Winter 12-14-2015

nD – PDPA: n Dimensional Probability Density Profile Analysis

Arjang Fahim
University of South Carolina

$nD - PDPA$: n Dimensional Probability Density Profile Analysis

by

Arjang Fahim

Bachelor of Science
Azad University of Tehran, 1998
Master of Science
University of South Carolina, 2014

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Computer Science

College of Engineering and Computing

University of South Carolina

2015

Accepted by:

Homayoun Valafar, Major Professor

John Rose, Committee Member

Marco Valtorta, Committee Member

Gabriel Terejanu, Committee Member

Mirko Hennig, External Committee Member

Lacy Ford, Senior Vice Provost and Dean of Graduate Studies

Dedication

This dissertation is dedicated to all the people who have brought so much into my life.

To my wife, without you this journey might not have been successful. Thanks for your emotional and intellectual support. With all my love, I dedicate this dissertation to you.

To my daughter, you have brought more into my life than you can imagine. I have become a better person, because of the reflection of myself I see in your eyes. To the day you will be a fabulous woman, I dedicate this dissertation to you.

To my mother, whom I missed her so much. You gave me the love and confidence to believe I could achieve anything. For always being in my heart, I dedicate this dissertation to you.

Abstract

Proteins are often referred as working molecule of a cell, performing many structural, functional and regulatory processes. Revealing the function of proteins still remains a challenging problem. Advancement in genomics sequence projects produces large protein sequence repository, but due to technical difficulty and cost related to structure determination, the number of identified protein structure is far behind. Novel structures identification are particularly important for a number of reasons: they generate models of similar proteins for comparison; identify evolutionary relationships; further contribute to our understanding of protein function and mechanism; and allow for the fold of other family members to be inferred. Considering the evolutionary mechanisms responsible for the generation of new structures in proteins, it has been speculated that there may be a limited number of unique protein folds as few as ten thousand families. Currently, the Protein Data Bank consists of nearly 113,000 protein structures, but less than 1,500 families are represented, and almost no new fold families have been reported since 2008. Ideally, solved protein structures for new protein families would be used as templates for in silico structure prediction methods, and the results of both solved and predicted structures would in turn be used to infer function. However, such an approach requires new, efficient and cost-effective computational methods for target selection and structure determination. Traditional characterization of a protein structure by NMR spectroscopy is expensive and time consuming regardless of the structural novelty of the target protein. In an effort to expand the applicability of NMR spectroscopy, the community is continually focused on the development of new and economical approaches that

enable the study of more challenging, or structurally novel proteins. While many advances have been made in this regard, very little attention has been made on reducing the cost of structural characterization of routine proteins.

Probability Density Profile Analysis (PDPA) has been previously introduced to directly addresses the economies of structure determination of routine proteins and subsequently, identification of novel structures from minimal sets of NMR data. The latest version of PDPA (2D-PDPA) has been successful in identifying the structural homologue of an unknown protein within a library of 1000 decoy structures. In order to further expand the selectivity and sensitivity of PDPA, incorporation of additional data is necessary. However, current PDPA approach is limited by its computational requirements, and its expansion to include additional data will render it computationally infeasible. Here we propose a new method and developments that eliminate PDPA's computational limitations and allow inclusion of Residual Dipolar Coupling (RDC) data from multiple vector types in multiple alignment media. Additionally nD-PDPA will be used to refine an unknown protein to obtain closer structure to the native in terms of bb-rmsd.

Table of Contents

List of Tables

List of Figures

Chapter 1

Proteins: The building block of life

1.1 Fundamentals of Protein Structures

Almost all biological processes involve the interaction of one or more proteins. These large molecules exhibit a remarkable versatility that allow them to perform a myriad of crucial activities and functions. Structure of proteins is not separate from their functionality. It has been shown that there is a relation between the protein structure and its functionality. Many reactions in biological systems are conducted by proteins, producing a sophisticated chemical reaction that an organism needs for its life. Moreover, proteins have the responsibility of transporting chemicals and regulating functions in organisms. Proteins are polymers constructed from a set of 20 *amino acids*. Polymers of amino acids are called *polypeptide*.A protein may consist of one or more polypeptides chain that are folded into a specific three-dimensional shape [50] [71].

1.1.1 The Primary Structure of Proteins: Sequence of Amino Acids

A protein is a linear combination of amino acids and this combination ultimately defines its three-dimensional shape. The sequence of amino acids is often called the *primary structure*. Amino acids are organic molecules consisting of both carboxyl(COOH) and amino groups(NH_2). Figure 1.1 illustrates the general chemical

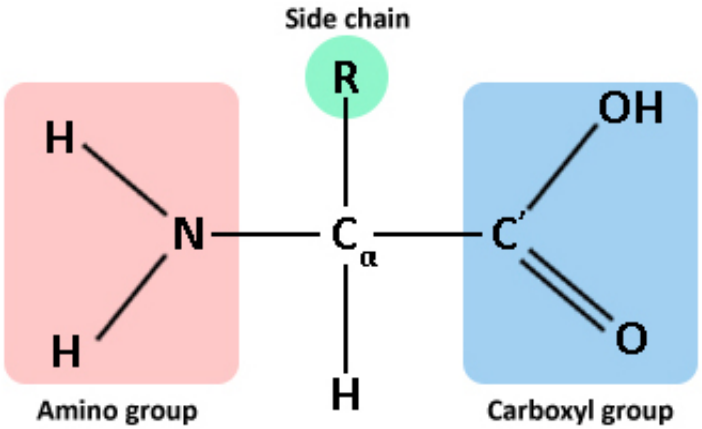


Figure 1.1: The Structure of a prototypical amino acid. The chemical groups bound to the central $\alpha - carbon$, are highlighted in the background. The R-group represents any of the possible 20 amino acid side chains.

formula of an amino acid. At the center of an amino acid, there is a carbon atom called $\alpha - carbon$. Surrounding the $\alpha - carbon$ are an amine group, carboxyl group, hydrogen atom and a variable group symbolized by the letter R. The R group is called the side chain and is different for every amino acid. There are nearly 20 amino acids that can be incorporated into a protein sequence. The resulting protein can use any number of 20 amino acids, in any order. Physical and chemical properties of the side chain determine the characteristic of an amino acid such as *hydrophobicity*, *hydrophilicity*, and *polarity* (Figure 1.2).

1.1.2 The Peptide Bond

When two amino acids are positioned in such a way that the carboxyl group of the first amino acid links with the amine group of the other, the result is a dehydration reaction where a water molecule is formed and removed from the reaction and the two amino acids come together to form a covalent bond called a *peptide bond* (Figure 1.3).A polypeptide is synthesized by linear formation of peptide bonds between two or more amino acids. An amino acid within a polypeptide chain may also be referred

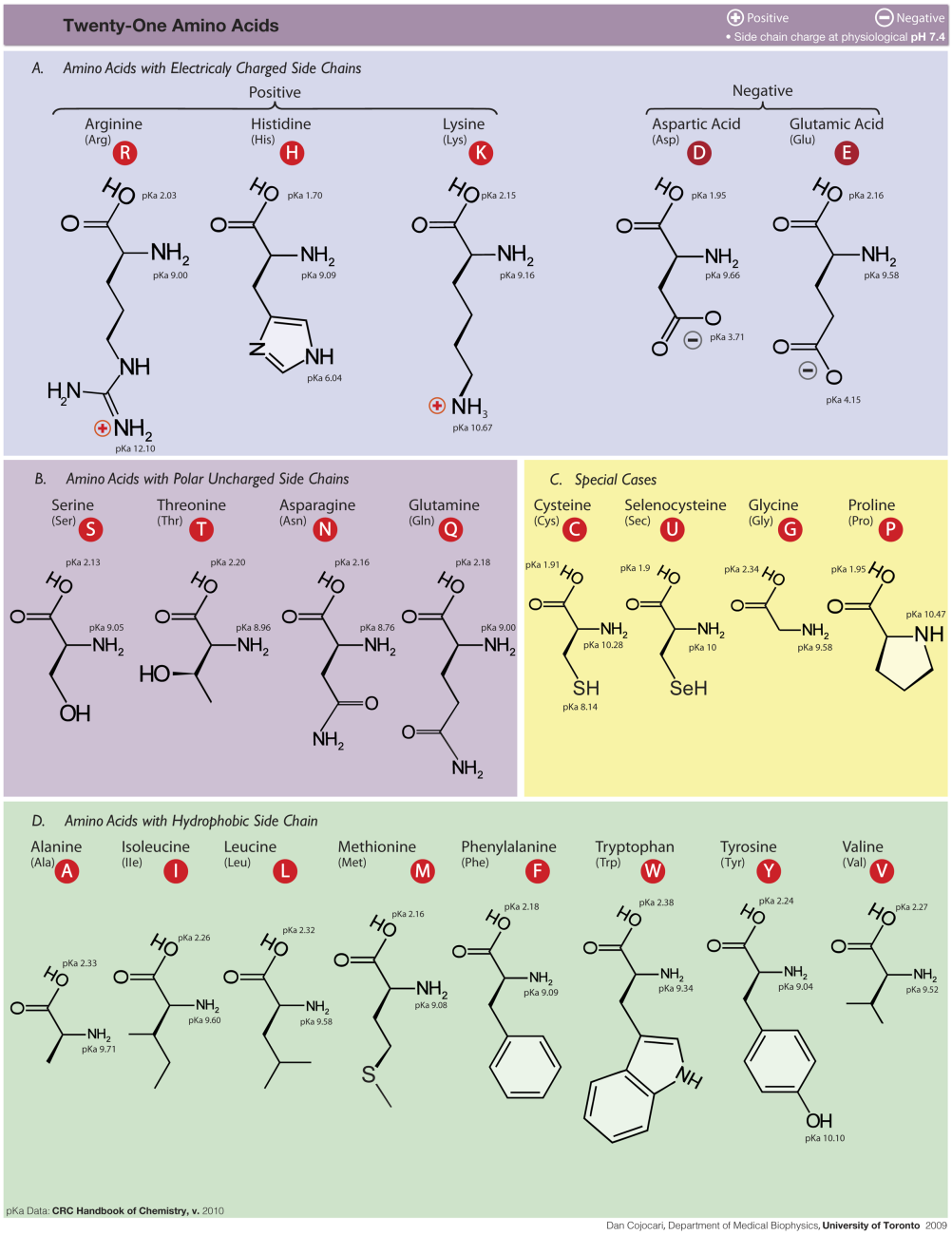


Figure 1.2: The twenty amino acids used in proteins. Each amino acid is labeled by its full name followed by three letters and one letter (in the red circle) abbreviations. Amino acids are grouped into negative or positive charges, hydrophobic or hydrophilic side chains [71].

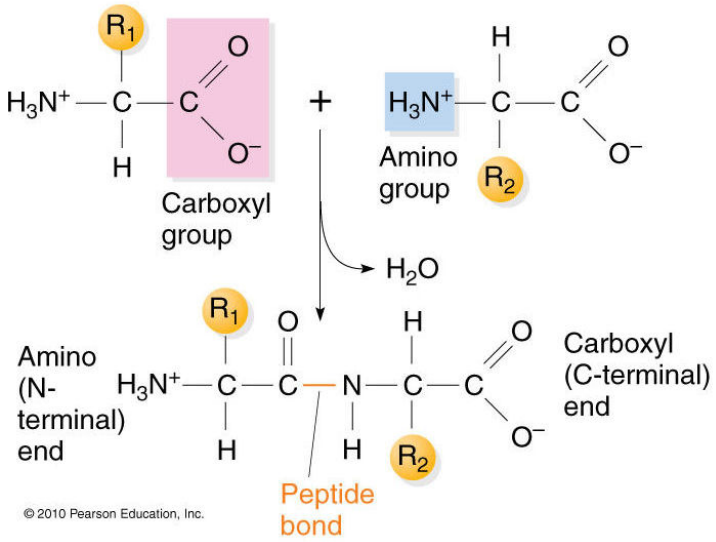


Figure 1.3: Peptide bond formation between successive amino acids. Amine group ends on the second (R_2) amino acid is added to the carboxyl end of the first (R_1) amino acids. The amino acid terminus of R_1 amino acid remains unchanged, end of polypeptide grows in the N to C direction. The repeating $N - C(R) - C$ subunit remaining after the dehydration is an amino acid residue [71].

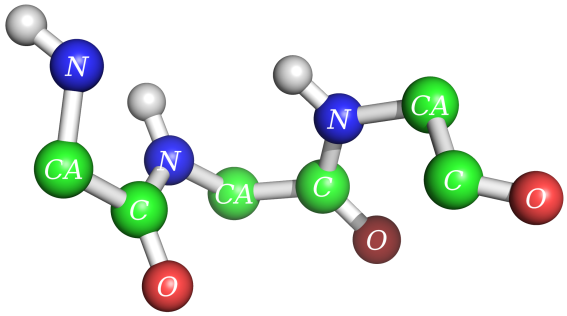


Figure 1.4: The backbone atoms of two joined amino acids.The green spheres denote the carbon atoms and the blue spheres are nitrogen atoms.

to as a *residue* and atoms on the peptide bonds along with the $\alpha - carbon$ atoms to which R-group are attached, are referred to as the *peptide backbone* (Figure 1.4).

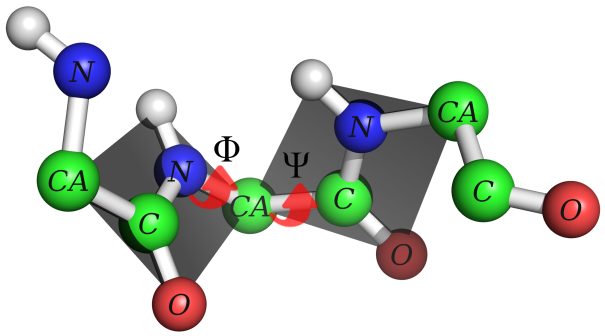


Figure 1.5: The planar characteristic of the peptide bond, and the rotation of the peptide backbone about the C_α atom. The two planar peptide bonds about the central $\alpha - carbon$, shown here as a ball-and-stick model. Rotation is only possible around ϕ and ψ angles.

1.1.3 Ramachandran plot

The peptide bonds have important effects on the three-dimensional structure of a protein. These bonds are labeled as ϕ (Phi) and ψ (Psi) angles (Figure 1.5). The peptide bonds give a polypeptide limited freedom to rotate only about the $\alpha - carbon$ bond. The limitation of the rotation of the $\phi(N - C_\alpha)$ and ψ $(C_\alpha - C)$ angles are due to steric hindrance between the side chain of the residue and the peptide backbone. A *Ramachandran Plot* (a plot of ϕ vs ψ angles) maps the entire allowed and disallowed conformational space of an amino acid. These restrictions were developed by G.N Ramachandran in the late 1960s based on studies of sterically allowed ϕ and ψ torsion angles (Figure 1.6). An amino acid with the simple structure in the side chain (e.g Glycine with a single hydrogen(Figure 1.2)) demonstrates less steric hindrance of ϕ and ψ which leads to expanding the conformational space. On the other hand, Proline (with a cyclic R group (Figure 1.2)) demonstrates less freedom of steric hindrance due to its cyclic structure (Figure 1.7).

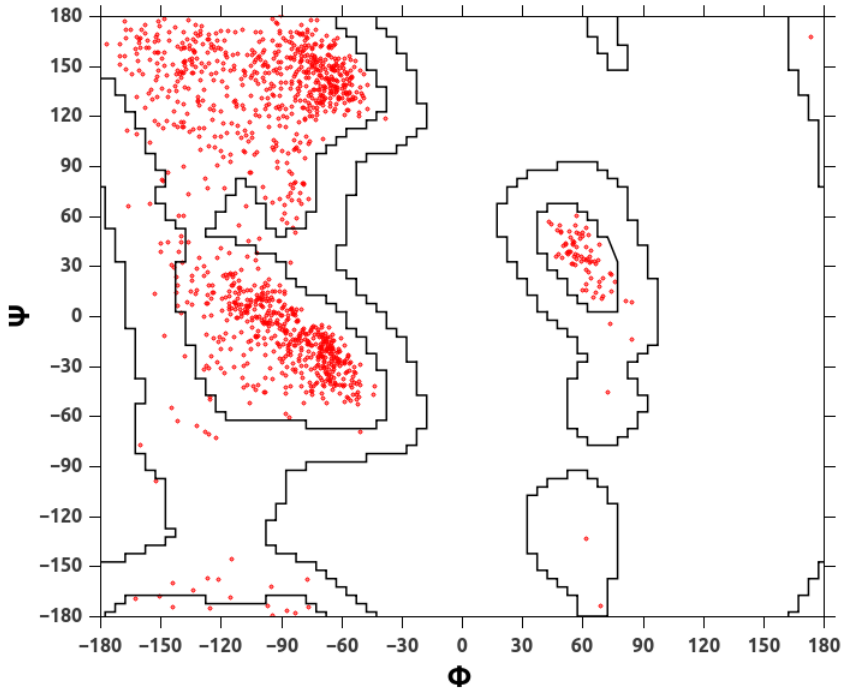


Figure 1.6: A schematic representation of a ramachandran plot (a plot of ϕ and ψ angles). The closed regions denotes valid regions for ϕ and ψ angles. The red dots are ϕ and ψ angles extracted from database of 50 structures. Data was taken from Richardson Lab (http://kinemage.biochem.duke.edu)

1.1.4 The secondary structure of a proteins: Local Three dimensional structures

The stability that is introduced by hydrogen bonds leads to locally stabilized conformations that are known as *Secondary Structures*. The secondary structure consists of polypeptide chains that repeatedly coils or folds into a pattern that contributes to a protein's overall conformation. The two types of secondary structure that are dominant in protein conformation are α-*helix* and β-*sheets*.

1.1.5 α-Helices

In the α-*helix* conformation, the H atom of residue i forms a hydrogen bond to the carbonyl O of residue $i + 4$ (Figure 1.8). Another symmetrical relationship between

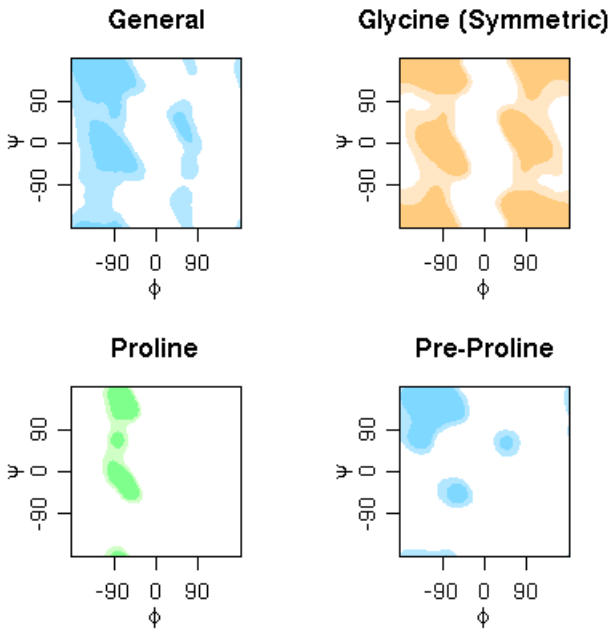


Figure 1.7: Comparison of Ramachandran plots for Proline and Glycine amino acids. The smaller side-chain of Glycine demonstrates the larger valid region ϕ and ψ in contrast to Proline and Pre-Proline.Generally the larger side-chain restricts backbone movements.

residues in an $\alpha - helix$ is the geometrical relationship. In an $\alpha - helix$ any residue $i + 1$ rotates approximately 100° rotation relative to the residue i around the helix axis. $\alpha - helices$ in protein, almost without exception, are right handed. If the chain is compressed more tightly than in an $\alpha - helix$, an alternative hydrogen-bonded structure can form, called the 3_{10} helix, where $N - H$ of residue i is hydrogen bonded to the carbonyl O of residue $i+3$. If the chain winds up less tightly than the $\alpha - helix$, it can form a $\pi - helices$, in which the $N - H$ of residue i is hydrogen bonded to the carbonyl O of residue $i+5$.

1.1.6 β-SHEETS

A $\beta - sheet$ is formed from two separate strands, which may arise from regions distant in the sequence. This arrangement produces a sheet that is pleated with the

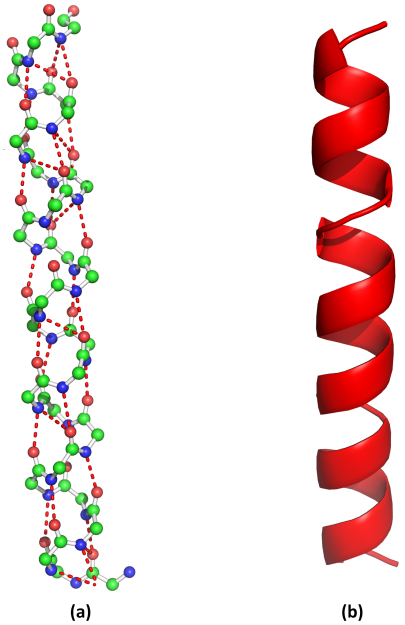


Figure 1.8: (a) Atomic formation of an $\alpha - helix$, the red dashed lines represent the hydrogen bonds that form the helix shape.(b) The cartoon view of the same helix.

residue side chains alternating position on the opposite sides of the sheet (Figure 1.9). The two possible arrangements for $\beta - sheets$ are parallel and anti-parallel. In parallel sheets, the strands are arranged in the same direction with respect to the amine terminal (N) and carboxyl terminal (C) ends. However, in the anti-parallel arrangement, the strands alternate the amino acid and carboxyl terminal ends in such a way that a given strand interacts with a strand in the opposite direction.

1.1.7 The tertiary structure of proteins: Global three-dimensional structure.

Tertiary structure of a protein is defined as the global three-dimensional structure of its polypeptide chain (Figure 1.10). Tertiary structure of a protein is the result of interaction between the side chains (R group) of the various amino acids. Consequently, the side chains in the tertiary structures of a protein play an important role in creating the final structure. In contrast, the backbone interaction is primary

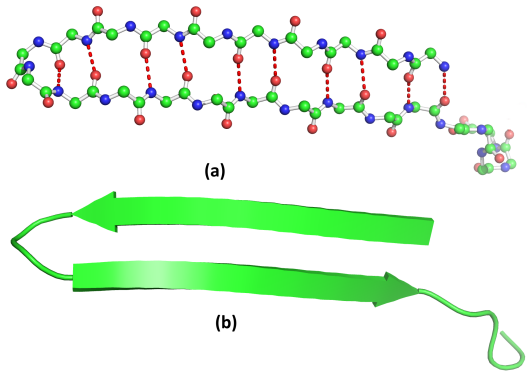


Figure 1.9: (a) Atomic formation of a $\beta - sheet$, red dashed lines represent the hydrogen bonds that form $\beta - sheet$.(b) The cartoon view of the same $\beta - sheet$.

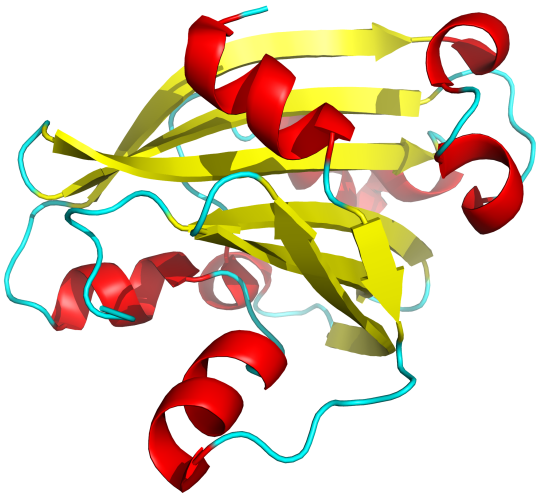


Figure 1.10: Tertiary representation of the protein 1G1B(164).

responsible for the generation of the secondary structure ($\alpha - helix$ and $\beta - sheet$).

1.1.8 Protein folding

The process of transferring the linear polypeptide chain to a three-dimensional structure is referred as protein folding. Protein folding is a complex process that is not completely understood yet. The major challenge is, how are the physiochemical

properties (such as hydrgenbonds, van der Waals interactions, backbone torsion angles preferences and etc) of linear set of amino acids translate to the three-dimensional native conformation of a protein and what forces drive amino acids chain into a folded structure [32]. The physical forces are described by *forcefields*. The forcefields utilize internal potential energy of proteins for computer simulations. Although computer aided simulation methods such as MD (Molecular Dynamic) Simulation are successful to address the folding problem, but so far, such a modeling succeeds on small and simple protein folds [67]. More complex protein structures require more computational power and speed, which is still out of reach of our computational capabilities. Most proteins probably go through many intermediate states to reach to the final stable folded stage; and looking at the mature conformation does not reveal the stage of the folding required to achieve the final conformation.

1.1.9 The Quaternary structure of proteins

Quaternary structure of a protein is the aggregation of two or more folded polypeptides into its functional macro-molecule. These proteins are also referred as multi-subunits. The subunits may be identical (homomeric) proteins or they can be constituted of different proteins subunits (hetromeric)(Figure 1.11).

1.2 Classification of Protein Structures

Protein structures can be categorized based on the similarity in sequence, topology or even in observable structural details. Protein sequence and topology similarity are two main features that are utilized in protein classification. As the number of characterized protein structures grew, the classification of these proteins became more difficult [64] [29] [30]. A protein fold family is a group of proteins that share common evolutionary origin, reflected by their related functions and similarities in sequence

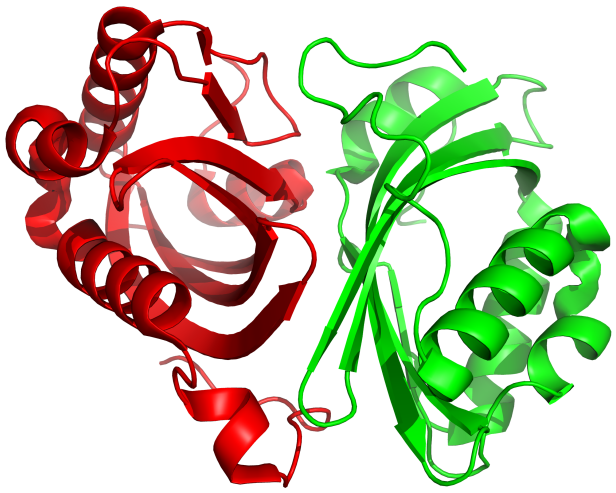


Figure 1.11: Homomeric quaternary representation of the protein 1NWW(149).

or structure. When a novel protein is identified, its functional properties can be potentially predicted based on the group to which it belongs. It is worthy to note that the classification of a novel protein solely based on the sequence may not always lead to a correct classification if three-dimensional conformation of the structure is not considered [50]. In the classification of protein, it is also important to study the biological evolutionary of the structure. The terms *super-family* (describing a large group of distantly related proteins) and *sub-family* (describing a small group of closely related proteins) are sometimes used in this context.

1.2.1 Comparison of proteins using sequences and structures

A common method to identify the similarities of proteins is comparing protein sequences. If the sequence of amino acids of two proteins aligns, then either the identical residues can be counted or a more subtle measure based on the index of

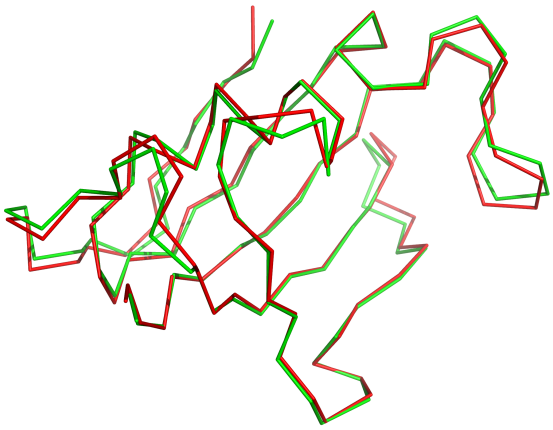


Figure 1.12: Superposition of human and the yeast FK506-binding proteins 2FKE(107)(red) and 1YAT(113)(green) the backbone RMSD score for the backbone atoms after alignment is 0.887 Å. These proteins have very similar structures.

similarity between amino acids can be used. The similarity between two sequences is then the sum of the value of the indices of similarity for each pair of aligned amino acids plus a correction to address for insertion and deletion of amino acids [51] [33].

