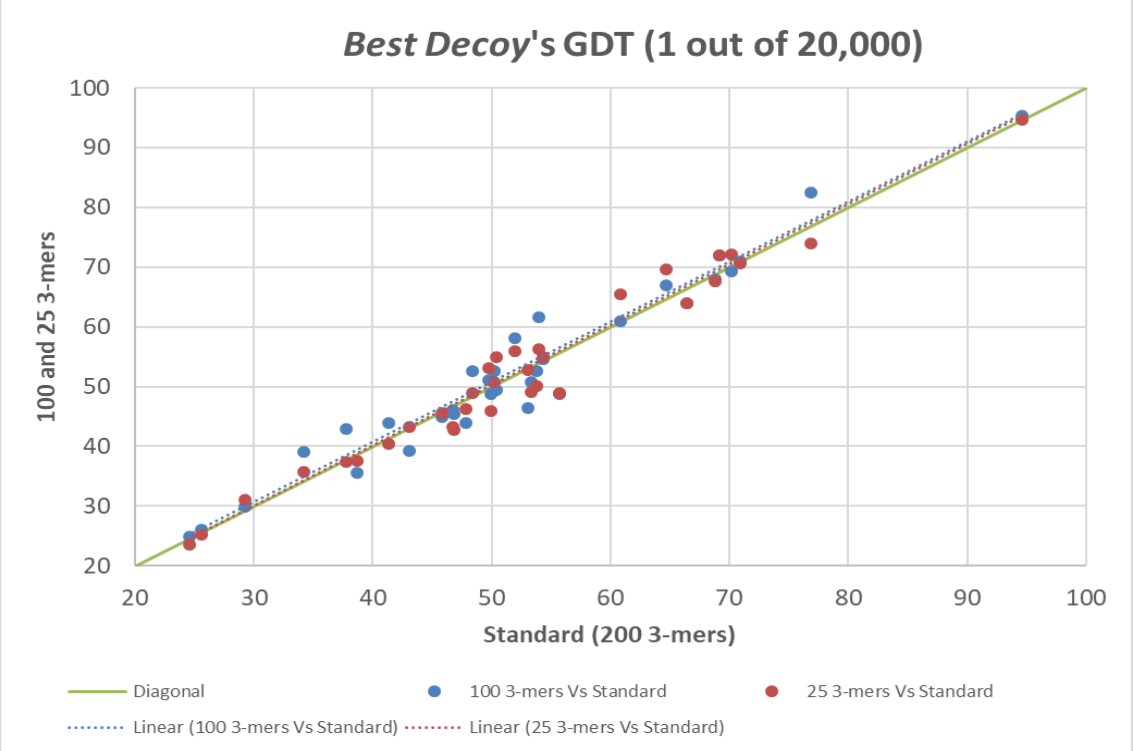

Figure 5.6: GDT of *best decoy* of the standard predictions versus those of predictions using 100 and 25 3-mer fragments only. The correlation coefficients are 0.97 and 0.98 respectively.

5.5 Discussion

Although Rosetta generates all-atom models, it relies on moderate coarse-grained protein modelling, where each amino acid is represented using C-alpha, C-Beta and the side chain's centroid. As a consequence, the energy landscape that Rosetta explores is expected to be quite smooth, as illustrated in Figure 5.7 (Kmiecik et al., 2016).

When the 9-mer insertions phase is performed, the energy landscape is explored through relatively "big jumps" corresponding to 9-mer substitutions. Consequently, the local minimum of a given funnel (position A in Figure 5.8-a) may not be reached, leading to a locally suboptimal conformation (position B in Figure 5.8-a). Then, during the 3-mer insertion phase, the dual role of correction and refinement is played. Whilst refinement allows moving a conformation deeper into the current funnel, correction permits investigating neighbouring funnels. On one hand, the more diverse the 3-mer library (e.g. 200), the larger and the wider the set of explored funnels is likely to be. On the other hand, less diversity (e.g. 100) is likely to reduce the size of the searched space allowing deeper exploration of the initially selected funnel (Figure 5.8-b).

This suggests that, when dealing with the easier targets, i.e. from alpha and alpha-beta classes, the 9-mer insertion phase tends to succeed in identifying a funnel close to the native area. As a consequence, usage of 3-mers with relatively low diversity, e.g. 100 fragments, is beneficial allowing exploration of that zone more in depth and eventually producing a more optimal conformation. Alternatively, for the harder targets, i.e. from the beta class, where the 9-mer insertion phase is less likely to have generated a conformation close to the native one, keeping a larger search space by using quite diverse fragments, e.g. 200, increases the probability of converging towards an acceptable conformation.

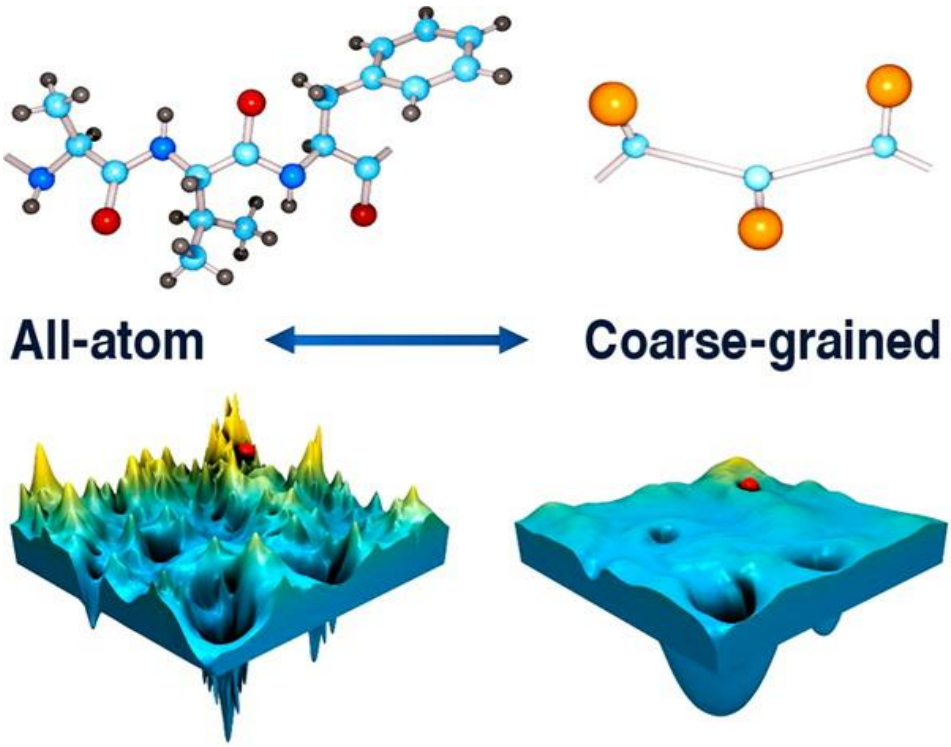

Figure 5.7: Energy landscape of all-atom versus coarse-grained protein modelling. Taken from (Kmiecik et al., 2016).

Usage of a reduced set of 3-mers for the easier targets is further supported by studies which demonstrated that native and native-like structures are likely to be found in the largest cluster/ broader funnel of decoys (Betancourt & Skolnick, 2001; Shortle, Simons, & Baker, 1998), see figure 5.9. Those observations have resulted in development of quality assessment prediction techniques, known as decoy clustering (Perez et al., 2014), to identify the "*Best model*" produced by *ab initio* methods that, like Rosetta, generate a large number of candidate structures known as decoys. Their strategy relies on, first, clustering those decoys according to some threshold similarity threshold, typically 3 to 4 Å, and, second, selecting the conformation with the lowest energy score from the largest cluster.

On the other hand, the outcome of this chapter is quite consistent with Baker and co-workers performed during the late 90's when they were thoroughly investigating the sequence-structure correlation (See Section 3.2 in Chapter 3). Without taking any specific secondary structure into account, their concluded that the sequence-structure correlation shows dramatic increasing relative entropy as the length goes from 3 to 8 amino acids, then a slow increase till 10 (the peak) followed by a slow decrease till 15. Moreover, those authors investigated this correlation for specific secondary structures and super-secondary structure motifs. Some "ideal" lengths of fragments were found as

follows: 13 and 15 for helix caps (helix-turn-helix motifs), 7 to 11 for helices, 7 and 9 for turn-to-sheet, 3 and 5 for β-strands, loops and turns. This conducted Baker and his team to adopt sizes of 9 and 3 in Rosetta. Results presented in Table 5.3 are coherent with the outcomes of that 20-year old study. While a large number of 3-mers would destroy helices and helix caps already built using fragments of size 9, i.e. fragments close to their ideal length for those structures, Beta strands are unlikely to display their correct configuration following the 9-mer insertions phase. Consequently the 3-mer phase is critical to model accurately Beta strands.

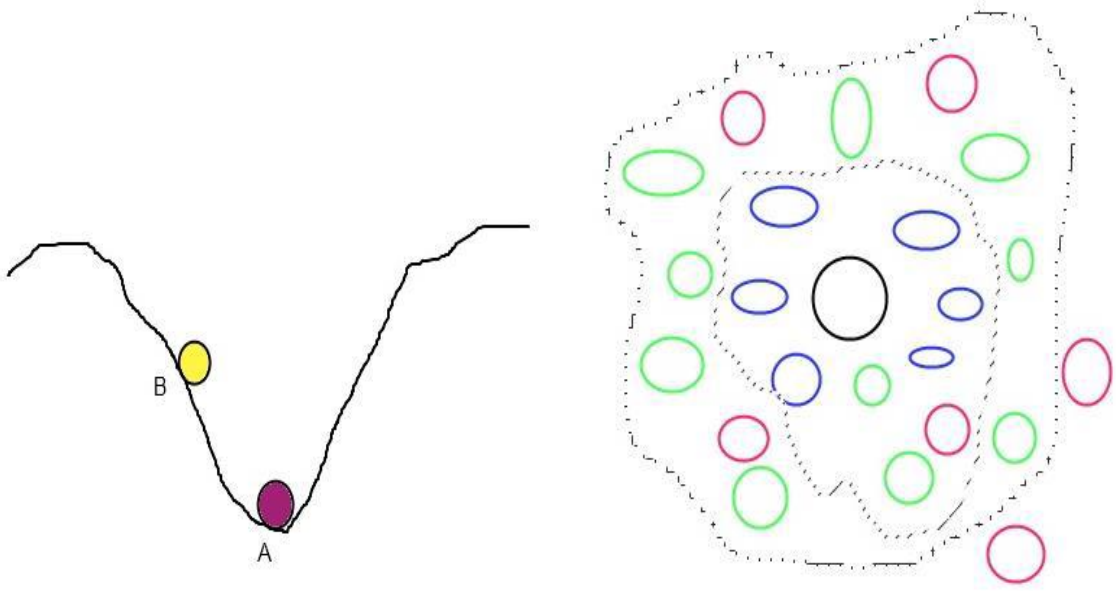

Figure 5.8: (a) Positions A and B illustrate the energy levels of the conformations resulting from the 3-mer and 9-mer insertion phases respectively. (b) The black circle represents the funnel which contains the conformation produced by the 9-mer insertion phase. Blue ellipses represent funnels that contain structures with good accuracy, whereas green and purple ones have worse accuracy. The inner, respectively outer, the dashed contour denotes the limit of the search space created by less, respectively more, diverse 3-mer insertions.

5.6 Conclusion

This chapter has presented a comprehensive study on the importance, role and effects of the fragments of size 3 in Rosetta protein structure prediction for the three main structural classes. Usage of the standard number of 3-mers for each position – i.e. 200 – has been shown to degrade alpha and alpha-beta protein conformations initially achieved by assembling 9-mers. Owing to the high accuracy of structural class prediction from sequence, a new Rosetta's pipeline dedicated to alpha and alpha beta proteins has been proposed where 3-mer diversity is reduced. Experimental results have

confirmed that a smaller number, namely 100, of less diverse 3-mers is more appropriate when predicting alpha and alpha-beta targets since it allows Rosetta focusing on the refinement of the initially generated conformations. In addition to produce better quality "*first models*", those models delivered by the proposed pipeline prove to be significantly closer to the actual "*Best model*", which is directly relevant to life scientists.

Based on the above, a potential future work may involve restricting the 3-mer insertion phase to specific regions where secondary structure predictions show high probability of beta strands, coils and turns. A more advanced idea would be to separate the role of each phase; the 9-mer phase would be responsible for the helices and helix caps whilst the next phase will take over the remaining regions

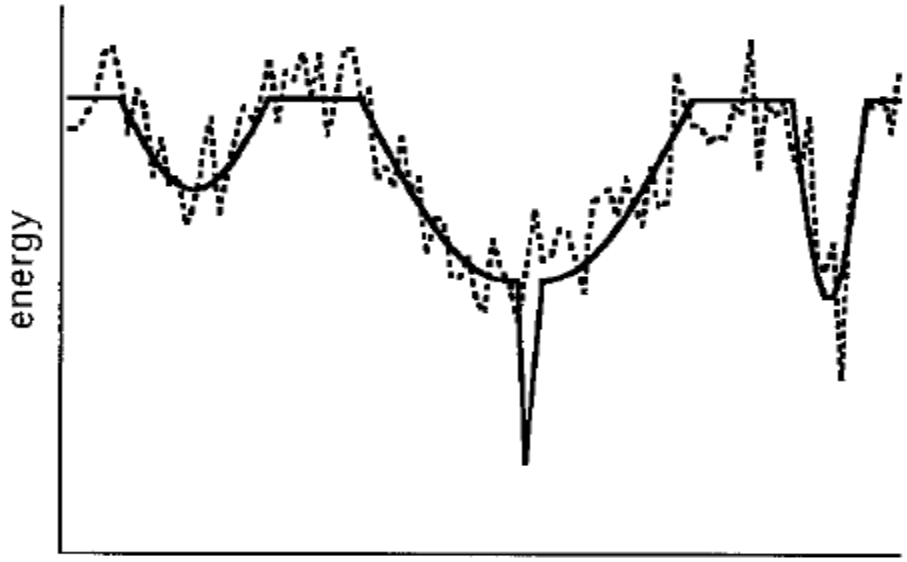

Figure 5.9: Hypothetical folding energy landscape. The solid and dashed lines represent the "real" energy and force field scores respectively (y axis) according to a generalised structure coordinate. This shows clearly that the broader funnel is the one that comprises the "best candidates" as they neighbour the native one. Taken from (Shortle et al., 1998).

5.7 CASP12 Results

By the time we were still performing the above experiments, CASP12 had been launched. We wanted to contribute in CASP12, however, with different approach than CASP11's. Due to time and computing resource limitations, and although we hadn't finished validating it, we decided to combine Protein structure prediction based on structural class annotations from previous chapter and reduction of the number of 3-mers. At the time, usage of 25 3-mers seemed the most promising. We contributed in a

total of 30 domains, however 8 of them were cancelled due to early release of their spatial coordinates. Our group showed better *first model*'s results in 4 domains only out of 22.

A concern was raised by David Baker in a 7-year old paper and was stated as follows "There is a tension between sampling too broadly (giving too diffuse a library) and sampling too narrowly (risking missing a critical set of torsion angles for a portion of the protein chain)" (Gront et al., 2011). Most likely, our combined methodology, i.e. reducing the database of template structures by up to 82% (previous chapter's technique) and decreasing the number of 3-mers by 87% has fallen in the second side of the Baker's concern; we sampled too narrowly.

6 Weighted Protein Regions for Fragment Cardinalities Based on Secondary Structure Prediction

6.1 Introduction

In our first contribution – Chapter 4 – we have created a customised fragments' library by restricting the set of structures where those fragments are extracted from. Such an experiment was able to improve quality of both decoys and *first models*. In the second contribution – Chapter 5 – we have introduced a methodology to decrease the number of 3-mers – that are responsible for both minor conformational changes and corrections for specific proteins that belong to certain structural classes. Likewise, a tangible improvement was shown, however, in terms of *first models* only as Rosetta wasn't able to discover new search area but rather focusing on a specific region where diversity of 3-mers is likely to ruin alpha and alpha-beta proteins' final conformations. In this chapter, we investigate further the optimisation of fragment usage. However, instead of changing the library of structures where fragments are selected from or the criteria of selection, here it is proposed to set the number of available candidate fragments for each position according to the secondary structure annotations that the fragment – either 9-mer or 3-mer – is likely to adopt. The proposed methodology takes advantage of strong empirical previous studies (presented in the next section); for instance, it has been shown that a set of alpha helix fragments is unlikely to comprise outliers; therefore, a subset of them or even a single randomly selected one should be "good enough" (de Oliveira et al., 2015). Thus, it is proposed to optimise the space search by dedicating more time on exploiting relatively small areas in the search space. This quite novel and simple idea does not require any resampling overhead as described below in similar implementations to exclude unnecessary fragments.

The rest of this chapter is organised as follows: the next section is dedicated to the background and related work to highlight the fact that some fragments used in Rosetta are not only unnecessary but also counterproductive. The third section entitled "Sequence-Structure Relationship for Different Secondary Structures" sheds the light on the diversity of fragments based on their secondary structure annotations. Sections 6.4 and 6.5 provide a detailed description of our proposed methodology as well as the dataset and performed experiments.

6.2 Background and Related Work

Resampling techniques have been widely adopted in Rosetta (Blum et al., 2010; T. Brunette & Brock, 2008; T. J. Brunette & Brock, 2005; Shrestha & Zhang, 2014; Simoncini et al., 2012; Simoncini & Zhang, 2013). This paradigm was able to narrow the search space by treating the first round of sampling as "draft" or training predictions, in order to detect "successful fragments" to be selected in the followed predictions. During the first set of experiments, data that reveal some clues to detect regions where near-native structures are likely to be are gathered and used to feed the next set(s) of experiments to focus/exploit deeper specific regions.

Using Rosetta, Brunette and Brock implemented their Model-Based Search (MBS) instead of Monte Carlo (MC) and showed a 14% improvement for the lowest energy models. After each iteration, they identify the regions that can be considered as funnels, then based on their shape, size and energy score, they select the most relevant ones which define the new search space. They are explored by reusing only the fragments from which conformations in those regions are formed (T. Brunette & Brock, 2008; T. J. Brunette & Brock, 2005).

An interesting piece of work in this context was carried out by the Rosetta team (Blum et al., 2010). In the first round of standard Rosetta predictions, structural and energy-based information such as torsion angles, secondary structures and beta pairings are gathered to be fed as restrictions to the subsequent rounds. At the second round, combination of the frequency of that information along with the low energy scores regions are used to change the way that the picking fragments process is carried out by "Fragment-picker" by using different criteria. Besides further exploiting promising funnels already explored in the first turn, the authors suggest their new methodology was able to even reach new region since new fragments have been used. Improvement averages of 1.7 Å and 0.4 Å were shown in terms of best prediction and best-of-five prediction respectively.

EdaFold, an advanced resampling technique based on Estimated Distribution Algorithm, has been successfully carried out on Rosetta in three releases (Simoncini et al., 2012, 2017; Simoncini & Zhang, 2013). The main goal of such approach is to assign and amend a probability value for each fragment in each subsequent iteration using the estimated probability mass functions (PMFs) by further focusing on the fragments found in the basins where the number of low-energy decoys lie. In their latest paper,

they introduced structural dissimilarity besides energy score as a second criterion for choosing guiding models – Fragments that have built those models will have their associated probability raised, consequently to be picked more often in the next conformation assembly iteration. This new addition takes advantage of as many "deep funnels" as possible as from each of those funnels, only one guiding model – the one having the lowest energy score - is chosen. They reported having more than 19% and 8% improvements in terms of best prediction and best-of-five prediction respectively (Decoys' quality was not shown).

These studies suggest that the standard number of fragments is much larger than what Rosetta needs to reach native-like conformations provided that the search space already explored in the first round contain such a region. All those ideas comprise at least a second customised round of sampling, which, therefore, requires more decoys, time and effort; for instance, EdaFold is 2.5 times slower than Rosetta. It is worth noting that the majority of the above-mentioned approaches do not make search trajectories go beyond the space reached in the first round during the subsequent rounds. As a consequence, they force the search to focus on smaller but presumably promising regions for better exploitation (Blum et al., 2010). We have the same objective in this chapter, however, without any additional data collection, and additional and dependent rounds.

6.3 Sequence-Structure Relationship for Different Secondary Structures

Sequence-structure mapping has been studied for a long time (Sibanda & Thornton, 1985; Vanhee et al., 2011). Although no accurate one-to-one function has been discovered even for quite small sequences, many studies suggest that the degree of complexity of such a mapping depends on the secondary structure; in increasing order it is as follows: alpha helices, beta strands, and then coils (Baldwin, 2013; Fiser et al., 2000). Particularly, fragments for alpha helices used for fragment-assembly protein structure prediction have shown a relatively limited diversity (de Oliveira et al., 2015). On the other hand, loops modelling and corresponding fragment prediction has attracted a huge amount of research due to the diversity of those structures as well as the influence of many factors such as length, anchor region – the secondary structures that limit the loop - and external interactions with the environment (Baldwin, 2013; Burke, Deane, & Blundell, 2000; Chothia et al., 1989; Donate, Rufino, Canard, & Blundell,

1996; Fernandez-Fuentes & Fiser, 2006; Fiser et al., 2000; Kwasigroch, Chomilier, & Mornon, 1996; Pardon et al., 1995).