Given the structures of two proteins, it is possible to superimpose the three-dimensional structures using computer tools to observe the similarities and differences of the structures. A commonly used mathematical measure of the difference between two structures is rmsd (root-mean-square deviation) in atomic position of the back bone atoms after optimal super position (Figure 1.12).

1.2.2 Classification of protein topologies

Classification based on the topology was first proposed by M. Levitt and C. Chothia [63]. This classification is based on the secondary and tertiary structures of *domains*. A domain is a distinct functional and structural unit of a protein. Usu-

ally domains are responsible for a particular function or interaction, contributing to the overall role of a protein. Classification of proteins based on the similarity of domain creates a very broad range of groups, protein structures sort themselves into distinct categories with noticeable different folding patterns. Within the sets of classification using topology there are families that share enough features to suggest evolutionary relationship. There are numerous databases available for classification of proteins. Protein Data Bank (PDB) [15] [14] contains more than 113672 protein structures and their toplogy information. CATH(Class, Architecture, Topology, Homology) [64] [29] [30] is a hierarchical domain classification of protein structures in the Protein Data Bank. Protein structures are classified using a combination of automated and manual procedures.

Chapter 2

Protein Structure Determination

2.1 Introduction

Despite the recent advances in various Structural Genomics Projects, a large gap remains between the number of sequenced and structurally characterized proteins. The reasons contribute to this inefficiency include technical difficulties, labor, and the cost related to structure characterization by experimental methods such as NMR spectroscopy. As of June 2014, UniPortKB contains more than 69 million protein sequences were deposited in the UniProtKB database [9](http://uniport.org). However, the number of protein structures in the Protein Data Bank (PDB) [14] [15] (http://www.rcsb.org) is only about 113,000; less than 1% of the protein sequences.

Protein structure determination is essential to understand its function and interaction, for important applications such as drug discovery and design. In principle, protein structure prediction methods can be grouped into two categories, experimental methods and computational methods. In the following two sections, the two major methods of determining protein structures are discussed.

2.2 Experimental Methods

X-ray crystallography and NMR spectroscopy are two methods of choice to determine protein structures experimentally. Based on the report from Protein Data Bank(PDB) (http://www.rcsb.org) [14] [15], about 88.6 percent of protein structures

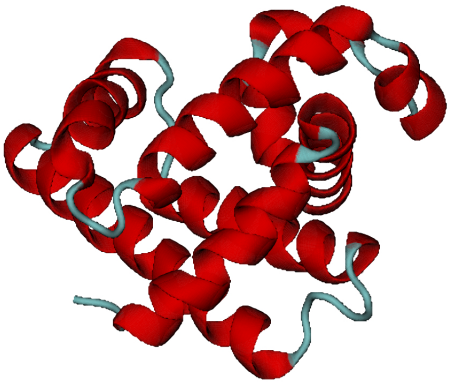


Figure 2.1: Myoglobin (PDBID:1MBN(153)). This protein is very common in muscle cells, and its function is to store Oxygen. The reserved Oxygen is used when muscle tissues are hard at work. It is characterized using X-ray crystallography in 1958.
.

are identified by X-ray crystallography and 10.3 percent by NMR spectroscopy and the rest of proteins are identified by other techniques. One reason for this disproportional contribution is due to the recent introduction of NMR spectroscopy as a routine method for structure determination.

2.2.1 X-ray Crystallography

Structural biology was born in 1958 with the utilization of X-ray technique to characterize the atomic structure of Myoglobin (PDBID:1MBN(153))(Figure 2.1) by John Kendrew [45]. By the early of 1970's, there were many proteins, characterized using the same technique and until now, X-ray Crystallography remained one of the major methods to study protein structures. X-ray Crystallography utilizes X-ray diffraction for a single protein crystal to determine the three-dimensional shape and structure of the molecules(Figure 2.2). The crystalline atoms cause a beam of X-ray to diffract in many directions. By measuring the angles and intensities of diffracted beams, a three-dimensional picture of the electrons density within the crystal can be produced. To crystallize a protein, the purified sample undergoes slow precipitation from an aqueous solution. As a result, individual protein molecules concentrated

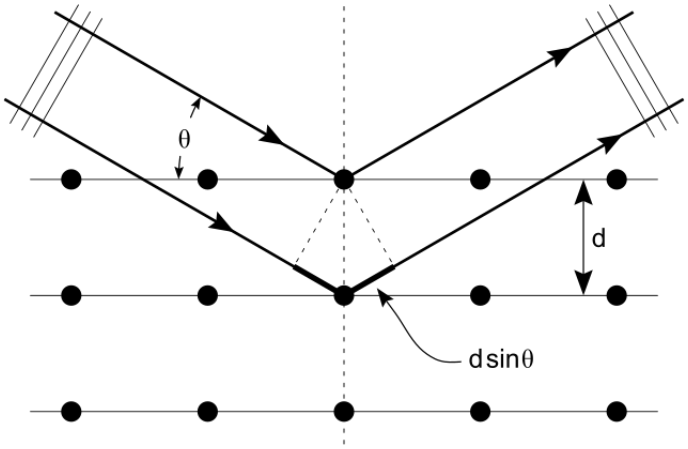


Figure 2.2: The illustration of diffraction of incoming beams when colliding with crystal points.([93])

.

and aligned themselves in a repeating series of "unit cells" by adopting consistent orientation [75]. X-ray Crystallography has several major drawbacks. The process of protein crystallization can be very time consuming since selected sample conditions and environment(such as varying PH, salt concentration, salt type, buffer type) need to be carefully explored for successful crystallization. Crystallization medium may introduce packing forces on the protein, which may alter the structure and internal dynamics of the protein. Additionally the X-ray diffraction phenomena known as radio damage has an effect on the protein structure. To reduce the radio damage, the sample is usually put into a very low-temperature environment. Under this condition, the internal dynamics of the molecule is suppressed. Obtaining results from X-ray Crystallography are relatively fast and require utilization of computer software.

2.2.2 Nuclear Magnetic Resonance

The first de-novo NMR (Nuclear Magnetic Resonance) [92] structure determination was completed in 1984 by Timothy F. Havel and Michael P. Williamson [54](Figure 2.3). Within five years, over 2,000 NMR structures have been deposited into newly established Protein Data Bank (PDB) [96]. NMR has a variety of applications in

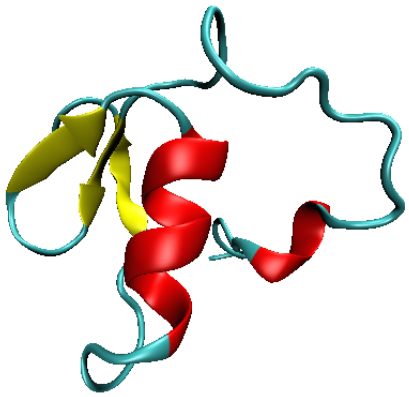


Figure 2.3: Protein BUS2(57) is known as the first de-novo protein characterized by NMR spectroscopy

.

physics, chemistry and biology. Specifically in biology, NMR spectroscopy is utilized to determine the protein structure by analyzing the magnetic properties of different nuclei under electric-magnetic stimuli. NMR provides structural information through measuring geometric restriction of a given structure. These restrictions are the distance between the different pair of atoms (NOE, Nuclear Overhauser Effect), the orientation of inter-nuclear vector (RDC, Residual Dipolar Coupling) or other relaxation properties of nuclei. The NOE is the transfer of nuclear spin polarization from one nuclear spin population to another via cross-relaxation. It is a common phenomenon observed by NMR spectroscopy. NOE provides information related to the inter-atomic distances within a short range. The distance information can be used to determine molecular structure based on the distance constraint. Figure 2.4 demonstrates a sample 2D-NOESY spectrum of the protein Ubiquitin(PDBID:1UBQ(76)). In this spectrum the intense regions refer to the inter-atomic distance between pair of atoms in particular frequencies. The magnitude of the NOE peaks exhibits an r^{-6} dependency with respect to the inter-atomic distance r. Therefore, NOE constraints are considered to provide short range distance information limited to no more than 5 Å. Although the NOE constraints are relatively easy to obtain. However, NOE based

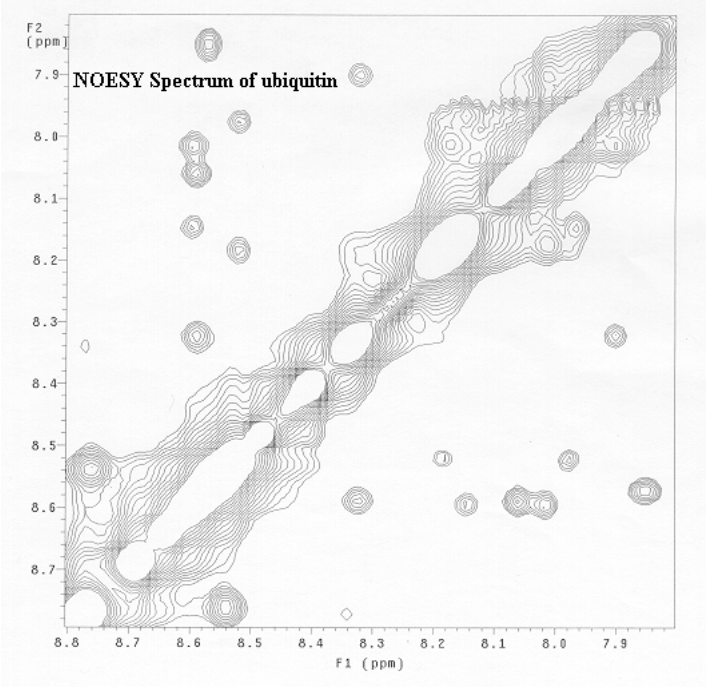


Figure 2.4: NOESY spectrum contour map of ubiquitin.[1]

structure determination is undermined by some significant limitations. For instance, molecular dynamics is hardly reflected by the inter-atomic distance within very short range. Therefore, NOE naturally is insensitive to internal motion. Furthermore, as the protein size grows the identification of NOEs between partners become a difficult task. Finally, the assignment process is time-consuming and error prone. Without assignment, NOE data provide the distance relation between chemical groups rather than residues. Residual Dipolar Coupling is the primary data source in this research. Next chapter provides more detail information about RDC and its application in this study.To fully understand the functionality of NMR experiments, detailed knowledge of subjects from quantum physics to chemistry and mathematics would be required. To avoid complexity of the subject and to maintain focus of our objective in this manuscript, NMR is treated as a *black-box*, providing information and data that is needed for proposed computational methods. A full exposition of the topic can be found in [25]. The main advantage of using NMR spectroscopy is the possibility of

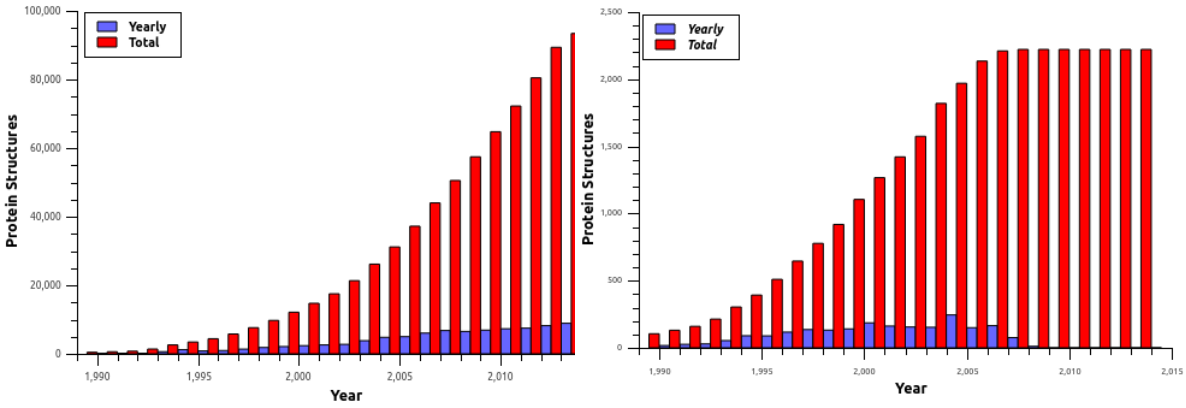
the study of the protein in its native environment. Study of a protein in its actual physiological conditions will provide better functional information while preserving the internal motions. The disadvantage of NMR spectroscopy is isotopic labeling of certain nuclei(such as ^{15}N , ^{13}C) and relatively long data acquisition periods.

2.3 Computational Methods

The core task of the computational methods for structure characterization is the prediction of protein fold (three-dimensional conformation) from a sequence of amino acids. Computational methods that are used routinely fall into three categories: *Template-Based Modeling*, *Homology-Based Modeling* and *De novo* or *Ab-initio* protein modeling.

2.3.1 Template-Based Modeling

If proteins of a similar structure are identified from the PDB library, the target model can be constructed by copying the framework of the solved proteins (templates). The procedure is called "Template Based Modelling (TBM) " [37] [95]. Although high-resolution models often can be generated by TBM, the procedure relies on completion of the protein database. Protein Data Bank contains 113,672 proteins. Several methods are developed to categorize proteins based on structural features and three-dimensional shapes often called folds or fold families[see section 1.2] [64]. Considering the intrinsic physical constraints of a sequence and the evolutionary mechanism responsible to generate a new protein structure, current estimates of the number of folds range from 1,000 to 10,000 depending on the models and approximation applied [41] [49]. Thus far based on CATH [29] [30] or SCOP [66] classifications the growth of unique fold per year indicates no significant change from 2008 to 2014 (Figure 2.5(a)). On the other hand, the yearly growth of novel protein structures



(a) (b)

Figure 2.5: Number of protein structures in PDB (a) unique folds reported by SCOP and (b) cumulative since 1992

shows a healthy trend of growth each year (Figure 2.5(b)). The main contributing factor to this inefficiency is the lack of any accurate method of target selection that is: a structure will be selected for analysis may not yield a novel structure. The current method for selection a structure is based on sequence homology analysis. Although homology method covers the available protein sequence space, it may not be an optimal method to cover protein three-dimensional structural space. Protein Data Bank contains protein structures with similar structures but different sequences. Thus, developing an efficient computer-based algorithm to predict three-dimensional structures from sequences is probably the only avenue to solve the problems.

2.3.2 Homology-Based Modeling

Homology modeling is based on the identification of one or more protein structures that are likely to resemble the structure of the query sequence. Sequencing method is used to align the query sequence against the accumulation of a known protein structure. Structural homologous can be identified using software such as BLAST [24] [19] or PSI-BLAST [6]. Once a protein with sufficient sequence identity has been found, it can be used as a template to predict the native structure or function of a

target protein.

2.3.3 Ab-initio Modeling

If protein templates are not available, the three-dimensional models are built from scratch. This procedure is known as ab initio modeling [52] or de-novo modeling [21]. Typically, ab-initio modeling conducts a conformational search under the guidance of a designed energy function. This procedure usually generates a number of possible conformations (structure decoys), and final models are selected from them. Therefore, a successful ab-initio modeling depends on three factors: (1) an accurate energy function with which the native structure of a protein corresponds to the most thermodynamically stable state, compared to all possible decoy structures; (2) an efficient search method that can quickly identify the low-energy states through conformational search; (3) selection of native-like models from a pool of decoy structures. The CASP [61] meeting has shown that, over the past few years, the most rapid development has been in the ability of ab-initio protein folding techniques to generate a reasonable structure for an arbitrary sequence. This has mostly been done using some form of statistical potential since our understanding of the physics of protein folding has not progressed anywhere near as much.While it is clear that the folding of proteins from first principles remains intractable and therefore outside of out computational abilities, the alternative approach of classification method is required.

2.4 Comparison of Experimental and Computational Methods - Summary of Current Method Limitations

X-ray crystallography and NMR spectroscopy methods provide very reliable and relatively accurate structural information. Generally the experimental methods suffer from three major setbacks: Cost, required analysis time and preparation of biological samples. The cost of producing a protein is generally near $1000,000 which on average, takes about one year of combined data acquisition and analysis. On the other hand, protein sample preparation for laboratory experimentation, in practice, becomes a major limitations factor. Most of the time protein extraction and purification is a difficult process. The conformation of proteins is often not preserved in chemicals environments other than their native solutions.

In summary, the production of a protein structure based on the conventional methods is slow and expensive. Computational methods produce a protein structure in a very cost efficient, and relatively fast and bypass the need for the physical existence of a biological sample. Although many advances have been made in the computational field, often the structures produced by this method contain considerable structural errors. Combining both experimental and computational methods can be a solution for aforementioned problems. Minimum data collected from NMR spectroscopy are often rapid and inexpensive. Combining these data with computational methods can produce more reliable structures. Such a hybrid method that combines minimal experimental data with computational methods is the topic of this study.

Chapter 3

Residual Dipolar Coupling - RDC

Residual Dipolar Coupling had been observed as early as 1963 [72] in a nematic crystal environment. In mid-1990s, a number of research reignited the usage of the RDC in the characterization of biomolecules [80] [78]. The usage of RDC in the analysis of the biological structures has expanded rapidly recently, ranging from automated backbone resonance assignment [90], structure determination [86], protein folding to ligand protein and protein-protein interactions [5] [89].

3.1 RDC Principles

The physical basis of RDC is the dipole-dipole interaction between two nuclear spins(Figure 3.1). In the presence of an external magnetic field the RDC between two spin $\frac{1}{2}$ nuclei i and j is given by Equation 3.1:

$$d_{ij} = \frac{-\mu_0 \gamma_i \gamma_j h}{8\pi^3} \langle \frac{3cos^2\theta(t) - 1}{2r(t)_{ij}^3} \rangle \tag{3.1}$$

where γ_i , γ_j are magnetogyric ratio of given nuclei, h is Plank's constant, r is the distance of two nuclei, and θ is the angle between internuclear vector and the external magnet field B_0. The angle brackets denotes the time average dependence of the RDC observable.

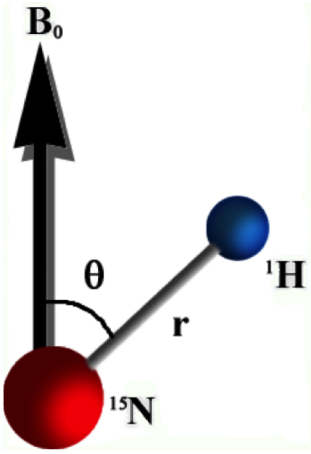


Figure 3.1: The dipolar coupling between two nuclei N and H that depends on the distance r and average orientation θ.

3.2 Alignment Media

The successful acquisition of RDC data depends on using a proper alignment media. To obtain RDC signal, the partial alignment of the molecule in the solution is required [69]. Alignment media help to restrict free tumbling protein. Therefore, the overall protein ensemble shows detectable net RDC signals. Alignment media are utilized routinely such as bicelles, bacteriophage and polyacrylamide gels [68]. The identification of suitable media for a protein is not necessarily trivial. The level of alignment media is an important factor. Alignment should be sufficient to produce measurable RDC, but not so large to introduce the spectral complexity [68].

3.3 RDC Assignment

The assignment of a set of RDC is to find the relationship between RDC data and corresponding protein residue. RDC data from the NMR device is unassigned. That means RDC data are not corresponding to the primary sequence of the structure. Assignment of RDC data can be difficult and time-consuming, depend on the size and complications of the protein.

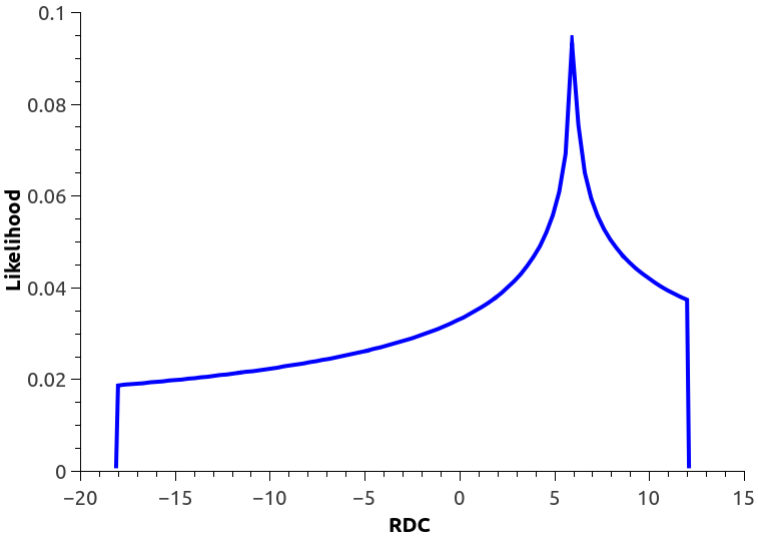


Figure 3.2: Sample powder pattern for the Residual Dipolar Coupling.

3.4 Powder Pattern

The distribution of the RDC data for the infinite number of isotropically distributed vectors in three-dimensional space will generate a *Powder Pattern* (Figure 3.2) [87]. A Powder Pattern is described by three parameters, S_{xx}, S_{yy} and S_{zz} where $S_{zz} = -S_{xx} - S_{yy}$ which are called Principle Order Parameters (POP). The three parameters demonstrate the alignment strength of the protein along the x, y and z axes. Two conditions are assumed to form a Powder Pattern from the distribution of RDC data. The first one is the number of internuclear vectors should be large enough, and the second is the distribution of the internuclear vectors data should be uniform in three-dimensional space. The distribution of an actual structure is highly non- random, depending on the shape of the protein (Figure 3.3). Therefore, the distribution of the RDC data contains structural information about the secondary and tertiary structure of the protein which is the fundamental ground for this research.

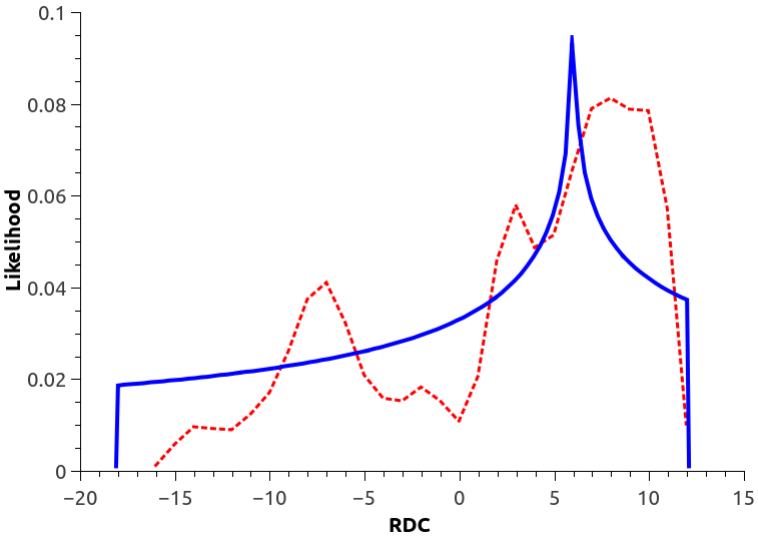


Figure 3.3: Distribution of simulated RDCs for protein 1A1Z(91) (in red-dotted color), with hypothetical order tensors. The horizontal axis represents value of RDC data and the vertical axis represents the likelihood of observing a given value of the RDC.

3.5 Order Tensor and Its Application in RDC Analysis

The averaging described in Equation 3.1 contains information about the internuclear with respect to the magnetic field. In the case of a macromolecule, the average can be described as the average orientation of the macromolecule with respect to the magnetic field and the orientation of the internuclear vector relative to the molecular frame. Let vector $\vec{B}(t) = B.[b_x(t)b_y(t)b_z(t)]$ be the magnetic field such that $B = |\vec{B}(t)|$ at time t, and let $\vec{R}_{ij}(t) = r_{ij}(t).[r_x r_y r_z]$ such that $r_{ij} = |\vec{R}_{ij}(t)|$. Substituting this time-dependent orientational information into Equation 3.1 yields Equations 3.2 to 3.9:

$$RDC^{ij} = \langle \frac{-\mu_0 \gamma_i \gamma_j h}{((2\pi) r_{ij}(t))^3} . (\frac{3([b_x(t)b_y(t)b_z(t)].[r_x r_y r_z])^2 - 1}{2}) \rangle \tag{3.2}$$

$$RDC^{ij} = \langle \frac{-\mu_0 \gamma_i \gamma_j h}{(2\pi)^3 \langle r_{ij}^3(t) \rangle} . (\frac{3\langle([b_x(t)b_y(t)b_z(t)].[r_x r_y r_z])^2\rangle - 1}{2}) \rangle \tag{3.3}$$

$$RDC^{ij} = \frac{-\mu_0\gamma_i\gamma_j h}{(2\pi)^3\langle r_{ij}^3(t)\rangle}\times(\frac{3}{2}.[r_x r_y r_z].\begin{bmatrix} \langle {b_x}^2(t)\rangle & \langle b_x(t)b_y(t)\rangle & \langle b_x(t)b_z(t)\rangle \\ \langle b_x(t)b_y(t)\rangle & \langle {b_y}^2(t)\rangle & \langle b_y(t)b_z(t)\rangle \\ \langle b_x(t)b_z(t)\rangle & \langle b_y(t)b_z(t)\rangle & \langle {b_z}^2(t)\rangle \end{bmatrix}.\begin{bmatrix} r_x \\ r_y \\ r_z \end{bmatrix}-\frac{1}{2}) \tag{3.4}$$

$$D_{max} = \frac{-\mu_0\gamma_i\gamma_j h}{(2\pi)^2} \tag{3.5}$$

$$r_{eff}^{ij} = \sqrt[3]{r_{ij}^3(t)} \tag{3.6}$$

$$v = [r_x r_y r_z] \tag{3.7}$$

$$S = \frac{3}{2}\begin{bmatrix} \langle {b_x}^2(t)\rangle & \langle b_x(t)b_y(t)\rangle & \langle b_x(t)b_z(t)\rangle \\ \langle b_x(t)b_y(t)\rangle & \langle {b_y}^2(t)\rangle & \langle b_y(t)b_z(t)\rangle \\ \langle b_x(t)b_z(t)\rangle & \langle b_y(t)b_z(t)\rangle & \langle {b_z}^2(t)\rangle \end{bmatrix} - \frac{1}{2}I \tag{3.8}$$

where I is the identity matrix.

$$RDC^{ij} = \frac{D_{max}}{(r_{eff}^{ij})^3}.vSv^T \tag{3.9}$$

As Equation 3.6 shows r_{eff}^{ij} is not the same as $\langle r_{ij}^3(t)\rangle$. This is because of the vibration of the particles that creates non-constant bonds length. In this manuscript we assume $r_{eff}^{ij} = \langle r_{ij}(t)\rangle$. In Equation 3.8, S is referred to as the *Saupe Order Tensor Matrix*, which in this manuscript will often be referred to as simply the Order Tensor Matrix (OTM). Equation 3.9 can be re-written as Equation 3.10

$$RDC^{ij} = (\frac{D_{max}}{(r_{eff}^{ij})^3})(S_{xx}x^2 + 2S_{xy}xy + 2S_{xz}xz + S_{yy}y^2 + 2S_{yz}yz + S_{zz}z^2) \tag{3.10}$$

where

$$S = \begin{bmatrix} S_{xx} & S_{xy} & S_{xz} \\ s_{xy} & s_{yy} & s_{yz} \\ s_{xy} & s_{yz} & s_{zz} \end{bmatrix} \tag{3.11}$$

$$v = [x\ y\ z] \tag{3.12}$$

$$|v| = 1 \tag{3.13}$$

3.6 Order Tensor Matrix Decomposition

Spectral Theorem of Linear Algebra states that every symmetric matrix has the factorization of $A = Q\Lambda Q^T$ with real eigenvalues in Λ and orthonormal eigenvector in Q [40]. Therefore, since every order tensor matrix is symmetric, there exists a decomposition of $S = RS'R^T$ for every order tensor matrix S such that S' is a diagonal matrix of the eigenvalues of S and R is rotation matrix whose columns are the eigenvectors of S. The rotation preserves the traceless property of a matrix, therefore S' is also Saupe Order Tensor Matrix (OTM). Equation 3.9 can be re-written as below:

$$RDC_i = (\frac{D_{max}}{{r_{eff}}^3}).v_i R S' R^T {v_i}^T \tag{3.14}$$

$$RDC_i = (\frac{D_{max}}{{r_{eff}}^3}).(v_i R) S' (v_i R)^T \tag{3.15}$$

Equation 3.15 explains that all vectors in the molecule can be rotate by R rotation matrix. Equation 3.15 can be re-written as below:

$$RDC_i = (\frac{D_{max}}{{r_{eff}}^3}).((x_i')^2 {S_{xx}}^2 + (y_i')^2 {S_{yy}}^2 + (z_i')^2 {S_{zz}}^2) \tag{3.16}$$

where

$$v_i R = v_i' = [x_i' y_i' z_i'] \tag{3.17}$$

$$|v_i| = 1 \tag{3.18}$$

We can assume R as an anchor frame that has the application of rotating of two domain of molecules with respect to each other from Residual Dipolar Coupling we used this property to generate RDC computationally two simulate medium alignments [84]. If $|S'_{xx}|$, $|S'_{yy}|$ and $|S'_{zz}|$ be the diagonal elements of S' , the relation between these elements are as follows:

$$|S'_{xx}| \leq |S'_{yy}| \leq |S'_{zz}| \tag{3.19}$$

Equation 3.17 states v_i' is in the *Principal Alignment Frame* (PAF) and the diagonal elements of matrix S' are known as *Principal Order Parameters* (POP) of S. Consequently the maximum and minimum of RDC value can be obtained by following equations:

$$RDC_{max} = \frac{D_{max}}{{r_{ref}}^3} S'_{zz} \qquad v' = [0 \quad 0 \quad \pm 1] \tag{3.20}$$

$$RDC_{min} = \frac{D_{max}}{{r_{ref}}^3} S'_{yy} \qquad v' = [0 \quad \pm 1 \quad 0] \tag{3.21}$$

Rotation matrix R can further be decomposed into three rotational matrix around z, y and z axes. Equations 3.22 to 3.25 demonstrate these rotational matrices:

$$R(\alpha, \beta, \gamma) = R_z(\alpha).R_y(\beta).R_z(\gamma) \tag{3.22}$$

$$R_y(\theta) = \begin{bmatrix} cos\theta & 0 & sin\theta \\ 0 & 1 & 0 \\ -sin\theta & 0 & cos\theta \end{bmatrix} \tag{3.23}$$

$$R_z(\theta) = \begin{bmatrix} cos\theta & -sin\theta & 0 \\ sin\theta & cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.24}$$

$$R(\alpha, \beta, \gamma) = \begin{bmatrix} cos\alpha cos\beta cos\gamma - sin\alpha sin\gamma & -cos\alpha cos\beta cos\gamma - sin\alpha cos\gamma & cos\alpha cos\beta \\ sin\alpha cos\beta cos\gamma + cos\alpha sin\gamma & -sin\alpha cos\beta sin\gamma + cos\alpha cos\gamma & sin\alpha sin\beta \\ -sin\beta cos\gamma & sin\beta sin\gamma & cos\beta \end{bmatrix} \tag{3.25}$$

3.7 Order Tensor Estimation

The core of the RDC analysis is the accurate estimation of order tensors, which provide valuable information about the alignment of the molecule. This information later can be used to study the structure and dynamics of a existing protein structure.