Very interesting research work, that happens to support ours, was carried out by Charlotte Deane and her co-workers at Oxford as they built a new fragment builder called "Flib" to replace Rosetta's and other state-of-the-art ones (de Oliveira et al., 2015). "Flib" creates fragments the length ranges of which are between 6 and 20. One of the main criteria they relied on to build was classifying fragments in four groups based on their dominant predicted secondary structures, namely majority α-helical, majority β-strand, majority loop and neutral. Authors showed that whenever fragments are classified as predominant predicted secondary structures, they tend to be more accurate: the impact is particularly strong for α-helical fragments followed by β-strands ones, whilst loop dominated fragments remain a real challenge as depicted in Figure 6.1. They compared their own tool Flib with HHFrag (Xu & Zhang, 2013) and NNMake -Rosetta's old tool. As shown in Figure 6.2, Flib performs better than those three competitors when dealing with fragments from the α-helical and β-strand classes.

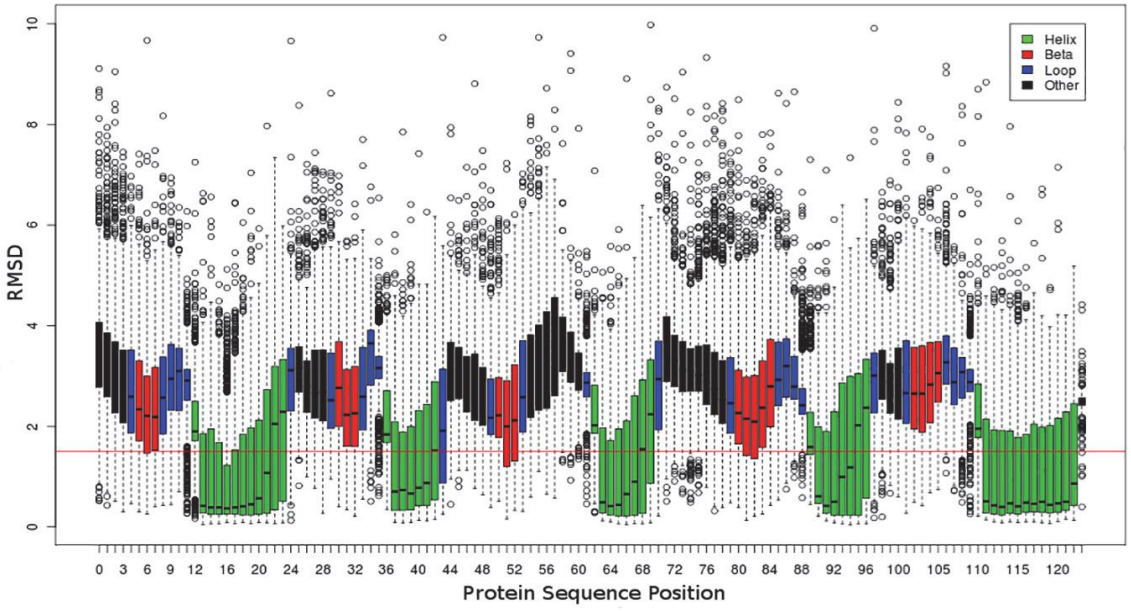

Figure 6.1: Boxplot of the top 200 fragments for the protein 1E6K. Four Different types of fragments are shown: majority α-helical (green), majority β-strand (red), majority loop (blue) and other (black). Taken from (de Oliveira et al., 2015).

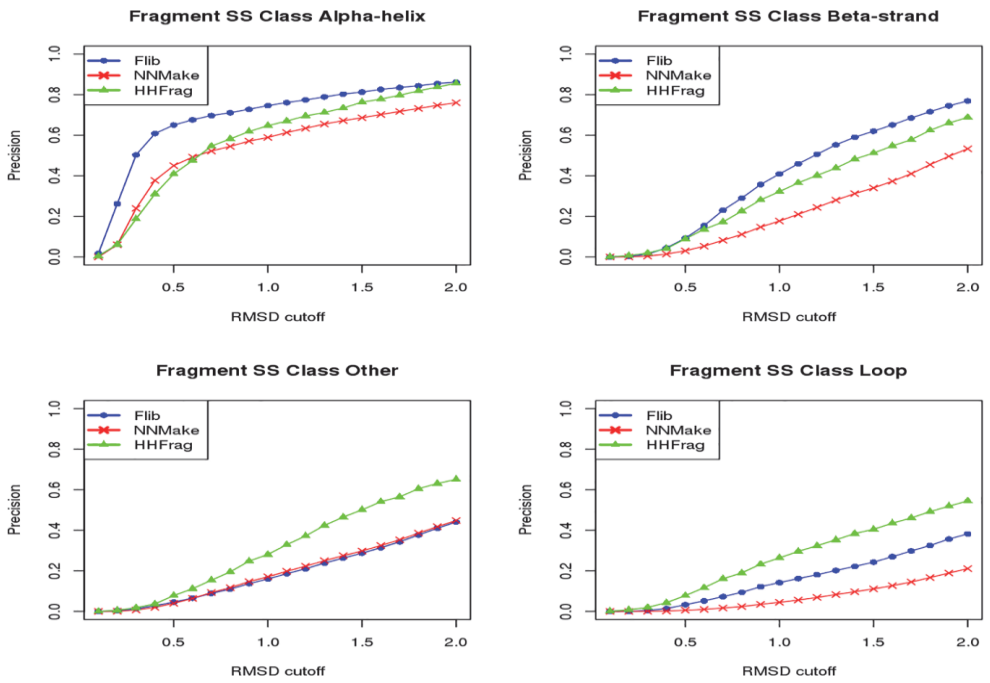

Figure 6.2: Comparison amongst three fragment libraries generators based on 41 structurally diverse targets. Precision is calculated as the proportion of good fragments in the library. Whereas on average, Flib and HHFrag generate 26 and 10 fragments of 7.4 and 9.1 length, NNMake's data is based on 200 9-mers. Taken from (de Oliveira et al., 2015).

6.4 Proposed Methodology

To our knowledge, in all studies, experiments and improvements conducted by researchers working on Rosetta, the number of fragments in each position, for both 3-mers and 9-mers, has been constant. We propose a different approach, where the fragment selection process is unchanged; however, the number of fragments per position varies based on its secondary structure prediction. Owing to the strong sequence-structure correlation for alpha helices and the loose one for loops (while beta strands lie in between) and the results of the works mentioned earlier, a novel approach is proposed so that the number of available fragments might vary at different positions.

In order to illustrate the standard Rosetta and the proposed strategies regarding the number of fragments used for each amino acid position, a protein, whose PDBID is 1CC8, has been chosen from the dataset (introduced later in Section 6.4). The choice of 1CC8 was random amongst the 5 proteins belonging to the Alpha-Beta structural class, which presents a relatively even distribution of the three secondary structures.

The standard Rosetta's fragments files' contents in terms of fragment number are depicted in Figures 6.4 and 6.5 for fragments of length 9 and 3 respectively. In the 9-mer insertion phase, 25 fragments are selected starting at the first amino acid till the

position: “size-9”. Similarly, in the 3-mer insertion phase 200 fragments are selected starting at the first amino acid till the position: “size-3”.

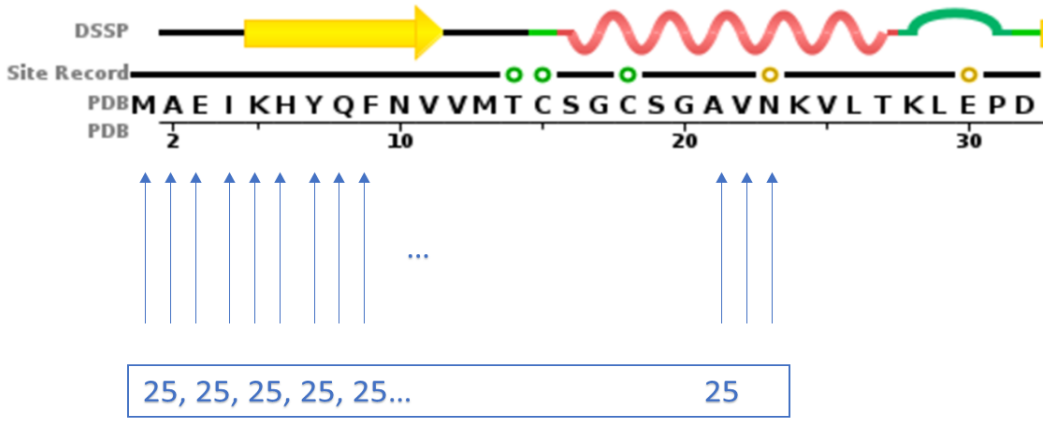

Figure 6.4: A pictorial representation of the contents of a Rosetta standard’s 9-mer file (the upper part of the figure was taken from the PDB, sequence tab of the Atx1 Metallochaperone Protein – PDBID: 1CC8). The blue arrows point to the positions where there is a set of 25 candidate fragments of length 9. In the example above, assuming that protein is of length 32, the 9-mer fragment library ends at position 23. The circles on site record line point to some residues that play important role in interacting with other macromolecules; in this study, such information were not taken into consideration in any way.

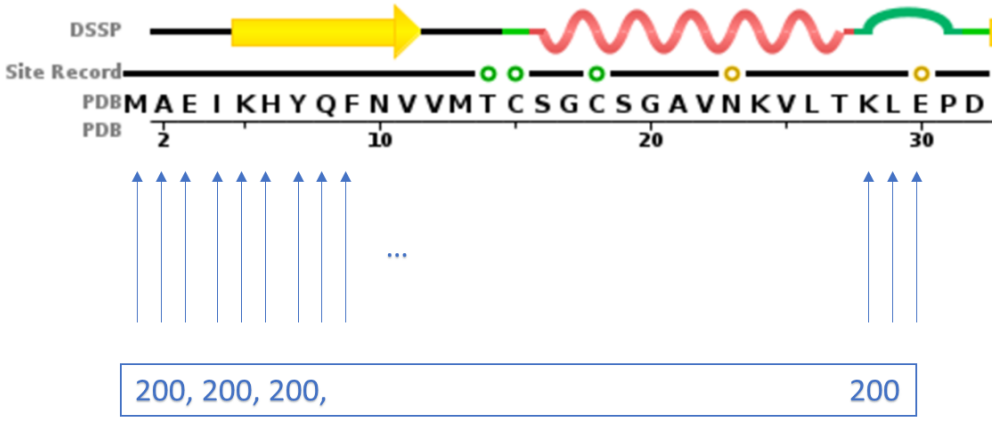

Figure 6.5: A pictorial representation of the contents of a Rosetta standard’s 3-mer file; 200 candidate fragments per position. Here the 3-mer fragment library ends at position 30.

The new methodology will be applied to both fragment files to dramatically reduce the number of candidates whenever the fragment is predicted to be either a helix or a beta strand. A pictorial description is shown in Figures 6.6 and 6.7. Our novel idea works as follows: in the 9-mer file, whenever a fragment that starts at position i is predicted to be a (part of a) helix till $i+8$, only one fragment is available to be inserted, whereas the standard number – i.e. 25 – is kept otherwise. A similar process is applied using the 3-mers file: whenever a fragment that starts at position i is predicted to be a (part of a) helix till $i+2$, only five fragments are to be inserted. Likewise, a fragment

window of length 3 that starts at position *i* is predicted to be a (part of a) beta strand, only 25 fragments will be considered. In all remaining positions, the standard number of available fragments – i.e. 200 – is taken into account. The rationale behind the choice of one, five and twenty-five fragments will be described in detail in the next paragraphs.

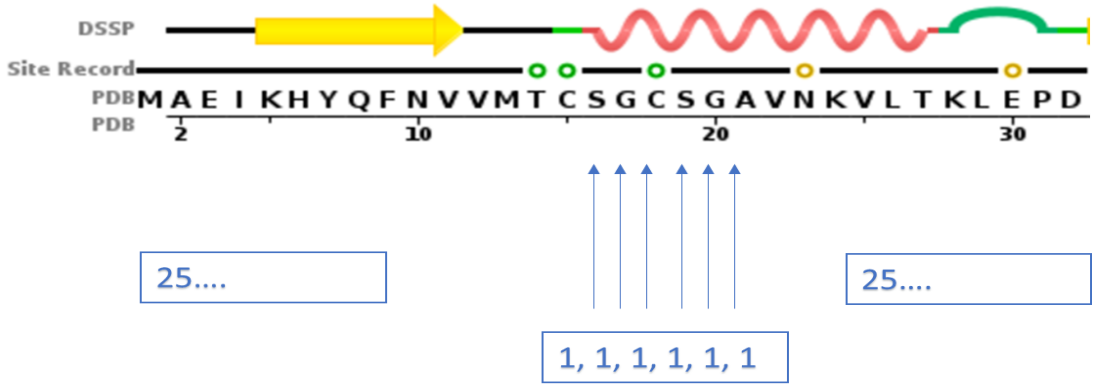

Figure 6.6: New approach to build the 9-mers file based on the secondary structure annotations of the protein in question. Since at positions 16, 17, 18, 19, 20 and 21 a helix of size 9 is predicted, only one fragment is used. The standard number of fragments – that is 25 – is kept at the remaining indexes.

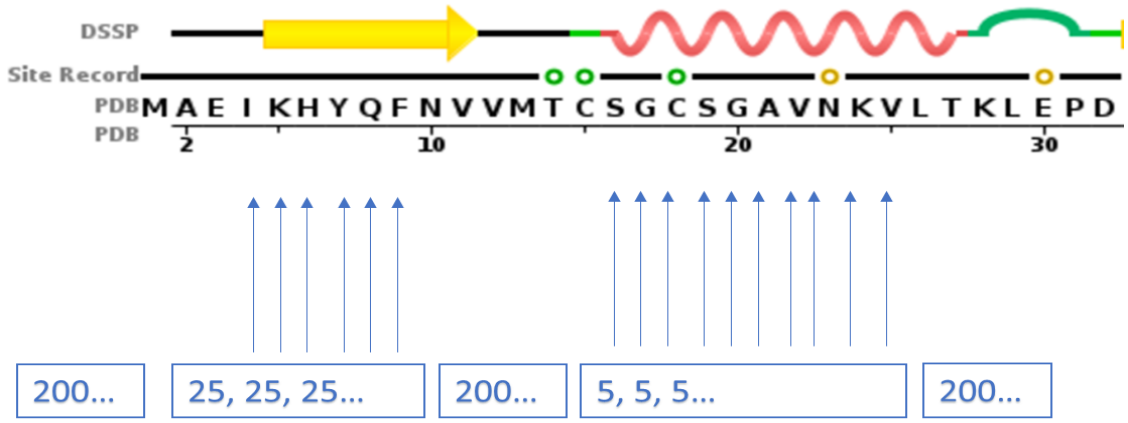

Figure 6.7: New approach to build the 3-mers file based on the secondary structure annotations of the protein in question. Since at positions 4 till 9 a strand of size 3 is predicted, only 25 fragments are used, and at positions 16 till 25 a helix of sixe 3 is predicted, only 5 fragments are used. The standard number of fragments – that is 200 – is kept at the remaining indexes.

In 1CC8, there are 9 positions where a 9-mer is supposed to be a pure alpha helical. For each of those 9 positions, the RMSD of the average, lowest (best), highest (worst) of the 25 fragments as well as the first one are plotted in Figure 6.8. It clearly shows that the first fragment in the set of 25 is always very close to be the best, better

than the average fragment and much better from the worst one. As a consequence, replacing the 25 fragments by the first one has the potential to lead to better structure prediction.

Similarly, the quality of 3-mers which are supposed to be pure alpha helical and a pure beta strand are studied in 1CC8, where they correspond to 21 and 17 positions respectively. In order to select an adequate minimum number of fragments at those positions, we have conducted a thorough study on how the RMSD of the best fragment out of 5, 10, 15, 20, 25, 30, 35 and 40 is being improved on average (Table 6.1).

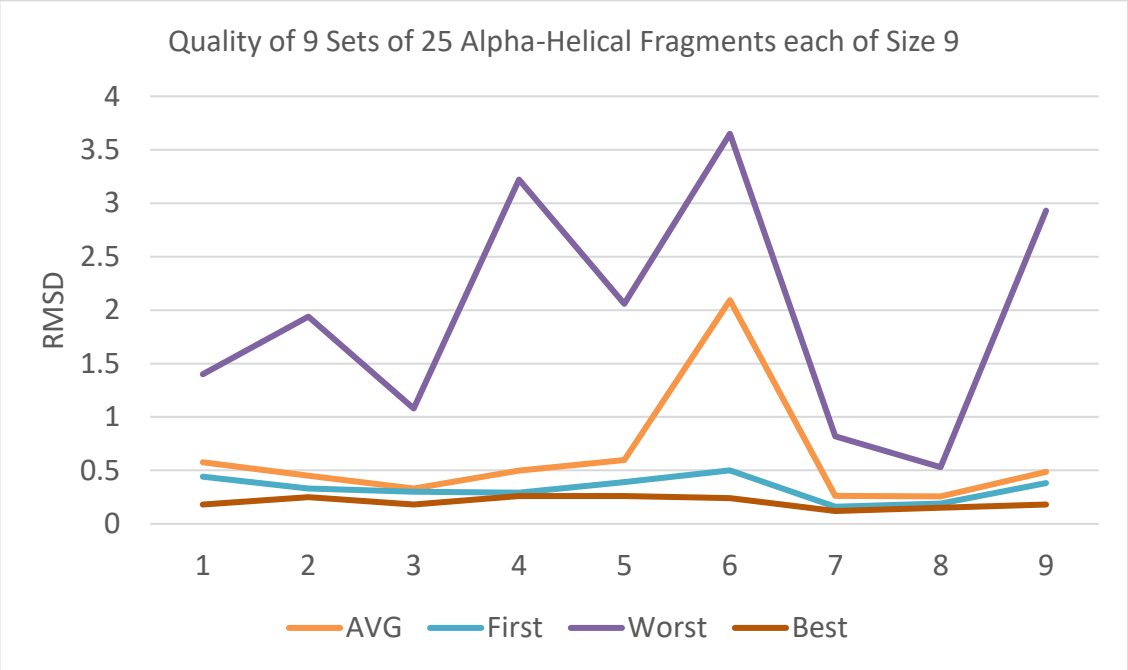

Figure 6.8: RMSD of 225 9-mers distributed amongst 9 positions as averages, lowest, highest and first ones.

Table 6.1: Study of improvements of 3-mers by taking into consideration the best fragment (lowest RMSD) in a set versus the best one in an extended group by 5 fragments.

		Best of the First n 3-mers vs Best of the First m 3-mers (Both out of 200)							
		n = 1 m= 5	n = 5 m= 10	n = 10 m = 15	n = 15 m = 20	n = 20 m = 25	n = 25 m = 30	n = 30 m = 35	n = 35 m = 40
Average of Improvements in Å	21 sets of 200 Alpha 3-mers	0.0390	0.0076	0.0038	0.0005	0.0024	0.0000	0.0000	0.0010
	17 sets of 200 Beta 3-mers	0.2347	0.0406	0.0029	0.0076	0.0059	0.0000	0.0012	0.0006

Results clearly show that usage of only the first fragment would not be a sensible choice since significant improvements are visible when increasing the number of fragments. This is particularly the case for the beta ones. Beyond the first 25 fragments, improvements become negligible (<0.2%) for both types of fragments. Moreover, in the case of alpha helical 3-mers, an improvement of less than 1.5% can be reached by including more than 5 fragments. As a consequence, this suggests that the number of fragments should be set at 5 for that category. On the other hand, the quality of beta 3-mers still improves significantly (+4%) with 10 fragments and keeps increasing (an additional 1.5%) until the top 25. Thus, the number of fragments for amino acids predicted to belong to beta sheets could be set at 25.

Further investigation has been conducted: out of the 200 3-mers, the lowest (best), highest (worst), average and the lowest RMSD of the first 5/25 fragments are plotted in Figure 6.9 and Figure 6.10 where a 3-mer is predicted to be an alpha helix/ beta strand respectively. The best fragment (out of 5 or 25) is very close to the whole set's best one (out of 200) and much better than the average. Accordingly, we have adopted the numbers of fragments of 5 and 25 for 3-mers that are predicted to be alpha helices and beta strands respectively.

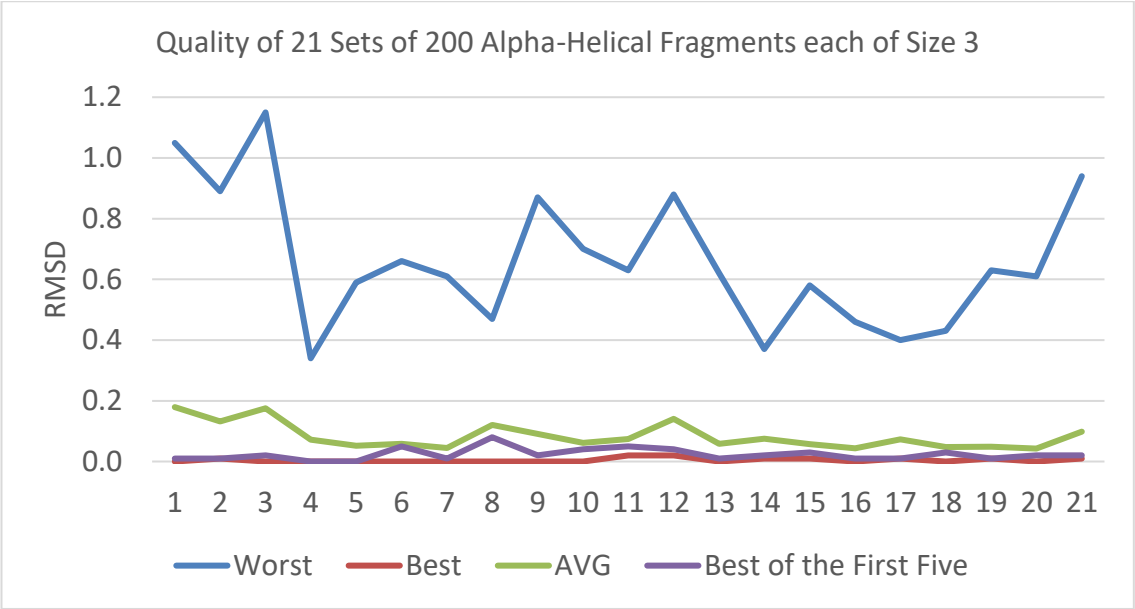

Figure 6.9: Quality of 4200 3-mers at 21 positions as RMSD of their averages, lowest, highest and first ones.

Conceptually, the above findings along with the proposed amendments would 'freeze' some regions whereas fragment insertion operation would be highly 'activated' in others. The low number of candidate fragments at more than the third of the target would force Rosetta to narrow the search space, i.e. dedicating more time on exploitation rather than exploration. Although experiments suggest that the best fragment have been "lost" in most cases, the probability of it to be chosen amongst a set of either 200 or even 25, as in the standard scheme, is quite low. Moreover, the standard scheme is at a higher risk of picking a worse fragment than the one selected in our new methodology. Furthermore, one could conclude that by preventing Rosetta from picking the most inaccurate fragments, the new search space would contain fewer "outliers" conformations, i.e. conformations that, although belonging to local minima, are far away from the native-like structure.

6.5 Experiments and Dataset

In order to validate the proposed methodology, Secondary Structure-based Rosetta (SS-Rosetta), where the number of fragments is customised for each amino acid, for each target, its performance is compared to the standard Rosetta. Whilst the whole fragment selection process has not been altered, the number of fragments at specific positions is changed. In other words, for a certain amino acid where for instance one 9-mer should be used instead of the set of 25, we have kept the fragment that has

the highest score according to fragment picker; a similar policy has been applied to keep the top 5 and 25 fragments in the 3-mer file.

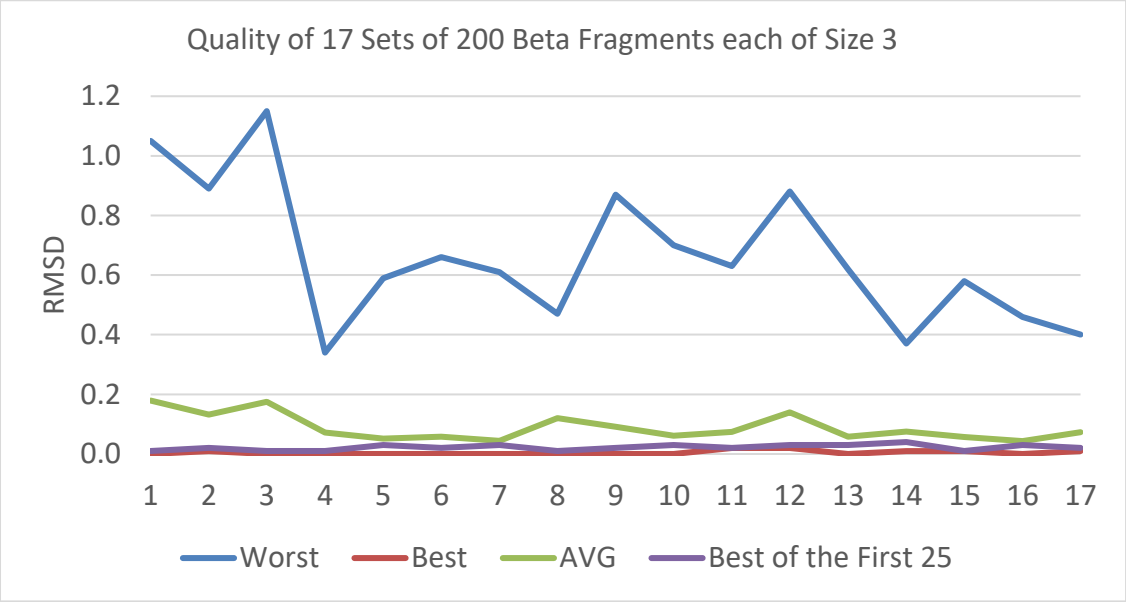

Figure 6.10: Quality of 3400 3-mers at 17 positions as RMSD of their averages, lowest, highest and first ones.

In principle, secondary structure prediction's information should be used since we aim at improving a protein's conformation prediction of unknown native structure. However, in this experiment, we have taken advantage of the secondary structure annotations of each target so that our proposal can be assessed without having to consider the choice of a predictor and its accuracy. Since secondary structure predictors are already very accurate - a range of 82% to 84% accuracy has been recorded (Q. Jiang, Jin, Lee, & Yao, 2017; Y. Yang et al., 2018) – and their accuracy is still improving, this simplification should not affect the nature of the lessons that will be learned from the experiment.

For this experiment, two sets of decoys are used: 20,000 and 2,000. The rationale behind using a large and a relatively small number of decoys is to investigate the exploration/exploitation compromise. Observing the behaviour of our proposed methodology using both sets is quite important to shed a light on the strength of limiting the number of fragments which in turn has supported Rosetta to "overcome" the problem of energy score inaccuracy.

Five similar-purpose publications - studying, enhancing and optimising fragments in Rosetta - conducted by EdaFold's and Rosetta's team (Blum et al., 2010; Kim, Blum, Bradley, & Baker, 2009; Simoncini et al., 2012, 2017; Simoncini & Zhang,

2013) used as a dataset of 20 proteins which shows diversity in terms of structure classes, percentages and length of helices and sheets. Based on that set, we created a dataset comprised of 24 proteins whose length varies from 56 to 149. Out of the 20-PDB set, one of the targets was removed since it was not relevant to our methodology; it belongs to small proteins structural class and consequently coils represent the majority of its structure. As a consequence, our methodology would not affect its processing compared to standard Rosetta. Moreover, a limitation of the 20-protein set was that it did not include too easy, too difficult or long targets. Whilst we agree that too easy and long targets (longer than 150) are likely not to reveal any improvements, we believe that the benchmark proteins should contain some hard targets, so a more general conclusion could be inferred from the outcome. Furthermore, since the number of all-alpha targets is only 3, for the sake of diversity in terms of structural classes, more targets of that class should be added. Accordingly, 5 CASP FM targets were added; those five proteins were carefully selected to belong to the larger part of the length range: from 56 to 149. The challenging nature of those 5 targets is demonstrated by the fact that the GDT of the *best decoy*s (out of 20,000) in standard predictions is significantly lower than the GDT of other targets having the same length. In addition, prediction of their structure during the CASP contest proved particularly difficult. The whole dataset is shown Table 6.2. The last 2 columns report the size of the new fragment file (corresponding to the number of fragments used for each target) as a percentage of the initial fragment file.

Table 6.2: The list of the 24 proteins that are included in the study.

#	PDB ID	Length	% of Helices	% of Strands	% of coils	New 9-mers File	New 3-mers File
1	2CI2_I	65	16	21	63	85.8%	51.7%
2	1CTF	68	51	24	25	85.2%	49.7%
3	1DI2	69	46	33	21	89.5%	47.8%
4	1SCJ_B	71	23	39	38	86.5%	49.5%
5	1HZ5	72	30	38	32	79.2%	37.0%
6	1CC8	72	28	35	37	88.2%	55.3%
7	3NZL	73	59	0	41	79.5%	41.4%
8	1DTJ	74	38	26	36	88.4%	51.8%
9	1IG5	75	61	5	34	83.5%	47.6%
10	1OGW	76	25	34	41	81.4%	47.9%
11	1DCJ	81	28	24	48	77.8%	40.5%
12	1TIG	88	32	32	36	50.0%	29.7%
13	1A19	89	43	17	40	92.9%	61.5%
14	1BM8	99	37	27	36	93.9%	58.4%
15	4UBP	100	54	17	29	77.2%	46.5%
16	1IIB	103	55	19	26	94.9%	73.4%
17	1M6T	106	77	0	23	84.8%	50.1%
18	1ACF	125	34	32	34	77.8%	45.1%
19	3CHY	128	45	17	38	73.9%	45.3%
20	2KDL*	56	62	0	38	87.7%	50.2%
21	2LR8*	70	57	0	43	74.0%	44.0%
22	4HLB*	95	28	24	48	78.2%	49.6%
23	2K4V*	125	28	32	40	72.1%	48.4%
24	2KY4*	149	59	1	40	87.9%	54.1%
Averages		**88.7**	**42.3**	**20.7**	**37.0**	**82.1%**	**49.0%**

*CASP targets.

6.6 Results and Discussion

Blind assessment mimics CASP's assessment as groups can submit up to 5 models and should designate one of them as the *first model*. The two main ranking scores adopted are the GDT of the *first model* as well as the *Best model* – the best of the 5 submitted structures. Furthermore, we introduce here a third criterion too - the average of the 5 models - to shed light on the quality of the models that lie in the lowest 5 positions in the search space. Figure – 6.8 shows the *first model*, *Best model* and average of the 5 models as a comparison between the standard and SS-Rosetta

predictions for both 20,000 and 2,000 decoys respectively. In Figure 6.8, 15 out of 24 targets have shown to have a better *first model*. Overall, an improvement of 6.3% has been recorded using the amended fragments files. Regarding the *Best model*, it is better in 12 targets than standard's whereas 3 targets have the same *Best model*; on average an improvement of 5% is on the favour of the SS-Rosetta pipeline. Finally, our novel idea shows also an overall improvement of 6.1% as 18 targets reach a higher GDT score with regards to the average of the best five models. However, moving to the 2,000-decoy experiment – Figure 6.9, *first model*'s improvement has interestingly reached 24%; approximately 4 times the improvement shown for 20,000-decoy's. Both scores are presented in Table 6.3.

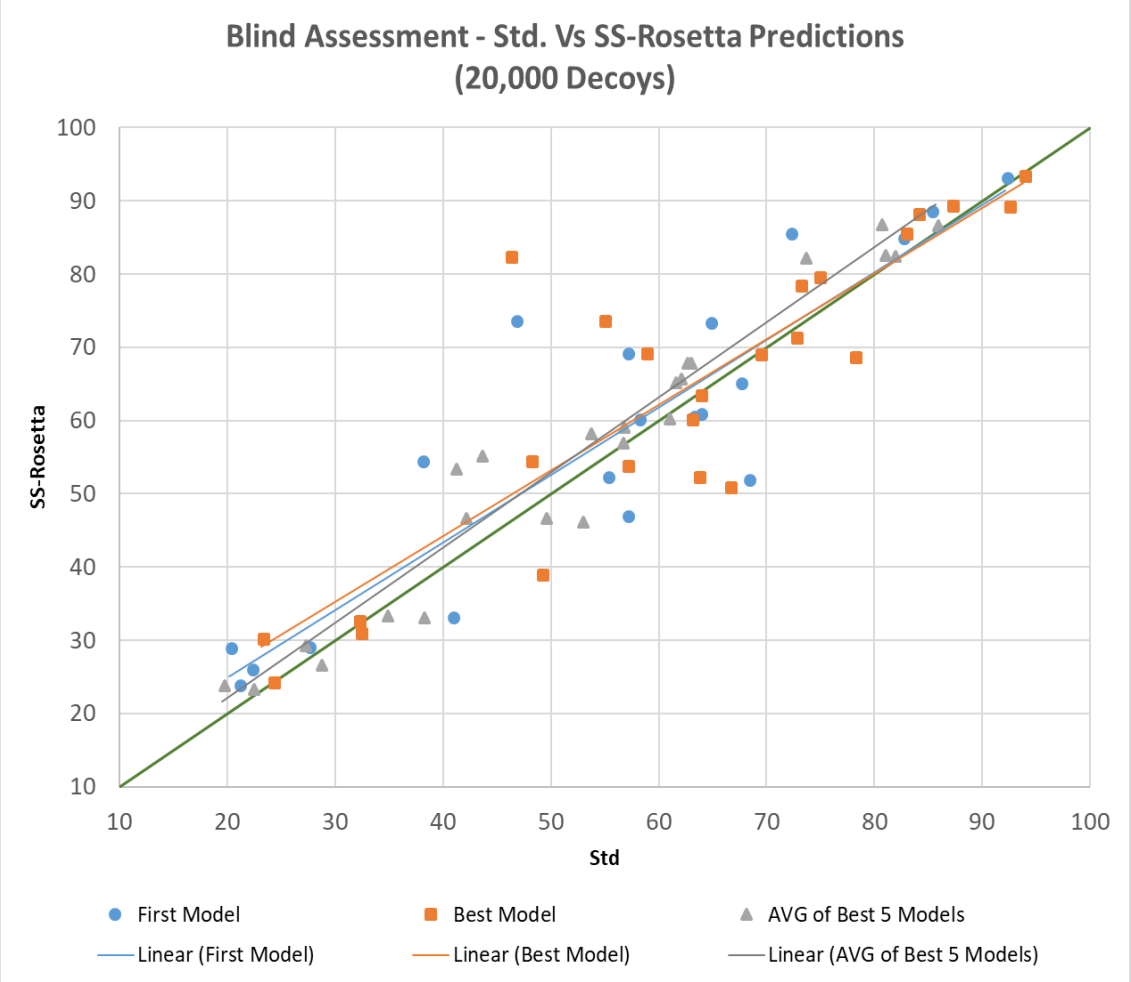

Figure 6.8: *First model*'s, *Best model*'s and average of *best 5 models*' GDT of standard prediction (denoted as "Std") versus the new paradigm (denoted as "SS-Rosetta") for 20,000 decoys each where correlation coefficients are 0.89, 0.88 and 0.97 respectively. Linear regression lines are shown for the three data sets.

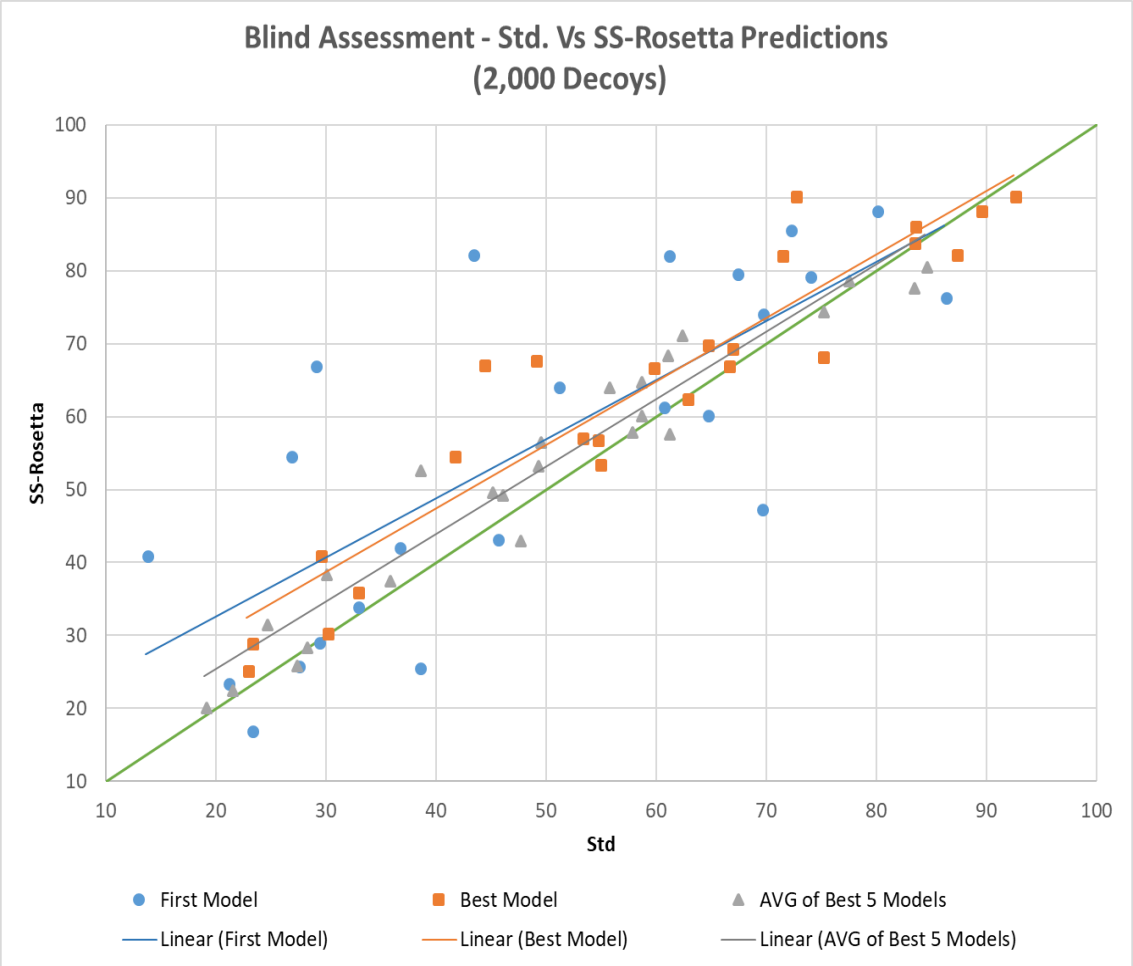

Figure 6.9: *First model*'s, *Best model*'s and average of *best 5 models*' GDT of standard prediction (denoted as "Std") versus the new paradigm (denoted as "SS-Rosetta") for 2,000 decoys each where correlation coefficients are 0.76, 0.94 and 0.97 respectively. Linear regression lines are shown for the three data sets.

Table 6.3: Blind assessment results against standard predictions: improvements in terms of number of better structures (out of 24) and GDT.

Std vs SS	*First model*	*Best model*	Average of the *Best 5 models*
20,000 Decoys	15/24 (63%) +6.3%	12/24 (50%) +5.0%	18/24 (75%) +3.1%
2,000 Decoys	16/24 (67%) +24.2%	18/24 (75%) +11.5%	18/24 (75%) +8.3%

The tangible difference between the 20,000-decoy and 2,000-decoy improvements is due to the fact for a small number of decoys, exploration gains more value than exploitation; the latter was made up by the lower number of fragments SS-

based predictions have used. In the same context, we have also conducted one additional study between the standard predictions for both 2,000 and 20,000, however it terms of *best decoy*. It shows that for all targets the 20,000-decoy experiments reached a more accurate decoy, but for no more than 6% on average (whilst for *Best model*, improvement was more than 26% as shown in Table 6.4). Whenever reliable fragments are found, increasing the number of decoys will have more effects on exploitation rather than exploration.

In addition, our experiments confirm Rosetta's rule of thumb that the higher number of decoys the better outcome. However, SS-based predictions have proved again that when it comes to increasing the number of decoys the new paradigm has delivered even more accurate results. Furthermore, a crucial outcome is presented in Table 6.5: generating 20,000 decoys through standard predictions produces similar results to generating 2,000 decoys through the new methodology. Such a finding is important in terms of usage of computational resources as 2,000 decoys can be easily carried out on a typical PC whereas a Rosetta's 20,000-decoy prediction usually requires supercomputer facilities; a hinder that prevents many users to use Rosetta to generate "acceptable" quality models.

Table 6.4: Performance comparison between generating 20,000 decoys versus 2,000 in both standard and SS-based predictions.

	First model	*Best model*	Average of the *Best 5 models*
Std. Predictions 2,000 vs 20,000 Decoys	17/24 (71%) +26.3%	19/24 (79%) +7.6%	19/24 (79%) +8.2%
SS-based Predictions 2,000 vs 20,000 Decoys	18/24 (75%) +30.2%	19/24 (79%) +11.7%	19/24 (79%) +14.7%

Table 6.5: Blind assessment results of standard predictions' 20,000 decoys against SS-based predictions' 2,000 decoys.