Order tensor can be estimated based on the assignment of resonance. This method however is costly and time-consuming and existence of high-resolution structure is required [84]. Other researches have developed methods to eliminate the need for assignment of resonance. These methods mainly use an unassigned collection of RDC sets from single medium and comparing it with an infinite number of uniformly distributed vectors (powder pattern) [91].In general principal order tensor generated by this method are accurate, furthermore it is mathematically impossible to obtain orientational information of the structure using this method, due to large searching space. Alternatively, a new method has been developed that combines the methods of estimating principle order parameter of order tensor from unassigned RDC data with a known structure to estimate the orientational components of the order tensor as well [12] [85].However the order tensor estimation may not be accurate due to the assumption of the adequate sampling of the RDC space.

Chapter 4

Protein Structure Analysis Using Unassigned RDC Data

4.1 Introduction

Residual Dipolar Coupling (RDC), provides useful orientational information for the inter-nuclear vectors within a molecule [80]. RDC data have been shown to be a very rich source of information about the structure and dynamics of proteins that can be acquired quickly on samples with more limited isotopic labeling. RDCs have been used in studies of carbohydrates [11] [1] [76], nucleic acids [79] [88] [2] [5] and proteins [10] [7] [31] [70] [83]. The use of RDCs as the main source of structural information has led to a significant reduction in data collection and analysis, while providing the possibility of resonance assignment [74] [44] [56] [48], and identification of dynamical regions [16] [20] [23]. Any distance-based constraints can be used only if they have been assigned(see Section 3.3). A given distance is called to be assigned, if the two atoms participating in the interaction within entire structure are known. RDC assignment process generally, is very time-consuming and it requires a large amount of experimental data that is often difficult to collect.Assigned RDC data have also been utilized in a number of instances for identification of homologous structures [31] [8] [58].

Another category of investigations focuses on development of simultaneous assignment and structure determination from RDC data [77] [57]. While these methods

help in extending the frontiers of science, they do not serve as an appropriate screening tool because they either rely on enormous amounts of RDC data acquired in multiple alignment media, or assist in assignment of RDCs to an a-priori known structure. Finally from the practical standpoint, acquisition of RDC data imposes the additional requirement for successful preparation of alignment media. This issue is continually mitigated through the introduction of new alignment media [69]. The large-scale applicability of RDC acquisition has been established by the Structural Genomics Centers (such as NESG http://spine.nesg.org/rdc.cgi) [17], where a large fraction of their target NMR proteins (if not all) have been subjected to RDC data acquisition. Relinquishing the need for assignment of NMR data significantly reduces the financial and temporal cost of data acquisition. Unassigned RDC data contain important structural information that can be extracted for the analysis of the protein structure. Our laboratory has successfully developed a *Probability Density Profile Analysis* (PDPA) method that utilizes unassigned RDC set to rapidly classify protein structures. PDPA also provides an optimal method of validating computationally structures using a minimal set of empirical data [70] [23]. Identifying a homologous structure for an unknown protein using PDPA should be of direct interest to structural biologists and pharmaceutical researchers, since they operate under the same general constraints as the structural genomic centers, which consist of reducing the cost of operation and increasing productivity. Rapid and cost-effective methods of identifying protein structures, which are truly novel, could also serve to increase the general efficiency of structure determination. Therefore, development of a method utilizing unassigned data is highly desirable. This chapter introduces PDPA method. First, the theory of the PDP is discussed, then one-dimensional PDPA and its application are demonstrated. Finally, the shortcomings of one-dimensional methods are brought to the reader's attention at the end of this chapter.

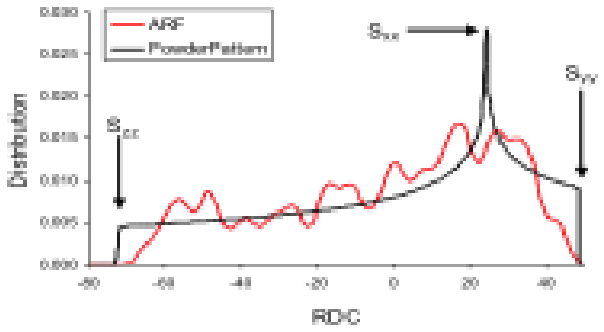


Figure 4.1: A powder pattern and the PDP for ARF (PDBID: 1HUR(180)) using principal order parameters of -71.1, 47.4 and 23.7 in units of Hz.

4.1.1 Theoretical Background

The proposed method for PDP analysis uses Unassigned Residual Dipolar Coupling (RDC) and Kernel Density Estimation (KDE) method for estimating probability density function of RDC data for a given protein. We first provide a brief description of KDE. It can be shown that the distribution of dipolar couplings for a large number of uniformly distributed vectors within a sphere will converge to a relatively featureless powder pattern shown in Figure 4.1. The theoretical basis of this behavior is well documented [13] [84] [53] [26] [87] [85] and an analytical form of this pattern has been derived [87] [85]. While this powder pattern does not contain any useful structural information, values for the principal order parameters (Sxx, Syy, Szz) can be obtained by examining the extreme points of this distribution [27](see Section 3.5). However, proteins appropriate in size for NMR spectroscopy neither contain a large number of vectors (of a specific type such as backbone $C_\alpha - H_\alpha$ or $N - H$) nor sample the entire space uniformly. The number of vectors in proteins (amenable for NMR spectroscopy) remain finite, and their distribution in space significantly departs from uniformity, dictated by the organization of vectors into groups established by the tertiary structure of a protein. Violation of both requirements leading to a featureless powder pattern results in a distribution of RDCs that is a direct function

of the tertiary structure of a protein. The black line in Figure 4.1 is an illustration of a powder pattern with the principal order parameters of 0.001, 0.002 and -0.003 (-71.1, 47.4 and 23.7 respectively in units of Hz for backbone $N-H$ vectors). The red line in this figure represents the distribution of the backbone, $N-H$ RDC data of a 20 kDa protein (the ADP ribosylating factor, PDBID:1HUR(180)) using the same principal order parameters. This deviation can be exploited in order to develop methods of identifying structural similarity based on the statistical profile of the Residual Dipolar Couplings. Here we define a probability density profile (PDP) as the estimated probability density of an observable set of unassigned RDC data originated from a set of inter-nuclear vectors. The resulting PDP can be viewed as a structural fingerprint. Comparison of the PDPs of two structures can provide information regarding their structural homology. To construct any statistical model based on empirical data, utilizing probability density function (PDF) is a necessary step. Parzen Density Estimation (PDE) [65] [36] is used as the main method of the probability density function. Here a kernel representing the local properties of each piece of datum is placed in the appropriate location. The final PDF can then be estimated by the simple summation of all local-likelihoods as shown in Figure 4.2. The appropriate choice for the kernel in this particular illustration is assumed to be Gaussian distributions. Equation 4.1 describes the method of calculating PDP using PDE and Equation 4.2 denotes the choice of Gaussian kernel.In Equation 4.1, K_D denotes individual kernel selected as a Gaussian function centered at the point D_i with standard deviation of σ (Equation 4.2). Also the variable n denoted the number of observed dipolar coupling and D_i is the value of the i-th dipolar coupling.

$$PDP(y) = \frac{1}{n}\sum_{i=1}^{n} K_D(y - D_i) \tag{4.1}$$

$$K(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{y^2}{2\sigma^2}} \tag{4.2}$$

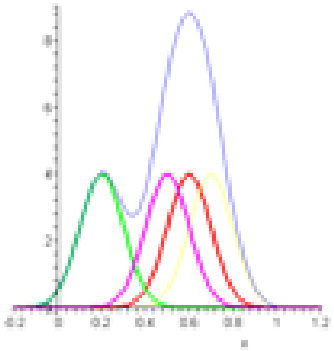


Figure 4.2: An example of Parzen density estimation using Gaussian kernels applied to four points.

4.1.2 Probability Density Profile Analysis

A PDP is defined as the distribution of observed (experimental) set of RDC data which contains structural information about a protein. The distribution of a RDC set usually does not contain large number of observable RDC vectors (such as $N - H$ or $C_\alpha - H_\alpha$) and acquired data are not distributed uniformly. Therefore, the distribution of the experimental data deviates significantly from the ideal featureless powder pattern (Figure 4.3). To facilitate further discussions, the concepts of *query* and *subject* protein is established. A query protein is a structure for which the experimental data have been obtained from NMR spectroscopy and is the subject of investigation. Although to establish correctness of our method initially we used several known protein structures, but we generally consider a query protein as a protein that does not have a previously determined structure. A subject protein is a structure that its structural information is already known. This information includes the coordinates of the atoms and therefore the fold family to which it belongs. The PDP of a query protein can be generated using experimental RDC data and is called *ePDP* and the PDP of subject protein is called *cPDP* that can be computed with given tensoral information from experimental RDC set. A comparison of two distributions (ePDP and cPDP) can be used to assess and quantify the similarity between two

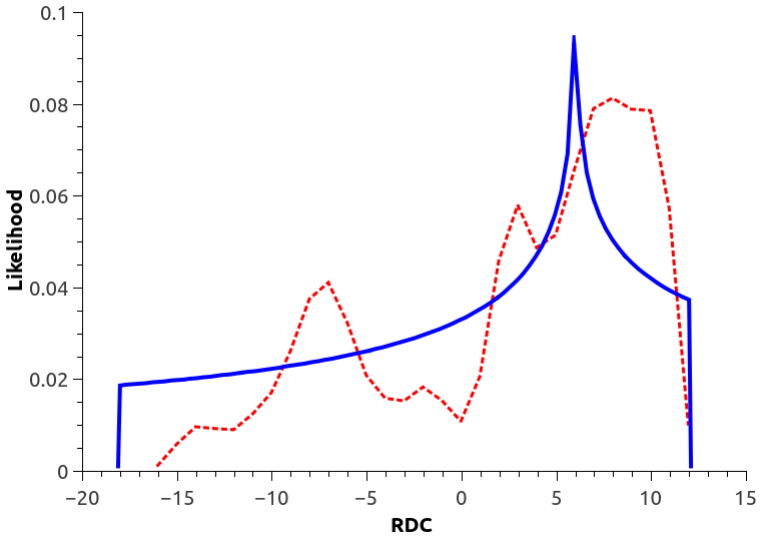


Figure 4.3: Distribution of simulated RDCs for protein 1A1Z(91) (in red-dotted color), with hypothetical order tensors. The horizontal axis in this figure represents value of RDC data and the vertical axis represents the likelihood of observing a given value of RDC.

structures. The entire process of utilizing ePDP and cPDP to measure(quantify) the similarity between two structures is referred to as Probability Density Profile Analysis(PDPA). Figure 4.4 illustrates the general approach used to implement the PDPA algorithm. To estimate principal order parameters (Sxx, Syy and Szz) the maxima of the distribution of the RDC data are taken to simplify the process. PDP distribution is based on the orientation of the protein structure, therefore it is possible that two identical protein structures, produce completely different PDP distribution. This problem can be addressed by an exhaustive search on all possible orientations of the subject protein to identify the best orientation. Selection of an appropriate scoring method to quantify the similarity of two PDP is important. Several methods are suggested in literature that potentially could be used as a suitable metric. χ^2 and *Manhattan* (or *city block*) metrics are selected to be utilized in PDPA scoring [36]. The conventional χ^2 is not appropriate because it does not produce a symmetric result of the distance between two patterns; that is for pattern A and B, $\chi^2(A, B) \neq$

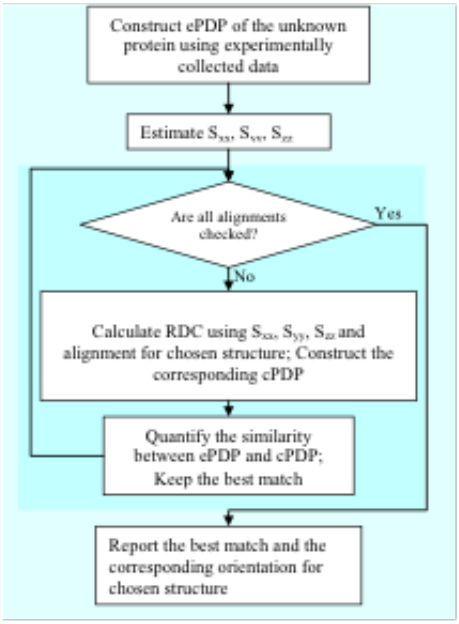


Figure 4.4: General flowchart of the PDPA algorithm.

$\chi^2(B, A)$. The main goal of our modification is to eliminate this lack of symmetry while reducing the harsh penalty due to missing data.

$$S(cPDP, ePDP) = \frac{1}{2}\sum_{i=1}^{M}[\chi^2(c_i, e_i) + \chi^2(c_i, e_i)] \tag{4.3}$$

$$\chi^2(c, e) = \begin{cases} \frac{(c-e)^2}{c} & \text{for} \quad c \neq 0 \\ 100e & \text{for} \quad c = 0 \end{cases} \tag{4.4}$$

The term $S(cPDP, ePDP)$ in Equations 4.3 denotes the final comparison score between cPDP and ePDP. The summation index M denotes the number of points that are sampled in comparing the two PDPs. c_i and e_i denote the values of computed and experimentally determined PDPs at the location i, respectively. The distance at any given position of two PDPs is determined by $\chi^2(c, e)$ as defined in Equation 4.4. Manhattan-block is another metric that is utilized in PDPA for comparison of cPDP and ePDP (Equation 4.5).

$$S(cPDP, ePDP) = \sum_{i=1}^{n}|c_i - e_i| \tag{4.5}$$

In Equation 4.5, c_i and e_i indicate the values of the computed and experimental PDPs at the location i.

4.2 1D-PDPA Method and Results

The implementation of one dimensional PDPA (1D-PDPA) [12] [85] utilizes one set of unassigned RDC data for the analysis. In this analysis, Parzen Density Dstimation(PDE) is used as the main method of Parzen Density Function (PDF) estimation [39] [18]. To demonstrate the sensitivity and selectivity of the PDPA method, a number of experiments are conducted and are listed below.

4.2.1 The application of the 1D-PDPA in Identification of Structural Homologous

The first evaluation of the PDPA was in application to simulated data without any noise. A library of 21 protein structures representing 9 fold families has been utilized in this exercise. The results of this experiment are listed in Table 4.1. The first two columns of Table 4.1, list the PDB code for each used structure and the PDB code of the family that is representative of each structure (reported by FSSP [43]), respectively. Furthermore, the "Repr"column lists the family fold representative for each structure and "Size"is the number of Amino Acids. The RMSD column is the backbone rmsd between each protein and its family representative as reported by FSSP. In this evaluation we treated the structure 1C99(79) as our query (unknown) structure and the remaining 20 structures as our subject structures. With the exception of the structure 1A91(79), all other structures exhibited no significant sequence similarity. The query structure 1C99(79) is a member of the family represented by 1A91(79). These two proteins share a 56% sequence identity and a 4.5 Å rmsd overlay of $\alpha - carbons$ over residues 2-70. Program PALES [99] was used to predict an order

Table 4.1: PDP analysis of the structure 1C99(79) with 20 different structures representing 9 family folds.

PDB Code	Repr.	Size (aa)	rmsd	Score
1A91	1A91	79	0	4.98
1C99	1A91	79	3.3	0
1CII	1A91	101	2.8	4.13
1CXZB	1A91	86	3.3	4.77
1FH1	1FH1	92	0	7.01
1A8O	1A8O	70	0	23.6
1ACP	1A8O	77	2.9	8.51
1A1Z	1A1Z	91	0	13.2
3CRD	1A1Z	100	2.5	17.7
1A32	1A32	88	0	15.1
1CXZA	1A32	86	2.3	25.5
1RB9	1RB9	53	0	40.7
1RDG	1RB9	52	0.7	27.09
1CC5	1CC5	83	0	47.8
451C	1CC5	82	2.7	35.6
1CTJ	1CC5	89	2.5	35.2
1YCC	1CC5	108	2.7	43.0
1BBZA	1BBZA	58	0	50.6
1GCPA	1BBZA	65	1.9	46.0

tensor for the query structure 1C99(79). This order tensor was used to produce simulated RDC data for the query protein using REDCAT [84]. While it is obvious that in the absence of any error, PDPA will succeed in identifying the original structure, the point of this evaluation remains to assess its ability in the identification of other members of the same protein fold family. After exploring all possible orientations of alignment, the best score (described by Equations 4.3 and 4.4) has been reported and is listed in the last column of Table 4.1. The results in this table indicate that PDPA succeeded in identifying the correct family of the unknown structure, based on unassigned RDC data. These results are encouraging but impractical, since the addition of noise can have a dramatic impact on the performance of any given method. During an extension of this evaluation, noise was added to the computed RDC data

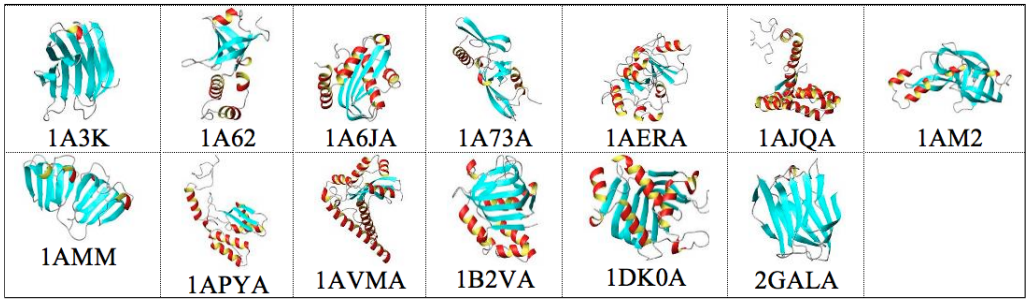


Figure 4.5: Structure of all 12 proteins used in the application of PDP analysis of Galectin3 (PDBID:1A3K(137)).

of 1C99(79). The results of this experiment (previously published, not shown here) demonstrated that through appropriate selection of the smoothing factor (variances) of kernel functions vital structural information can be successfully extracted [85]. The final evaluation of 1D-PDPA is the test of this approach in application to experimental data. Here, backbone RDC data from Galectin-3 (PDBID:1A3K(137)) have been utilized [68]. This set of data presented the most realistic scenario. Here only 80% of the backbone resonances had been observed. Furthermore, the experimentally collected data contained as much as 50% error for some individual data points (15 Hz error). Application of 1D-PDPA succeeded in selecting the correct structure form a library of 12 other structures (Figure 4.5). In addition, the representative of Galectin-3 family (PDBID:2GAL(135)) was also identified as a significantly close structural relative. These results are listed in Table 4.2.

4.3 Limitation of 1D-PDPA Method

1D-PDPA illustrates a great potential for establishing structural similarity of an unknown protein using unassigned RDC data therefore based on easily obtainable backbone N-H RDC data, one can ascertain the structural novelty of an unknown protein. Structural templates can be selected for the purpose of threading using PDPA independent of sequence homology. However, there are a number of impediments rooted

Table 4.2: Results of PDP analysis to experimental data collected from Galentic3 (PDBID:1A3K(137))

Structure	Size	Score
1A3K	127	4.27
1A62	115	24.27
1A6JA	143	29.64
1A73A	148	21.18
1AERA	186	63.00
1AJQA	195	53.31
1AM2	169	56.75
1AMM	165	26.95
1APYA	150	42.68
1AVMA	191	16.25
1B2VA	169	23.12
1DK0A	169	22.63
2GALA	121	1.10

in properties of RDC data that need to be addressed to improve the PDPA selectivity and sensitivity. Generally small proteins do not contain enough RDC vectors, therefore comparison of the RDC distribution of the small proteins using one RDC set, often lacks sensitivity. Also, the fact that experimental RDC data contain device errors in a certain range, potentially have an effect on the distribution of the vectors and consequently on PDPA analysis. It is also possible that two completely different structures produce identical PDPs if the structural relationship between the two structures perfectly coincides with symmetric properties of the alignment, such as inversion [4] [85]. To address these issues, collection of RDC data from two or more independent alignment media, which is simple to obtain experimentally, should differ between two structures exhibiting a close distribution of RDCs in perspective of one alignment medium. While it is possible that the second alignment medium to share the structural degeneracies, occurrence of this phenomenon for both datasets in the same region should be unlikely if both alignment media differ from each other by more than a simple scaling factor. Therefore utilizing second RDC set, can po-

tentially be a solution for the limitations of 1D-PDPA listed above. Although the expansion of 1D-PDPA to 2D-PDPA is straight forward, its expansion into higher dimensions become computationally intractable. Therefore our challenges includes:

1. Expansion of existing 1D-PDPA technology to 2D-PDPA.
2. Development of meaningful interpretation of the results.
3. Re-engineering of the computational core to enable expansion into higher dimensions.

The expansion of PDPA that utilizes RDC sets from multiple alignment media, is subject of this research that is discussed in detail in the next chapter.

Chapter 5

2D-PDPA: Two Dimensional Probability Density Profile Analysis

5.1 Introduction

In the following sections, the 2D-PDPA method, an extension of the 1D-PDPA is discussed. Then the results of several experiments are presented. The results include the performance of the 2D-PDPA method utilizing simulated RDC data modeled from ideal conditions to prove the concepts, and experimental data that reflect more pragmatic conditions. In this extent the focus is more on protein structures for which both NMR and X-ray structures are known and their experimental RDC data are available. Finally, the results of 2D-PDPA in ranking of computationally modeled structures for a target protein with no known structure is presented.

5.2 Expansion of one dimensional to two dimensional of PDPA method

1D-PDPA method utilizes one RDC dataset from one alignment medium, and it has demonstrated limited success when applied to large proteins due to the incapability of resolving different internuclear vectors with similar RDC values. Moreover, NMR data acquisition error reduces the sensitivity of the data analysis, and can produce an inaccurate RDC distribution. On the other hand, the RDC data from

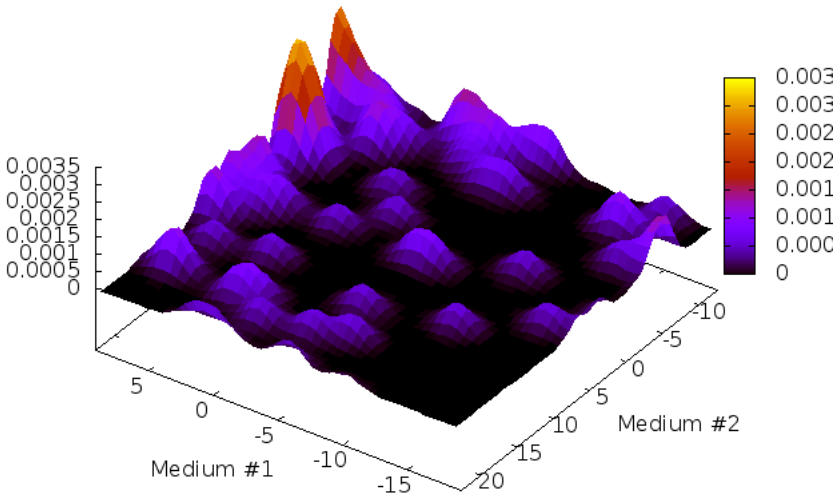


Figure 5.1: An example of a 2D-PDP map generated using kernel density estimation. This 2D-PDP can serve as a structural fingerprint.

multiple alignment media are relatively easy to obtain and previous works have shown that utilizing RDC sets from two alignment media, significantly reduces the degeneracy problem and increases the information content, sensitivity and selectivity [3] [12]. 2D-PDPA is an extension of the 1D-PDPA method that allows simultaneous analysis of RDC data from a second alignment medium. The overall principle of 2D-PDPA method is the same as 1D-PDPA that is two similar structures must exhibit a similar distribution of RDC data as shown in Figure 5.1. In this figure, the distributions of RDC points is a function of the protein structure and can be used as a structural fingerprint of an unknown protein. Therefore the measure of the similarity between two distributions of RDC data can be interpreted as a measure of structural similarity. Overall operations of 2D-PDPA proceed in three main stages as shown in Figure 5.2. During the first stage, experimental RDC data are analyzed to estimate seven of the ten needed parameters [62] [97] that are used to back-calculate RDC data from any given structure in two alignment media. During this stage, the scattering of

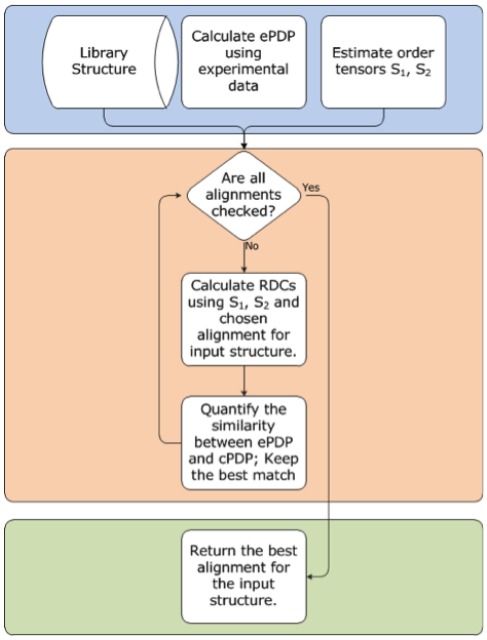


Figure 5.2: Operational schematic of the 2D-PDPA method illustrated in three main phases.

the RDC data in two alignment media is converted to a distribution function using Kernel Density Estimation [39] [42] [85]. This distribution is constructed through superposition of Gaussian kernels that are centered at each RDC data point Figure 5.1 illustrates an example distribution map that is denoted as ePDP throughout this report. An ePDPA is referred to as a distribution map (or a fingerprint) that is generated from experimental data. During the second phase of 2D-PDPA, a similar map is created based on the back-calculated RDC data from each of the protein structures available in the library of structures using the same Kernel Density Estimation procedure. The computed maps are denoted as the cPDPs. For each structure in the database, a cPDP is created for each possible rotation of the structure in a grid search over the Euler angles (α, β, γ) at the resolution of $5\,^{\circ}$. Each of these cPDPs is compared to the ePDP and the best score as well as the corresponding Euler angles are recorded for each structure in the database. These 46,656 ($36 \times 36 \times 36$ rotations over $\alpha, \beta, and\, \gamma$) alternate cPDPs are created as a result of a $5\,^{\circ}$ grid search over the

three remaining parameters that are needed for back-calculation of the RDC data. These three remaining parameters essentially represent all possible orientations of any given structure. RDCs are insensitive to 180 ° rotations; hence the search space can be reduced to a range of $[0\,^\circ - 180\,^\circ]$ in increments of 5 ° for each parameter. The best matching score and its corresponding three search parameters are recorded for the third and final stage of 2D-PDPA. During the concluding stage of the 2D-PDPA, all of the proteins in the library of structures are ranked based on their 2D-PDPA fitness score, which was measured during the previous stage, and the results are reported.

5.3 Scoring and Interpretation of 2D-PDPA Raw Scores

In contrast to 1D-PDPA [85] [12] that utilizes χ^2 metric [42] of comparison, 2D-PDPA employs a more intuitive Manhattan (or City-Block) metric for comparison of cPDP and ePDP. Equation 5.1 describes the Manhattan distance that is computed by 2D-PDPA. In this equation B denotes the 2D-PDPA′s raw score (Block score), the summation indices i and j traverse the entire range of RDCs over the two alignment media M_1 and M_2, and σ_i and σ_j denote the step size of uniform grid sampling along each of the RDC dimensions.In this equation $cPDP_{ij}$ and $ePDP_{ij}$ represent the likelihood reported by each PDP set at locations i and j. Since the cPDP and ePDP are normalized to be a qualified probability density functions, their integral over the entire range of RDCs equates to one. Therefore the block-score will have an effective range of $[0 - 2]$, where a score of 0 indicates 100% similarity and a score of 2 indicates 0% similarity between the two structures.

$$B_{2D\_PDPA} = \sum_{i=Min(M_1)}^{Max(M_1)} \sum_{j=Min(M_2)}^{Max(M_2)} |cPDP_{ij} - ePDP_{ij}|.\sigma_i.\sigma_j \tag{5.1}$$

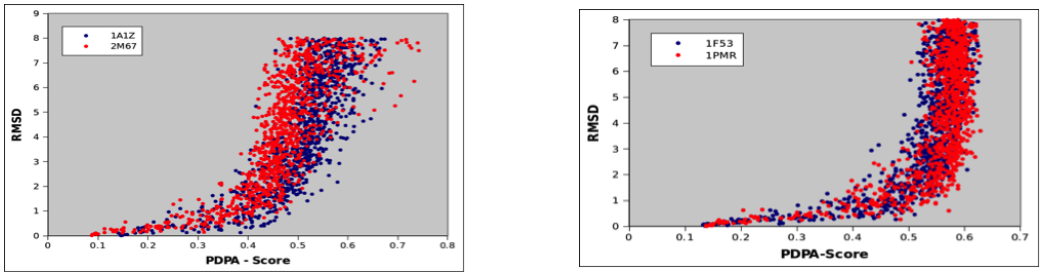
In instances where meaningful bb-rmsd values can be calculated between the members of the search database and the unknown protein, a more informative relationship between the 2D-PDPA's B-score (see Section 5.3) and the expected bb-rmsd can be established. Such interpretation patterns can be created based on the following observations:

1. Interpretations patterns are primarily a function of class of protein structure (α and β protein) and protein size
2. Interpretation patterns depend on completeness of data
3. Interpretation patterns exhibit a dependency on quality of experimental data, and more directly on the quality of the two estimated order tensors

The latter dependency is intuitive and is investigated in the literature [85] [12] [62] [97] and it is therefore not discussed further in this report. We demonstrate the first and second above dependencies by generating a scatter plot of bb-rmsd versus their corresponding B-score for 1000 derivative structures. These derivative structures were generated by randomly altering backbone dihedral angles of the native structure for a given protein. The ensemble of altered structures was used to compute a B-score and bb-rmsd with respect to the native structure. In this exercise, we have used two sample $\alpha-proteins$ (1A1Z and 2M67) and two sample $\beta-proteins$ (1F53 and 1PMR) that are approximately of equal sizes. Table 5.1 shows the detailed information for each of these four proteins. It is important to note that the two proteins in each structural class are unrelated. Figure 5.3 illustrates the interpretation patterns for each of the two classes. The two patterns are remarkably well conserved between the two proteins from the same structural class. Any noted differences are due to random sampling of the RDC space and will be resolved by increasing the number of random sampling. These interpretation patterns also exhibit a very predictable behavior as a function of missing data. To illustrate this point, we performed a similar

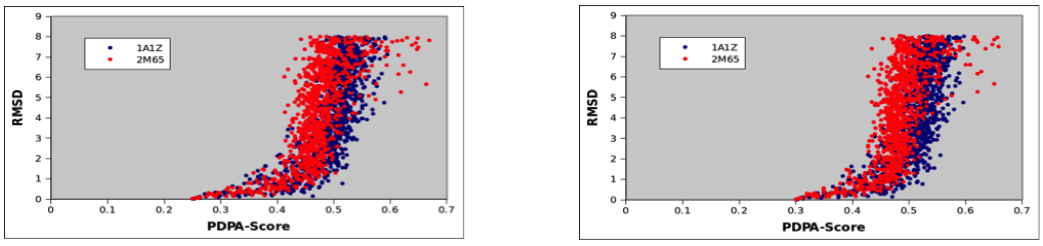
Table 5.1: List of four proteins that are used in establishing the properties of 2D-PDPA bb-rmsd interpretation patterns

Protein PDBID	Protein Size	CATH Classification	Number of Secondary Structural Elements
1A1Z	83	1.10.533.10	11 $\alpha - helices$
2M67	81	Not available yet	6 $\alpha - helices$
1F53	84	2.60.20.30	6 $\beta - strands$
1PMR	80	2.40.50.100	6 $\beta - strands$



(a) (b)

Figure 5.3: Sensitivity of 2D-PDPA analysis as a function of bb-rmsd when applied to (a) two unrelated α-proteins and (b) two unrelated b-proteins. Simulations included addition of ± 0.5 Hz of uniformly distributed noise.



(a) (b)

Figure 5.4: Sensitivity of 2D-PDPA analysis as a function of bb-rmsd on two α-proteins 1A1Z(91) and 2M67(81) (a) with 25% of the data randomly removed and (b) with 30% of the data randomly removed.

exercise as above on the $\alpha - protein$ set 1A1Z, 2M65 by randomly removing 25% and 30% of the data. The final results are shown in Figure 5.4 and as expected, the lowest scores correspond to the percentage of missing data (shown in Equation 5.2). This exercise was repeated for a number of other proteins with very similar results (not shown here). Based on this observation, a corrected score can be computed by subtracting the fraction of missing data from the raw score. This correction eliminates the contribution of missing data and allows for easier comparison of 2D-PDPA's scoring mechanism across different instances of analyses:

$$CorrectedScore = RawScore - MissingData \tag{5.2}$$

The noted properties of the 2D-PDPA's Block scoring mechanism enables the creation of an interpretation pattern from another protein with similar structural attributes as the target protein. The resultant interpretation pattern can then be used to establish the quality of the 2D-PDPA's selected structure.

5.4 2D-PDPA Results and Discussion

5.4.1 Structure Identification from Simulated RDC Data

2D-PDPA was validated using synthetic data generated from eleven different protein structures (listed in Table 5.2) to represent a spectrum of sizes and structure types. Data from each protein structure was used to identify the correct structure from a library of 619 decoy representative structures. In each test case, the decoy structures that were not within ±20 percent size of the target structure were eliminated from the pool of potential candidates. This filtering mechanism reduced the list of possible structural candidates to within 100 for proteins with less than 120 residues in length, and around 20% for larger proteins (more than 250 residues). The

Table 5.2: Results of structure identification using simulated data.

Target Structure	Size(# of NH Vectors)	Error Added	Rank
1BRF	46	±1 hz	1
1P7E	55	±1 hz	1
1SF0	67	±1 hz	1
1BQZ	75	±1 hz	1
110M	149	±1 hz	1
1NCX	160	±1 hz	1
$1QHS_A$	172	±1 hz	1
3FIB	241	±1 hz	1
$16VP_A$	289	±1 hz	1
$1VSG_A$	353	±1 hz	1
$1A4A_A$	445	±1 hz	1

identification results of 2D-PDPA on the eleven randomly selected test proteins are shown in Table 1. The first column of this table lists the PDB-ID of each protein, followed by the protein size (based on the number of $N-H$ vectors), the magnitude of the uniformly added noise, and the ranking of each protein by 2D-PDPA.

5.4.2 Structure Identification Using Experimental RDC Data

A search through the BMRB [35] database resulted in three proteins with backbone RDC data from two or more alignment media. These three proteins consisted of 1P7E [81], 1D3Z [28] and 1RWD [77] with backbone $N-H$ RDC data from two alignment media. Structural homologous (both NMR and X-ray when possible) were added to our existing database of 619 decoy structures to examine 2D-PDPA's ability to identify the actual or any homologous structures. Table 5.3 shows the results for the protein structure 1P7E. The structure 1P7E was identified as the highest plausible structure by the 2D-PDPA as expected. Of even more interest, however, are 2^{nd} and 3^{rd} place rankings, which consisted of 1IGD and 1P7F. These are the structural

Table 5.3: Results of structure identification from unassigned experimental RDC data for the protein PDBID:1P7E.

Library Structure	Size(# of NH Vectors)	Rank	Raw Score
1P7E	55	1 hz	0.45
1IGD	59	2 hz	0.47
1P7F	55	3 hz	0.48

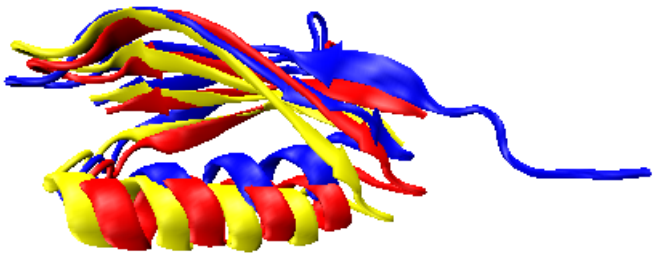


Figure 5.5: Cartoon representation of proteins 1P7E(56) (yellow), 1IGD(61) (blue) and 1P7F(56) (red).

homologous added to the library, and are ranked 2nd and 3rd respectively. The structures 1P7E, 1IGD and 1P7F exhibit around 1.0 Å of difference measured over the backbone atoms as shown in Figure (5.5). These results exhibit 2D-PDPA's ability to identify not only the identical structure from a library of decoys, but also other homologous structures. Of even more importance is the fact that this experiment was performed with relatively small amounts (43 RDCs from 55 residues, 78%) of experimental data.

For 1RWD (results shown in 5.4), its X-Ray determined homologous 1BRF (bb-rmsd of 1.8 Å with respect to 1RWD as shown in Figure 5.6) ranked first. The 1RWD structure ranked second behind 1BRF. At first it may seem odd that the X-Ray structure outranked the NMR structure. However, although 2D-PDPA ranks 1BRF as the better suited structure, the ranking score of 1BRF is negligibly better than 1RWD. Furthermore, it is generally accepted that X-Ray structures fit RDC

Table 5.4: Results of structure identification from unassigned experimental RDC data for the protein PDBID:1RWD.

Library Structure	Size(# of NH Vectors)	Rank	Raw Score
1BRF	46	1	0.661
1RWD	43	2	0.667

Figure 5.6: Cartoon representation of the superimposed structures 1BRF(53) (yellow) and 1RWD(53) (blue).

data better than NMR structures. This experiment once again demonstrates 2D-PDPA's success in finding structural homologous within a large library of possible structures.

5.4.3 Computationally Modeled Structure of PF2048.1

PF2048.1 is a 9.16 kDa, 78 residues; (including His-tag) monomeric protein with less than 26% sequence identity to any structurally characterized protein. An ensemble consisting of ten modeled structures from ROBETTA [46] [12] and five modeled structures from I-TASSER [98] for the unknown protein PF2048.1 were obtained (superimposed structures shown in Figure 5.7). Table 5.5 lists the results for an exhaustive pairwise comparison of the ensemble of fifteen structures measured over the

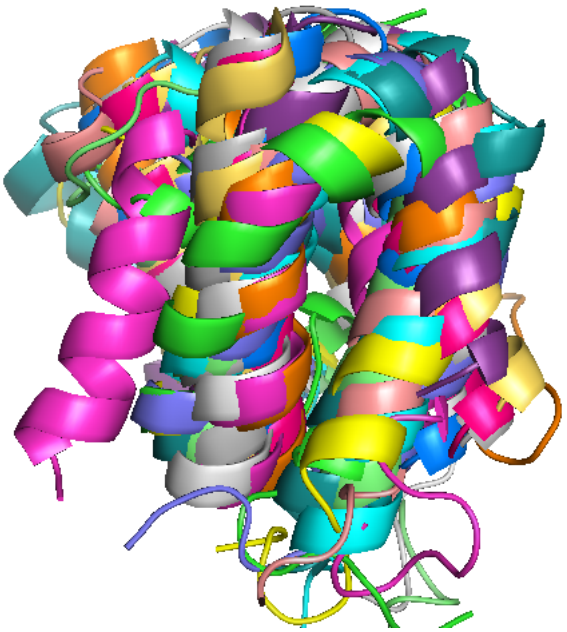


Figure 5.7: Fifteen modeled structures of PF2048.1 by ROBETTA and I-TASSER.

backbone atomic positions. In this table, structures R1-R10 and I1-I5 correspond to the ROBETTA and I-TASSER structures respectively. The areas of this table that are shaded in green or yellow correspond to the intra-modeling distances, while the dark-blue areas correspond to the inter-modeling distances. Based on these results, structures modeled by ROBETTA exhibit structural similarity in the range of 2.91Å-7.83Å while structures modeled by I-TASSER exhibit more convergence with structural similarity in the range of 1.21Å-3.62Å. It is clear from this exercise that both methods have been successful in producing a reasonable model of the structure since all of them consist of a bundle of four helices. It is also clear that in the absence of a-priori knowledge of the protein's structure, selection of the most suitable structure would have not been possible. Due to the general lack of convergence in the modeled structures, arbitrary selection of a model could lead to an erroneous structure.

Table 5.5: Pairwise bb-rmsd of the ten structures modeled by ROBETTA and five structures modeled by I-TASSER.

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	I1	I2	I3	I4	I5
R1	0	6.51	2.93	3.01	2.95	3.39	4.33	3.37	2.73	5.42	4.43	5.14	4.82	4.25	3.48
R2	6.51	0	7.32	7.52	6.62	6.44	7.04	8.05	7.83	6.29	7.49	7.51	7.93	7.69	7.71
R3	2.93	7.32	0	4.8	5.17	3.08	6.19	4.06	3.69	4.11	6.45	7.21	7.06	6.21	5.29
R4	3.01	7.52	4.8	0	3.34	5.26	2.91	3.75	2.68	7.31	3.08	3.78	3.85	3.23	2.79
R5	2.95	6.62	5.17	3.34	0	4.72	3.1	4.04	3.94	6.81	3.32	3.78	3.66	2.96	2.83
R6	3.39	6.44	3.08	5.26	4.72	0	5.75	5.69	3.75	3.49	6.52	7	6.84	6.18	5.56
R7	4.33	7.04	6.19	2.91	3.1	5.75	0	5.45	4.2	7.73	2.89	3.02	3.56	3.22	3.43
R8	3.37	8.05	4.06	3.75	4.04	5.69	5.45	0	4.36	7.06	4.87	5.77	5.74	4.61	3.73
R9	2.73	7.83	3.69	2.68	3.94	3.75	4.2	4.36	0	6	4.92	5.57	5.25	4.77	4.04
R10	5.42	6.29	4.11	7.31	6.81	3.49	7.73	7.06	6	0	8.48	8.85	8.86	8.23	7.6
I1	4.43	7.49	6.45	3.08	3.32	6.52	2.89	4.87	4.92	8.48	0	1.21	2.75	1.31	1.91
I2	5.14	7.51	7.21	3.78	3.78	7	3.02	5.77	5.57	8.85	1.21	0	2.55	1.89	2.89
I3	4.82	7.93	7.06	3.85	3.66	6.84	3.56	5.74	5.25	8.86	2.75	2.55	0	3.07	3.62
I4	4.25	7.69	6.21	3.23	2.96	6.18	3.22	4.61	4.77	8.23	1.23	1.89	3.07	0	1.44
I5	3.48	7.71	5.29	2.74	2.83	5.56	3.43	3.73	4.04	7.6	1.91	2.89	3.62	1.44	0

5.4.4 2D-PDPA Ranking of the Modeled Structures

2D-PDPA was applied to the ensemble of ten modeled structures of PF2048.1 by ROBETTA and five models by I-TASSER. Due to experimental conditions only 49 RDC data points were obtained from this protein in two alignment media. Considering the size of the PF2048.1 protein (79 residues), 49 RDC data points constitutes only 62% of the complete dataset (38% missing data). The relative order tensors describing the alignment of this protein in each of the media were determined using the previously reported 2D-RDC [62] method ($\lambda - map$ shown in Figure 5.8) and are listed in Table 5.6. Results of the 2D-PDPA ranking of ROBETTA and I-TASSER structures are shown in Table 5.7 and Table 5.8 respectively. The three columns in these tables list the structural identifiers, 2D-PDPA's raw score for each structure, and the corrected scores respectively. The corrected scores are based on contribution of the percentage missing data on the raw score and are computed as shown in Equation 5.2. By selecting a reasonably stringent raw score of 0.8 (corrected score of 0.42) as the cutoff threshold for structural quality, the list of fifteen structures can

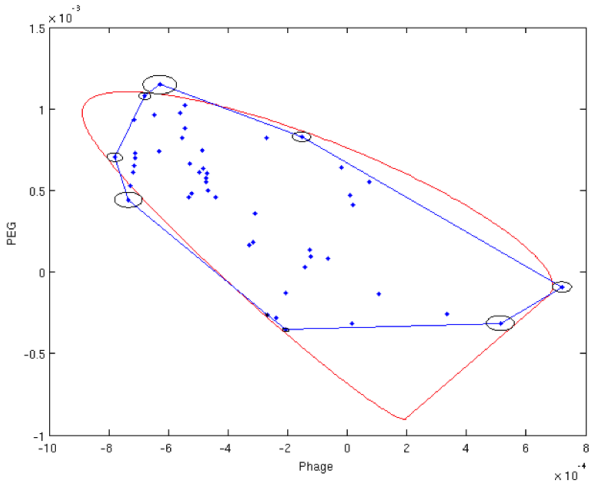


Figure 5.8: Results of 2D-RDC analysis based on unassigned data from PF2048.1 obtained in Phage and PEG alignment media. The blue lines indicate the convex hull of the 2D-RDC dataset determined from the experimental data and the red line indicates the convex hull of the distribution of 2D-RDC data points for the order tensor estimate.

Table 5.6: Order tensors of PF2048.1 estimated from 2D-RDC analysis using unassigned RDC data from two alignment media (Phage and PEG).

	S_{xx}	S_{xy}	S_{xz}	S_{yy}	S_{yz}	Da
M1(Phage)	2.04E-04	0.00E+00	0.00E+00	7.11E-04	0.00E+00	-10.8
M1(PEG)	-9.14E-04	1.71E-05	1.61E-04	-8.55E-04	3.89E-04	-13.03

be reduced to five; R5 and R1 of the ROBETTA structures, and I5, I4, and I2 of the I-TASSER. Figure 8 illustrates the superposition of these five structures with an average bb-rmsd of 2.53 Å. The emergence of structural convergence among the top five selected structures signifies the systematic selection mechanism of 2D-PDPA. It is important to note that 2D-PDPA's selection mechanism is exclusively based on fitness to the experimental data and not simply based on clustering of the bb-rmsd data shown in Table 5.5. This independent and yet consistent selection between 2D-PDPA and bb-rmsd provides a strong evidence for accuracy of the top five structures.

Table 5.7: 2D-PDPA scores for the ten ROBETTA structures.

Modeled structure	2D-PDPA raw score	2D-PDPA corrected score
R5	0.74	0.36
R1	0.79	0.41
R8	0.81	0.43
R4	0.81	0.43
R7	0.82	0.44
R2	0.82	0.44
R6	0.83	0.45
R10	0.84	0.46
R9	0.85	0.47
R3	0.87	0.49

Table 5.8: 2D-PDPA scores for the five I-TASSER structures.

Modeled structure	2D-PDPA raw score	2D-PDPA corrected score
I5	0.73	0.35
I4	0.76	0.38
I2	0.78	0.40
I3	0.81	0.43
I1	0.82	0.44

5.4.5 Interpretation of 2D-PDPA Results for Modeled Structures of PF2048.1

Results listed in Table 5.7 and Table5.8 rank the fitness of the modeled structures. However these results do not provide any information regarding the accuracy of the modeled structures with respect to the solution state structure of this protein. This information can be retrieved from further analysis of the raw scores that are provided by 2D-PDPA. To interpret the results of 2D-PDPA meaningfully, a simulation exercise has been conducted to relate the PDPA fitness score to backbone RMSD. Here we have utilized protein 1A1Z (83 residues) as a comparable structure to PF2048.1 on the

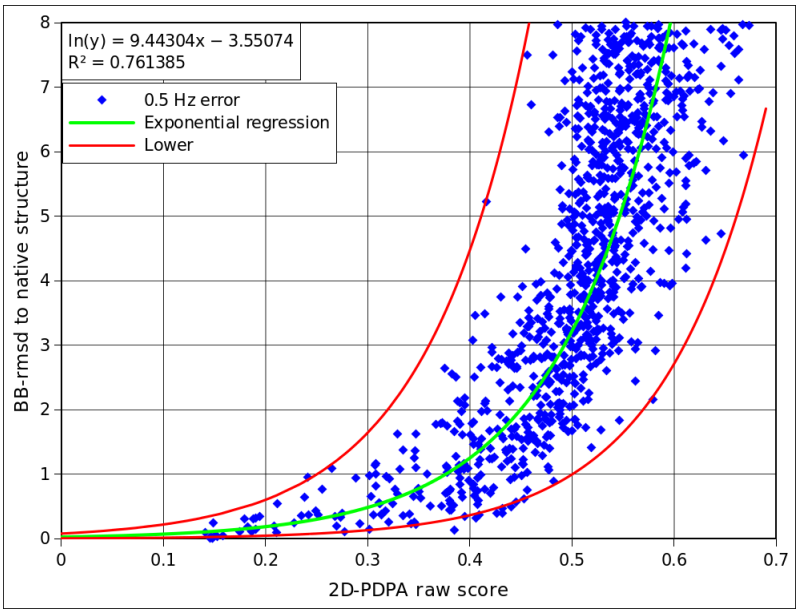


Figure 5.9: An interpretation patter for the protein PF2048.1, which illustrates the relationship between 2D-PDPA's score and structural quality in bb-rmsd.

basis of its size and $\alpha - helical$ nature. RDC data have been computed for these two proteins using typically observed order tensors as shown in Table 5.6.Each dataset has been corrupted through the addition of ± 0.5 Hz of uniformly distributed noise. One thousand derivative structures have been generated from the native structure by randomly perturbing the backbone dihedral angles (ϕ, ψ). The set of derivative structures provided a sampling of the bb-rmsd in the range of 0-8Å with respect to the starting structure. The 2D-PDPA procedure was then applied to the set of 1000 sample structures. Figure 5.9 shows the scatter plot of 2D-PDPA scores versus the backbone rmsd's. This figure is very valuable in establishing the operational limits of 2D-PDPA as a function of data quality, and help in interpreting the results shown in Table 5.7 and 5.8. Based on the extrapolated upper and lower boundaries, the scores of 2D-PDPA can be converted to a range of bb-rmsd with respect to the solution state structure of the PF2048. Table5.9 lists the lower and upper estimates of bb-rmsd for each of the top five modeled structures. Therefore it can be concluded with high certainty that the R5 and the I5 structures are within 3Å of the solution state

Table 5.9: Results of 2D-PDPA analysis of modeled structures for PF2048.1 with the estimated range of bb-rmsd to the solution state structure using 1A1Z(91) as a template for the interpretation pattern.

Modeled structure	2D-PDPA corrected score	Lower bb-rmsd(Å)	Upper bb-rmsd(Å)
I5	0.35	0.22	2.72
R5	0.36	0.25	3.00
I4	0.38	0.30	3.67
I2	0.40	0.37	4.48
R1	0.41	0.41	4.95

structure of the PF2048.1.

Chapter 6

$nD-PDPA$: $n-Dimensional$ Probability Density Profile Analysis

6.1 Introduction

In principle PDPA does not have to be limited to two sets (as in 2D-PDPA) or homogeneous data types (e.g. using only $C_\alpha - H_\alpha$ RDC sets). During RDC data acquisition, additional relevant data (such as $C_\alpha - H_\alpha$) are oftentimes available and will add little impact on data acquisition time. A minimal extension of data acquisition time can provide significantly more experimental data such as two or more RDC data sets(from N-H, $C_\alpha - H_\alpha$). Integration of additional data sets is predicted to substantially increase the information content and therefore significantly improve the sensitivity and robustness of the PDPA method. However, the inclusion of more RDC sets significantly increases the computational time of the analysis, which would render it computationally intractable. Moreover, 2D-PDPA program lacks the capability of utilization of more than two RDC sets. To fulfill these requirements, the development of an improved PDPA engine is required. In the following sections, nD-PDPA method is described then the result of 2D-PDPA and nD-PDPA are compared for accuracy, sensitivity and measure of execution time. Finally, the nD-PDPA method is validated by utilizing protein structures varying in size and secondary structures with synthetic data.