	First model	*Best model*	Average of the *Best 5 models*
20,000 Decoys (Std. Predictions) vs 2,000 Decoys (SS-based. Predictions)	12/24 (50%) +0.4%	13/24 (54%) +4.6%	10/24 (42%) +0.3%

Narrowing the search space by reducing the number of fragments at certain positions has proven to minimise the number of local minima explored and consequently made the process of choosing the lowest energy model(s) a less "arbitrary" procedure. As mentioned by Brunette and Brock (T. Brunette & Brock, 2008), not all funnels are relevant even if they are as deep as the others. Inaccurate fragments, if chosen, will most probably lead to the generation of such funnels. Since the fragments we have kept are "good enough", missing some regions and consequently some funnels and local minima standard predictions explored does not affect the overall quality of the decoys. Some evidence is shown in Figure 6.10 that depicts a comparison between the *best decoy* in both standard and SS-based Rosetta predictions; generally, the best standard decoys is not as good as the *best decoy* produced by our new approach (SS). Therefore, if an optimal quality assessment tool was found, our approach would also provide better results.

We have proved that the fragments of Rosetta's standard predictions contain some redundancy and somewhat similar contents at some positions which make the trajectory paths explore a large space on the energy landscape, however minimising the exploitation procedure, therefore, less ability to reach better positions in a certain funnel. In addition, although the standard approach explores a larger search space, it is not able – regardless of the energy score - to locate better conformations than those produced by the proposed approach.

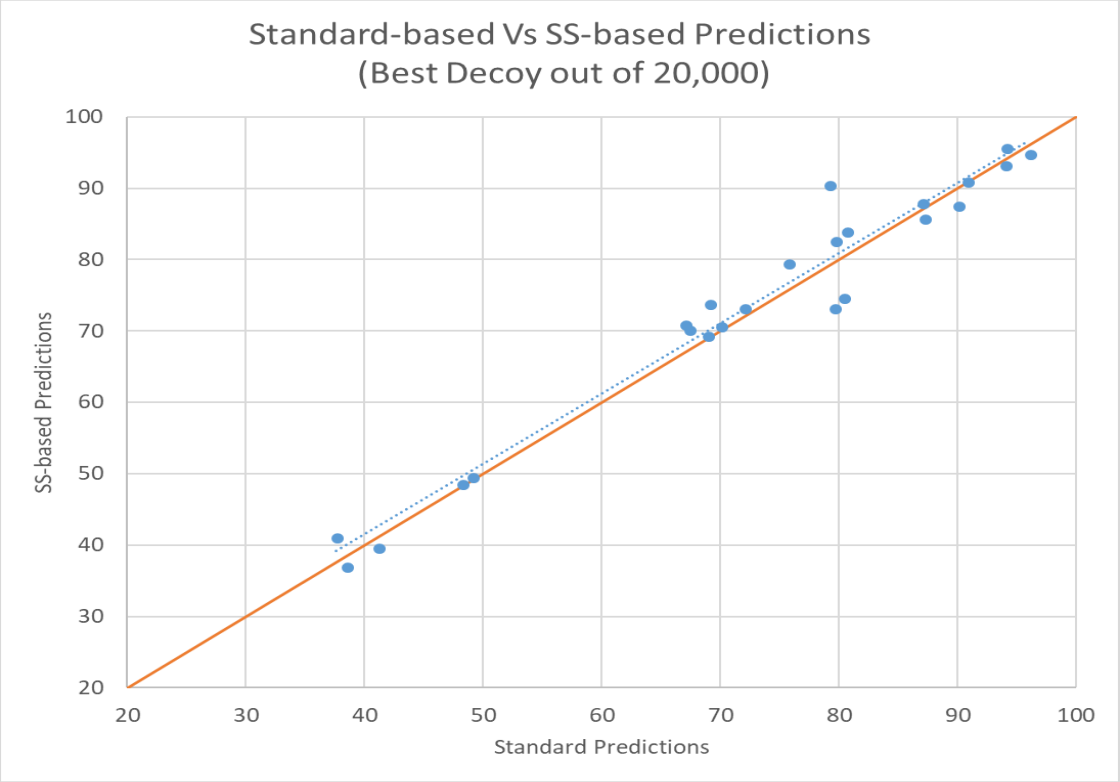

Figure 6.10: Comparison between the standard predictions' and SS-based predictions' *best decoy*. SS-based was able to reach a higher GDT score for 16 out of 24 targets with an overall improvement of 1.5%; correlation coefficient records a value of 0.98. The dotted line represents the linear regression.

6.7 Conclusion

This chapter has shown again that the usage of the sequence-structure correlation for different secondary structures could be of benefit in fragment assembly protein structure prediction. The fragment insertion process should not be treated evenly on all regions of the conformation being built; since some parts are known to have a loose subsequence-substructure liaison, one would expect a variety of candidate fragments. On the other hand, a very limited number of substructures could be sufficient for "easy" regions. Whereas, previous works have focused on the fragment picking phase, this chapter's findings have introduced a new concept during fragment-assembly: the allocation of the number of fragment insertion operations according to the dominant secondary structure of the parts where fragments are to be inserted.

7 Conclusion

This chapter summarises the findings of this thesis, namely the three main contributions we have made, their results and suggests future lines of investigation. The next section presents an overview of the contributions. Sections 7.2 and 7.3 discuss the results and future work respectively. Closing remarks are presented at the end.

7.1 Summary of Contributions

In this thesis, we have studied thoroughly fragment-based protein structure prediction, using Rosetta as an example, by exploring additional critical roles secondary structure predictions can play.

State-of-the-art tools that rely on fragment assembly methodology use a unique template structures set that represents the PDB by eliminating some degree of homology and non-accurate conformations. Such a large number of very diverse structures have made the picking of fragments a challenging job despite the usage of many criteria such as sequence and secondary structure similarity and knowledge-based properties such as Ramachandran map propensities. Neither selection of a large set of candidates nor usage of stringent standards for choosing fragments have so far proved satisfactory: in many cases the ultimate goal of not missing any good substructure and picking the best ones amongst a large number is not fulfilled. We have shown in chapter 4 that predicting the structural class of a target and customising the template structures set accordingly is an approach leading us closer to the compromise researchers have been trying to reach. CASP11's results represented some evidence of the effectiveness of our novel approach as we were able to submit *first models* of higher accuracy than Rosetta's formal group in 6 out of 14 domains.

We have introduced a "preliminary step" that makes the whole fragment-based prediction process converges towards more accurate results: by predicting the structural class a target is likely to belong to – a task which has now become accurate -, we have proposed to alter the route followed during the structure prediction process. This novel approach not only was able to create a customised and relevant template structure set (up to 80% reduction of the original size of Rosetta's) for each of the main structural classes, but also has enabled to determine the appropriate "amount of required corrections" in the final phase of the structure prediction process. We have shown that, even when using Rosetta's default template structures set, setting a relative amount of fine tunings to each structural class's targets has led to tangible improvements.

Further usage of secondary structure predictions has been introduced for the first time when assigning the number of candidate fragments at each position in the sequence of interest. Current fragment assembly methods deal with all positions equally whenever it comes to fragment insertions. For instance, Rosetta employs 25 and 200 9-mers and 3-mers respectively to be chosen randomly for each amino acid; the choice of an amino acid amongst all possible locations is in its turn chosen randomly. This somehow contradict the rule of sequence-structure correlation that has been widely accepted for more than two decades that states that among the secondary structures, alpha helices, beta strands and coils are of increasing diversity. Taking into advantage of this, we have allocated a number of candidate fragments that is customised according to the secondary structure that is likely to be adopted at each position in the target. This has led to significant acceleration of convergence towards native-like structures, since the generation of a 2,000-decoy set using this strategy produced models of similar quality as those generated by 20,000 decoys using the standard prediction process.

7.2 Discussion

Despite the improvements we have presented in this research work, structures predicted by Rosetta, as a cutting-edge fragment-based proteins structure prediction, are still considered as putative protein structures.

First, although in principle the “vall” – Rosetta’s default template structures set – comprises all the 9-mers and 3-mers that are needed to build any conformation, some chosen fragments are not of acceptable quality. Moreover, the ranking of the fragments based on their scores is not relevant in many regions, and consequently, picking fragments function’s terms and their corresponding weights need to be reassessed. Even when we customised the “vall” based on the main structural classes, accelerating convergence, there is still a need to generate a huge amount of decoys. The new search space explored (chapter 4) is undoubtedly more relevant but most probably doesn’t comprise the ideal funnel for some targets.

Second, there is no clear clue where the first phase of conformational sampling, i.e. 9-mers insertion, should “stop” in order to let the correction phase, i.e. 3-mers insertion, to take place. The exploration-exploitation compromise’s threshold is still undetermined. The new methodology we presented (chapter 5) has indeed paved the way for a novel approach in this regard, however, further and thorough studies should be conducted.

Third, combining the customised fragments libraries and reduced amount of corrections (Chapters 4 and 5) has led to very limited success over standard predictions (CASP12's results), although both approaches work well separately. It seems that conformational sampling is such a complicated task to the extent that the act of "narrowing" this process needs an extensive empirical analysis. It seems that some unpromising regions need to be sampled, probably as a temporary phase in the route towards the native-like funnels.

Fourth, studies to determine the ideal number and length of candidate fragments have failed to identify a specific threshold, especially that this issue is highly related to the effectiveness of the functions that score fragments to be selected. Our contribution in chapter 6 has shown that complexity of a sequence region, e.g. "easy", "moderate" and "hard" region, should be taken into consideration.

Finally, the gap between the real search space – assuming using an optimal energy function – and the one which is explored using various energy functions has been investigated in its own research area, i.e. decoys quality assessment which remains very challenging.

7.3 Future Work

The preliminary step we introduced, i.e. to take advantage of the prediction of a target's structural class, can be used not only to customise the template structure set and the amount of correction but also to introduce restrictions regarding the fold a conformation can adopt for each of the structural classes. Although predicting the second level in the hierarchy of both SCOP and CATH, i.e. Fold and Architecture, is still lacking in accuracy, but the number of folds a mainly-alpha target, for instance, can adopt is quite limited and therefore such a limitation could be used as a structural constraint. This would prevent irrelevant decoys to be generated and consequently make the energy score and conformation's accuracy correlation tighter.

In the same context, a new customisation can be taken into account with respect to the energy functions based on the result of a target's structural class predictions. As stated in chapters 2 and 3, there have been many types of energy functions, even for Rosetta, three updates have taken place in the past 4 years; the ultimate goal may be to design a balancing energy function that is "fair" in the sense of taking into account all forces that take place amongst atoms and amino acids, however to be evaluated within a reasonable amount of processing time. Each structural class, especially mainly-alpha

and mainly-beta, has some dominant forces that should be calculated more accurately. For example, in mainly-beta proteins, hydrogen bonds are of higher importance to determine the regions where beta sheets are likely to be formed. Accordingly, a customised energy function associated with the main structural classes may have potential.

Although we have distributed the number of candidate fragments amongst various regions based on the secondary structure prediction, easy regions, i.e. pure alpha helices, still "consume" time by allowing for minor changes (neighbouring fragments may affect them) and reusing (the sole 9-mer at a certain position is likely to be chosen again and again). Further experiments can be conducted by totally "freezing" regions where a pure alpha helix 9-mer has been inserted. This would allow Rosetta to dedicate more time and therefore insertions for the remaining challenging regions.

Based on the findings revealed by the Barker's group with regard to the size of fragments where the correlation between sequences and sub-structures, including alpha helices, helix caps, beta strands, loops and turns, reaches the highest entropy; see Sections 3.2., a more efficient and promising idea would be to separate the role of each prediction phase. More specifically, the 9-mer phase would be only responsible, for example, for the helices and helix caps, whilst the 3-mer phase would be dedicated to refine other regions such as beta strands, loops and turns. In addition, the distribution of the "weight" of each phase – currently 28,000 and 6,000 insertion attempts respectively – could be adjusted according to the percentage of each set of secondary structure. This would yield not only decoys with higher accuracy but also reducing computational cost by preventing unhelpful and even sometimes harmful additional corrections.

7.4 Closing Remarks

This research has investigated a state-of-the-art fragment-assembly tool, Rosetta and contributed to the improvement process. We have shown the extent to which the usage of secondary structure predictions can improve fragment-based protein structure prediction by proposing novel ideas that either directly or indirectly use such predictions. We believe that this study will pave the way for many more research discoveries; it is a remarkable step in the one-thousand-mile journey to "decipher the holy grail of microbiology".

References

Abbasi, E., Ghatee, M., & Shiri, M. E. (2013). FRAN and RBF-PSO as two components of a hyper framework to recognize protein folds. *Computers in Biology and Medicine*, *43*, 1182–91.

Abbass, J., & Nebel, J.-C. (2015). Customised fragments libraries for protein structure prediction based on structural class annotations. *BMC Bioinformatics*, *16*, 136.

Abbass, J., & Nebel, J.-C. (2017). Reduced Fragment Diversity for Alpha and Alpha-Beta Protein Structure Prediction using Rosetta. *Protein & Peptide Letters*, *24*, 215–222.

Abbass, J., Nebel, J.-C., & Mansour, N. (2013). Ab Initio Protein Structure Prediction: Methods and challenges. In M. Elloumi & A. Y. Zomaya (Eds.), *Biological Knowledge Discovery Handbook* (pp. 703–724). Hoboken, New Jersey: John Wiley & Sons, Inc.

Agarwala, R., Batzoglou, S., Dancík, V., Decatur, S. E., Hannenhalli, S., Farach, M., … Skiena, S. (1997). Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *4*, 275–296.

Ahmad, S., Singh, Y. H., Paudel, Y., Mori, T., Sugita, Y., & Mizuguchi, K. (2010). Integrated prediction of one-dimensional structural features and their relationships with conformational flexibility in helical membrane proteins. *BMC Bioinformatics*, *11*, 533.

Alder, B. J., & Wainwright, T. E. (1957, November 13). Phase transition for a hard sphere system. *The Journal of Chemical Physics*. American Institute of Physics.

Alder, B. J., & Wainwright, T. E. (1959). Studies in molecular dynamics. I. General method. *The Journal of Chemical Physics*, *31*, 459–466.

Alexander, P., He, Y., Chen, Y., Orban, J., & Bryan, P. N. (2009). A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 21149–54.

Alford, R. F., Koehler Leman, J., Weitzner, B. D., Duran, A. M., Tilley, D. C., Elazar, A., & Gray, J. J. (2015). An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLOS Computational Biology*, *11*, e1004398.

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., … Gray, J. J. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, *13*, 3031–3048.

Altschul, S. F. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, *215*, 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997, September 1). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*.

Amani, S., & Naeem, A. (2013, November 1). Understanding protein folding from globular to amyloid state: Aggregation: Darker side of protein. *Process Biochemistry*. Elsevier.

Amir, E. A. D., Kalisman, N., & Keasar, C. (2008). Differentiate, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. *Proteins: Structure, Function and Genetics*, *72*, 62–73.

Anand, A., Pugalenthi, G., & Suganthan, P. N. (2008). Predicting protein structural class by SVM with class-wise optimized features and decision probabilities. *Journal of Theoretical Biology*, *253*, 375–380.

Andersen, C. A. F., Palmer, A. G., Brunak, S., & Rost, B. (2002). Continuum secondary structure captures protein flexibility. *Structure*, *10*, 175–184.

Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S. E., Hubbard, T. J. P., Chothia, C., & Murzin, A. G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research*, *36*, D419-25.

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., & Murzin, A. G. (2014). SCOP2 prototype: A new approach to protein structure mining. *Nucleic Acids Research*, *42*.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, *181*, 223–230.

Anfinsen, C. B., Haber, E., Sela, M., & White, F. H. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, *47*, 1309–1314.

Arab, S., Sadeghi, M., Eslahchi, C., Pezeshk, H., & Sheari, A. (2010). A pairwise residue contact area-based mean force potential for discrimination of native protein structure. *BMC Bioinformatics*, *11*, 16.

Arnautova, Y. A., Jagielska, A., & Scheraga, H. A. (2006). A new force field (ECEPP-05) for peptides, proteins, and organic molecules. *Journal of Physical Chemistry B*, *110*, 5025–5044.

Arnautova, Y. A., Vorobjev, Y. N., Vila, J. A., & Scheraga, H. A. (2009). Identifying native-like protein structures with scoring functions based on all-atom ECEPP force fields, implicit solvent models and structure relaxation. *Proteins: Structure, Function and Bioinformatics*, *77*, 38–51.

Baeten, L., Reumers, J., Tur, V., Stricher, F., Lenaerts, T., Serrano, L., … Schymkowitz, J. (2008). Reconstruction of protein backbones from the BriX collection of canonical protein fragments. *PLoS Computational Biology*, *4*, e1000083.

Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Research*, *28*, 304–305.

Baker, D. (2014). Protein folding, structure prediction and design. *Biochemical Society Transactions*, *42*, 225–9.

Baldwin, L. (2013, February 12). *How long is a piece of Silk ? PeerJ*. Retrieved from https://peerj.com/articles/1

Barth, P., Schonbrun, J., & Baker, D. (2007). Toward high-resolution prediction and design of transmembrane helical protein structures. *Proceedings of the National Academy of Sciences*, *104*, 15682–15687.

Barth, P., Wallner, B., & Baker, D. (2009). Prediction of membrane protein structures with complex topologies using limited constraints. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 1409–14.

Bates, P. A., Kelley, L. A., MacCallum, R. M., & Sternberg, M. J. (2001). Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins*, *Suppl 5*, 39–46.

Ben-David, M., Noivirt-Brik, O., Paz, A., Prilusky, J., Sussman, J. L., & Levy, Y. (2009). Assessment of CASP8 structure predictions for template free targets. *Proteins: Structure, Function and Bioinformatics*, *77*, 50–65.

Berg, B. A., & Neuhaus, T. (1992). Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Physical Review Letters*, *68*, 9–12.

Berg, J. M., Tymoczko, J., & Stryer, L. (2002). *Biochemistry*. *Biochemistry*.

Berger-Tal, O., Nathan, J., Meron, E., Saltz, D., & Houston, A. (2014). The Exploration-Exploitation Dilemma: A Multidisciplinary Framework. *PLoS ONE*, *9*, e95693.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*, 235–242.

Bernsel, A., Viklund, H., Falk, J., Lindahl, E., von Heijne, G., & Elofsson, A. (2008). Prediction of membrane-protein topology from first principles. *Proceedings of the National Academy of Sciences*, *105*, 7177–7181.

Bernsel, A., Viklund, H., Hennerdal, A., & Elofsson, A. (2009). TOPCONS: Consensus prediction of membrane protein topology. *Nucleic Acids Research*, *37*, 465–468.

Betancourt, M. R., & Skolnick, J. (2001). Finding the Needle in a Haystack: Educing Native Folds from Ambiguous Ab Initio Protein Structure Predictions. *Journal of Computational Chemistry*, *22*, 339–353.

Bhattacharya, D., Adhikari, B., Li, J., & Cheng, J. (2016). FRAGSION: Ultra-fast protein fragment library generation by IOHMM sampling. *Bioinformatics*, *32*, 2059–2061.

Bhattacharyya, S., & Varadarajan, R. (2013). Packing in molten globules and native states. *Current Opinion in Structural Biology*.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., … Schwede, T. (2014). SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, *42*, W252–W258.

Bill, R. M., Henderson, P. J. F., Iwata, S., Kunji, E. R. S., Michel, H., Neutze, R., … Vogel, H. (2011). Overcoming barriers to membrane protein structure determination. *Nat Biotech*, *29*, 335–340.

Blum, B., Jordan, M. I., & Baker, D. (2010). Feature space resampling for protein conformational search. *Proteins*, *78*, 1583–93.

Boas, F. E., & Harbury, P. B. (2007, April). Potential energy functions for protein design. *Current Opinion in Structural Biology*.

Bowie, J. U., & Eisenberg, D. (1994). An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proceedings of the National Academy of Sciences of the United States of America*, *91*, 4436–40.

Bradley, P., Malmström, L., Qian, B., Schonbrun, J., Chivian, D., Kim, D. E., … Baker, D. (2005). Free modeling with Rosetta in CASP6. *Proteins: Structure, Function and Genetics*, *61*, 128–134.

Bragg, W. H., & Bragg, W. L. (1913). The Reflection of X-rays by Crystals. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *88*, 428–438.

Bromberg, S., & Dill, K. A. (1994). Side-chain entropy and packing in proteins. *Protein Science*, *3*, 997–1009.

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

Browne, W., North, A., Phillips, D., Brew, K., Vanaman, T., & Hill, R. (1969). A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. *Journal of Molecular Biology*, *42*, 65–86.

Brunette, T., & Brock, O. (2008). Guiding conformation space search with an all-atom energy potential. *Proteins: Structure, Function, and Bioinformatics*, *73*, 958–972.

Brunette, T. J., & Brock, O. (2005). Improving protein structure prediction with model-based search. *Bioinformatics*, *21*, 66–74.

Bukau, B., Weissman, J., & Horwich, A. (2006). Molecular Chaperones and Protein Quality Control. *Cell*.

Burke, D. F., Deane, C. M., & Blundell, T. L. (2000). Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. *Bioinformatics*, *16*, 513–519.

Bystroff, C., & Baker, D. (1997). Blind predictions of local protein structure in CASP2 targets using the I-sites library. *Proteins: Structure, Function and Genetics*, *29*, 167–171.

Bystroff, C., Simons, K. T., Han, K. F., & Baker, D. (1996). Local sequence-structure correlations in proteins. *Current Opinion in Biotechnology*.

Bystrov, V. F., Portnova, S. L., Tsetlin, V. I., Ivanov, V. T., & Ovchinnikov, Y. a. (1969). Conformational studies of peptide systems. The rotational states of the NHCH fragment of alanine dipeptides by nuclear magnetic resonance. *Tetrahedron*, *25*, 493–515.

Cai, Y. D., Feng, K. Y., Lu, W. C., & Chou, K.-C. (2006). Using LogitBoost classifier to predict protein structural classes. *Journal of Theoretical Biology*, *238*, 172–176.

Cao, R., & Cheng, J. (2016). Protein single-model quality assessment by feature-based probability density functions. *Scientific Reports*, *6*, 23990.

Cao, R., Wang, Z., Wang, Y., & Cheng, J. (2014). SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. *BMC Bioinformatics*, *15*, 120.

Cao, Y., Liu, S., Zhang, L., Qin, J., Wang, J., & Tang, K. (2006). Prediction of protein structural class with Rough Sets. *BMC Bioinformatics*, *7*, 20.

Cavanagh, J., Fairbrother, W. J., Palmer III, A. G., Rance, M., & Skelton, N. J. (1996). *Protein NMR spectroscopy*. *Protein NMR Spectroscopy*.

Cheatham, T. E., & Young, M. A. (2000). Molecular dynamics simulation of nucleic acids: Successes, limitations, and promise. *Biopolymers*, *56*, 232–256.

Chen, K. E., Kurgan, L. A., & Ruan, J. (2008). Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *Journal of Computational Chemistry*, *29*, 1596–1604.

Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., … Grishin, N. V. (2014). ECOD: An Evolutionary Classification of Protein Domains. *PLoS Computational Biology*, *10*, e1003926.

Cheng, J., Eickholt, J., Wang, Z., & Deng, X. (2012). Recursive protein modeling: a divide and conquer strategy for Protein Structure Prediction and its case study in CASP9. *Journal of Bioinformatics and Computational Biology*, *10*, 1242003.

Cheng, Y., LeGall, T., Oldfield, C. J., Dunker, A. K., & Uversky, V. N. (2006). Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry*, *45*, 10448–10460.

Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*.

Chivian, D., Kim, D. E., Malmström, L., Schonbrun, J., Rohl, C. A., & Baker, D. (2005). Prediction of CASP6 structures using automated Robetta protocols. In *Proteins: Structure, Function and Genetics* (Vol. 61, pp. 157–166).

Chothia, C. (1992, June 18). One thousand families for the molecular biologist. *Nature*. Nature Publishing Group.

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, *5*, 823–826.

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., … Poljak, R. J. (1989). Conformations of immunoglobulin hypervariable regions. *Nature*, *342*, 877–883.

Chou, K.-C. (1995). A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Computer-Aided Drug Discovery*, *4*, 319–344.

Chou, K.-C. (2000). Prediction of protein structural classes and subcellular locations. *Current Protein & Peptide Science*, *1*, 171–208.

Chou, K.-C. (2005a). Progress in protein structural class prediction and its impact to bioinformatics and proteomics. *Current Protein & Peptide Science*, *6*, 423–436.

Chou, K.-C. (2005b). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, *21*, 10–19.

Chou, K.-C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology*.

Chou, K.-C., & Cai, Y. D. (2005). Prediction of membrane protein types by incorporating amphipathic effects. *Journal of Chemical Information and Modeling*, *45*, 407–413.

Chou, K.-C., & Elrod, D. W. (1999). Prediction of membrane protein types and subcellular locations. *Proteins: Structure, Function and Genetics*, *34*, 137–153.

Chou, K.-C., Liu, W. M., Maggiora, G. M., & Zhang, C. T. (1998). Prediction and classification of domain structural classes. *Proteins: Structure, Function and Genetics*, *31*, 97–103.

Chou, K.-C., & Zhang, C. T. (1995). Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology*, *30*, 275–349.

Chou, P. (1989). Prediction of Protein Structural Classes from Amino Acid Compositions. In G. Fasman (Ed.), *Prediction of Protein Structure and the Principles of Protein Conformation SE - 12* (pp. 549–586). Springer US.

Chowdhury, S., Lee, M. C., Xiong, G., & Duan, Y. (2003). Ab initio folding simulation of the Trp-cage mini-protein approaches NMR resolution. *Journal of Molecular Biology*, *327*, 711–717.

Christen, M., & Van Gunsteren, W. F. (2008). On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. *Journal of Computational Chemistry*, *29*, 157–166.

Clarke, N. D., Ezkurdia, L., Kopp, J., Read, R. J., Schwede, T., & Tress, M. L. (2007a). Domain definition and target classification for CASP7. *Proteins: Structure, Function and Genetics*, *69*, 10–18.

Clarke, N. D., Ezkurdia, L., Kopp, J., Read, R. J., Schwede, T., & Tress, M. L. (2007b). Domain definition and target classification for CASP7. In *Proteins: Structure, Function and Genetics* (Vol. 69, pp. 10–18).

Claros, M. G., & Heijne, G. Von. (1994, December). Toppred II: An improved software for membrane protein structure predictions. *Bioinformatics*.

Cooper, S., Baker, D., Popović, Z., Treuille, A., Barbero, J., Leaver-Fay, A., … Salesin, D. (2010). The challenge of designing scientific discovery games. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games - FDG '10* (pp. 40–47). New York, New York, USA: ACM Press.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., … players, F. (2010). Predicting protein structures with a multiplayer online game. *Nature*, *466*, 756–760.

Coutsias, E. A., Seok, C., & Dill, K. A. (2004). Using quaternions to calculate RMSD. *Journal of Computational Chemistry*, *25*, 1849–1857.

Crippen, G. M. (1978). The tree structural organization of proteins. *Journal of Molecular Biology*, *126*, 315–332.

Cross, T. A., Sharma, M., Yi, M., & Zhou, H. X. (2011). Influence of solubilizing environments on membrane protein structures. *Trends in Biochemical Sciences*.

Csaba, G., Birzele, F., & Zimmer, R. (2009). Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. *BMC Structural Biology*, *9*, 23.

Das, R., & Baker, D. (2008). Macromolecular modeling with rosetta. *Annual Review of Biochemistry*, *77*, 363–382.

Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., … Baker, D. (2007). Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. In *Proteins: Structure, Function and Genetics* (Vol. 69, pp. 118–128).

De Jong, D. H., Singh, G., Bennett, W. F. D., Arnarez, C., Wassenaar, T. A., Schäfer, L. V., … Marrink, S. J. (2013). Improved parameters for the martini coarse-grained protein force field. *Journal of Chemical Theory and Computation*, *9*, 687–697.

de Oliveira, S. H. P., Shi, J., & Deane, C. M. (2015). Building a Better Fragment Library for De Novo Protein Structure Prediction. *Plos One*, *10*, e0123998.

Dehzangi, A., Paliwal, K., Lyons, J., Sharma, A., & Sattar, A. (2014). Proposing a highly accurate protein structural class predictor using segmentation-based features. *BMC Genomics*, *15 Suppl 1*, S2.

Dehzangi, A., Paliwal, K., Sharma, A., Dehzangi, O., & Sattar, A. (2013). A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM*, *10*, 564–75.

Deng, X., & Cheng, J. (2014). Enhancing HMM-based protein profile-profile alignment with structural features and evolutionary coupling information. *BMC Bioinformatics*, *15*, 252.

Deschavanne, P., & Tufféry, P. (2008). Exploring an alignment free approach for protein classification and structural class prediction. *Biochimie*, *90*, 615–625.

di Luccio, E., & Koehl, P. (2011). A quality metric for homology modeling: the H-factor. *BMC Bioinformatics*, *12*, 48.

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

Dill, K. A., & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nature Structural Biology*, *4*, 10–19.

Dill, K. A., & MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science (New York, N.Y.)*, *338*, 1042–6.

Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). The protein folding problem. *Annual Review of Biophysics*, *37*, 289–316.

Ding, S., Li, Y., Shi, Z., & Yan, S. (2014). A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile. *Biochimie*, *97*, 60–65.

Ding, Y.-S., Zhang, T.-L., & Chou, K.-C. (2007). Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. *Protein and Peptide Letters*, *14*, 811–815.

Dixon, J. S. (1997). Evaluation of the CASP2 docking section. *Proteins: Structure, Function, and Genetics*, *29*, 198–204.

Dobson, C. M. (2001). The structural basis of protein folding and its links with human disease. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, *356*, 133–145.

Donate, L. E., Rufino, S. D., Canard, L. H. J., & Blundell, T. L. (1996). Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: A database for modeling and prediction. *Protein Science*, *5*, 2600–2616.

Dong, L., Yuan, Y., & Cai, Y. (2006). Using Bagging classifier to predict protein domain structural class. *Journal of Biomolecular Structure & Dynamics*, *24*, 239–242.

Downing, A. K. (2004). *Protein NMR Techniques*. *Methods in molecular biology (Clifton, N.J.)*.

Duan, Y., & Kollman, P. A. (1998). Pathways to a protein folding intermediate observed in a 1 ms simulation in aqueous solution. *Science*, *282*, 740–744.

Dunbrack, R. L. (2002, August). Rotamer libraries in the 21st century. *Current Opinion in Structural Biology*.

Dunbrack, R. L., & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science*, *6*, 1661–1681.

Dunker, A. K., Babu, M. M., Barbar, E., Blackledge, M., Bondos, S. E., Dosztányi, Z., … Uversky, V. N. (2013). What's in a name? Why these proteins are intrinsically disordered. *Intrinsically Disordered Proteins*, *1*, e24157.

Dunker, A. K., & Kriwacki, R. W. (2011). The orderly chaos of proteins. *Scientific American*, *304*, 68–73.

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., … Obradovic, Z. (2001). Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, *19*, 26–59.

Dyson, H. J., & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*.

Eiben, C. B., Siegel, J. B., Bale, J. B., Cooper, S., Khatib, F., Shen, B. W., … Baker, D. (2012). Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat Biotech*, *30*, 190–192.

Eisenhaber, F., Frömmel, C., & Argos, P. (1996). Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. *Proteins: Structure, Function and Genetics*, *25*, 169–179.

Elofsson, A., Joo, K., Keasar, C., Lee, J., Maghrabi, A. H. A., Manavalan, B., … Wallner, B. (2017). Methods for estimation of model accuracy in CASP12. *Proteins: Structure, Function, and Bioinformatics*.

Engh, R. A., & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallographica Section A*, *47*, 392–400.

Engin, F., & Hotamisligil, G. S. (2010). Restoring endoplasmic reticulum function by chemical chaperones: An emerging therapeutic approach for metabolic diseases. *Diabetes, Obesity and Metabolism*, *12*, 108–115.

Evers, A., & Klebe, G. (2004). Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *Journal of Medicinal Chemistry*, *47*, 5381–5392.

Feng, K. Y., Cai, Y. D., & Chou, K.-C. (2005). Boosting classifier for predicting protein domain structural class. *Biochemical and Biophysical Research Communications*, *334*, 213–217.

Fernandez-Fuentes, N., & Fiser, A. (2006). Saturating representation of loop conformational fragments in structure databanks. *BMC Structural Biology*, *6*, 15.

Fiser, A., Do, R. K., Sali, A., Fiser, A., Kinh, R., Do, G., & Andrej, S. (2000). Modeling of loops in protein structures [In Process Citation] Modeling of loops in protein structures. *Protein Science*, *9*, 1753–1773.

Floudas, C. A. (2007, June 1). Computational methods in protein structure prediction. *Biotechnology and Bioengineering*.

Floudas, C. A., Fung, H. K., McAllister, S. R., Mönnigmann, M., & Rajgaria, R. (2006). Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science*, *61*, 966–988.

Forrest, L. R., Tang, C. L., & Honig, B. (2006). On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophysical Journal*, *91*, 508–517.

Fraenkel, A. S. (1993). Complexity of protein folding. *Bulletin of Mathematical Biology*, *55*, 1199–1210.

Galiez, C., & Coste, F. (2015). Amplitude spectrum distance: measuring the global shape divergence of protein fragments. *BMC Bioinformatics*, *16*, 256.

Galli, S. (2014, December 9). X-ray crystallography: One century of nobel prizes. *Journal of Chemical Education*. American Chemical Society and Division of Chemical Education, Inc.

Gilski, M., Kazmierczyk, M., Krzywda, S., Zábranská, H., Cooper, S., Popović, Z., … Jaskolski, M. (2011). High-resolution structure of a retroviral protease folded as a monomer. *Acta Crystallographica Section D*, *67*, 907–914.

Ginalski, K., Elofsson, A., Fischer, D., & Rychlewski, L. (2003). 3D-Jury: A simple approach to improve protein structure predictions. *Bioinformatics*, *19*, 1015–1018.

Ginalski, K., & Rychlewski, L. (2003). Detection of reliable and unexpected protein fold predictions using 3D-Jury. *Nucleic Acids Research*, *31*, 3291–3292.

Giorgetti, A., Raimondo, D., Miele, A. E., & Tramontano, A. (2005). Evaluating the usefulness of protein structure models for molecular replacement. *Bioinformatics (Oxford, England)*, *21 Suppl 2*, ii72-i76.

Govindarajan, S., Recabarren, R., & Goldstein, R. A. (1999). Estimating the total number of protein folds. *Proteins: Structure, Function and Genetics*, *35*, 408–414.

Gribskov, M., McLachlan, A. D., & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences*, *84*, 4355–4358.

Gront, D., & Kolinski, A. (2005). A new approach to prediction of short-range conformational propensities in proteins. *Bioinformatics*, *21*, 981–987.

Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. M., & Baker, D. (2011). Generalized fragment picking in Rosetta: design, protocols and applications. *PloS One*, *6*, e23294.

Guntas, G., Purbeck, C., & Kuhlman, B. (2010). Engineering a protein–protein interface using a computationally designed library. *Proceedings of the National Academy of Sciences*, *107*, 19296–19301.

Guyon, F., & Tufféry, P. (2010). Assessing 3D scores for protein structure fragment mining. *Open Access Bioinformatics*, *2*, 67–77.

Guyon, F., & Tufféry, P. (2014). Fast protein fragment similarity scoring using a Binet–Cauchy kernel. *Bioinformatics*, *30*, 784–791.

Hadley, C., & Jones, D. T. (1999). A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure (London, England : 1993)*, *7*, 1099–112.

Hagler, A. T., Huler, E., & Lifson, S. (1974). Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *Journal of the American Chemical Society*, *96*, 5319–5327.

Han, K. F., & Baker, D. (1995). Recurring Local Sequence Motifs in Proteins. *Journal of Molecular Biology*, *251*, 176–187.

Han, K. F., & Baker, D. (1996). Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 5814–8.

Handl, J., Knowles, J., Vernon, R., Baker, D., & Lovell, S. C. (2012). The dual role of fragments in fragment-assembly methods for de novo protein structure prediction. *Proteins: Structure, Function and Bioinformatics*, *80*, 490–504.

Hansmann, U. H. E., & Okamoto, Y. (1999, April). New Monte Carlo algorithms for protein folding. *Current Opinion in Structural Biology*.

Hart, W. E., & Istrail, S. (2000). Invariant patterns in crystal lattices: Implications for protein folding algorithms. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1075*, 288–303.

Hart, W. E., & Newman, A. (1997). Lattice and off-lattice side chain models of protein folding: linear time structure prediction better than 86% of optimal. *Journal of Computational Biology*, *4*, 137–146.

Hart, W. E., & Newman, A. (2006). Protein structure prediction with lattice models. In *Handbook of Molecular Biology* (pp. 1–24). Computer and Information Science Series. CRC Press.

Hartl, F. U., & Hayer-Hartl, M. (2002). Protein folding. Molecular chaperones in the cytosol: From nascent chain to folded protein. *Science*.

Hatahet, F., Nguyen, V. D., Salo, K. E. H., & Ruddock, L. W. (2010). Disruption of reducing pathways is not essential for efficient disulfide bond formation in the cytoplasm of E. coli. *Microbial Cell Factories*, *9*, 67.

Hauptman, H. A. (2015). History of X-Ray Crystallography. In *Science of Crystal Structures* (pp. 19–22). Cham: Springer International Publishing.

Hayat, M., & Khan, A. (2012a). Mem-PHybrid: Hybrid features-based prediction system for classifying membrane protein types. *Analytical Biochemistry*, *424*, 35–44.

Hayat, M., & Khan, A. (2012b). MemHyb: Predicting membrane protein types by hybridizing SAAC and PSSM. *Journal of Theoretical Biology*, *292*, 93–102.

Hayat, M., Khan, A., & Yeasin, M. (2012). Prediction of membrane proteins using split amino acid and ensemble classification. *Amino Acids*, *42*, 2447–2460.

He, Y., Chen, Y., Alexander, P., Bryan, P. N., & Orban, J. (2008). NMR structures of two designed proteins with high sequence identity but different fold and function. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 14412–7.

He, Y., Mozolewska, M. A., Krupa, P., Sieradzan, A. K., Wirecki, T. K., Liwo, A., … Scheraga, H. A. (2013). Lessons from application of the UNRES force field to predictions of structures of CASP10 targets. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 14936–14941.

Helles, G. (2008). A comparative study of the reported performance of ab initio protein structure prediction algorithms. *Journal of the Royal Society, Interface / the Royal Society*, *5*, 387–96.

Heun, V. (1999). Approximate protein folding in the hp side chain model on extended cubic lattices. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 1643, pp. 212–223).

Himu, F. A., Jahangir, K. B., Ridoy, M. Z., Dhar, S., & Shatabda, S. (2015). A new effective algorithm for protein chain lattice fit problem. In *2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)* (pp. 515–518). IEEE.

Hockenmaier, J., Joshi, A. K., & Dill, K. A. (2007). Routes are trees: the parsing perspective on protein folding. *Proteins*, *66*, 1–15.

Holland, J. H. (John H., & H., J. (1992). *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence*. MIT Press.

Hopkins, A. L., & Groom, C. R. (2002). The druggable genome. *Nature Reviews Drug Discovery*, *1*, 727–730.

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., … MacKerell, A. D. (2016). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, *14*, 71–73.

Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J. F., Honig, B., Shaw, D. E., & Friesner, R. A. (2004). A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins: Structure, Function and Genetics*, *55*, 351–367.

Jahandideh, S., Abdolmaleki, P., Jahandideh, M., & Asadabadi, E. B. (2007). Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. *Biophysical Chemistry*, *128*, 87–93.

Jahnke, W., & Widmer, H. (2004). Protein NMR in biomedical research. *Cellular and Molecular Life Sciences*, *61*, 580–599.

Jana, N. D., Sil, J., & Das, S. (2017). Selection of appropriate metaheuristic algorithms for protein structure prediction in AB off-lattice model: a perspective from fitness landscape analysis. *Information Sciences*, *391–392*, 28–64.

Jaroszewski, L., Godzik, A., & Rychlewski, L. (2000). Improving the quality of twilight-zone alignments. *Protein Science*, *9*, 1487–1496.

Jauch, R., Yeo, H. C., Kolatkar, P. R., & Clarke, N. D. (2007). Assessment of CASP7 structure predictions for template free targets. *Proteins: Structure, Function, and Bioinformatics*, *69*, 57–67.

Jiang, Q., Jin, X., Lee, S. J., & Yao, S. (2017, September 1). Protein secondary structure prediction: A survey of the state of the art. *Journal of Molecular Graphics and Modelling*. Elsevier.

Jiang, T., Cui, Q., Shi, G., & Ma, S. (2003). Protein folding simulations of the hydrophobic–hydrophilic model by combining tabu search with genetic algorithms. *The Journal of Chemical Physics*, *119*, 4592–4596.

Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, *11*, 431.

Jones, D. T. (1997). Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins*, *Suppl 1*, 185–91.

Jones, D. T. (1998). Chapter 13 THREADER: protein sequence threading by double dynamic programming. *New Comprehensive Biochemistry*, *32*, 285–311.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, *292*, 195–202.

Jones, D. T. (2001). Predicting novel protein folds by using FRAGFOLD. *Proteins: Structure, Function and Genetics*, *Suppl 5*, 127–132.