6.2 Expansion of 2D-PDPA to $nD - PDPA$

The details of the PDPA method are described previously [38], [85], [12] (for more information also see Chapters 4 and 5 of this manuscript). In this manuscript we provide a brief overview of the PDPA method and focus primarily on the new additions and improvements of the nD-PDPA. The core principle of the PDPA method is based on the fact that two similar structures should produce the same distribution patterns of RDCs, and can be used as a structural fingerprint. Therefore, measurement of the similarity between two distributions can be interpreted as similarity of two structures. The nD-PDPA algorithm is encapsulated in three functional layers. In the first layer, the experimental RDC sets are used to estimate the relative order tensors [60], [62]. The number of parameters needed in this stage is a function of the number of alignment media in which RDC data are acquired. Generally for RDC data from n alignment media, 5n-3 parameters are required to describe the relative order tensors [97]. The estimated order tensor parameters are utilized to back calculate the RDC data for a given structure. Then the n-dimensional Kernel Density Estimation is utilized to construct the distribution map for both experimental and calculated RDC sets. The kernel Density Distribution is calculated based on a hyper dimensional Gaussian Kernel function (described in Equation 6.1) that is located at the center of each RDC data point (Figure 6.1). In this equation X denotes independent function parameters and M denotes the vector of RDCs that defines the center of the kernel and Σ is covariance matrix. Both X and M vectors are of size k while the Σ is a matrix of size k x k:

$$N(X \bigvee M, \Sigma) = (2\pi)^{-k} \|\Sigma\|^{-k} exp[-\frac{1}{2}(X - M)'\Sigma^{-1}(X - M)] \qquad (6.1)$$

The orientation of the anchor alignment medium [38], [85], [12]is exhaustively searched with respect to the reference structure. Therefore in the second stage the PDP map is calculated for the subject structure in every possible orientation using a grid

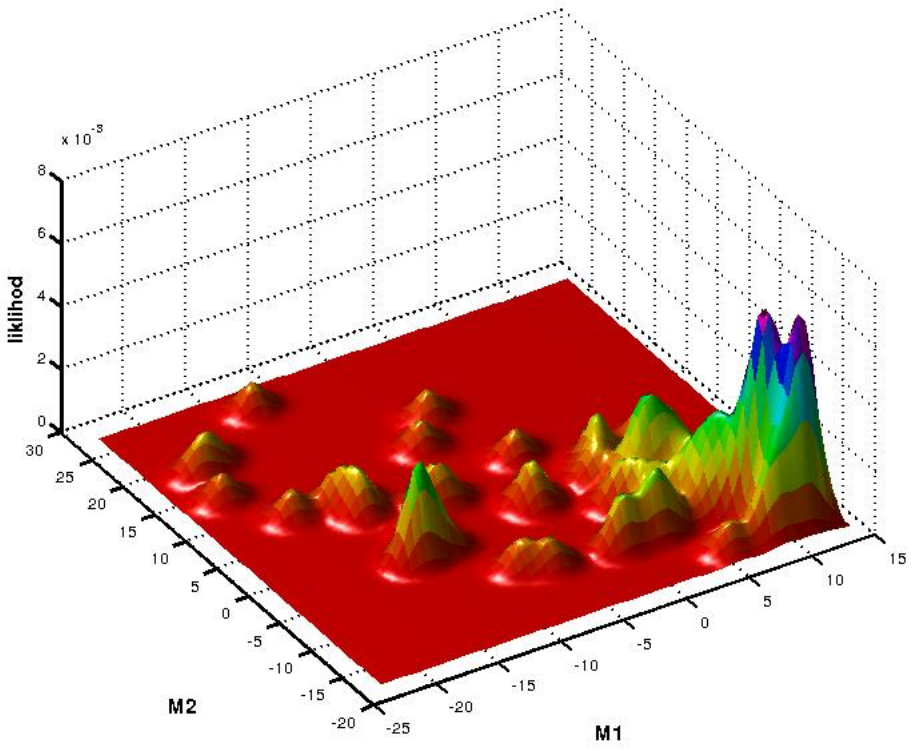


Figure 6.1: An example of a 2D-PDP map, using kernel density estimation. This 2D-PDP can serve as a structural finger print.M1 and M2 denote RDC sets from two alignment media.

search over the Euler angles (α, β, γ) at the resolution of 5deg. The best score is calculated from comparison of the experimental and calculated RDC distributions for all orientations and this score is reported as the final result in the third stage. The process is repeated for every structure in a given library of structures.

6.3 Scoring of the nD-PDPA vs. 2D-PDPA

Manhattan (City Block) metric [42] (shown in Equation 6.2) had been utilized in 2D-PDPA to compare calculated and experimental probability density distributions. In this equation B_{score} denotes the 2D-PDPA score; the summation indexes i and

j cover the entire range of RDCs over alignment media M1 and M2; and $cPDP_{ij}$ (calculated PDP) and $ePDP_{ij}$ (experimental PDP) denote the likelihood of the RDC values at the location i and j. 2D-PDPA utilizes the locations of i,j that are represented by a 64 $\times$ 64 grid. This grid is constructed by uniformly sampling the entire range of both RDC sets for both ePDP and cPDP. Utilization of the grid guarantees the similar intervals and range (begins with minimum RDC and ends with maximum RDC values) for both cPDP and ePDP.

$$B_{score} = \sum_{Min(M1)}^{Max(M1)} \sum_{Min(M2)}^{Max(M2)} |cPDP_{ij} - ePDP_{ij}| \tag{6.2}$$

In order to be qualified for probability density functions, the summation of ePDP and cPDP for the entire range of RDC values are normalized to be zero. Therefore the B_{score} ranges from [0-2]. The B_{score} of 2 refers to completely dissimilar structures and B_{score} of 0 refers to 100% structural similarity. The other factors such as RDC error and availability of data also effects in the B_{score}. Comparison of ePDP and cPDP in a grid fashion is the main contributing factor for the exponential time-complexity of this approach. In that sense, expansion of ePDP and cPDP patterns to n-dimensions requires an exponentially increasing number of grid points (64^n if 64 points along each dimension) to serve as the location of comparisons by a factor of grid size. This quickly becomes a limiting factor for n $>$ 2. Moreover, since RDC data are not uniformly distributed [85], any PDP distribution will contain large areas with a likelihood of zero or near zero. Incision of these unimportant regions for comparison of two PDPAs consumes unnecessary computational time (Figure6.2). The score in nD-PDPA is on the other hand, calculated by comparison of only the information rich regions within the distributions. By using this approach the regions with likelihood of zero or close to zero are not considered for calculation and therefore exponential contribution of the grid size is eliminated.

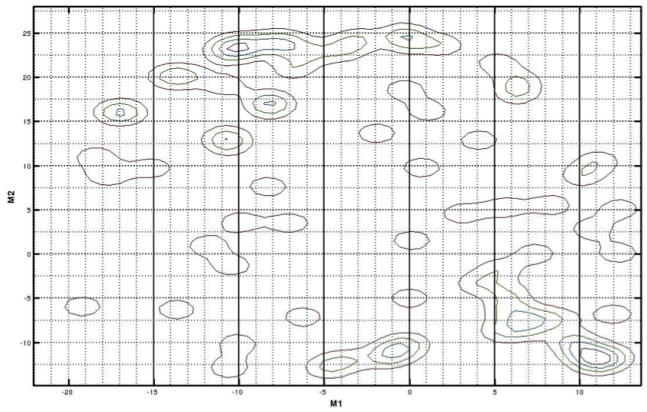


Figure 6.2: 2D-PDPA utilizes a 64 by 64 grid for both computational and experimental RDC sets for scoring. The out of boundaries area are unnecessary for calculation.

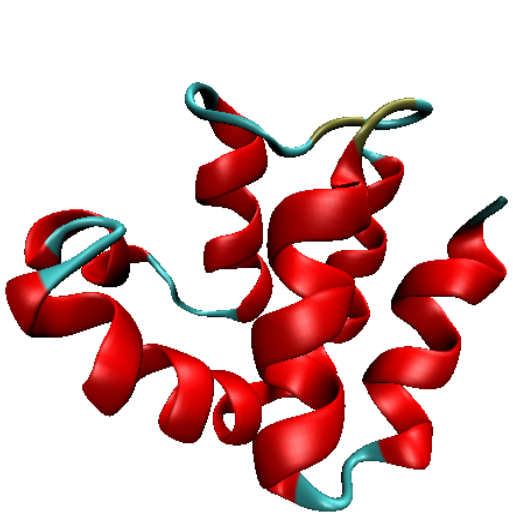
6.4 Results and Discussion

In this section the results of the experiments conducted on various protein structures using nD-PDPA are demonstrated. The listed experiments are categorized into two major groups of experiments utilizing synthetic data and experiments utilizing experimental data. In each category, the preparation of the corresponding data and the objective of the experiment are explained in details.

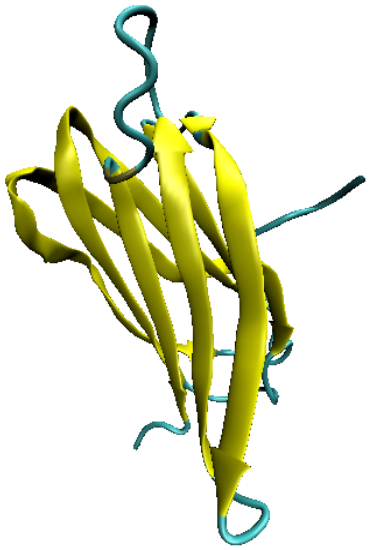
6.5 $nD-PDPA$ analysis utilizing synthetic RDC datasets

6.5.1 Data Preparation

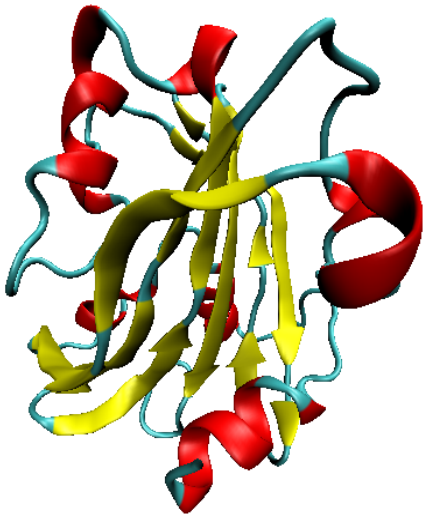
Three structures are shown in table 6.1 are used throughout our experiment. These proteins have been selected on the basis of secondary structure representing distinct secondary structure categories α, β and α/β (Figure 6.3). The atomic coordinates of the structures are obtained from Protein Data-Bank [14]. Three sets of RDCs including N-H vectors representing first alignment medium and N-H and



(a) 1A1Z(83)



(b) 1OUR(114)



(c) 1G1B(164)

Figure 6.3: cartoon representation of the proteins used in the experiment

Table 6.1: List of the protein structures that are used for the experiment.These structures are obtained from Protein Data Bank.

Protein	Secondary Structure	Number of Residues	CATH Classification
1A1Z	α	83	1.10.553.10
1OUR	β	114	2.60.120.40
1G1B	α /β	164	3.40.1410.10

Table 6.2: List of initial order parameters generated by REDCAT that are used to calculate RDC sets.

	Sxx	Syy	Szz	α	β	γ
M1	3e-4	5e-4	-8e-4	0	0	0
M2	-4e-4	-6e-4	10e-4	10	20	30

Cα-Hα vectors representing second alignment medium are generated under two conditions: 1- Ideal RDC sets containing no error, 2- corrupted RDC sets through the addition of $\pm$1Hz of uniformly distributed noise with 25% of RDCs randomly eliminated from each set to better represent pragmatic conditions. The first set (no error) that represents the ideal conditions is utilized for demonstrating the proof of concept and the second set represents a more realistic conditions. To generate synthetic RDC sets the software REDCAT [84] was used with the initial relative order tensors listed in Table 6.2. Upon completion of the data generation, the assignment information is discarded before utilization of the synthetic RDC data in nD-PDPA. To back calculate the order tensors two approaches were used: First the optimal order tensor is calculated using REDCAT and second estimation of the order tensors were conducted using approximation method as described previously [62]. Estimation of the order tensor parameters is of central importance for PDPA analysis in the absence of atomic coordinates of a structure. 2D and 3D approximation software were used [60] [62] to estimate relative order tensors that can be employed in PDPA analysis. In nD-PDPA

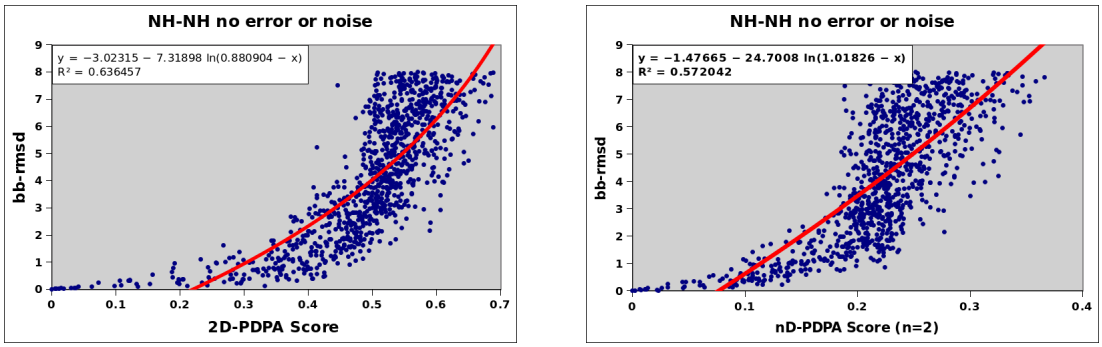
Table 6.3: Order Tensor parameters estimation using 2D-Approx software for the data listed in Table 6.2. The data is corrupted by ±1Hz of error and 25% of the RDCs are removed from datasets.

	Sxx	Sxy	Sxz	Syy	Syz
M1	0.00028	-5.38e-07	4.71e-07	0.00045	1.86e-07
M2	0.00043	-0.00042	0.00023	0.00073	-0.00013

experiments, relative order tensors are estimated utilizing approximation when the data are synthetically corrupted by adding errors (Table 6.3).

6.5.2 The comparison of 2D-PDPA and $nD - PDPA$ results

The objective of this experiment is to establish the relation between nD-PDPA score and bb-rmsd. To accomplish this objective 1000 decoy structures were generated from the native structure by randomly altering the ϕ and ψ angles to generate structures with bb-rmsd in the range of 0-8Å of the native structure. The entire ensemble of the decoy structures was then subjected to evaluation by nD-PDPA. Finally, the scatter plot of bb-rmsd versus nD-PDPA scores were used to observe any significant patterns. Previously the universal funneling effect of such an exercise had been demonstrated [38] for proteins regardless of their structural characteristics. In this experiment, we repeat the previous exercise using nD-PDPA engine and compare some of the results with the previous PDPA program (2D-PDPA). To conduct this experiment, the protein 1A1Z was selected with RDC data generated in REDCAT [84] with no added error. Figure6.4(a) shows the relationship between 2D-PDPA score and bb-rmsd for one thousand decoy structures generated from reference structure 1A1Z. The same experiment was conducted using nD-PDPA engine in Figure6.4(b). As it is mentioned in the previous section, in nD-PDPA engine the comparison for calculated and experimental PDP is based on the RDC points and not by 64 by 64 grid as it is



(a) (b)

Figure 6.4: (a) Calculated nD-PDPA scores vs. bb-rmsd for protein 1G1B using two N-H RDC sets. (b) Calculated nD-PDPA scores vs. bb-rmsd for protein 1G1B using two N-H and Cα-Hα RDC sets.

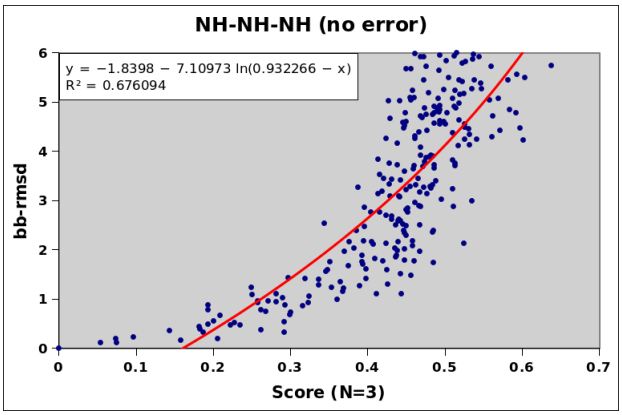
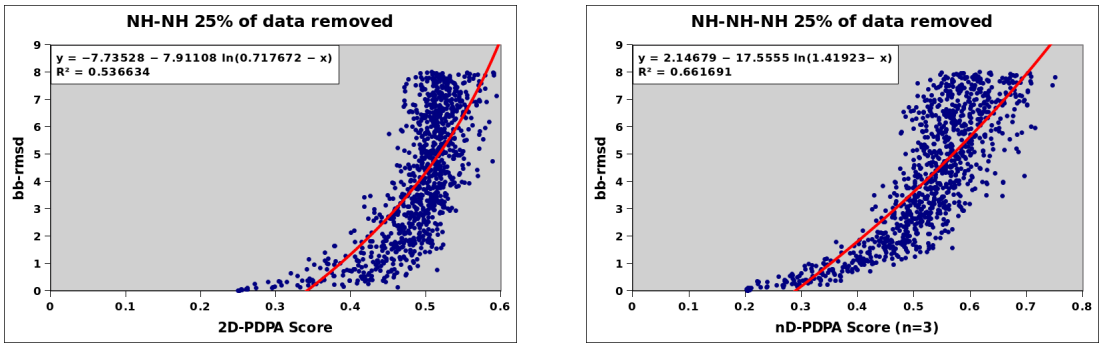


Figure 6.5: The funneling pattern of nD-PDPA score and bb-rmsd of 1000 decoy structures for protein 1A1Z(83). Three N-H RDC sets were used to conduct this experiment.

performed in 2D-PDPA.

This may reduce the sensitivity of the scoring in nD-PDPA compared to the 2D-PDPA (the R^2 fitness for 2D-PDPA is slightly better than R^2 in nD-PDPA analysis in Figure6.4). The lack of sensitivity can be addressed by adding more RDC sets improving the information content of the nD-PDPA analysis and nD-PDPA score fitness. Figure6.5 shows the nD-PDPA analysis using three N-H RDC sets from three alignment media. The R^2 shows improvement, compared to both 2D-PDPA and nD-PDPA analysis using two RDC sets (Figure6.4(a) and (b)). In the previous



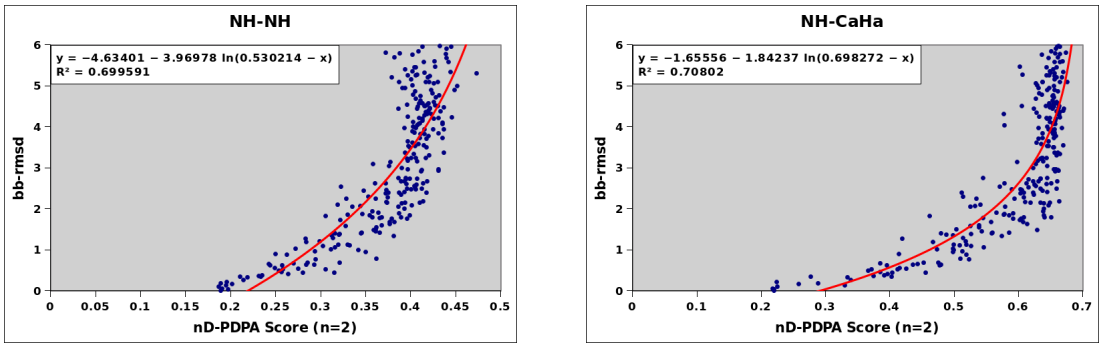
(a) (b)

Figure 6.6: 250 decoy structures (1A1Z as reference) with (a) two N-H RDC sets (2D-PDPA) (b) three N-H RDC sets (nD-PDPA). 25% of the data were randomly removed from each set.

experiment, the improvement of R^2 in ideal conditions by adding more RDC data was demonstrated. It is useful to utilize RDC data that is closer to the experimental conditions. The Previous experiment was repeated by randomly removing 25% of the RDC values for protein 1A1Z. In Figure 6.6(a) and (b), the results of 2D-PDPA and nD-PDPA (using three sets of N-H RDC vectors) analysis are demonstrated. The score and bb-rmsd fitness score shows a better correlation in the case of nD-PDPA (n=3) analysis.

6.5.3 The improvement of the fitness score by adding extra RDC sets

In the previous experiment, protein 1G1B(164) was used, and RDC datasets were corrupted by the addition of ±1Hz of uniformly distributed error and randomly removing of 25% of the RDC values from each set. 250 structures were generated by altering backbone torsion angles range from 0-6Å from 1G1B. In Figure6.7(a) nD-PDPA analysis was conducted by using two N-H RDC sets from two alignment media and in Figure6.7(b) the analysis was conducted by utilizing two non-homogeneous RDC sets, N-H and Cα-Hα from two alignment media respectively. R^2 values for both exper-



(a) (b)

Figure 6.7: (a) Calculated nD-PDPA scores vs. bb-rmsd for protein 1G1B using two N-H RDC sets. (b) Calculated nD-PDPA scores vs. bb-rmsd for protein 1G1B using two N-H and Cα-Hα RDC sets.

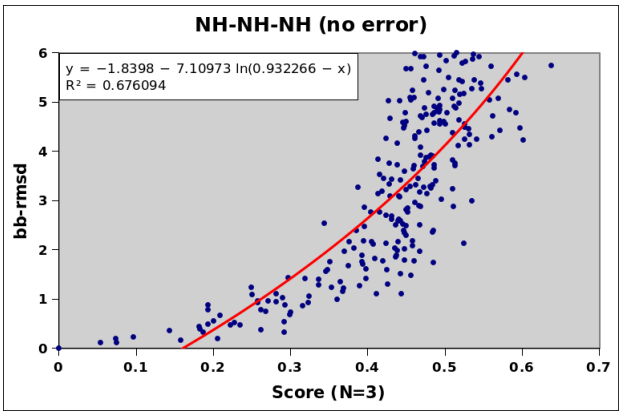


Figure 6.8: Calculated nD-PDPA scores vs. bb-rmsd for protein 1G1B utilizing three sets of RDC. Two of which are N-H sets and the third one is Cα-Hα. R^2 value is improved in comparison to utilization of two RDC sets.

iments are approximately similar (approximately 0.7). Figure6.8 shows the plot of nD-PDPA analysis for protein 1G1B by adding Cα-Hα as the third RDC set to the collection of two N-H RDC sets. The $R^2 = 0.8146$ indicates the improvement of nD-PDPA analysis fitness compared to two sets of RDCs (Figure6.6(a) and (b)). The experiment once again confirms the improvement of the bb-rmsd and PDPA score fitness by adding more RDC data. Tables 6.4 and 6.5 demonstrate the R^2 value for two of the proteins listed in Table 6.3. In Tables 6.4 and 6.5 in all instances the value of the R^2 decreases by introducing error to the data (compare second and third col-

Table 6.4: Summary of the results of 1OUR using RDC sets without any error (second column) and RDC sets with ±1Hz error and 25% of RDC values randomly removed(third column)

Protein 1OUR	R^2 no error	R^2 error
NH-NH	0.5781	0.4522
NH-CαHα	0.6797	0.4935
NH-NH-NH	0.7273	0.7026
NH-NH-CαHα	0.6573	0.6547

Table 6.5: Summary of the results of 1G1B using RDC sets without any error (second column) and RDC sets with ±1Hz error and 25% of RDC values randomly removed(third column)

Protein 1G1B	R^2 no error	R^2 error
NH-NH	0.6924	0.6995
NH-CαHα	0.7204	0.7080
NH-NH-NH	0.7494	0.7575
NH-NH-CαHα	0.8029	0.8146

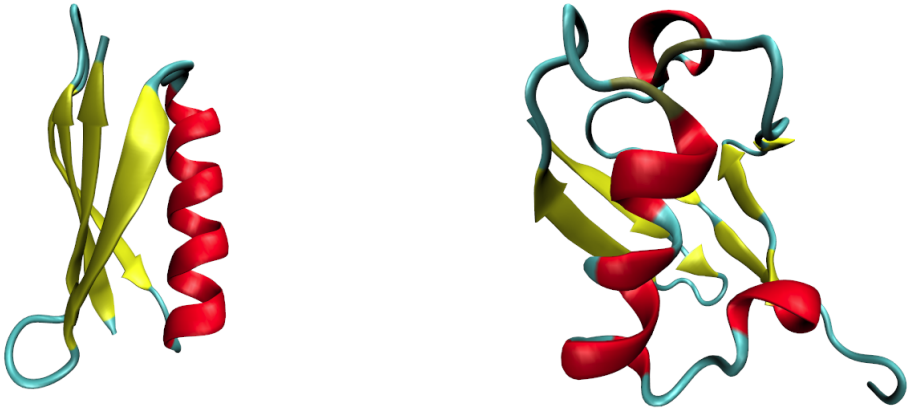
umn of Tables 6.4 and 6.5). Moreover, the value of R^2 increases by adding an extra set of RDC. The improvement is manifested in the greater degree in the erroneous data sets. For example in Table 6.4 R^2 for three N-H RDC sets demonstrates about 0.35 improvement, compared to two N-H RDC sets (See the darker background cells in Table 6.4).

6.6 $nD-PDPA$ ANALYSIS UTILIZING EXPERIMENTAL RDC DATA SETS

Two structures shown in table 6.6 and 6.9 are used throughout our experiment. These proteins have been selected on the basis of the availability of the experimental data in BMRB database [82] [35] with backbone RDC data from two or more alignment media.

Table 6.6: Protein structures that are obtained from BMRB database based on availability of experimental RDC.

Protein	Secondary Structure	Number of Residues	CATH Classification
1P7E	α/β	56	3.10.20.10
1D3Z	α/β	76	3.10.20.90



(a) 1P7E(56)

(b) 1D3Z(76)

Figure 6.9: Cartoon representation of the proteins used in the experiment

Table 6.7: This table shows the QFactor for 5 N-H RDC sets for protein 1P7E.

Alignment Media	QFactor
M1	0.0232
M2	0.0286
M3	0.0393
M4	0.0259
M5	0.0325

6.6.1 Data Preparation for Protein 1P7E

Five sets of N-H RDC sets are available for protein 1P7E. Only 71% of the data are presented in each set.The QFactor of each set of RDC are listed in Table 6.7.

In order to estimate the order tensor values, 2D and 3D approximation software (2Dapprox [62] [97]) are used. Table 6.8 shows the relative Order Tensor estimated

Table 6.8: List of relative Order Tensor values estimated for M1 and M2 with respect to M1. Also M3 and M4 with respect to M3 for protein 1P7E.

	Sxx	Sxy	Sxz	Syy	Syz
M1	7.75E-05	0	0	7.76E-05	0
M2	-2.55E-05	7.09E-05	3.56E-04	4.68E-04	9.28E-06
M2	2.18E-04	0	0	2.18E-04	0
M3	-4.61E-04	2.02E-04	1.24E-05	-1.50E-04	6.76E-04

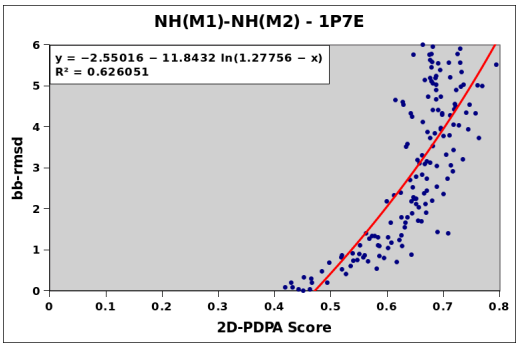
Table 6.9: The estimation of the Order Tensor values for three sets of experimental N-H RDC using 3DApprox software for protein 1P7E.