Jones, D. T. (2007). Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, *23*, 538–544.

Jones, D. T., Bryson, K., Coleman, A., McGuffin, L. J., Sadowski, M. I., Sodhi, J. S., & Ward, J. J. (2005). Prediction of novel and analogous folds using fragment assembly and fold recognition. *Proteins*, *61 Suppl 7*, 143–51.

Jones, D. T., & McGuffin, L. J. (2003). Assembling novel protein folds from super-secondary structural fragments. *Proteins*, *53 Suppl 6*, 480–5.

Jones, D. T., Moody, C. M., Uppenbrink, J., Viles, J. H., Doyle, P. M., Harris, C. J., … Thornton, J. M. (1996). Towards meeting the paracelsus challenge: The design, synthesis, and characterization of paracelsin-43, an α-helical protein with over 50% sequence identity to an all-β protein. *Proteins: Structure, Function, and Genetics*, *24*, 502–513.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1994). A Model Recognition Approach to the Prediction of All-Helical Membrane Protein Structure and Topology. *Biochemistry*, *33*, 3038–3049.

Jonic, S., & Vénien-Bryan, C. (2009). Protein structure determination by electron cryo-microscopy. *Current Opinion in Pharmacology*, *9*, 636–642.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, *32*, 922–923.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, *34*, 827–828.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*, 2577–2637.

Kachlishvili, K., Maisuradze, G. G., Martin, O. A., Liwo, A., Vila, J. A., & Scheraga, H. A. (2014). Accounting for a mirror-image conformation as a subtle effect in protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 8458–63.

Käll, L., Krogh, A., & Sonnhammer, E. L. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. *Nucleic Acids Research*, *35*, 429–432.

Kalman, M., & Ben-Tal, N. (2010). Quality assessment of protein model-structures using evolutionary conservation. *Bioinformatics (Oxford, England)*, *26*, 1299–307.

Kamisetty, H., Ovchinnikov, S., & Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 15674–9.

Kandathil, S. M., Handl, J., & Lovell, S. C. (2016). Toward a detailed understanding of search trajectories in fragment assembly approaches to protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, *84*, 411–426.

Karplus, K. (2009). SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Research*, *37*, W492–W497.

Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L., & Hughey, R. (1999). Predicting protein structure using only sequence information. *Proteins: Structure, Function, and Genetics*, *37*, 121–125.

Karplus, K., Barrett, C., & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics (Oxford, England)*, *14*, 846–856.

Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M., & Hughey, R. (2003). Combining Local-Structure, Fold-Recognition, and New Fold Methods for Protein Structure Prediction. In *Proteins: Structure, Function and Genetics* (Vol. 53, pp. 491–496).

Kavousi, K., Moshiri, B., Sadeghi, M., Araabi, B. N., & Moosavi-Movahedi, A. A. (2011). A protein fold classifier formed by fusing different modes of pseudo amino acid composition via PSSM. *Computational Biology and Chemistry*, *35*, 1–9.

Kendrew, J. C., Bodo, G., DINTZIS, H. M., PARRISH, R. G., WYCKOFF, H., PHILLIPS, D. C., & Philips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, *181*, 662–666.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., & Shore, V. C. (1960). Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 Å. Resolution. *Nature*, *185*, 422–427.

Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popovic, Z., … Players, F. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*, *108*, 18949–18953.

Khor, B. Y., Tye, G. J., Lim, T. S., & Choong, Y. S. (2015). General overview on structure prediction of twilight-zone proteins. *Theoretical Biology and Medical Modelling*, *12*. Retrieved from http://dx.doi.org/10.1186/s12976-015-0014-1

Khoury, G. A., Smadbeck, J., Kieslich, C. A., & Floudas, C. A. (2014, February 1). Protein folding and de novo protein design for biotechnological applications. *Trends in Biotechnology*. Springer.

Kihara, D., Lu, H., Kolinski, A., & Skolnick, J. (2001). TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proceedings of the National Academy of Sciences*, *98*, 10125–10130.

Kim, D. E., Blum, B., Bradley, P., & Baker, D. (2009). Sampling Bottlenecks in De novo Protein Structure Prediction. *Journal of Molecular Biology*, *393*, 249–260.

Kim, D. E., Chivian, D., & Baker, D. (2004a). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, *32*, W526–W531.

Kim, D. E., Chivian, D., & Baker, D. (2004b). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res*, *32*.

Kinch, L. N., Li, W., Schaeffer, R. D., Dunbrack, R. L., Monastyrskyy, B., Kryshtafovych, A., & Grishin, N. V. (2016). CASP 11 Target Classification. *Proteins: Structure, Function, and Bioinformatics*, *84*, 20–33.

Kinch, L. N., Qi, Y., Hubbard, T. J. P., & Grishin, N. V. (2003). CASP5 Target Classification. *Proteins: Structure, Function and Genetics*, *53*, 340–351.

Kinch, L. N., Shi, S., Cheng, H., Cong, Q., Pei, J., Mariani, V., … Grishin, N. V. (2011). CASP9 target classification. *Proteins: Structure, Function, and Bioinformatics*, *79*, 21–36.

Kinch, L. N., Yong Shi, S., Cong, Q., Cheng, H., Liao, Y., & Grishin, N. V. (2011). CASP9 assessment of free modeling target predictions. *Proteins: Structure, Function and Bioinformatics*, *79*, 59–73.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science (New York, N.Y.)*, *220*, 671–680.

Kleffner, R., Flatten, J., Leaver-Fay, A., Baker, D., Siegel, J. B., Khatib, F., & Cooper, S. (2017). Foldit Standalone: a video game-derived protein structure manipulation interface using Rosetta. *Bioinformatics (Oxford, England)*, *33*, 2765–2767.

Klein, P., & Delisi, C. (1986). Prediction of protein structural class from the amino acid sequence. *Biopolymers*, *25*, 1659–1672.

Klepeis, J. L., & Floudas, C. A. (2003a). ASTRO-FOLD: a combinatorial and global optimization framework for Ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophysical Journal*, *85*, 2119–2146.

Klepeis, J. L., & Floudas, C. A. (2003b). Prediction of β-sheet topology and disulfide bridges in polypeptides. *Journal of Computational Chemistry*, *24*, 191–208.

Klepeis, J. L., Pieja, M. J., & Floudas, C. A. (2003). Hybrid global optimization algorithms for protein structure prediction: Alternating hybrids. *Biophysical Journal*, *84*, 869–882.

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., & Kolinski, A. (2016, July 27). Coarse-Grained Protein Models and Their Applications. *Chemical Reviews*. American Chemical Society.

Kneller, D. G., Cohen, F. E., & Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *Journal of Molecular Biology*, *214*, 171–182.

Knowles, T. P. J., Vendruscolo, M., & Dobson, C. M. (2014). The amyloid state and its association with protein misfolding diseases. *Nature Reviews Molecular Cell Biology*, *15*, 384–396.

Kohn, J. E., Millett, I. S., Jacob, J., Zagrovic, B., Dillon, T. M., Cingel, N., … Plaxco, K. W. (2004). Random-coil behavior and the dimensions of chemically unfolded proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 12491–6.

Kolata, G. (1986). Trying to crack the second half of the genetic code. *Science*, *233*, 1037–1039.

Koldsø, H., Jensen, M. Ø., Jogini, V., & Shaw, D. E. (2017). Permeation, Gating, and Modulation of the TRPA1 Channel in Long-Timescale Molecular Dynamics Simulations. *Biophysical Journal*, *112*, 466a.

Kolinski, A. (2004). Protein modeling and structure prediction with a reduced representation. In *Acta Biochimica Polonica* (Vol. 51, pp. 349–371).

Kolinski, A., & Skolnick, J. (2004). Reduced models of proteins and their applications. *Polymer*, *45*, 511–524.

Kolodny, R., Koehl, P., Guibas, L., & Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. *Journal of Molecular Biology*, *323*, 297–307.

König, R., & Dandekar, T. (2001). Solvent entropy-driven searching for protein modeling examined and tested in simplified models. *Protein Engineering*, *14*, 329–35.

Konopka, B. M., Nebel, J.-C., & Kotulska, M. (2012). Quality assessment of protein model-structures based on structural and functional similarities. *BMC Bioinformatics*, *13*, 242.

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

Kortemme, T., Morozov, A. V., & Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *Journal of Molecular Biology*, *326*, 1239–1259.

Kosciolek, T., & Jones, D. T. (2014). De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PloS One*, *9*, e92197.

Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I., & Wright, P. E. (1996). Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proceedings of the National Academy of Sciences*, *93*, 11504–11509.

Krogh, A., Larsson, B., Von Heijne, G., & Sonnhammer, E. L. . (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, *305*, 567–580.

Krupa, P., Mozolewska, M. A., Joo, K., Lee, J., Czaplewski, C., & Liwo, A. (2015). Prediction of Protein Structure by Template-Based Modeling Combined with the UNRES Force Field. *Journal of Chemical Information and Modeling*, *55*, 1271–1281.

Kruse, A. C., Hu, J., Pan, A. C., Arlow, D. H., Rosenbaum, D. M., Rosemond, E., … Kobilka, B. K. (2012). Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature*, *482*, 552–556.

Kryshtafovych, A., Fidelis, K., & Tramontano, A. (2011). Evaluation of model quality predictions in CASP9. *Proteins: Structure, Function and Bioinformatics*, *79*, 91–106.

Kryshtafovych, A., Monastyrskyy, B., & Fidelis, K. (2016). CASP11 statistics and the prediction center evaluation system. *Proteins: Structure, Function and Bioinformatics*, 9–13.

Kufareva, I., & Abagyan, R. (2012). Methods of protein structure comparison. *Methods in Molecular Biology*, *857*, 231–257.

Kunz, H. (2002). Emil Fischer--unequalled classicist, master of organic chemistry research, and inspired trailblazer of biological chemistry. *Angewandte Chemie (International Ed. in English)*, *41*, 4439–51.

Kurgan, L. A., & Chen, K. (2007). Prediction of protein structural class for the twilight zone sequences. *Biochemical and Biophysical Research Communications*, *357*, 453–460.

Kurgan, L. A., Cios, K., & Chen, K. (2008). SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. *BMC Bioinformatics*, *9*, 226.

Kurgan, L. A., & Homaeian, L. (2006). Prediction of structural classes for protein sequences and domains-Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. *Pattern Recognition*, *39*, 2323–2343.

Kurgan, L. A., Zhang, T., Zhang, H., Shen, S., & Ruan, J. (2008). Secondary structure-based assignment of the protein structural classes. *Amino Acids*, *35*, 551–564.

Kwasigroch, J. M., Chomilier, J., & Mornon, J. P. (1996). A global taxonomy of loops in globular proteins. *Journal of Molecular Biology*, *259*, 855–872.

Ladd, M., & Palmer, R. (2013). *Structure Determination by X-ray Crystallography*. *Structure Determination by X-ray Crystallography*. Boston, MA: Springer US. Retrieved from http://link.springer.com/10.1007/978-1-4614-3954-7

Laidig, K. E., & Daggett, V. (1996). Molecular dynamics simulations of apocytochrome b562- The highly ordered limit of molten globules. *Folding and Design*, *1*, 335–346.

Lam, S. D., Das, S., Sillitoe, I., & Orengo, C. A. (2017). An overview of comparative modelling and resources dedicated to large-scale modelling of genome sequences. *Acta Crystallographica Section D: Structural Biology*, *73*, 628–640.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*, 2947–2948.

Lathrop, R. H. (1994). The protein threading problem with sequence amino acid interaction preferences is np-complete. *Protein Engineering, Design and Selection*, *7*, 1059–1068.

Lau, K. F., & Dill, K. A. (1989). A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins. *Macromolecules*, *22*, 3986–3997.

Lazaridis, T., & Karplus, M. (1999). Effective energy function for proteins in solution. Proteins. *Proteins*, *35*.

Leaver-Fay, A., O'Meara, M. J., Tyka, M., Jacak, R., Song, Y., Kellogg, E. H., … Kuhlman, B. (2013). Scientific benchmarks for guiding macromolecular energy function improvement. *Methods in Enzymology*, *523*, 109–143.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., … Bradley, P. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, *487*, 545–74.

Lee, B. T., & Richards, F. M. (1971). The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*.

Lee, J. (1993). New Monte Carlo algorithm: Entropic sampling. *Physical Review Letters*, *71*, 211–214.

Lee, J., Kim, S.-Y., Joo, K., Kim, I., & Lee, J. (2004). Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. *Proteins: Structure, Function, and Bioinformatics*, *56*, 704–714.

Lee, J., Liwo, A., Ripoll, D. R., Pillardy, J., Saunders, J. A., Gibson, K. D., & Scheraga, H. A. (2000). Hierarchical energy-based approach to protein-structure prediction: Blind-test evaluation with CASP3 targets. *International Journal of Quantum Chemistry*, *77*, 90–117.

Lee, J., Wu, S., & Zhang, Y. (2009). Ab initio protein structure prediction. In *From Protein Structure to Function with Bioinformatics* (pp. 3–25). Dordrecht: Springer Netherlands.

Lee, S., Tran, A., Allsopp, M., Lim, J. B., Hénin, J., & Klauda, J. B. (2014). CHARMM36 United Atom Chain Model for Lipids and Surfactants. *The Journal of Physical Chemistry B*, *118*, 547–556.

Lei, H., & Duan, Y. (2007). Two-stage Folding of HP-35 from Ab Initio Simulations. *Journal of Molecular Biology*, *370*, 196–206.

Lei, H., Wu, C., Liu, H., & Duan, Y. (2007). Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, *104*, 4925–4930.

Leman, J. K., Mueller, R., Karakas, M., Woetzel, N., & Meiler, J. (2013). Simultaneous prediction of protein secondary structure and transmembrane spans. *Proteins: Structure, Function and Bioinformatics*, *81*, 1127–1140.

Lemmon, G., & Meiler, J. (2012). Rosetta ligand docking with flexible XML protocols. *Methods in Molecular Biology*, *819*, 143–155.

Lesh, N., Mitzenmacher, M., & Whitesides, S. (2003). A complete and effective move set for simplified protein folding. In *Proceedings of the seventh annual international conference on Computational molecular biology - RECOMB '03* (pp. 188–195). New York, New York, USA: ACM Press.

Levinthal, C. (1968). Are there pathways for protein folding? *Journal de Chimie Physique*, *65*, 44–45.

Levitt, M., & Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, *261*, 552–558.

Levy-Moonshine, A., Amir, E. A. D., & Keasar, C. (2009). Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. *Bioinformatics*, *25*, 2639–2645.

Li, L., Cui, X., Yu, S., Zhang, Y., Luo, Z., Yang, H., … Zheng, X. (2014). PSSP-RFE: Accurate prediction of protein structural class by recursive feature extraction from PSI-BLAST profile, physical-chemical property and functional annotations. *PLoS ONE*, *9*.

Li, S. C., Bu, D., Gao, X., Xu, J., & Li, M. (2008). Designing succinct structural alphabets. *Bioinformatics*, *24*, i182-9.

Li, S. C., Bu, D., Xu, J., & Li, M. (2008). Fragment-HMM: a new approach to protein structure prediction. *Protein Science : A Publication of the Protein Society*, *17*, 1925–34.

Li, Z.-C., Zhou, X.-B., Lin, Y.-R., & Zou, X.-Y. (2008). Prediction of protein structure class by coupling improved genetic algorithm and support vector machine. *Amino Acids*, *35*, 581–590.

Li, Z., & Scheraga, H. A. (1987). Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences*, *84*, 6611–6615.

Lichtenthaler, F. W. (2002). Emil Fischer, his personality, his achievements, and his scientific progeny. *European Journal of Organic Chemistry*.

Liebert, M. A. (2000). Structure Comparison and Structure Patterns. *J Comput Biol*, *7*, 685–716.

Lindorff-Larsen, K., Maragakis, P., Piana, S., & Shaw, D. E. (2016). Picosecond to Millisecond Structural Dynamics in Human Ubiquitin. *Journal of Physical Chemistry B*, *120*, 8313–8320.

Lindorff-Larsen, K., Piana, S., Dror, R. O., & Shaw, D. E. (2011). How fast-folding proteins fold. *Science (New York, N.Y.)*, *334*, 517–520.

Liu, T., Geng, X., Zheng, X., Li, R., & Wang, J. (2012). Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. *Amino Acids*, *42*, 2243–2249.

Liu, T., & Jia, C. (2010). A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. *Journal of Theoretical Biology*, *267*, 272–275.

Liu, T., Zheng, X., & Wang, J. (2010). Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile. *Biochimie*, *92*, 1330–4.

Liwo, A., Baranowski, M., Czaplewski, C., Gołaś, E., He, Y., Jagieła, D., … Zaborowski, B. (2014). A unified coarse-grained model of biological macromolecules based on mean-field multipole-multipole interactions. *Journal of Molecular Modeling*, *20*, 2306.

Liwo, A., Czaplewski, C., Pillardy, J., & Scheraga, H. A. (2001). Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *Journal of Chemical Physics*, *115*, 2323–2347.

Liwo, A., Khalili, M., & Scheraga, H. A. (2005). Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 2362–7.

Lo Conte, L., Brenner, S. E., Hubbard, T. J. P., Chothia, C., & Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Research*, *30*, 264–267.

Lu, W., & Liu, H. (2007). Correlations Between Amino Acids at Different Sites in Local Sequences of Protein Fragments with Given Structural Patterns. *Chinese Journal of Chemical Physics*, *20*, 71.

Luo, L. (2014). Quantum theory on protein folding. *Science China Physics, Mechanics and Astronomy*, *57*, 458–468.

Lyskov, S., Chou, F. C., Conchúir, S. Ó., Der, B. S., Drew, K., Kuroda, D., … Das, R. (2013). Serverification of Molecular Modeling Applications: The Rosetta Online Server That Includes Everyone (ROSIE). *PLoS ONE*, *8*, 5–7.

Maccallum, J. L., Pérez, A., Schnieders, M. J., Hua, L., Jacobson, M. P., & Dill, K. A. (2011). Assessment of protein structure refinement in CASP9. *Proteins: Structure, Function and Bioinformatics*, *79*, 74–90.

MacKerell, A. D., Wiórkiewicz-Kuczera, J., Karplus, M., & MacKerell, A. D. (1995). An All-Atom Empirical Energy Function for the Simulation of Nucleic Acids. *Journal of the American Chemical Society*, *117*, 11946–11975.

Maher, B., Albrecht, A. a, Loomes, M., Yang, X.-S., & Steinhöfel, K. (2014). A firefly-inspired method for protein structure prediction in lattice models. *Biomolecules*, *4*, 56–75.

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

Mandell, D. J., Coutsias, E. A., & Kortemme, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature Methods*, *6*, 551–552.

Mao, B., Tejero, R., Baker, D., & Montelione, G. T. (2014). Protein NMR Structures Refined with Rosetta Have Higher Accuracy Relative to Corresponding X-ray Crystal Structures. *Journal of the American Chemical Society*, *136*, 1893–1906.

Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics (Oxford, England)*, *29*, 2722–8.

Marti-Renom, M. A., Madhusudhan, M. S., & Sali, A. (2004). Alignment of protein sequences by their profiles. *Protein Science : A Publication of the Protein Society*, *13*, 1071–87.

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., & Šali, A. (2000). Comparative Protein Structure Modeling of Genes and Genomes. *Annual Review of Biophysics and Biomolecular Structure*, *29*, 291–325.

McCammon, J. A., Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. *Nature*, *267*, 585–590.

McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, *16*, 404–5.

Melamud, E., & Moult, J. (2003). Evaluation of disorder predictions in CASP5. *Proteins*, *53 Suppl 6*, 561–5.

Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science*, *15*, 125–130.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, *21*, 1087.

Metropolis, N., & Ulam, S. M. (1949). The {Monte Carlo} Method. *Journal of the American Statistical Association*, *44*, 335–341.

Michel, M., Hayat, S., Skwark, M. J., Sander, C., Marks, D. S., & Elofsson, A. (2014). PconsFold: improved contact predictions improve protein models. *Bioinformatics*, *30*, i482–i488.

Michel, M., Menéndez Hurtado, D., Uziela, K., & Elofsson, A. (2017). Large-scale structure prediction by improved contact predictions and model quality assessment. *Bioinformatics*, *33*, i23–i29.

Michie, A. D., Orengo, C. A., & Thornton, J. M. (1996). Analysis of domain structural class using an automated class assignment protocol. *Journal of Molecular Biology*, *262*, 168–185.

Mihăşan, M. (2010). Basic protein structure prediction for the biologist: A review. *Archives of Biological Sciences*, *62*, 857–871.

Milner-White, E. J. (1990). Situations of gamma-turns in proteins. Their relation to alpha-helices, beta-sheets and ligand binding sites. *Journal of Molecular Biology*, *216*, 385–397.

Mizianty, M. J., & Kurgan, L. A. (2009). Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. *BMC Bioinformatics*, *10*, 414.

Modi, V., & Dunbrack, R. L. (2016). Assessment of refinement of template-based models in CASP11. *Proteins: Structure, Function and Bioinformatics*, *84*, 260–281.

Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A., & Fidelis, K. (2014). Assessment of protein disorder region predictions in CASP10. *Proteins*, *82 Suppl 2*, 127–37.

Montelione, G. T., Zheng, D., Huang, Y. J., Gunsalus, K. C., & Szyperski, T. (2000). Protein NMR spectroscopy in structural genomics. *Nature Structural Biology*, *7*, 982–985.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP) - round x. *Proteins: Structure, Function and Bioinformatics*, *82*, 1–6.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2018). Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins: Structure, Function and Bioinformatics*, *86*, 7–15.

Moult, J., Fidelis, K., Zemla, A., & Hubbard, T. (2001). Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins*, *Suppl 5*, 2–7.

Moult, J., Pedersen, J. T., Judson, R., & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, *23*, ii--iv.

Murata, K., & Wolf, M. (2018). Cryo-electron microscopy for structural analysis of dynamic biological. *BBA - General Subjects*, *1862*, 324–334.

Murzin, a G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, *247*, 536–40.

Nakashima, H., Nishikawa, K., & Ooi, T. (1986). The folding type of a protein is relevant to the amino acid composition. *Journal of Biochemistry*, *99*, 153–162.

Nannenga, B. L., & Gonen, T. (2014). Protein structure determination by MicroED. *Current Opinion in Structural Biology*, *27*, 24–31.

Ngo, J. T., & Marks, J. (1992). Computational complexity of a problem in molecular structure prediction. *Protein Engineering*, *5*, 313–321.

Ngom, A. (2006). Parallel evolution strategy on grids for the protein threading problem. *Journal of Parallel and Distributed Computing*, *66*, 1489–1502.

Nugent, T., Cozzetto, D., & Jones, D. T. (2014). Evaluation of predictions in the CASP10 model refinement category. *Proteins: Structure, Function and Bioinformatics*, *82*, 98–111.

O'Meara, M. J., Leaver-Fay, A., Tyka, M. D., Stein, A., Houlihan, K., Dimaio, F., … Kuhlman, B. (2015). Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *Journal of Chemical Theory and Computation*, *11*, 609–622.

Oldziej, S., Czaplewski, C., Liwo, A., Chinchio, M., Nanias, M., Vila, J. A., … Scheraga, H. A. (2005). Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. *Proceedings of the National Academy of Sciences*, *102*, 7547–7552.

Olechnovič, K., Kulberkyte, E., & Venclovas, Č. (2013). CAD-score: A new contact area difference-based function for evaluation of protein structural models. *Proteins: Structure, Function and Bioinformatics*, *81*, 149–162.

Olson, B., Molloy, K., Hendi, S. F., & Shehu, A. (2012). Guiding probabilistic search of the protein conformational space with structural profiles. *Journal of Bioinformatics and Computational Biology*, *10*, 1242005.

Ordway, G. A. (2004). Myoglobin: an essential hemoprotein in striated muscle. *Journal of Experimental Biology*, *207*, 3441–3446.

Orengo, C. A., Bray, J. E., Hubbard, T., LoConte, L., & Sillitoe, I. (1999). Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins: Structure, Function and Genetics*, *37*, 149–170.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure (London, England : 1993)*, *5*, 1093–1108.

Ott, H. (1927). Structure analysis. *Z. f. Krist*, *66*, 136–153.

Ovchinnikov, S., Kim, D. E., Wang, R. Y. R., Liu, Y., Dimaio, F., & Baker, D. (2016). Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins: Structure, Function and Bioinformatics*, 1–9.

Ovchinnikov, S., Park, H., Kim, D. E., Liu, Y., Wang, R. Y. R., & Baker, D. (2016). Structure prediction using sparse simulated NOE restraints with Rosetta in CASP11. *Proteins: Structure, Function and Bioinformatics*.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., … Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science (New York, N.Y.)*, *355*, 294–298.

Pande, V. S., & Rokhsar, D. S. (1998). Is the molten globule a third phase of proteins? *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 1490–1494.

Pande, V. S., & Rokhsar, D. S. (1999). Folding pathway of a lattice model for proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *96*, 1273–8.

Pandit, S. B., Zhang, Y., & Skolnick, J. (2006). TASSER-Lite: an automated tool for protein comparative modeling. *Biophysical Journal*, *91*, 4180–90.

Pardon, E., Haezebrouck, P., De Baetselier, A., Hooke, S. D., Fancourt, K. T., Desmet, J., … Joniau, M. (1995). A Ca(2+)-binding chimera of human lysozyme and bovine alpha-lactalbumin that can form a molten globule. *The Journal of Biological Chemistry*, *270*, 10514–24.

Park, J., & Saitou, K. (2014). ROTAS: a rotamer-dependent, atomic statistical potential for assessment and prediction of protein structures. *BMC Bioinformatics*, *15*, 307.

Park, S.-J. (2005). A study of fragment-based protein structure prediction: biased fragment replacement for searching low-energy conformation. *Genome Informatics. International Conference on Genome Informatics*, *16*, 104–13.

Paterson, M., & Przytycka, T. M. (1996). On the Complexity of String Folding (pp. 658–669). Springer, Berlin, Heidelberg.

Pauling, L., Corey, R. B., & Branson, H. R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, *37*, 205–11.

Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., DeBolt, S., … Kollman, P. (1995). 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. *Computer Physics Communications*, *91*, 1–41.

Pechkova, E., & Nicolini, C. (2003a). Proteomics and Nanocrystallography. doi:10.1007/978-1-4615-0041-4

Pechkova, E., & Nicolini, C. (2003b). State of the Art in Proteomics, Crystallography and Nanobiotechnology. In *Proteomics and Nanocrystallography* (pp. 9–61). Boston, MA: Springer US.

Peng, J., & Xu, J. (2011). Raptorx: Exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function and Bioinformatics*, *79*, 161–171.

Perez, A., Morrone, J. A., & Dill, K. A. (2017). Accelerating physical simulations of proteins by leveraging external knowledge. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1–15.

Perez, A., Yang, Z., Bahar, I., Dill, K. A., & MacCallum, J. L. (2012). FlexE: Using Elastic network models to compare models of protein structure. *Journal of Chemical Theory and Computation*, *8*, 3985–3991.

Perez, A., Yang, Z., Bahar, I., Dill, K. A., Maccallum, J. L., Olechnovič, K., … Elofsson, A. (2014). Calibur: a tool for clustering large numbers of protein decoys. *Bioinformatics*, *30*, 256.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., & North, A. C. (1960). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature*, *185*, 416–22.

Petrey, D., Xiang, Z., Tang, C. L., Xie, L., Gimpelev, M., Mitros, T., … Honig, B. (2003). Using Multiple Structure Alignments, Fast Model Building, and Energetic Analysis in Fold Recognition and Homology Modeling. In *Proteins: Structure, Function and Genetics* (Vol. 53, pp. 430–435).

Petsko, G. A., & Ringe, D. (2004). *Protein structure and function*. New Science Press.

Piana, S., Klepeis, J. L., & Shaw, D. E. (2014, February). Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations. *Current Opinion in Structural Biology*.

Pillardy, J., Czaplewski, C., Liwo, A., Lee, J., Ripoll, D. R., Kazmierkiewicz, R., … Scheraga, H. A. (2001). Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proceedings of the National Academy of Sciences*, *98*, 2329–2333.

Pillardy, J., Czaplewski, C., Wedemeyer, W. J., & Scheraga, H. A. (2000). Conformation-Family Monte Carlo (CFMC): An Efficient Computational Method for Identifying the Low-Energy States of a Macromolecule. *Helvetica Chimica Acta*, *83*, 2214–2230.

Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C. J., … Tosatto, S. C. E. (2017). DisProt 7.0: A major update of the database of disordered proteins. *Nucleic Acids Research*, *45*, D219–D227.

Pitera, J. W., & Swope, W. (2003). Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 7587–7592.

Pollack, V., Scheiber, K., Pfaller, W., & Schramek, H. (1997). Loss of cytokeratin expression and formation of actin stress fibers in dedifferentiated MDCK-C7 cell lines. *Biochemical and Biophysical Research Communications*, *241*, 541–547.

Ponder, J. W., Wu, C., Ren, P., Pande, V. S., Chodera, J. D., Schnieders, M. J., … Head-Gordon, T. (2010). Current status of the AMOEBA polarizable force field. *Journal of Physical Chemistry B*, *114*, 2549–2564.

Proteomics. (2007). In *Bioinformatics* (pp. 261–298). Berlin, Heidelberg: Springer Berlin Heidelberg.

Prusiner, S. B. (1998). Nobel Lecture: Prions. *Proceedings of the National Academy of Sciences*, *95*, 13363–13383.

Przybylski, D., & Rost, B. (2004). Improving fold recognition without folds. *Journal of Molecular Biology*, *341*, 255–269.

Ptitsyn, O. B. (1995). Molten Globule and Protein Folding. *Advances in Protein Chemistry*, *47*, 83–229.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., … Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Research*, *40*, D290-301.

Punta, M., Forrest, L. R., Bigelow, H., Kernytsky, A., Liu, J., & Rost, B. (2007). Membrane protein prediction methods. *Methods (San Diego, Calif.)*, *41*, 460–474.

Rahman, A. (1964). Correlation in the Motion of Atoms in Liquid Argon. *Phys. Rev.*, *136*, 405–411.

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

Ramachandran, G. N., & Sasisekharan, V. (1968). Conformation of Polypeptides and Proteins. *Advances in Protein Chemistry*, *23*, 283–437.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., … Baker, D. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins*, *77 Suppl 9*, 89–99.

Ramelot, T. a, Raman, S., Kuzin, A. P., Xiao, R., Ma, L.-C., Acton, T. B., … Kennedy, M. a. (2009). Improving NMR protein structure quality by Rosetta refinement: a molecular replacement study. *Proteins*, *75*, 147–67.

Ramirez-Alvarado, M., Kelly, J. W., & Dobson, C. M. (2010). *Protein Misfolding Diseases: Current and Emerging Principles and Therapies*. *Protein Misfolding Diseases: Current and Emerging Principles and Therapies*. John Wiley and Sons.

Regan, L. (2003). Molten globules move into action. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 3553–4.

Rehm, T., Huber, R., & Holak, T. A. (2002). Application of NMR in structural proteomics: Screening for proteins amenable to structural analysis. *Structure*, *10*, 1613–1618.

Reichel, K., Fisette, O., Braun, T., Lange, O. F., Hummer, G., Sch?fer, L. V., & Schäfer, L. V. (2017). Systematic evaluation of CS-Rosetta for membrane protein structure prediction with sparse NOE restraints. *Proteins: Structure, Function, and Bioinformatics*, *85*, 812–826.

Remmert, M., Biegert, A., Hauser, A., & Soding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Meth*, *9*, 173–175.

Renner, C., & Holak, T. A. (2001). NMR 15N relaxation of the insulin-like growth factor (IGF)-binding domain of IGF binding protein-5 (IGFBP-5) determined free in solution and in complex with IGF-II. *European Journal of Biochemistry*, *268*, 1058–1065.

Ripoll, D. R., & Scheraga, H. A. (1988). On the multiple-minima problem in the conformational analysis of polypeptides. II. An electrostatically driven Monte Carlo method--tests on poly(L-alanine). *Biopolymers*, *27*, 1283–1303.

Ripoll, D. R., & Scheraga, H. A. (1989). The multiple-minima problem in the conformational analysis of polypeptides. III. An electrostatically driven Monte Carlo method: tests on enkephalin. *Journal of Protein Chemistry*, *8*, 263–87.

Rohl, C. A., Strauss, C. E. M., Chivian, D., & Baker, D. (2004). Modeling structurally variable regions in homologous proteins with rosetta. *Proteins: Structure, Function and Genetics*, *55*, 656–677.

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

Rondeau, J. M., & Schreuder, H. (2015). Protein Crystallography and Drug Discovery. In *The Practice of Medicinal Chemistry: Fourth Edition* (pp. 511–537).

Rose, G. D. (1979). Hierarchic organization of domains in globular proteins. *Journal of Molecular Biology*, *134*, 447–470.

Rose, G. D., & Creamer, T. P. (1994, May). Protein folding: Predicting predicting. *Proteins: Structure, Function and Genetics*.

Rose, P. W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., … Burley, S. K. (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Research*, *45*, D271.

Rossi, P., Swapna, G. V. T., Huang, Y. J., Aramini, J. M., Anklin, C., Conover, K., … Montelione, G. T. (2010). A microscale protein NMR sample screening pipeline. *Journal of Biomolecular NMR*, *46*, 11–22.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering Design and Selection*, *12*, 85–94.

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, *5*, 725–38.

Rule, G. S., & Hitchens, T. K. (2006). *Fundamentals of Protein NMR Spectroscopy* (Vol. 5). Springer.

Ruoppolo, M., Vinci, F., Klink, T. A., Raines, R. T., & Marino, G. (2000). Contribution of individual disulfide bonds to the oxidative folding of ribonuclease A. *Biochemistry*, *39*, 12033–12042.

Rychlewski, L., Li, W., Jaroszewski, L., & Godzik, A. (2008). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science*, *9*, 232–241.

Saibil, H. R. (2000). Macromolecular structure determination by cryo-electron microscopy. *Acta Crystallographica Section D: Biological Crystallography*, *56*, 1215–1222.

Sali, A., & Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, *234*, 779–815.

Samudrala, R., & Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology*, *275*, 895–916.

Sborgi, L., Verma, A., Piana, S., Lindorff-Larsen, K., Cerminara, M., Santiveri, C. M., … Muñoz, V. (2015). Interaction Networks in Protein Folding via Atomic-Resolution Experiments and Long-Time-Scale Molecular Dynamics Simulations. *Journal of the American Chemical Society*, *137*, 6506–6516.

Scheraga, H. A. (2015). My 65 years in protein chemistry. *Quarterly Reviews of Biophysics*, *48*, 117–177.

Schmid, N., Eichenberger, A. P., Choutko, A., Riniker, S., Winger, M., Mark, A. E., & Van Gunsteren, W. F. (2011). Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *European Biophysics Journal*, *40*, 843–856.

Schrödinger, LLC. (2015). *The {PyMOL} Molecular Graphics System, Version~1.8*.

Schwede, T., Kopp, J., Guex, N., & Peitsch, M. C. (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Research*, *31*, 3381–5.

Seshadri, S., Salma, P., & Chhatbar, C. (2009). Intrinsically Unstructured Proteins: Potential Targets for Drug Discovery. *American Journal of Infectious Diseases, 5*, 133–141.

Shapovalov, M. V., & Dunbrack, R. L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure, 19*, 844–858.

Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., … Wriggers, W. (2010). Atomic-level characterization of the structural dynamics of proteins. *Science (New York, N.Y.), 330*, 341–346.

Shen, M.-Y., & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science, 15*, 2507–24.

Sherwood, D., & Cooper, J. (2011). *Crystals, X-rays and Proteins: Comprehensive Protein Crystallography*. *Crystals, X-rays and Proteins: Comprehensive Protein Crystallography*. doi:10.1093/acprof:oso/9780199559046.001.0001

Shi, J., Blundell, T. L., & Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology, 310*, 243–257.

Shi, S., Pei, J., Sadreyev, R. I., Kinch, L. N., Majumdar, I., Tong, J., … Grishin, N. V. (2009). Analysis of CASP8 targets, predictions and assessment methods. *Database : The Journal of Biological Databases and Curation, 2009*, bap003.

Shi, Y. (2014, November 20). A glimpse of structural biology through X-ray crystallography. *Cell*.

Shirts, M., & Pande, V. S. (2000). COMPUTING: Screen Savers of the World Unite! *Science (New York, N.Y.), 290*, 1903–4.

Shmygelska, A., & Hoos, H. H. (2003). An Improved Ant Colony Optimisation Algorithm for the 2D HP Protein Folding Problem (pp. 400–417). Springer, Berlin, Heidelberg.

Shmygelska, A., & Hoos, H. H. (2005). An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. *BMC Bioinformatics, 6*, 30.

Shortle, D., Simons, K. T., & Baker, D. (1998). Clustering of low-energy conformations near the native structures of small proteins. *Proceedings of the National Academy of Sciences of the United States of America, 95*, 11158–11162.

Shrestha, R., & Zhang, K. Y. J. (2014). Improving fragment quality for de novo structure prediction. *Proteins: Structure, Function, and Bioinformatics, 82*, 2240–2252.

Shuker, S. B., Hajduk, P. J., Meadows, R. P., & Fesik, S. W. (1996). Discovering high-affinity ligands for proteins: SAR by NMR. *Science, 274*, 1531–1534.

Sibanda, B. L., & Thornton, J. M. (1985). β-hairpin families in globular proteins. *Nature, 316*, 170–174.

Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., … Dunker, A. K. (2007). DisProt: The database of disordered proteins. *Nucleic Acids Research*, *35*. doi:10.1093/nar/gkl893

Siew, N., Elofsson, A., Rychlewski, L., & Fischer, D. (2000). MaxSub: An automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, *16*, 776–85.

Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., … Orengo, C. A. (2015). CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, *43*, D376–D381.

Simmerling, C., Strockbine, B., & Roitberg, A. E. (2002). All-atom structure prediction and folding simulations of a stable protein. *Journal of the American Chemical Society*, *124*, 11258–11259.

Simoncini, D., Berenger, F., Shrestha, R., & Zhang, K. Y. J. (2012). A probabilistic fragment-based protein structure prediction algorithm. *PloS One*, *7*, e38799.

Simoncini, D., Schiex, T., & Zhang, K. Y. J. (2017). Balancing exploration and exploitation in population-based sampling improves fragment-based de novo protein structure prediction. *Proteins: Structure, Function and Bioinformatics*, *85*, 852–858.

Simoncini, D., & Zhang, K. Y. J. (2013). Efficient Sampling in Fragment-Based Protein Structure Prediction Using an Estimation of Distribution Algorithm. *PLOS ONE*, *8*, 1–10.

Simons, K. T., Bonneau, R., Ruczinski, I., & Baker, D. (1999). Ab Initio Protein Structure Prediction of CASP III Targets Using ROSETTA. *Proteins: Structure, Function and Genetics*, *37*, 171–176.

Simons, K. T., Kooperberg, C., Huang, E., & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, *268*, 209–25.

Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C., & Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, *34*, 82–95.

Sippl, M. J., & Flöckner, H. (1996). Threading thrills and threats. *Structure*, *4*, 15–19.

Sircar, A., Chaudhury, S., Kilambi, K. P., Berrondo, M., & Gray, J. J. (2010). A generalized approach to sampling backbone conformations with RosettaDock for CAPRI rounds 13-19. *Proteins: Structure, Function and Bioinformatics*, *78*, 3115–3123.

Skolnick, J. (2006, April). In quest of an empirical potential for protein structure prediction. *Current Opinion in Structural Biology*.

Skolnick, J., Kihara, D., & Zhang, Y. (2004). Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins*, *56*, 502–518.

Skwark, M. J., Raimondi, D., Michel, M., & Elofsson, A. (2014). Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLoS Computational Biology*, *10*, e1003889.

Sloan, C. A., Chan, E. T., Davidson, J. M., Malladi, V. S., Strattan, J. S., Hitz, B. C., … Cherry, J. M. (2016). ENCODE data at the ENCODE portal. *Nucleic Acids Research*, *44*, D726–D732.

Smyth, M. S., & Martin, J. H. (2000a). x ray crystallography. *Molecular Pathology : MP*, *53*, 8–14.

Smyth, M. S., & Martin, J. H. J. (2000b). x Ray crystallography. *Journal of Clinical Pathology - Molecular Pathology*.

Snow, C. D., Nguyen, H., Pande, V. S., & Gruebele, M. (2002). Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, *420*, 102–106.

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, *21*, 951–960.

Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, *33*, W244–W248.

Song, Y., Dimaio, F., Wang, R. Y. R., Kim, D. E., Miles, C., Brunette, T., … Baker, D. (2013). High-resolution comparative modeling with RosettaCM. *Structure*, *21*, 1735–1742.

Song, Y., Tyka, M. D., Leaver-fay, A., Thompson, J., Baker, D., Simons, K. T., … Jordan, M. I. (2005). Toward High-Resolution de Novo Structure Prediction for Small Proteins. *Science*, *309*, 1868–1871.

Sorin, E. J., & Pande, V. S. (2005). Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophysical Journal*, *88*, 2472–2493.

Srinivasan, R., Fleming, P. J., & Rose, G. D. (2004). Ab initio protein folding using LINUS. *Methods in Enzymology*, *383*, 48–66.

Srinivasan, R., & Rose, G. D. (1995). LINUS: A hierarchic procedure to predict the fold of a protein. *Proteins: Structure, Function, and Bioinformatics*, *22*, 81–99.

Srinivasan, R., & Rose, G. D. (2002). Ab initio prediction of protein structure using LINUS. *Proteins*, *47*, 489–95.