	Sxx	Sxy	Sxz	Syy	Syz
M1	0.0007	0	0	0.001	0
M2	0.0002909	6.21E-05	0.0003189	0.000408	5.44E-05
M3	0.0005063	0.0004637	-0.0002655	0.0002289	-0.0002664

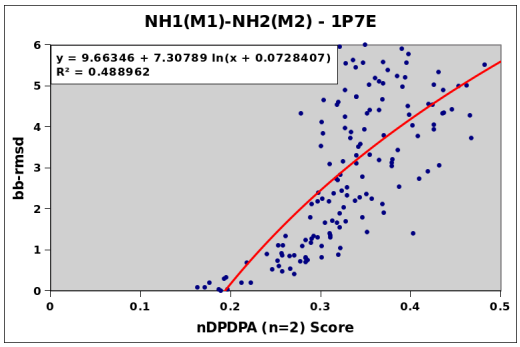
for two alignment media M1 and M2. In this case M1 is considered as reference frame which coincides with Molecular Frame(MF) therefor it is diagonalized and M2 is estimated with respect to M1 molecular frame. Likewise M2 is considered as reference frame and M3 is estimated with respect to M2. Table6.9 listed the estimated relative Order Tensor values for M1 and M2 and M3 using 3DApprox software. The Order Tensor values for M2 and M3 are calculated with respect of M1 where M1 is assumed to be coincided with Molecular Frame(MF).

6.6.2 The improvement of fitness score by adding extra RDC sets for protein 1P7E

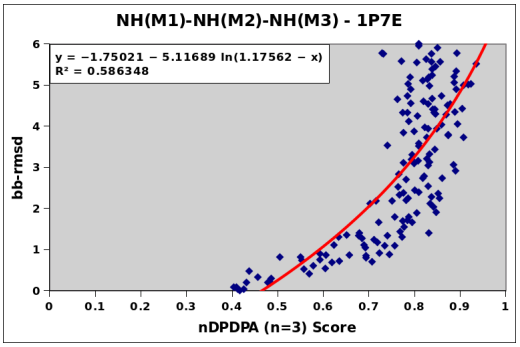
The objective of this experiment is to confirm the relation between nD-PDPA score and bb-rmsd and to compare the result of 2D-PDPA with nD-PDPA utilizing experimental data. Moreover, the improvement of the nD-PDPA analysis by adding an extra set of RDC is investigated.



(a)



(b)

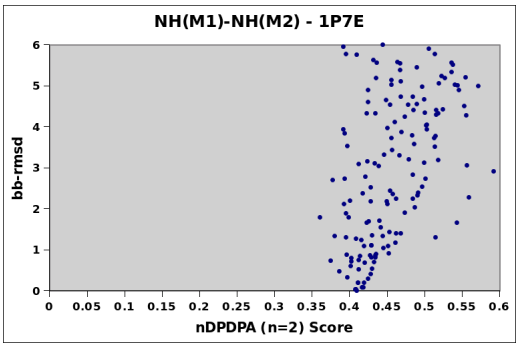


(c)

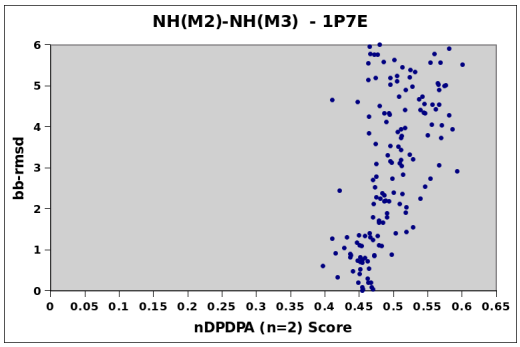
Figure 6.10: The plots of nD-PDPA and 2D-PDPA analysis using 250 decoy structures for protein 1P7E. All RDC sets are experimental and Order Tensor values are calculated using REDCAT software.(a) 2D-PDPA analysis using { NH, NH } vectors from two alignment media;(b)nD-PDPA analysis using { NH, NH } vectors from two alignment media;(c)nD-PDPA analysis using { NH, NH, NH } vectors from three alignment media(M1, M2 and M3)

A dataset of 250 decoy structures is generated by altering ϕ and ψ back bone torsion angles in the range of 0 to 6Å.The generated datasets are utilized throughout all experiments described here.

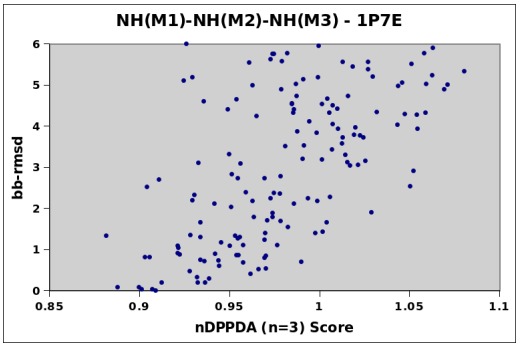
In Figure 6.10(a) shows the relationship between 2D-PDPA score and bb-rmsd of decoy structures. The same experiment was conducted using nD-PDPA engine in 6.10(b). Generally the fitness score of nD-PDPA for n=2 does indicate any better score than the 2D-PDPA, this is due to lack of sensitivity in the lower dimension by not utilizing the 64 by 64 grid in nD-PDPA. However, addition of extra RDC sets improve the fitness as it is shown in 6.10(c) Figure 6.11 demonstrates the relationship between nD-PDPA score and bb-rmsd for 250 decoy structures generated from refer-



(a)



(b)



(c)

Figure 6.11: The plots of nD-PDPA analysis utilizing 250 decoy structures for protein 1P7E. All RDC sets are experimental and Order Tensor values are estimated using 2D and 3D approximation software. (a){ NH-NH } RDC vectors from two alignment media (M1 and M2);(b){ NH-NH } RDC vectors from two alignment media (M2 and M3);(c){ NH-NH } RDC vectors from three alignment media (M1, M2 and M3);

ence structure 1P7E. The utilized RDC sets are experimental and the Order Tensors are estimated using 2D and 3D Approx software. It is clear that the funneling effect that is expected for such analysis does not exist in any of the experiments in Figure 6.11. The experimental error and inaccurate estimation of the Order Tensors could possibly contribute in this . However, closer investigation reveals more information.

Tables 6.10 and 6.11 show ten best structures ranked by nD-PDPA score for two N-H vectors in Figure 6.11(a) and three N-H vectors in Figure 6.11(c) experiments.In Table 6.10, nD-PDPA results indicate seven structures out of ten under 3Å and one structure is 5.9Å away from reference structure. In Table 6.11 the improvement of the analysis is clear by adding an additional set of RDC resulting all first ten structures below 3Å with respect to the subject protein 1P7E.

Table 6.10: The nD-PDPA scores for the first ten structures of Figure 6.11(b). The RDC sets that are used in this experiment are M1 and M2. Both RDC sets are N-H vectors.

		α	β	γ		score	bbrmsd
1	*struct*_33	160	165	95	z	0.3606	1.778
2	*struct*_95	110	145	50	z	0.3752	0.733
3	*struct*_62	35	160	160	0	0.3778	2.7034
4	*struct*_138	125	130	70	z	0.3803	1.335
5	*struct*_124	160	20	135	y	0.3870	0.472
6	*struct*_54	85	110	70	x	0.3916	5.949
7	*struct*_10	170	25	115	x	0.3916	3.925
8	*struct*_100	85	125	20	0	0.3934	2.118
9	*struct*_84	115	140	40	z	0.3939	2.732
10	*struct*_87	120	140	20	z	0.3942	3.828

Table 6.11: The nD-PDPA scores for the first ten structures 6.11(c). The RDC sets are used in this experiment are M1, M2 and M3. All RDC sets are N-H vectors.

		α	β	γ		score	bbrmsd
1	*struct*_138	170	25	175	x	0.8817	1.335
2	*struct*_1	150	155	25	x	0.8880	0.079
3	*struct*_97	150	155	25	x	0.8997	0.085
4	*struct*_110	80	5	95	y	0.9014	0.034
5	*struct*_145	115	165	120	z	0.9031	0.805
6	*struct*_34	70	15	170	0	0.9041	2.511
7	*struct*_78	135	165	10	x	0.9056	0.805
8	*struct*_17	80	5	95	y	0.9073	0.031
9	$1P7E$	80	5	95	y	0.9091	0
10	$struct_6 2$	170	170	55	x	0.9109	2.704

Table 6.12: List of relative Order Tensor values estimated for M1 and M2 with respect to M1. Also M3 and M4 with respect to M3 for protein 1P7E.

	Sxx	Sxy	Sxz	Syy	Syz
N-H(M1)	0.00028	0	0	0.00045	0
N-H(M2)	0.00043	-0.00042	0.00023	0.00073	-0.00013
Cα-Hα(M2)	-0.00045	0.00046	-0.00027	-0.00084	0.00016
N-C(M2)	0.00053	-0.00052	0.00027	0.00089	-0.00015

6.6.3 Data Preparation for Protein 1D3Z

For protein 1D3Z, available RDC vectors are C-N, N-H, C-H and Cα-Hα from two alignment media (total of 8 RDC sets). Table6.12 lists relative Order Tensors estimation where N-H RDC set from first alignment medium(M1) is considered to be Molecular Frame(MF). Only 71% of the RDC data are available and the analysis of the RDC sets indicate up to $\pm$3.7Hz of noise.

6.6.4 The improvement of fitness score by adding extra RDC sets for protein 1D3Z

In the following experiments protein 1D3Z is used as reference structure. A dataset of 250 decoy structures is generated by altering ϕ and ψ back bone torsion angles in the range of 0 to 6Å.The generated datasets are utilized throughout all experiments described here. The availability of variety of RDC sets (see the previous Section) allow us to produce combination of RDC sets in the following experiments. Figure 6.12 demonstrates the result of nD-PDPA experiments on protein 1D3Z. Figure 6.12(a) shows the R^2 of 0.4 for two N-H RDC vectors. Combination of N-H, Cα-Hα in Figure 6.12(b) and N-H, C-N in Figure 6.12(c) demonstrates slightly improved R^2 comparing with N-H, N-H vectors, specially for the structures below 2Å away from 1D3Z. However, in Figure 6.12(d) the value of R^2 demonstrates the fitness score, comparing with other three experiments. This result once again confirms the improvement of the

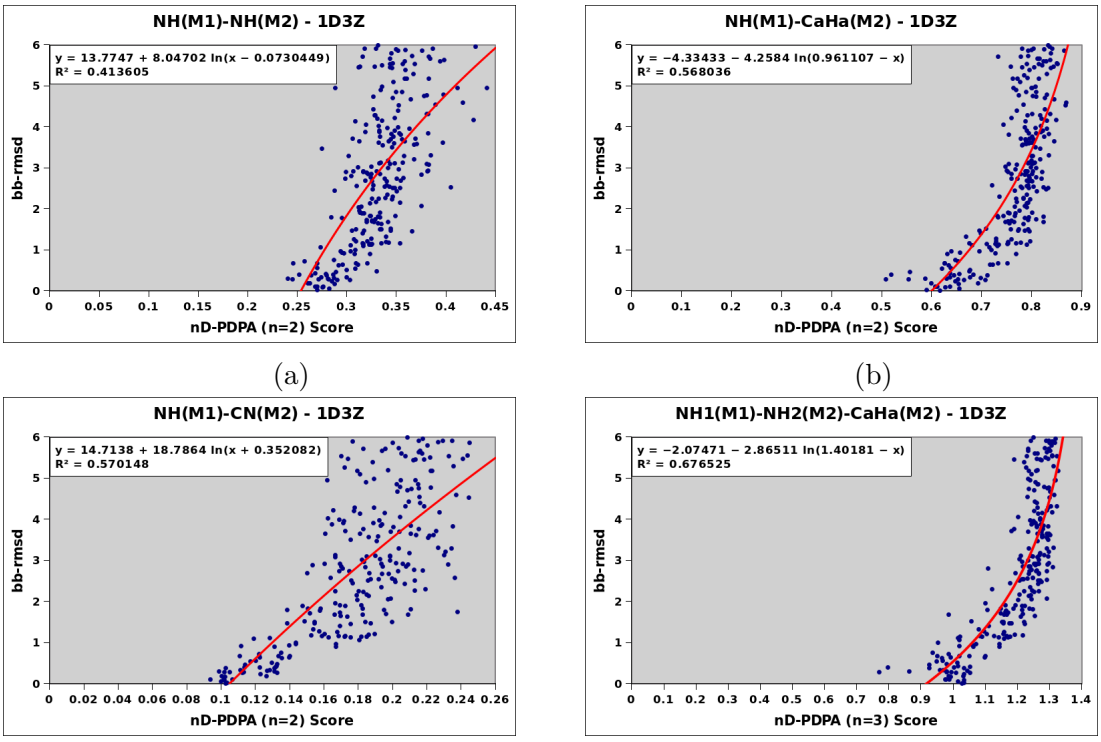


Figure 6.12: The plots of nD-PDPA analysis utilizing 250 decoy structures for protein 1D3Z. All RDC sets are experimental and Order Tensor values are estimated using 2D and 3D approximation software (a) { NH-NH } RDC vectors from two alignment media;(b){ NH, CαHα } RDC vectors from two alignment media;(c) { NH, CN } RDC vectors from two alignment media;(d){ NH, NH, CαHα } from two alignment media.

fitness score of nD-PDPA by adding extra set of RDC as were demonstrated before for synthetic datasets.

Chapter 7

Structure refinement using $nD - PDPA$

7.1 Introduction

The process of correcting the structural discrepancies and improving the overall structural qualities of a modeled structure to bring it as close as possible to its native structure is known as protein structure refinement. Practically, the process of refinement includes simultaneous improvement in protein backbone geometry, irregular hydrogen bonds, atomic collisions, irregular bond length, erroneous torsion angles, side-chain displacement. To satisfy the most of these restraints (if not all), often a considerable amount of experimental data are required. In this section, the application of the nD-PDPA method in the refinement of modeled protein structure is discussed. Since our refinement method is based on the nD-PDPA engine, unassigned RDC sets are the only requirement and the primary focus of the method is on back-bone refinement. The application of the method is tested on both synthetic and experimental datasets. Also, to measure the performance and the accuracy of the nD-PDPA refinement method, the results are compared to Xplor-NIH software.

7.2 Refinement process utilizing $nD - PDPA$ engine

The capability of PDPA to rank a decoy set of structures relative to an unknown protein is discussed previously(For more information see Section 5.4.5).

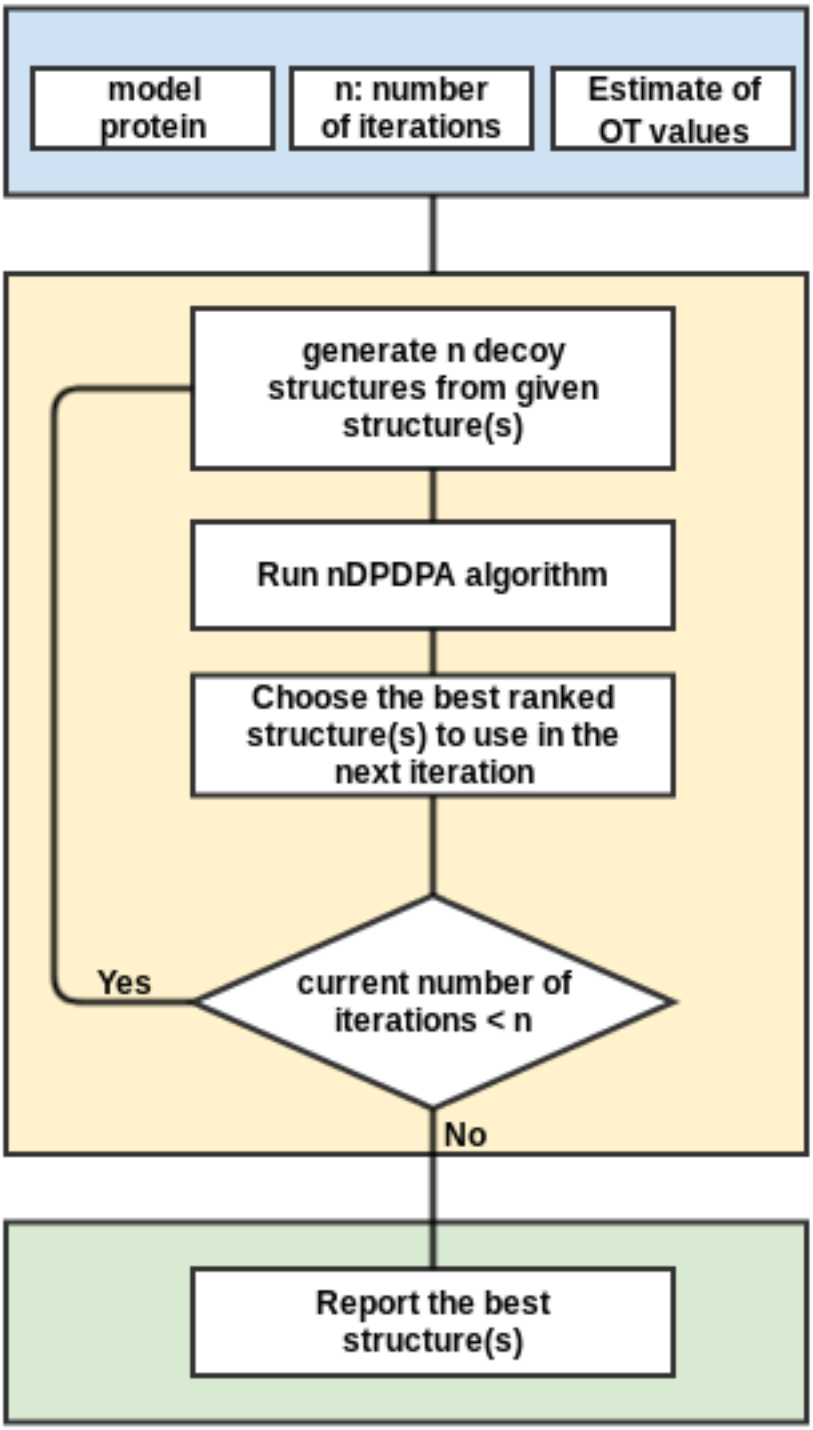


Figure 7.1: Operation schematic of refinement illustrated in three main stages.

The overall operation of the refinement proceeds in three main stages as shown in Figure 7.1. In the first stage, the relative order tensor parameters are calculated. The program utilizes a modeled protein (the protein that needs to be refined) and the number of iterations (variable n) as the input parameters (the blue box in Figure 7.1). In the next stage, decoy structures are generated (See Appendix A) from the modeled structure . Then, the decoy structures set is ranked by the nD-PDPA. The structure with the lowest PDPA score is selected and used as the reference structure for the next iteration. This process is repeated until the iteration count reaches to n (the yellow box in Figure 7.1). In the last stage, the final structure in each iteration is reported as the potential refined structures(the green box in Figure 7.1).

7.3 Results and Discussion

7.3.1 Refinement of structure using synthetic data

The protein 1A1Z is used to conduct the first experiment (for more information about this protein see Section 6.4 and Table 6.1). The objective of the following experiment is to track of the nD-PDPA score for each iteration to establish the feasibility of the nD-PDPA refinement method. The utilized RDC sets consist of two sets of N-H RDC vectors with no error added to the set to accommodate the ideal condition.

Figure 7.2 shows the refinement process for the protein 1A1Z. The red dots denote the best ranked structure in each iteration. Then the best structure is used as the reference to generate decoy set for the next iteration. This process is repeated for six iterations. In this experiment, a modeled structure (from the decoy structure dataset) that is 2.847Å away from 1A1Z is selected to serve as starting refinement point. Table 7.1 shows the details information recorded for each round of refinement. The bb-rmsd is improved from 2.847Å in Run0 to 1.953Å in Run5. The final Run (Run5) indicates

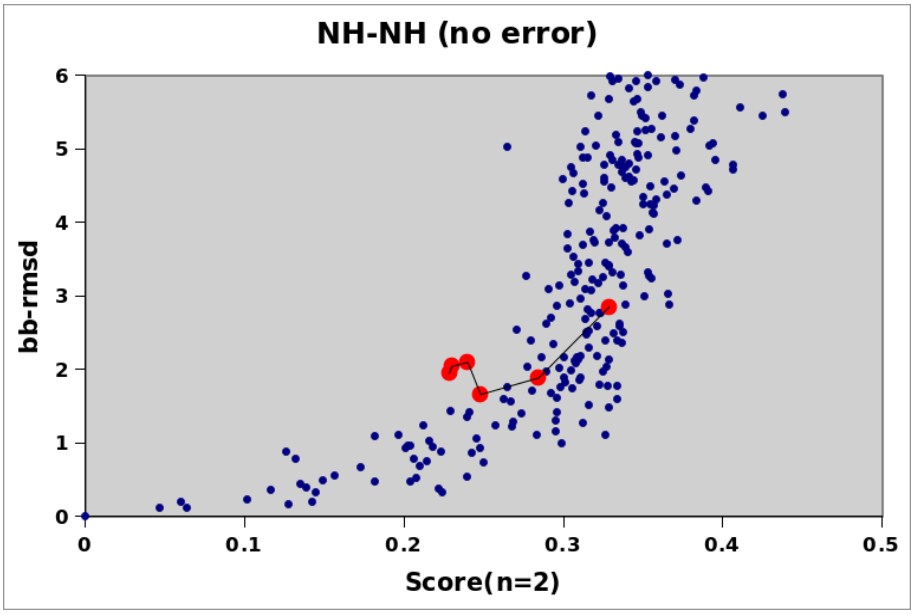


Figure 7.2: The refinement process of a modeled structure that is 2.847Å away from 1A1Z. The red dots denotes the structure with the best nD-PDPA score at each round. Totally this refinement ran in six iterations. The final structure is approximately 1.9Å away from 1A1Z. Two N-H RDC sets with no error is used for this experiment.

Table 7.1: The result of refinement from Figure 7.2 is listed here. The second column shows the iteration (Run) number. 7th column shows the nD-PDPA score for the best structure in each iteration.

		α	β	γ		score	bbrmsd
1	Run0	55	10	140	z	0.3286	2.847
2	Run1	130	5	60	z	0.2845	1.884
3	Run2	60	5	130	z	0.2480	1.658
4	Run3	40	5	150	z	0.2400	2.095
5	Run4	40	5	150	z	0.2299	2.041
6	Run5	40	5	150	z	0.2284	1.953

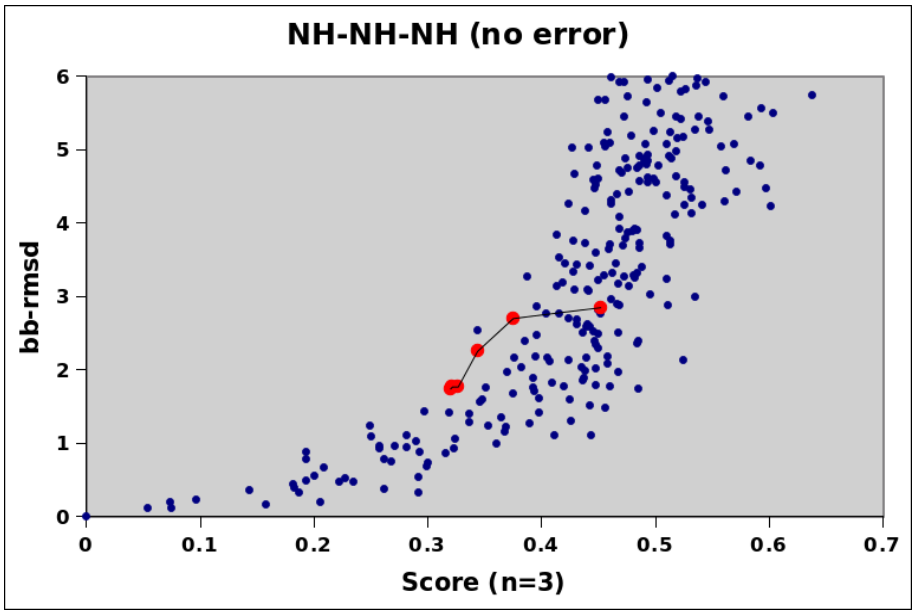


Figure 7.3: The refinement process of a modeled structure that is 2.847Å away from 1A1Z. The red dots denotes the best structure at each iteration. Totally this refinement ran in 6 iterations.Final structure is about 1.7Å away from 1A1Z. Three N-H RDC sets with no error is used for this experiment.

Table 7.2: The detail refinement process result from Figure 7.3. Column one denotes the iterations number followed by the best candidate structures name in column two. Columns three to six indicate the orientation of the molecule and the rotation axis in which the best nD-PDPA score produced. Column seven is nD-PDPA score for each iteration and the last column is the bb-rmsd with respect to 1A1Z protein.

		α	β	γ		score	bbrmsd
1	Run0	75	15	125	z	0.4517	2.847
2	Run1	60	15	135	z	0.3751	2.699
3	Run2	75	5	115	z	0.3436	2.251
4	Run3	75	10	115	z	0.3266	1.776
5	Run4	75	10	115	z	0.3217	1.767
6	Run5	75	10	115	z	0.3201	1.741

the improvement of the nD-PDPA score (about 0.1) and about 1Å of the bb-rmsd respectively. However, the middle iterations exhibit inconsistency in nD-PDPA score in relation with bb-rmsd. For example in Run2 to Run3 the bb-rmsd decreases while the nD-PDPA score improves for both runs. A number of reasons explain this

anomaly. First, it is possible the structures with the same bb-rsmd produce different nD-PDPA score.This fact was observed previously in the funneling pattern of bb-rmsd and nD-PDPA(Figure 7.2 and see Section 5.3). Second, generating structures randomly (see Appendix A to get more information) to form a decoy dataset, do not guarantee to produce the better quality structure than the previous iteration. The second reason suggests, if the iterations are repeated sufficiently large, there will be a higher probability of producing the high-quality structures. The same experiment has been repeated for the protein 1A1Z using NH(M1)-NH(M2)-NH(M3) RDC vectors with no error introduced to the data. The first reference structure is 2.847Å away from 1A1Z (the same structure as the previous experiment). Figure 7.3 demonstrates the plot of each iteration, and Table 7.2 shows the same results in the numerical format. The final result after five iterations indicates the improvement in both bb-rmsd from 2.847Å to 1.741Å and nD-PDPA score about 0.13 respectively. The addition of an extra set of RDC reduced the bb-rmsd of the final refined structure to 1.741Å that demonstrates about 0.2Å of improvement compare to the previous experiment.