Staunton, D., Owen, J., & Campbell, I. D. (2003). NMR and Structural Genomics. *Accounts of Chemical Research*, *36*, 207–214.

Stillinger, F. H., & Rahman, A. (1974). Improved simulation of liquid water by molecular dynamics. *The Journal of Chemical Physics*, *60*, 1545–1557.

Su, X. D., Zhang, H., Terwilliger, T. C., Liljas, A., Xiao, J., & Dong, Y. (2015, January 2). Protein crystallography from the perspective of technology developments. *Crystallography Reviews*.

Subramani, A., Wei, Y., & Floudas, C. A. (2012). ASTRO-FOLD 2.0: An enhanced framework for protein structure prediction. *AIChE Journal*, *58*, 1619–1637.

Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Science : A Publication of the Protein Society*, *2*, 762–785.

Tai, C. H., Bai, H., Taylor, T. J., & Lee, B. (2014). Assessment of template-free modeling in CASP10 and ROLL. *Proteins: Structure, Function and Bioinformatics*, *82*, 57–83.

Tanaka, S., & Scheraga, H. A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, *9*, 945–50.

Taylor, T. J., Tai, C.-H., Huang, Y. J., Block, J., Bai, H., Kryshtafovych, A., … Lee, B. (2014). Definition and classification of evaluation units for CASP10. *Proteins: Structure, Function, and Bioinformatics*, *82*, 14–25.

Teichert, F., Minning, J., Bastolla, U., & Porto, M. (2010). High quality protein sequence alignment by combining structural profile prediction and profile alignment using SABERTOOTH. *BMC Bioinformatics*, *11*, 251.

Teso, S., Di Risio, C., Passerini, A., & Battiti, R. (2010). An On/Off Lattice Approach to Protein Structure Prediction from Contact Maps (pp. 368–379). Springer, Berlin, Heidelberg.

Thachuk, C., Shmygelska, A., & Hoos, H. H. (2007). A replica exchange Monte Carlo algorithm for protein folding in the HP model. *BMC Bioinformatics*, *8*, 342.

Thomas, P. J., Qu, B. H., & Pedersen, P. L. (1995). Defective protein folding as a basis of human disease. *Trends in Biochemical Sciences*.

Tian, L., Wu, A., Cao, Y., Dong, X., Hu, Y., & Jiang, T. (2011). NCACO-score: An effective main-chain dependent scoring function for structure modeling. *BMC Bioinformatics*, *12*, 208.

Tramontano, A., & Morea, V. (2003). Assessment of Homology-Based Predictions in CASP5. In *Proteins: Structure, Function and Genetics* (Vol. 53, pp. 352–368).

Tress, M. L., Ezkurdia, I., & Richardson, J. S. (2009, January 1). Target domain definition and classification in CASP8. *Proteins: Structure, Function and Bioinformatics*. Wiley Subscription Services, Inc., A Wiley Company.

Trevizani, R., Custódio, F. L., Dos Santos, K. B., & Dardenne, L. E. (2017). Critical Features of Fragment Libraries for Protein Structure Prediction. *PloS One*, *12*, e0170131.

Tsirigos, K. D., Peters, C., Shu, N., Käll, L., & Elofsson, A. (2015). The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Research*, *43*, W401–W407.

Tu, B. P., & Weissman, J. S. (2004). Oxidative protein folding in eukaryotes: Mechanisms and consequences. *Journal of Cell Biology*.

Tusnády, G. E., & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics*, *17*, 849–850.

Tyka, M. D., Keedy, D. A., André, I., Dimaio, F., Song, Y., Richardson, D. C., … Baker, D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. *Journal of Molecular Biology*, *405*, 607–618.

Unger, R., & Moult, J. (1993a). Finding the lowest free energy conformation of a protein is an NP-hard problem: Proof and implications. *Bulletin of Mathematical Biology*, *55*, 1183–1198.

Unger, R., & Moult, J. (1993b). Genetic Algorithms for Protein Folding Simulations. *Journal of Molecular Biology*, *231*, 75–81.

Unwin, P. N. T., & Henderson, R. (1975). Molecular structure determination by electron microscopy of unstained crystalline specimens. *Journal of Molecular Biology*, *94*, 425–440.

Uziela, K., & Wallner, B. (2016). ProQ2: Estimation of model accuracy implemented in Rosetta. *Bioinformatics*, *32*, 1411–1413.

Vanhee, P., Stricher, F., Baeten, L., Verschueren, E., Lenaerts, T., Serrano, L., … Schymkowitz, J. (2009). Protein-Peptide Interactions Adopt the Same Structural Motifs as Monomeric Protein Folds. *Structure*, *17*, 1128–1136.

Vanhee, P., Verschueren, E., Baeten, L., Stricher, F., Serrano, L., Rousseau, F., & Schymkowitz, J. (2011). BriX: a database of protein building blocks for structural analysis, modeling and design. *Nucleic Acids Research*, *39*, 435–442.

Vanommeslaeghe, K., & Mackerell, A. D. (2015, May). CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochimica et Biophysica Acta - General Subjects*.

Venko, K., Choudhury, A. R., & Novič, M. (2017). Computational Approaches for Revealing the Structure of Membrane Transporters: Case Study on Bilitranslocase. *Computational and Structural Biotechnology Journal*, *15*, 232–242.

Verschueren, E., Vanhee, P., van der Sloot, A. M., Serrano, L., Rousseau, F., & Schymkowitz, J. (2011). Protein design with fragment databases. *Current Opinion in Structural Biology*, *21*, 452–459.

Vitkup, D., Melamud, E., Moult, J., & Sander, C. (2001). Completeness in structural genomics. *Nature Structural Biology*, *8*, 559–565.

Voelz, V. A., Bowman, G. R., Beauchamp, K., & Pande, V. S. (2010). Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *Journal of the American Chemical Society*, *132*, 1526–1528.

Voelz, V. A., & Dill, K. A. (2007). Exploring zipping and assembly as a protein folding principle. *Proteins*, *66*, 877–88.

von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *Journal of Molecular Biology*, *225*, 487–494.

Wahid, H., Ahmad, S., Nor, M. A. M., & Rashid, M. A. (2017). Prestasi kecekapan pengurusan kewangan dan agihan zakat: perbandingan antara majlis agama islam negeri di Malaysia. *Jurnal Ekonomi Malaysia*, *51*, 39–54.

Wang, L. P., McKiernan, K. A., Gomes, J., Beauchamp, K. A., Head-Gordon, T., Rice, J. E., … Pande, V. S. (2017). Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. *Journal of Physical Chemistry B*, *121*, 4023–4039.

Wang, T., Yang, Y., Zhou, Y., & Gong, H. (2016). LRFragLib: an effective algorithm to identify fragments for de novo protein structure prediction. *Bioinformatics*, btw668.

Wang, Z. X. (1996). How many fold types of protein are there in nature? *Proteins: Structure, Function and Genetics*, *26*, 186–191.

Wang, Z. X. (1998). A re-estimation for the total numbers of protein folds and superfamilies. *Protein Engineering*, *11*, 621–626.

Webb, B., & Sali, A. (2014). Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinform/Editoral Board, Andreas D Baxevanis [et Al]*, *47*.

Webb, B., & Sali, A. (2016). Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics*, *2016*, 5.6.1-5.6.37.

White, S. H. (2009). Biophysical dissection of membrane proteins. *Nature*.

Wollacott, A. M., Zanghellini, A., Murphy, P., & Baker, D. (2007). Prediction of structures of multidomain proteins from structures of the individual domains. *Protein Science : A Publication of the Protein Society*, *16*, 165–175.

Woolfson, M. M. (1954). The statistical theory of sign relationships. *Acta Crystallographica*, *7*, 61–64.

Wroblewska, L., Jagielska, A., & Skolnick, J. (2008). Development of a Physics-Based Force Field for the Scoring and Refinement of Protein Models. *Biophysical Journal*, *94*, 3227–3240.

Wu, H. (1995). Studies on denaturation of proteins xiii. a theory of denaturation. *Advances in Protein Chemistry*, *46*, 6–26.

Wu, S., Skolnick, J., & Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biology*, *5*, 17.

Wu, S., Szilagyi, A., & Zhang, Y. (2011). Improving protein structure prediction using multiple sequence-based contact predictions. *Structure (London, England : 1993)*, *19*, 1182–91.

Xu, D., & Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, *80*, 1715–35.

Xu, D., & Zhang, Y. (2013). Toward optimal fragment generations for ab initio protein structure assembly. *Proteins*, *81*, 229–39.

Yan, R., Xu, D., Yang, J., Walker, S., & Zhang, Y. (2013). A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction, *3*, 2619.

Yang, J.-Y., Peng, Z.-L., & Chen, X. (2010). Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. *BMC Bioinformatics*, *11 Suppl 1*, S9.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: Protein Structure and Function Prediction. Nat Meth, 12, 7–8.Nat Meth, 12*, 7–8.

Yang, Y., Faraggi, E., Zhao, H., & Zhou, Y. (2011). Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics*, *27*, 2076–2082.

Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K., & Zhou, Y. (2018). Sixty-five years of the long march in protein secondary structure prediction: The final stretch? *Briefings in Bioinformatics*, *19*, 482–494.

Yang, Y., & Zhou, Y. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins: Structure, Function and Genetics*, *72*, 793–803.

Yarov-Yarovoy, V., Schonbrun, J., & Baker, D. (2006). Multipass membrane protein structure prediction using Rosetta. *Proteins*, *62*, 1010–1025.

Young, J. C., Agashe, V. R., Siegers, K., & Hartl, F. U. (2004). Pathways of chaperone-mediated protein folding in the cytosol. *Nature Reviews Molecular Cell Biology*, *5*, 781.

Zagrovic, B., Snow, C. D., Shirts, M. R., & Pande, V. S. (2002). Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *Journal of Molecular Biology*, *323*, 927–937.

Zemla, a. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research*, *31*, 3370–3374.

Zhang, C., & DeLisi, C. (1998). Estimating the number of protein folds. *Journal of Molecular Biology*, *284*, 1301–1305.

Zhang, C., Srinivasan, Y., Arlow, D. H., Fung, J. J., Palmer, D., Zheng, Y., … Kobilka, B. K. (2012). High-resolution crystal structure of human protease-activated receptor 1. *Nature*, *492*, 387–92.

Zhang, C. T. (1997). Relations of the numbers of protein sequences, families and folds. *Protein Engineering, Design and Selection*, *10*, 757–761.

Zhang, J., Wang, Q., Barz, B., He, Z., Kosztin, I., Shang, Y., & Xu, D. (2010). MUFOLD: A new solution for protein 3D structure prediction. *Proteins*, *78*, 1137–52.

Zhang, S., Ye, F., & Yuan, X. (2012). Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM. *Journal of Biomolecular Structure and Dynamics*, *29*, 1138–1146.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, *9*, 40.

Zhang, Y. (2014). Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins*, *82 Suppl 2*, 175–87.

Zhang, Y., Arakaki, A. K., & Skolnick, J. (2005). TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins*, *61 Suppl 7*, 91–8.

Zhang, Y., Kihara, D., & Skolnick, J. (2002). Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins: Structure, Function and Genetics*, *48*, 192–201.

Zhang, Y., Kolinski, A., & Skolnick, J. (2003). TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophysical Journal*, *85*, 1145–64.

Zhang, Y., & Skolnick, J. (2004a). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 7594–9.

Zhang, Y., & Skolnick, J. (2004b). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function and Genetics*, *57*, 702–710.

Zhang, Y., & Skolnick, J. (2004c). SPICKER: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry*, *25*, 865–871.

Zhang, Y., & Skolnick, J. (2005). The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 1029–34.

Zhou, C., Zheng, Y., & Zhou, Y. (2004). Structure prediction of membrane proteins. *Genomics, Proteomics & Bioinformatics / Beijing Genomics Institute*.

Zhou, H., Pandit, S. B., Lee, S. Y., Borreguero, J., Chen, H., Wroblewska, L., & Skolnick, J. (2007). Analysis of TASSER-based CASP7 protein structure prediction results. *Proteins*, *69 Suppl 8*, 90–97.

Zhou, H., & Skolnick, J. (2007). Ab initio protein structure prediction using chunk-TASSER. *Biophysical Journal*, *93*, 1510–1518.

Zhou, H., & Skolnick, J. (2011). GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophysical Journal*, *101*, 2043–2052.

Zhou, H., & Zhou, Y. (2004). Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins: Structure, Function and Genetics*, *55*, 1005–1013.

Zhou, H., & Zhou, Y. (2005). Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins: Structure, Function and Genetics*, *58*, 321–328.

Zhou, R., Maisuradze, G. G., Suñol, D., Todorovski, T., Macias, M. J., Xiao, Y., … Liwo, A. (2014). Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 18243–8.

Zimmerman, M. I., & Bowman, G. R. (2015). FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *Journal of Chemical Theory and Computation*, *11*, 5747–5757.

Zwanzig, R., Szabo, A., & Bagchi, B. (1992). Levinthal's paradox. *Proceedings of the National Academy of Sciences of the United States of America*, *89*, 20–22.

Appendix

Table 1: Results of CASP12's 96 targets/domains.

Rank	Predictors	Number of targets	SUM Zscore (>-2.0) of GDT_TS
1	Zhang-Server	96	98.70
2	QUARK	96	93.50
3	BAKER-ROSETTASERVER	96	92.36
4	GOAL	94	79.60
5	RaptorX	96	67.25
6	ToyPred_email	95	63.85
7	MULTICOM-CONSTRUCT	96	36.39
8	MULTICOM-CLUSTER	96	34.28
9	MULTICOM-NOVEL	96	33.22
10	IntFOLD4	96	26.39
11	Seok-server	96	22.70
12	HHGG	96	19.27
13	HHPred0	96	17.69
14	HHPred1	96	17.58
15	FALCON_TOPO	96	13.47
16	FALCON_TOPOX	96	12.71
17	FFAS-3D	96	12.17
18	RBO_Aleph	93	-14.78
19	tsspred2	96	-14.89
20	Distill	92	-19.93
21	BhageerathH-Plus	96	-20.66
22	chuo-u2	96	-29.76
23	chuo-u-server	96	-29.76
24	YASARA	92	-32.01
25	ZHOU-SPARKS-X	85	-32.37
26	myprotein-me	92	-33.16
27	MUfold1	96	-33.26
28	MUfold2	87	-34.21
29	slbio	91	-34.51
30	Atome2_CBS	85	-46.01
31	FFAS03	80	-48.21
32	FLOUDAS_SERVER	95	-55.07
33	RaptorX-Contact	93	-61.04
34	Pareto-server	94	-65.45
35	PhyreTopoAlpha	96	-68.37
36	Pcons-net	71	-91.64
37	MULTICOM-REFINE	96	-92.27
38	Seok-assembly	45	-94.44
39	GAPF_LNCC_SERVER	91	-102.69
40	M4T-SmotifTF	63	-118.89
41	ACOMPMOD	89	-130.89
42	GOAL_COMPLEX	14	-161.23
43	Seok-naive_assembly	16	-174.63

Table 2: Results of CASP12's 39 FM targets/domains.

Rank	Predictors	Number of targets	SUM Zscore (>-2.0) of GDT_TS
1	BAKER-ROSETTASERVER	39	45.83
2	Zhang-Server	39	45.44
3	QUARK	39	43.82
4	GOAL	37	33.85
5	RaptorX	39	31.48
6	ToyPred_email	38	27.23
7	MULTICOM-NOVEL	39	8.31
8	RaptorX-Contact	38	5.45
9	MULTICOM-CONSTRUCT	39	5.44
10	Seok-server	39	3.35
11	MULTICOM-CLUSTER	39	3.06
12	FFAS-3D	39	1.20
13	FALCON_TOPO	39	0.24
14	FALCON_TOPOX	39	-0.69
15	RBO_Aleph	36	-2.11
16	IntFOLD4	39	-3.83
17	chuo-u-server	39	-4.50
18	chuo-u2	39	-4.50
19	HHGG	39	-7.47
20	HHPred1	39	-7.83
21	HHPred0	39	-7.83
22	BhageerathH-Plus	39	-8.92
23	MULTICOM-REFINE	39	-9.44
24	PhyreTopoAlpha	39	-10.02
25	MUfold1	39	-11.39
26	Distill	36	-12.83
27	Pareto-server	39	-13.56
28	tsspred2	39	-16.73
29	ZHOU-SPARKS-X	33	-16.78
30	YASARA	36	-17.48
31	GAPF_LNCC_SERVER	36	-19.02
32	myprotein-me	35	-21.05
33	MUfold2	34	-21.33
34	Atome2_CBS	32	-23.15
35	Pcons-net	24	-24.83
36	slbio	34	-29.94
37	FFAS03	27	-38.47
38	FLOUDAS_SERVER	39	-39.02
39	Seok-assembly	19	-39.40
40	ACOMPMOD	36	-42.99
41	M4T-SmotifTF	17	-60.92
42	GOAL_COMPLEX	1	-75.90
43	Seok-naive_assembly	1	-78.00

Table 3: Results of CASP12's 38 TBM targets/domains.

Rank	Predictors	Number of targets	SUM Zscore (>-2.0) of GDT_TS
1	BAKER-ROSETTASERVER	38	32.11
2	Zhang-Server	38	29.90
3	GOAL	38	29.67
4	QUARK	38	28.43
5	ToyPred_email	38	23.45
6	RaptorX	38	23.05
7	HHGG	38	22.00
8	HHPred0	38	20.91
9	HHPred1	38	20.80
10	MULTICOM-CLUSTER	38	19.59
11	MULTICOM-CONSTRUCT	38	18.88
12	IntFOLD4	38	17.26
13	MULTICOM-NOVEL	38	13.86
14	Seok-server	38	13.46
15	FALCON_TOPOX	38	7.38
16	FALCON_TOPO	38	7.21
17	FFAS-3D	38	3.79
18	tsspred2	38	3.27
19	BhageerathH-Plus	38	-0.85
20	Distill	37	-1.75
21	YASARA	37	-3.55
22	slbio	38	-3.76
23	FFAS03	36	-3.93
24	MUfold2	36	-4.02
25	MUfold1	38	-4.12
26	myprotein-me	38	-8.10
27	ZHOU-SPARKS-X	33	-10.96
28	Fastro	37	-11.93
29	chuo-u2	38	-12.08
30	chuo-u-server	38	-12.08
31	Atome2_CBS	37	-12.81
32	RBO_Aleph	38	-15.93
33	M4T-SmotifTF	34	-30.19
34	Seok-assembly	18	-34.01
35	Pareto-server	36	-36.69
36	PhyreTopoAlpha	38	-44.28
37	GOAL_COMPLEX	11	-51.57
38	Pcons-net	33	-56.05
39	RaptorX-Contact	37	-57.37
40	Seok-naive_assembly	13	-60.48
41	ACOMPMOD	35	-61.70
42	MULTICOM-REFINE	38	-67.02
43	GAPF_LNCC_SERVER	36	-67.71

Table 4: Results of CASP12's 19 FM/TBM targets/domains.

Rank	Predictors	Number of targets	SUM Zscore (>-2.0) of GDT_TS
1	Zhang-Server	19	23.36
2	QUARK	19	21.24
3	GOAL	19	16.07
4	BAKER-ROSETTASERVER	19	14.42
5	ToyPred_email	19	13.17
6	IntFOLD4	19	12.96
7	RaptorX	19	12.72
8	MULTICOM-CONSTRUCT	19	12.07
9	MULTICOM-CLUSTER	19	11.63
10	MULTICOM-NOVEL	19	11.05
11	FFAS-3D	19	7.18
12	FALCON_TOPO	19	6.03
13	FALCON_TOPOX	19	6.01
14	Seok-server	19	5.89
15	HHGG	19	4.74
16	HHPred0	19	4.61
17	HHPred1	19	4.61
18	RBO_Aleph	19	3.27
19	slbio	19	-0.81
20	tsspred2	19	-1.44
21	myprotein-me	19	-4.01
22	FLOUDAS_SERVER	19	-4.12
23	ZHOU-SPARKS-X	19	-4.64
24	Distill	19	-5.34
25	FFAS03	17	-5.81
26	MUfold2	17	-8.85
27	RaptorX-Contact	18	-9.12
28	Atome2_CBS	16	-10.05
29	Pcons-net	14	-10.76
30	BhageerathH-Plus	19	-10.89
31	YASARA	19	-10.97
32	chuo-u-server	19	-13.18
33	chuo-u2	19	-13.18
34	PhyreTopoAlpha	19	-14.06
35	Pareto-server	19	-15.21
36	MULTICOM-REFINE	19	-15.81
37	GAPF_LNCC_SERVER	19	-15.96
38	MUfold1	19	-17.75
39	Seok-assembly	8	-21.03
40	ACOMPMOD	18	-26.21
41	M4T-SmotifTF	12	-27.78
42	GOAL_COMPLEX	2	-33.76
43	Seok-naive_assembly	2	-36.15