7.3.2 Refinement of structure using experimental data

The similar datasets that have been used in Section 6.5.1 are utilized in this experiment (see Table 6.1). Figure 7.4 shows the plot of iteration steps for the protein 1D3Z using NH(M1)-NH(M2)-CH(M2) RDC datasets. The order tensor values are estimated using 3DApprox software as we described in Section 6.6.3. In this experiment, the refinement iterations were repeated 22 times. Table7.3 demonstrates the nD-PDPA scores along with bb-rmsd with respect to protein 1D3Z. The refinement process produced structural improvement of about 0.7Å from 2.834Å to 2.18Å. There is not a significant improvement from Run14 to the end of Run22. A number of reasons contribute to this inefficiency. First, the utilized experimental RDC sets are unassigned and contain error and missing data. The RDC error affects the order

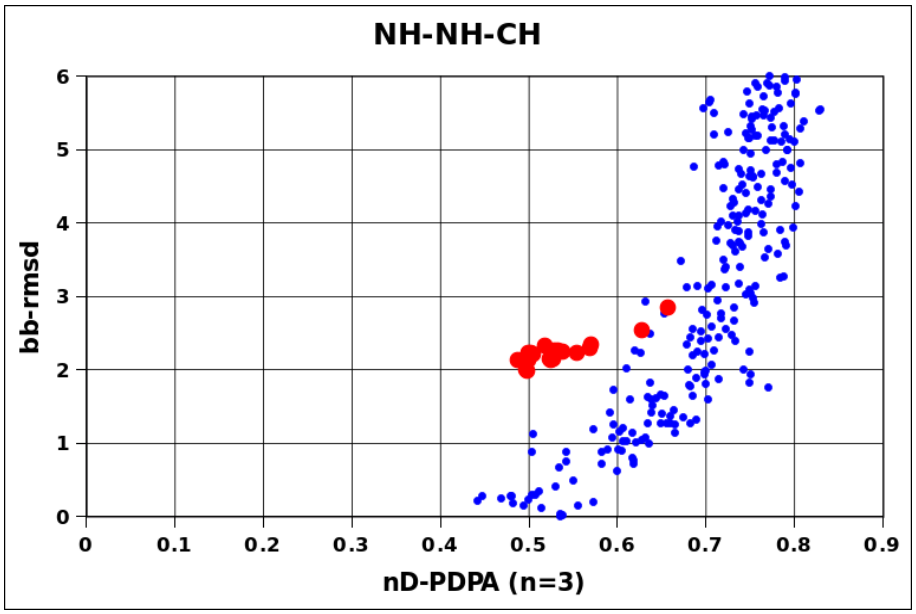


Figure 7.4: The refinement process of a structure that is 2.843Å away from protein 1D3Z. The experiment is repeated 22 times using experimental RDC sets.

tensor estimation in the absence of the subject structure. Second, there is no guarantee to produce higher quality structures that exhibit better RDC fitness in a decoy structure set in the process of decoy structure generation. Third, the selection of only one candidate for the next iteration from the ranked structures list is not an efficient way while it is possible that a structure with the lower bb-rmsd stays in the higher ranking score. Such a problem is addressed in the next section, by the selection of the n-Best structures from the ranked structures list. In the next experiment protein, 1P7E is selected. In this experiment three N-H RDC vectors from three alignment media are utilized.

Figure 7.5 demonstrates the refinement steps (red dots) and Table 7.4 shows the detail of the steps. The first structure is 2.911Å away from 1P7E. In the Run1, there is a leap of the bb-rmsd to 3.034 but as the iterations go forward, the bb-rmsd shows improvement. In the Run9 and 10 the results are the same (gray shaded rows) indicating no improvement in the Run10. The lowest bb-rmsd is 1.989Å (Run11). In all rows, the nD-PDPA score is improved. This indicates a constant improvement

Table 7.3: The detail refinement process result from Figure 7.4. Column one denotes the iterations number followed by the best candidate structures name in column two. Columns three to six indicate the orientation of the molecule and the rotation axis in which the best nD-PDPA score produced. Column seven is the nD-PDPA score for each iteration and the last column is the bb-rmsd with respect to 1D3Z protein.

		α	β	γ		score	bbrmsd
1	Run1	65	160	55	y	0.6568	2.843
2	Run2	85	155	70	y	0.5689	2.297
3	Run3	85	155	70	y	0.5541	2.22
4	Run4	85	155	70	y	0.5391	2.246
5	Run5	85	155	70	y	0.5311	2.246
6	Run6	85	155	70	y	0.5284	2.148
7	Run7	85	155	70	y	0.5255	2.135
8	Run8	85	155	70	y	0.5237	2.146
9	Run9	85	155	70	y	0.6284	2.537
10	Run10	85	155	70	y	0.5702	2.337
11	Run11	85	155	70	y	0.5326	2.258
12	Run12	85	155	70	y	0.5290	2.25
13	Run13	85	155	70	y	0.5186	2.329
14	Run14	85	155	70	y	0.5052	2.203
15	Run15	85	155	70	y	0.5027	2.221
16	Run16	85	155	70	y	0.4998	2.221
17	Run17	85	155	70	y	0.4879	2.125
18	Run18	85	155	70	y	0.4997	2.204
19	Run19	85	155	70	y	0.4999	2.12
20	Run20	85	155	70	y	0.4988	1.988
21	Run21	85	155	70	y	0.4977	1.991
22	Run22	85	155	70	y	0.4757	2.18

in RDC fitness to the produced structure that necessarily do not improve the bb-rmsd. Figure 7.6 demonstrates superimpose of four structures from Table 7.4. The structures include Run0 in blue, Run6 in red, Run8 in purple and Run10 in green. The reference structure, protein 1P7E also is added to this image in cyan. In each iteration, by improvement the nD-PDPA score the secondary structures are improved as it is clear in Run10(green).

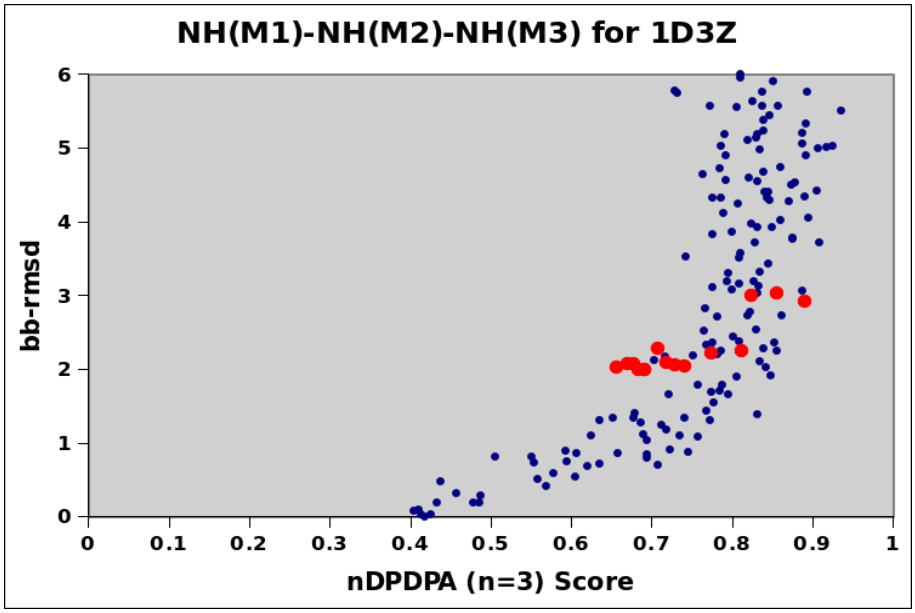


Figure 7.5: The refinement plot of a structure that is 2.911Å away from the protein 1P7E. The experiment is repeated 15 iterations.

Table 7.4: The detail refinement process result from Figure 7.5. Column 1 denotes the iterations number followed by the run number in column 2. Columns 3 to 4 indicate the orientation angles of the molecule and the rotation axis in which the best nD-PDPA score produced. Column 5 is the nD-PDPA score for each iteration and the last column is the bb-rmsd with respect to 1P7E protein.

		α	β	γ		score	bbrmsd
1	Run0	55	125	35	0	0.8906	2.911
2	Run1	60	145	70	0	0.8551	3.034
3	Run2	30	130	50	x	0.8232	2.993
4	Run3	130	50	130	z	0.8111	2.24
5	Run4	55	130	135	z	0.7734	2.218
6	Run5	125	45	140	z	0.7409	2.044
7	Run6	125	45	140	z	0.7292	2.051
8	Run7	165	45	105	z	0.7187	2.085
9	Run8	165	45	105	z	0.7072	2.274
10	Run9	165	45	105	z	0.6907	1.993
11	Run10	165	45	105	z	0.6907	1.993
12	Run11	165	45	105	z	0.6832	1.989
13	Run12	165	45	105	z	0.6776	2.077
14	Run13	165	45	105	z	0.6691	2.076
15	Run14	165	45	105	z	0.6562	2.03

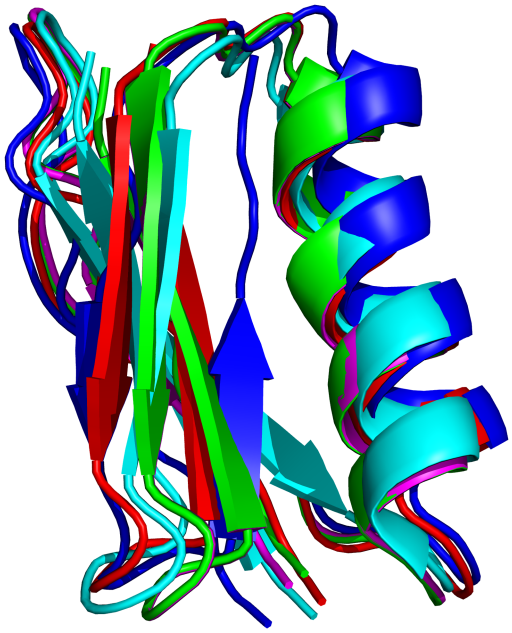


Figure 7.6: The superimpose of four structures from Table7.4.The structures include Run0 in blue, Run6 in red, Run8 in purple and Run10 in green. The reference structure, protein 1P7E also is added in cyan.

7.3.3 $n - best$ STRUCTURE REFINEMENT

In the refinement process, a structure with the lowest nD-PDPA score is considered to be the best structure and is used as the reference to generate the decoy structures for the next iteration. However, often the selection of one structure as the reference for the next round is not the best choice for a number of reasons. First, it is possible that structures with the higher nD-PDPA score exhibit, the lower bb-rmsd with respect to subject structure. Second, erroneous data is the source of inaccurate estimation of order tensor that affects the nD-PDPA scores and, therefore ranking. Moreover, it is possible in the nD-PDPA ranking list, the difference of nD-PDPA score for two high ranked structures is minuscule (for example the difference of 0.01) and two structures exhibit approximately similar secondary structure quality. In such

Table 7.5: The result of ranking of an refinement iterations. Run1 shows 3.051Å while Run2 shows the better bb-rmsd. In this case structures from row2 to row 12 are selected as reference structures for the next refinement iteration.

		α	β	γ		score	bbrmsd
1	Run1	100	160	125	y	0.3539	3.051
2	Run2	90	165	115	y	0.3572	2.898
3	Run3	95	160	115	y	0.3574	2.631
4	Run4	95	160	115	y	0.3589	2.495
5	Run5	95	160	115	y	0.3633	2.702
6	Run6	95	160	115	y	0.3646	2.704
7	Run7	95	160	115	y	0.3676	2.546
8	Run8	95	160	115	y	0.3687	2.485
9	Run9	95	160	115	y	0.3689	2.678
10	Run10	95	160	115	y	0.3697	2.693
11	Run11	95	160	120	y	0.3697	2.892
12	Run12	95	160	115	y	0.3698	2.675

a case considering the second structure as a potential candidate along with the first one will improve the likelihood of generating the better quality decoy structure set.

To achieve this goal, n is defined as the rank of the reference structure in a refinement process. For example in Table 7.5, the reference structure is in rank 13 therefore n = 13(not shown in the table). In this case, the first 12 structures can be utilized as a reference set for the next iteration since all of them have the better nD-PDPA score compare to the reference structure in row 13. This approach improves the process of nD-PDPA refinement by offering more valid structures as references to generate decoy set in the next iteration. Table 7.6 demonstrates the results of the nD-PDPA refinement for the protein 1D3Z using the similar data to the experiment shown in Table 7.3. In this experiment, a set of the best structures is selected as references for the following iterations. For example, all structures labeled as Run1 are the best score structures up to the reference in Run1 and are the subject of generating decoy structure for the Run2 (next iteration). The gray shaded rows denote the best score for each set that are superimposed in Figure 7.7. The result exhibits

Table 7.6: The result of nD-PDPA refinement using n-Best structures selection method. Each group is separated by a blank line, indicate a refinement iteration. The gray shaded structures indicate the best structure among n best structure in each iteration. The structure refinement shows improvement of 1.5Å.

		α	β	γ		score	bbrmsd
1	Run0	65	160	55	y	0.6568	2.843
2	Run1	85	165	70	y	0.6216	2.508
3	Run1	85	165	75	y	0.6263	2.799
4	Run1	70	155	60	y	0.6412	2.321
5	Run1	85	160	70	y	0.6513	2.36
6	Run1	55	160	35	y	0.6354	2.795
7	Run1	85	165	75	y	0.6565	2.651
8	Run2	75	155	70	y	0.6146	2.057
9	Run2	85	165	70	y	0.6167	2.534
10	Run3	85	150	80	y	0.6006	1.499
11	Run3	80	155	70	y	0.6049	2.261
12	Run3	85	165	70	y	0.6100	2.516
13	Run3	80	155	75	y	0.6120	2.127
14	Run3	75	155	70	y	0.6126	2.468
15	Run4	80	150	75	y	0.5710	1.499
16	Run4	85	150	80	y	0.5914	1.87
17	Run4	85	155	80	y	0.5931	1.561
18	Run4	75	155	70	y	0.5946	2.343
19	Run4	85	150	80	y	0.5972	1.505
20	Run4	85	150	80	y	0.5990	1.505
21	Run5	80	150	80	y	0.5379	1.371
22	Run5	75	155	75	y	0.5597	1.658
23	Run5	75	150	75	y	0.5600	1.906
24	Run5	85	150	85	y	0.5679	1.57
25	Run5	85	150	85	y	0.5709	1.95
26	Run6	80	150	80	y	0.5344	1.372
27	Run7	80	150	80	y	0.5282	1.358
28	Run7	80	150	80	y	0.5288	1.358
29	Run7	80	150	80	y	0.5290	1.386
30	Run7	80	150	80	y	0.5327	1.371
31	Run7	80	150	80	y	0.5340	1.372
32	Run7	80	150	80	y	0.5341	1.365
33	Run8	80	150	80	y	0.5282	1.358

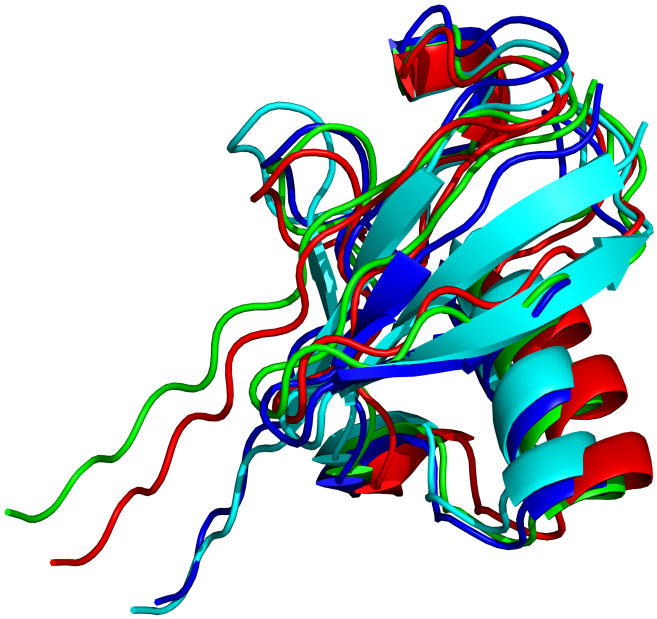


Figure 7.7: Three out of nine shaded structures from Table 7.6 and 1D3Z are superimposed. 1D3Z is in cyan, the reference (row 1) is in red, the row 2 is in green and the row 8 is in blue.

1.488 improvement in the nD-PDPD score (about 1.5Å in bb-rmsd). This also demonstrates the improvement of about 1.317 in bb-rmsd compared to the previous experiment (Table 7.5).

7.3.4 Comparison of the $nD-PDPA$ refinement to Xplor-NIH

Several computational techniques have been developed to address the refinement issue [55], [98], [22]. Most of these methods relies on the extensive utilization of the experimental data such as NOEs [22].

The PDPA method utilizes unassigned RDC datasets to conduct the refinement. At

Table 7.7: The result of the comparison of the nD-PDPA with the xPlor-NIH. The Starting Structure column denotes the bb-rmsd of the structure with respect to protein 1A1Z. And Refined Structure column shows the bb-rmsd of the refined structures with respect to 1A1Z.

Method	Starting StructureÅ	Refined Structure Å
nD-PDPA	2.847	2.195
xPlor-NIH	2.847	1.940

Table 7.8: The result of comparison of nD-PDPA with xPlor-NIH. The Starting Structure column denotes the bb-rmsd of the structure in Å with respect to protein 1D3Z. And Refined Structure column shows the bb-rmsd of the refined structures with respect to 1D3Z.

Method	Starting StructureÅ	Refined Structure Å
nD-PDPA	2.843	2.180
xPlor-NIH	2.843	2.177

the time of writing this manuscript, we are not aware of any software using only unassigned RDC as the primary source of the data for the refinement a structure. Xplor-NIH is a software package for Molecular Dynamic Simulation [22]. Xplor-NIH is capable of using RDC sets to fold a protein. However, the RDC must be assigned. In Table 7.7 the result of comparison of the nD-PDPA method to the xPlor-NIH is demonstrated. The target protein is 1A1Z, and three synthetic N-H RDC sets is used with $\pm$1Hz of error uniformly added to the sets and 25% of the data are removed randomly. The Xplor-NIH exhibits slightly better result (about 0.3Å). The nD-PDPA ran only for 6 runs (Table 7.2). As it is shown previously, more iterations potentially generate the higher quality structure. In the second experiment (Table 7.8) protein 1D3Z is utilized with experimental RDC sets as listed in Section 6.6.3. The result of nD-PDPA refinement after running for 20 iterations is 2.180Å and xPlor-NIH refinement result is 2.177Å respectively. The results of both experiments suggest

the nD-PDPA refinement process is as capable as of xPlor-NIH. Moreover, nD-PDPA utilizes unassigned RDC that makes the method suitable to initiate the refinement process as early as data acquisition time. From this standpoint, the nD-PDPA's refinement functionality and practicality are unique.

Chapter 8

Time Complexity and Software Engineering of $nD - PDPA$

8.1 Introduction

Today's scientific researches are heavily dependent on software programs. Utilization of software programs in any research laboratory is undeniable for a number of reasons. Reasonable cost for high-speed computers that can process millions of data in the short time that once was difficult if it was not impossible. Moreover, the development of novel algorithm in the response to needs for precise calculations for the scientific purpose also contributes to the vast usability of the modern software programs [94]. The process of development of scientific software can be different from standard software. It requires more attention in the certain aspect of the software development such as calculation precision. Moreover, any computational error has an impact on scientific discovery or publication [59]. In this section, the development of the nD-PDPA software from software engineering point of view is discussed. Then, the running time of the nD-PDPA software and 2D-PDPA is investigated.

8.2 The Development of $nD - PDPA$

The nD-PDPA is developed based on the 2D-PDPA software engine. However, two significant changes needed to be developed in nD-PDPA that were not in the 2D-PDPA (See chapter7 for more details). The first transformation of the PDPA is

```
AlignMedCount = 2
!Dataset1
        DataType1 = RDC
        RDCType1 = NH
        ExprRDCFile1 = medium/NH1.rdc
        Sxx1 = 0.0003
        Sxy1 = 0
        Sxz1 = 0
        Syx1 = 0
        Syy1 = 0.0005
        Syz1 = 0
        Szx1 = 0
        Szy1 = 0
        Szz1 = 0

!Dataset2
        DataType2 = RDC
        RDCType2 = CaHa
        ExprRDCFile2 = medium/CaHa1.rdc
        Sxx2 = -2.97422e-05
        Sxy2 = 0.000408305
        Sxz2 = 0.000627309
        Syx2 = 0
        Syy2 = -6.06816e-05
        Syz2 = 0.000439771
        Szx2 = 0
        Szy2 = 0
        Szz2 = 0
```

Figure 8.1: A fragment of the configuration file for nD-PDPA. The alignment media count and the information about the RDC type and order tensor values are shown.

about the capability of utilizing more than two different types of RDC sets. The second change is about using RDC values as the points of calculation of likelihood instead of using 64 by 64 grid in 2D-PDPA.

Figures 8.1 and 8.2 demonstrate a fragment of the nD-DPDA's configuration file. The configurations include the number and type of RDC sets , the relative order tensor values, start and end values for rotational angles and other variables. These variables are designed to offer more flexibility and ability to run the software program by manipulating the values of the variables for the research purpose. The software generates two primary results as output. The first output is the result of the nD-PDPA score, along with the best rotational angles. The score usually is redirected to a result file and if a library of the structure is examined all results are redirected to a single output file. The second output file is the calculated RDC and distribution probability (PDP) for the best orientation of the subject protein. The nD-PDPA software is written in

```
!IO
#This is for calculated RDC.
ProteinLibDir = /home/fahim/Project/Data/Struct5/redcat_template
PdpFile = pdp
OutFile = /home/fahim/Project/nDPDPA/Data/Struct5/result/result.nd.txt
BestStructDir = pdp

!Other
Increment = 5
Sigma = 1
Threshold = 3
ExprRDCSize = 83

!QFactor = 1 calculates QFactor, otherwise not
!This parameter works only for the decoy structures
!( structures with the same size and assigned)
QFactor = 1

!angles
a_start = 0
a_end = 180
b_start = 0
b_end = 180
g_start = 0
g_end = 180

! Number between 0 and 1
AvailableRDCPercent = 1
```

Figure 8.2: A fragment of the configuration file for nD-PDPA is demonstrated. Setting information for Kernel calculation such as sigma and start and end rotational angles are shown.

C++ programming language, utilizing Object Oriented software design technology. It is compiled, optimized and tested for Linux Operating System (OS). The software is free for downloading from our laboratory's server http://ifestos.cse.sc.edu .

8.3 Software Testing Strategies

While different methods are used for testing commercial software, many scientific software programs lack utilizing efficient testing methods for the quality assurance and reliability. For the evaluation of the software, systematic test cases were required. To evaluate the nD-PDPA software the strategies below is utilized: 1. To prove the concept, synthetic data is used for a known structure. In this case, non-uniform RDC sets such as $\{ N - H \}$ and $\{ C\alpha\text{-}H\alpha \}$ are calculated. This experiment can prove the validity of the method by ranking the best structures among a pool of decoy

or database of homologous structures. Since both query and subject structures are known, other measurement methods such as bb-rmsd can be used for correctness and accuracy of the results. 2. For further expansion of the method, experimental data is obtained from online resources such as BMRB [82], [34] for the structures that are characterized before. The experimental data add more ambiguity in terms of experimental error to the nD-PDPA analysis. This test cases, confirms the error tolerance of the proposed method. 3. Using experimental data is one of the main objective of this research to proof the capability of the nD-PDPA method for characterizing of an unknown protein. To facilitates such an experiment, our laboratory established collaboration with MUSC. The purpose of this partnership is to conduct experiments to characterize the structure of a novel protein PF2048.1. In addition, this experiment potentially can be extended for structure refinement to obtain the better quality structure to the native one.

8.4 $nD - PDPA$ Algorithm Analysis and Running Time

In the nD-PDPA analysis, there are two major bottlenecks affect on the running time of the program:

8.4.1 The calculation of Euler angles for cPDP

It is mentioned previously that for a given structure, a cPDP (calculated PDP) is created for each possible rotation of the structure in a grid search over the Euler angles (α, β, γ) at the resolution of 5° (for more information please see Section 5.2). Since the rotation of the subject protein in an alignment medium is unknown, therefore this grid search is unavoidable. The grid search performs 46,656 (36 x 36 x 36 rotation) as a result of 5° intervals. While the improvement of the grid search is still under

investigation, it is possible to manipulate some of the variables related to the grid search and Euler angles. Figure 8.2 listed some of the parameters that can be used to adjust the running time. "increment" parameter denotes the Euler angles interval (default is 5°). Assigning a larger number (for example 10°), finishes the process quicker with the cost of the lower resolution in the search of the best PDP. Also the range of Euler angles can be adjusted in the configuration file if the range of the angles is known prior to the process. This is useful for the refinement process, when after some iteration the structures merge toward one point and generally produce similar Euler angles (See Figure 7.4).

8.4.2 Calculation of PDPs

The calculation of the PDP and subsequently the scoring procedure is the major contribution to the running time of the PDPA method. Algorithm 1 describes the process of calculating PDP when two set of RDC are used (in 2D-PDPA). Algorithm accepts two sets of RDC M1 and M2 as input. Two RDC sets are in the same size. In lines 7 and 9 two loops run from 1 to 64. These loops construct a 64 by 64 grid. The third loop in line 12 calculates the distance of the current grid cell (x and y are calculated in lines 8 and 10 respectively) from any RDC located in the index k. The summation of the likelihood for each point x and y are reported and saved into a matrix (lines 18 to 20). The running time of the algorithm depends on three for loops in lines 7, 9 and 12. It is clear that the running time of the algorithm is a function of $O(S64^2)$. Since the user can manipulate the dimension length (64 for example) as an input variable, therefore, the previous function in general form is $O(SC^2)$, where S denotes the size of the RDC set and C denotes the desired dimension (for example, a C by C grid). Note that, the addition of dimensions to the procedure (n > 2) increases the complexity of the algorithm exponentially and is given by $O(SC^n)$. Therefore, the Algorithm 1 is unfavorable in higher dimensions of the PDPA experiment. In

Algorithm 1 CALC-2DPDP(RDCSet $M1$, RDCSet $M2$)

```
1: Let (x, y, z) ← 0
2: Let index ← 0
3: Let (delta1, delta2) ← 0
4: Let S ← sizeof(M1)                    ▷ Note: Both M1 and M2 are in the same size.
5: Let M be a matrix of (64 x 64) rows by 3 columns
6:
7: for i ← 1 to 64 do
8:     x ← i * delta1 + start1
9:     for j ← 1 to 64 do
10:        y ← j * detla2 + start2
11:        z ← 0
12:        for k ← 1 to S do
13:            t1 ← abs( x - M1[k])
14:            t2 ← abs( y - M2[k])
15:            do temp ← 2DGaussian(t1, t2)
16:            z ← z + temp
17:        end for
18:        M(index, 0) ← x
19:        M(index, 1) ← y
20:        M(index, 2) ← z
21:        index ← index + 1
22:    end for
23: end for
24:
25: return M
```

Algorithm 2, the input is a matrix of RDCs and the columns of this matrix denote the number of RDC sets and the rows denote the RDC sets size. Algorithm consists of two main loops in lines 5 and 7. The summation of the likelihood for each row (containing n RDC set) is calculated in line 9 and 10. The time complexity of the algorithm is $O(S^2)$ where S denotes the size of the sets.

Algorithm 2 CALC-NDPDP(RDCSets RDC, Size S)

```
1: Let index ← 0
2: Let (delta1, delta2) ← 0
3: Let M be a vector of size S
4:
5: for i ← 1 to S do
6:     row1 ← RDC[i]
7:     for j ← 1 to S do
8:         row2 ← RDC[j]
9:         do temp ← nDGaussian(row1, row2)
10:        z ← z + temp
11:    end for
12:    M[i] ← z
13:    z ← 0
14: end for
15:
16: return M
```

8.5 THE RUNNING TIME OF $nD - PDPA$ VS. 2D-PDPA

To benchmark the performance of the nD-PDPA for two or more sets of RDC, protein 1A1Z was selected, and the results were compared to 2D-PDPA results. Four synthetic N-H RDC sets were generated in an ideal condition. Both programs were executed on a Linux desktop with Intel Core i7, 2.67 GHz processor and 8 MB of memory. Table 6 shows the results of execution time for 2D-PDPA and nD-PDPA. In the Section 8.4.2, it is shown that the asymptotic execution time of 2D-PDPA is a function of $O(C^n)$ while the execution time complexity of nD-PDPA is a function of $O(nC^2)$. The 2D-PDPA is incapable of incorporating more than two RDC sets hence the 2D-PDPA running times for the dimensions n $>$ 2 were approximated using the 2D-PDPA asymptotic function. For two RDC sets the running time was measured by executing the 2D-PDPA software. The running time for one RDC set was collected from 1D-PDPA version of the software and was assumed both 2D-PDPA and

Table 8.1: The execution time for $nD - PDPA$ and 2D-PDPA

# of available RDC sets	$nD - PDPA$ required time(seconds)	2D-PDPA required time(seconds) Å
1	20	20
2	323	363
3	484	6859
4	906	130321

nD-PDPA consume the same execution time. The results indicate tremendous time reduction in nD- PDPA engine especially for n>=3(Table 8.1). It is worthy to note that the listed running times are only for one structure. Usually a PDPA experiment utilizes a library of structures that is indeed impossible to be finished in a reasonable time in the case of 2D-PDPA method.

Chapter 9

Conclusion and Future work

9.1 Conclusion

In this study, PDPA is introduced as a novel software tool that can be utilized in protein structure identification and classification. The 2D-PDPA uses two sets of homogeneous RDC sets. The application of the 2D-PDPA to identify an unknown structure in a database is demonstrated successfully. Moreover, the 2D-PDPA is utilized to identify the closest structure to the native structure from a set of 15 modeled structures.

The limitations of the 2D-PDPA are addressed by development of nD-PDPA method. The transition from 2D-PDPA to nD-PDPA can be deemed advantageous for a number of reasons. First, based on availability, additional RDC datasets can be combined from multiple alignment media to increase the information content without imposing a substantial penalty in the execution time. Second, nD-PDPA's scope of RDC analysis no longer limited to just N-H RDC data. The new improvements enable flexible inclusion of RDC data from the same or different alignment media. For example N-H, Ca-Ha data from one alignment medium may be combined with N-H of the second and C-N of the third alignment medium for a total of four-dimensional analysis of PDPA. This flexible inclusion of any available datasets from any number of alignment media can increase the information content significantly leading to a more improved sensitivity and selectivity performance of nD-PDPA. Elimination of the exponential time-complexity and translation of the algorithm into a polynomial

time-complexity is a major achievement with clear consequences in the execution time of the algorithm. Moreover, the application of the nD-PDPA for structure refinement is introduced. The immediate advantage of nD-PDPA refinement method laid on the utilization of unassigned RDC set. In that sense, the nD-PDPA refinement method can be used as early as data acquisition time with the requirement of minimum two RDC sets.

9.2 Future works

This section explains some potential avenues of future research.

9.2.1 Improvement of $nD-PDPA$ refinement method

The refinement method can be improved in a number of ways. First, the generation of decoy structure using potential energy of protein. Incorporation of this method prevents producing of structures with the atomic collision. Moreover, the PDPA score can be combined with other protein energy terms for RDC based refinement. Second, the refinement of a portion of a protein such as loops is suggested. The decoy structure generator software can be improved to perturb only a portion of a structure while the rest of the structure remains intact (rigid body).

9.2.2 Incorporation of any paired vector base data

The nD-PDPA utilizes RDC sets from different alignment media and vector type. The insensitivity of RDC data in translation in space and utilization of the unassigned data reduce the robustness of the nD-PDPA analysis regardless of the amount of data. To increase the robustness of the nD-PDPA incorporation of other distance based data such as JCouplings and Residual Chemical Shifts are suggested.

Bibliography

[1] J Adeyeye, H F Azurmendi, C J M Stroop, S Sozhamannan, A L Williams, A M Adetumbi, J A Johnson, and C A Bush. Conformation of the hexasaccharide repeating subunit from the Vibrio cholerae O139 capsular polysaccharide. *Biochemistry*, 42(13):3979–3988, 2003.

[2] H M Al-Hashimi, A Gorin, A Majumdar, Y Gosser, and D J Patel. Towards structural Genomics of RNA: Rapid NMR resonance assignment and simultaneous RNA tertiary structure determination using residual dipolar couplings. *J Mol Biol*, 318(3):637–649, 2002.

[3] H M Al-Hashimi and D J Patel. Residual dipolar couplings: Synergy between NMR and structural genomics. *Journal of Biomolecular NMR*, 22(1):1–8, 2002.

[4] H M Al-Hashimi, H Valafar, M Terrell, E R Zartler, M K Eidsness, and J H Prestegard. Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 143(2):402–6, April 2000.

[5] Hashim M Al-Hashimi, Yuying Gosser, Andrey Gorin, Weidong Hu, Ananya Majumdar, and Dinshaw J Patel. Concerted motions in HIV-1 TAR RNA may allow access to bound state conformations: RNA dynamics from NMR residual dipolar couplings. *Journal of molecular biology*, 315(2):95–102, January 2002.

[6] S F Altschul, T L Madden, A A Schaffer, J H Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[7] M Andrec, P C Du, and R M Levy. Protein backbone structure determination using only residual dipolar couplings from one ordering medium. *Journal of biomolecular NMR*, 21(4):335–347, December 2001.

[8] Michael Andrec, Yuichi Harano, Matthew P Jacobson, Richard A Friesner, and Ronald M Levy. Complete protein structure determination using backbone residual dipolar couplings and sidechain rotamer prediction. *Journal of structural and functional genomics*, 2(2):103–11, January 2002.

[9] Rolf Apweiler, Alex Bateman, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Yasmin Alam-Faruque, and Emanuele Alpi. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 42(November 2013):191–198, 2014.

[10] Michael Assfalg, Ivano Bertini, Paola Turano, A Grant Mauk, Jay R Winkler, and Harry B Gray. 15N-1H Residual dipolar coupling analysis of native and alkaline-K79A Saccharomyces cerevisiae cytochrome c. *Biophysical journal*, 84(6):3917–23, June 2003.

[11] Hugo F Azurmendi and C Allen Bush. Conformational studies of blood group A and blood group B oligosaccharides using NMR residual dipolar couplings. *Carbohydrate Research*, 337(10):905–915, 2002.

[12] Sonal Bansal, Xijiang Miao, Michael WW. Adams, James H. Prestegard, and Homayoun Valafar. Rapid classification of protein structure models using unassigned backbone rdcs and probability density profile analysis (pdpa). *Magnetic Resonance*, 192(1):60–68, May 2008.

[13] A Bax, G Kontaxis, and N Tjandra. Dipolar couplings in macromolecular structure determination. *Methods in enzymology*, 339:127–74, January 2001.

[14] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242, January 2000.

[15] Helen M. Berman, Gerard J. Kleywegt, Haruki Nakamura, and John L. Markley. How community has shaped the protein data bank. *Structure*, 21(9):1485–1491, 2013.

[16] Pau Bernadó and Martin Blackledge. Local dynamic amplitudes on the protein backbone from dipolar couplings: toward the elucidation of slower motions in biomolecules. *Journal of the American Chemical Society*, 126(25):7760–1, June 2004.

[17] P. Bertone. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Research*, 29(13):2884–2898, July 2001.

[18] Christopher M Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[19] Grzegorz M Boratyn, Alejandro A Schäffer, Richa Agarwala, Stephen F Altschul, David J Lipman, and Thomas L Madden. Domain enhanced lookup time accelerated BLAST, 2012.

[20] Guillaume Bouvignies, Pau Bernadó, Sebastian Meier, Kyuil Cho, Stephan Grzesiek, Rafael Brüschweiler, and Martin Blackledge. Identification of slow correlated motions in proteins using residual dipolar and hydrogen-bond scalar couplings. *Proc Natl Acad Sci U S A*, 102(39):13885–13890, September 2005.

[21] Philip Bradley, Kira M S Misura, and David Baker. Toward high-resolution de novo structure prediction for small proteins. *Science (New York, N.Y.)*, 309(2005):1868–1871, 2005.

[22] Axel Brünger. XPLOR Interface Manual, 2011.

[23] Michael Bryson, Fang Tian, James H Prestegard, and Homayoun Valafar. REDCRAFT: a tool for simultaneous characterization of protein backbone structure and motion from RDC data. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 191(2):322–34, April 2008.

[24] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC bioinformatics*, 10:421, 2009.

[25] J Cavanagh, Wayne j. Fairbrother, Arthur G. Palmer, Mark Rance, and Nicholas Skelton. *Principle and Practice Protein NMR Spectroscopy Second Edition*. Elsevier, 2007.

[26] G M Clore, a M Gronenborn, and a Bax. A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 133(1):216–21, July 1998.

[27] G Marius Clore, Angela M Gronenborn, and Nico Tjandra. Direct Structure Refinement against Residual Dipolar Couplings in the Presence of Rhombicity of Unknown Magnitude. *Journal of Magnetic Resonance, 1998*, 162(131):159–162, 1998.

[28] Gabriel Cornilescu, John L Marquardt, Marcel Ottiger, and Ad Bax. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *Journal of American Chemical Society*, 120:6836–6837, June 1998.

[29] Alison I. Cuff, Ian Sillitoe, Tony Lewis, Oliver C. Redfern, Richard Garratt, Janet Thornton, and Christine a. Orengo. The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research*, 37(November 2008):310–314, 2009.

[30] Alison L. Cuff, Ian Sillitoe, Tony Lewis, Andrew B. Clegg, Robert Rentzsch, Nicholas Furnham, Marialuisa Pellegrini-Calace, David Jones, Janet Thornton, and Christine a. Orengo. Extending CATH: Increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Research*, 39(November 2010):420–426, 2011.

[31] Frank Delaglio, Georg Kontaxis, and Ad Bax. Protein Structure Determination Using Molecular Fragment Replacement and NMR Dipolar Couplings. *Journal of the American Chemical Society*, 122(9):2142–2143, March 2000.

[32] Ken a Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *Science (New York, N.Y.)*, 338(6110):1042–6, November 2012.

[33] P.A. DiMaggio Jr and C.A. Floudas. A mixed-integer optimization framework for de novo peptide identification. *AIChE journal. American Institute of Chemical Engineers*, 53(1):160, 2007.

[34] J F Doreleijers, M L Raves, T Rullmann, and R Kaptein. Completeness of NOEs in protein structure: A statistical analysis of NMR data. *J. Biomol. NMR*, 14:123–132, 1999.

[35] Jurgen F Doreleijers, Steve Mading, Dimitri Maziuk, Kassandra Sojourner, Lei Yin, Jun Zhu, John L Markley, and Eldon L Ulrich. BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J Biomol NMR*, 26(2):139–146, June 2003.

[36] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification Second Edition*. Wiley, 2006.

[37] Narayanan Eswar, Ben Webb, Marc a Marti-Renom, M S Madhusudhan, David Eramian, Min-Yi Shen, Ursula Pieper, and Andrej Sali. Comparative protein structure modeling using MODELLER. *Current protocols in protein science / editorial board, John E. Coligan ... [et al.]*, Chapter 2:Unit 2.9, 2007.

[38] Arjang Fahim, Rishi Mukhopadhyay, Ryan Yandle, James H Prestegard, and Homayoun Valafar. Protein Structure Validation and Identification from Unas-

signed Residual Dipolar Coupling Data Using 2D-PDPA. *Molecules (Basel, Switzerland)*, 18(9):10162–88, January 2013.

[39] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, Incorporated, San Diego, first edition, 1990.

[40] Strang Gilbert. *Introduction to Linear Algebra Forth Edition*. Pearson, 2009.

[41] Alastair Grant, David Lee, and Christine Orengo. Progress towards mapping the universe of protein folds. *Genome Biol*, 5(5):107, 2004.

[42] Neil A Greshenfeld. *The Nature of Mathematical Modeling*. Cambridge University Press, Cambridge, UK, 1998.

[43] L Holm and C Sander. The FSSP database of structurally aligned protein fold families. *Nucleic acids research*, 22(17):3600–9, September 1994.

[44] Young-Sang S Jung, Mukesh Sharma, and Markus Zweckstetter. Simultaneous assignment and structure determination of protein backbones by using NMR dipolar couplings. *Angewandte Chemie (International Ed. in English)*, 43(26):3479–81, June 2004.

[45] J. C. Kendrew, G. Bodo, H. M. Dintzis, H. Wyckoff R.G. Parrish, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtain by X-ray analysis. *Nature*, 181:662–666, 1958.

[46] David E Kim, Dylan Chivian, and David Baker. Protein structure prediction and analysis using the Robetta server. *Nucleic acids research*, 32(Web Server issue):W526–31, July 2004.

[47] R Kordani, M Billeter, and K Wuthrich. Molmol: a program for display and analysis of macromolecular structures. *J Mol*, 14:29–32, 1996.

[48] Christopher James Langmead and Bruce Randall Donald. High-throughput 3D structural homology detection via NMR resonance assignment. *Proceedings / IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference*, pages 278–289, 2004.

[49] Hadas Leonov, Joseph S B Mitchell, and Isaiah T Arkin. Monte Carlo estimation of the number of possible protein folds: effects of sampling bias and folds distributions. *Proteins*, 51(October 2002):352–359, 2003.

[50] Arthur M. Lesk. *Introduction to protein architecture*. Oxford University Press, 2001.

[51] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, 2010.

[52] Adam Liwo, Mey Khalili, and Harold a Scheraga. 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*, 102:2362–2367, 2005.

[53] J A Losonczi, M Andrec, M W Fischer, and J H Prestegard. Order matrix analysis of residual dipolar couplings using singular value decomposition. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 138(2):334–42, June 1999.

[54] K. Wuthrich M. P. Williamson, T. F. Havel. Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. *Journal of biomolecular biology*, 182(2):295–315, 1985.

[55] J MacCallum, A Perez, M Schnieders, L Hua, M Jacobson, and K Dill. Assessment of protein structure refinement in CAPS9. *Bioinformatics*, 79:74–90, 2011.

[56] F M Marassi and S J Opella. Simultaneous resonance assignment and structure determination in the solid-state NMR spectrum of a membrane protein in lipid bilayers. *Biophys J*, 82(1):467A–467A, 2002.

[57] F. M. Marassi and S. J. Opella. Simultaneous assignment and structure determination of a membrane protein from NMR orientational restraints. *Protein Science*, 12(3):403–411, 2003.

[58] J Meiler, W Peti, and C Griesinger. DipoCoup: A versatile program for 3D-structure homology comparison based on residual dipolar couplings and pseudo-contact shifts. *Journal of biomolecular NMR*, 17(4):283–94, August 2000.

[59] Zeena Merali. Why Scientific Programming Does Not Compute. *Nature*, 467(10):775âĂŞ-777, 2010.

[60] Xijiang Miao, Peter J Waddell, and Homayoun Valafar. TALI: local alignment of protein structures using backbone torsion angles. *Journal of bioinformatics and computational biology*, 6(1):163–81, February 2008.

[61] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction (CASP)–round x. *Proteins*, 82 Suppl 2(October):1–6, 2014.

[62] Rishi Mukhopadhyay, Xijiang Miao, Paul Shealy, and Homayoun Valafar. Efficient and accurate estimation of relative order tensors from lambda-maps. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 198(2):236–47, June 2009.

[63] Alexey G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[64] C A Orengo, A D Michie, S Jones, D T Jones, M B Swindells, and J M Thornton. CATH - a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, August 1997.

[65] Emanuel Parzen. On estimation of probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, September 1962.

[66] Ralph B Pethica, Michael Levitt, and Julian Gough. Evolutionarily consistent families in SCOP: sequence, structure and function. *BMC Structural Biology*, 12(1):27, 2012.

[67] J. W. Ponder and D. A. Case. Force Fields for Protein Simulations. 66:27–85, 2003.

[68] J H Prestegard. New techniques in structural NMR–anisotropic interactions. *Nature structural biology*, 5 Suppl:517–22, July 1998.

[69] J H Prestegard, C M Bougault, and a I Kishore. Residual dipolar couplings in structure determination of biomolecules. *Chemical reviews*, 104(8):3519–40, August 2004.

[70] J H Prestegard, K L Mayer, H Valafar, and G C Benison. Determination of protein backbone structures from residual dipolar couplings. *Methods in enzymology*, 394:175–209, January 2005.

[71] Campbell Reece. *Biology Seventh Edition*. Pearson, 2005.

[72] A Saupe and G Englert. High-Resolution Nuclear Magnetic Resonance Spectra of Orientated Molecules. *Physical Review Letters*, 11(10):462–464, November 1963.

[73] Chris Schmidt, Stephanie J Irausquin, and Homayoun Valafar. Advances in the REDCAT software package. *BMC bioinformatics*, 14(1):302, October 2013.

[74] Paul Shealy, Yizhou Liu, Mikhail Simin, and Homayoun Valafar. Backbone resonance assignment and order tensor estimation using residual dipolar couplings. *Journal of biomolecular NMR*, 50(4):357–69, August 2011.

[75] M S Smyth and J H Martin. X Ray Crystallography. *Molecular pathology : MP*, 53:8–14, 2000.

[76] F Tian, H M Al-Hashimi, J L Craighead, and J H Prestegard. Conformational analysis of a flexible oligosaccharide using residual dipolar couplings. *Journal of the American Chemical Society*, 123(3):485–492, 2001.

[77] F Tian, H Valafar, and J H Prestegard. A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. *Journal of the American Chemical Society*, 123(47):11791–6, November 2001.

[78] N Tjandra, J G Omichinski, A M Gronenborn, G M Clore, and A Bax. Use of dipolar H-1-N-15 and H-1-C-13 couplings in the structure determination of magnetically oriented macromolecules in solution. *Nature Structural Biology*, 4(9):732–738, 1997.

[79] N Tjandra, S Tate, A Ono, M Kainosho, and A Bax. The NMR structure of a DNA dodecamer in an aqueous dilute liquid crystalline phase. *J. Am. Chem. Soc.*, 122(26):6190–6200, 2000.

[80] J R Tolman, J M Flanagan, M A Kennedy, J H Prestegard, and J R Tolman Flanagan, J M, Kennedy, M A & Prestegard, JH. Nuclear Magnetic Dipole Interactions in Field-Oriented Proteins - Information for Structure Determination in Solution. *Proc Natl Acad Sci U S A*, 92(20):9279–9283, September 1995.

[81] Tobias S Ulmer, Benjamin E Ramirez, Frank Delaglio, and Ad Bax. Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. *Journal of the American Chemical Society*, 125(30):9179–91, July 2003.

[82] Eldon L Ulrich, Hideo Akutsu, Jurgen F Doreleijers, Yoko Harano, Yannis E Ioannidis, Jundong Lin, Miron Livny, Steve Mading, Dimitri Maziuk, Zachary Miller, Eiichi Nakatani, Christopher F Schulte, David E Tolmie, R Kent Wenger,

Hongyang Yao, and John L Markley. BioMagResBank. *Nucleic acids research*, 36(Database issue):D402–8, January 2008.

[83] H Valafar, K Mayer, C Bougault, P LeBlond, F E Jenney, P S Brereton, M Adams, and J H Prestegard. Backbone solution structures of proteins using residual dipolar couplings: application to a novel structural genomics target. *J Struct Funct Genomics*, 5:241–254, 2005.

[84] H. Valafar and J. H. Prestegard. Redcat: a residual dipolar coupling analysis tool. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 167(2):228–41, April 2004.

[85] Homayoun Valafar and James H Prestegard. Rapid classification of a protein fold family using a statistical analysis of dipolar couplings. *Bioinformatics (Oxford, England)*, 19(12):1549–55, August 2003.

[86] Homayoun Valafar, Mikhail Simin, and Stephanie Irausquin. A Review of REDCRAFT: Simultaneous Investigation of Structure and Dynamics of Proteins from RDC Restraints. *Annual Reports on NMR Spectroscopy*, 76:23–66, 2012.

[87] Sj Varner, Rl Vold, and Gl Hoatson. An Efficient Method for Calculating Powder Patterns. *Journal of magnetic resonance. Series A*, 123(1):72–80, November 1996.

[88] Annaleen Vermeulen, Hongjun Zhou, and Arthur Pardi. Determining DNA Global Structure and DNA Bending by Application of NMR Residual Dipolar Couplings. *Journal of the American Chemical Society*, 122(40):9638–9647, October 2000.

[89] Lincong Wang and Bruce Randall Donald. Exact solutions for internuclear vectors and backbone dihedral angles from NH residual dipolar couplings in two media, and their application in a systematic search algorithm for determining protein backbone structure. *Journal of biomolecular NMR*, 29(3):223–42, July 2004.

[90] Xingsheng Wang, Brian Tash, John M Flanagan, and Fang Tian. RDC derived protein backbone resonance assignment using fragment assembly. *Journal of biomolecular NMR*, 49(2):85–98, February 2011.

[91] J J Warren and P B Moore. A maximum likelihood method for determining D(a)(PQ) and R for sets of dipolar coupling data. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 149(2):271–5, April 2001.

[92] H Widmer and W Jahnke. Protein NMR in biomedical research. *Cellular and molecular life sciences : CMLS*, 61:580–599, 2004.

[93] Wikipedia. X-ray crystallography — Wikipedia, the free encyclopedia, 2014. [Online; 16-July-2014].

[94] Greg Wilson, D a Aruliah, C Titus Brown, Neil P Chue Hong, Matt Davis, Richard T Guy, Steven H D Haddock, Kathryn D Huff, Ian M Mitchell, Mark D Plumbley, Ben Waugh, Ethan P White, and Paul Wilson. Best practices for scientific computing. *PLoS biology*, 12(1):e1001745, January 2014.

[95] Sitao Wu and Yang Zhang. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins: Structure, Function and Genetics*, 72:547–556, 2008.

[96] Kurt Wüthrich. The way to NMR structures of proteins. *Nature Publishing Group*, 8(11):923–925, 2001.

[97] Homayoun Valafar Xijiang Miao, Rishi Mukhopadhyay. Estimation of Relative Order Tensors, and Reconstruction of Vectors in Space using Unassigned RDC Data and its Application. *JMR*, 194(2):202–211, 2009.

[98] Yang Zhang. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9:40, 2008.

[99] Markus Zweckstetter and Ad Bax. Prediction of Sterically Induced Alignment in a Dilute Liquid Crystalline Phase: Aid to Protein Structure Determination by NMR. *Journal of the American Chemical Society*, 122(15):3791–3792, April 2000.

Appendix A

Generating of Decoy Structures

A.1 Introduction

Producing decoy structures sets are of the central importance of the PDPA process. Generation of the decoy structures is used in various experiments in the PDPA method such as the refinement experiment. In this section, the algorithm for generating a set of decoy structure is explained.

A.2 Utilization of software MolMol for decoy structures generation

MolMol is a protein structure visualization software that was developed in 1990's [47]. Although unfortunately MolMol no longer receives any technical support or update from original authors, still is one of the popular protein visualization software in the community. MolMol consists of a rich API set that performs a variety of the operation through command-line on protein coordinate file (PDB).

Several wrapper programs in Perl were developed to utilize MolMol's APIs capabilities. Figure A.1 demonstrates the overall algorithm for generation of decoy structures.The algorithm uses the size of the decoy library and the maximum desired bb-rmsd in Å as input variables. In the first step, the algorithm generates a random number between 0 to 360° to be added to each back-bone ϕ and ψ angles to generate a new perturbed structure. After producing a new structure, the bb-rmsd of the new

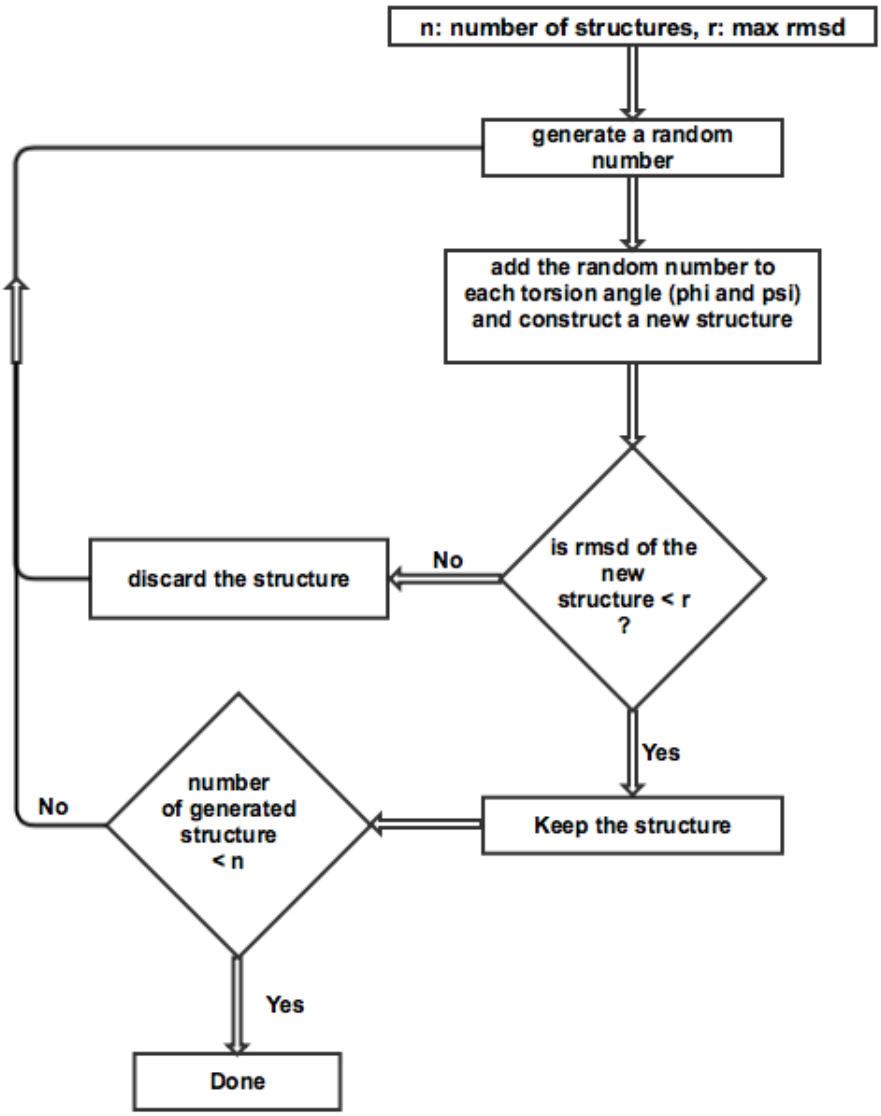


Figure A.1: The flowchart of the decoy structures generator program.

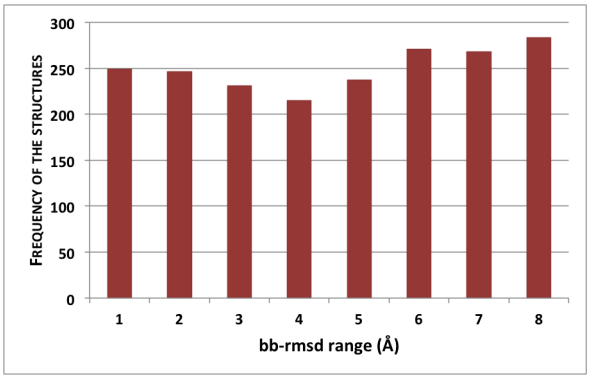


Figure A.2: The distribution of the bb-rmsd for 1000 decoy structures from protein 1A1Z.

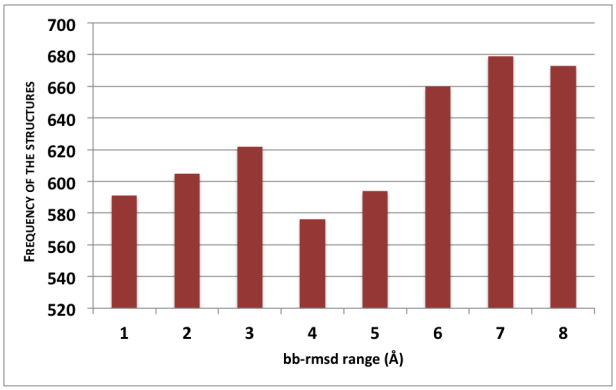


Figure A.3: The distribution of the bb-rmsd for 5000 decoy structures from protein 1A1Z.

structure with respect to the reference structure, is calculated. If the back-bone is lesser than the given maximum bb-rmsd by the user (value of r in Figure A.1) then the algorithm accepts the structure otherwise not. The process is repeated until the number of generated structures reaches to the desired number of proteins given by the user (value of n in Figure A.1).

Figures A.2 and A.3 demonstrate the frequency distribution of the bb-rmsd of two decoy structure sets for 1000 and 5000 structures for protein 1A1Z. Usually before any experiment such an observation is conducted to ensure that the generated structures are distributed uniformly over desired bb-rmsd range.