Spring 2022

Concurrent Identification, Characterization, and Reconstruction Of Protein Structure and Mixed-Mode Dynamics From RDC Data Using Redcraft

Hanin Rafiq Omar

CONCURRENT IDENTIFICATION, CHARACTERIZATION, AND RECONSTRUCTION OF PROTEIN STRUCTURE AND MIXED-MODE DYNAMICS FROM RDC DATA USING REDCRAFT

By

Hanin Rafiq Omar

Bachelor of Science
University of Jordan, 2003

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 and Engineering

College of Engineering and Computing

University of South Carolina

2022

Accepted by:

Homayoun Valafar, Major Professor

Marco Valtorta, Committee Member

Stephen A. Fenner, Committee Member

Jijun Tang, Committee Member

Joseph Flora, Committee Member

Tracey L. Weldon, Interim Vice Provost and Dean of the Graduate School

Acknowledgments

I want to express my sincere gratitude to my advisor Prof. Homayoun Valafar for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me throughout my research and writing of this thesis. I could not have imagined having a better advisor and mentor.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Marco Valtorta, Prof. Joseph Flora, Prof. Jijun Tang, and Prof. Stephen Fenner, for their insightful comments and encouragement.

I thank my colleagues for their support and engorgement.

Last but not least, I would like to thank my family for all the support throughout my Ph.D.

Abstract

A complete understanding of the structure-function relationship of proteins requires an analysis of their dynamic behaviors and the static structure. However, all current approaches to studying dynamics in proteins have their shortcomings. A conceptually attractive and alternative approach simultaneously characterizes a protein's structure and its intrinsic dynamics. Ideally, such an approach could solely rely on RDC data-carrying both structural and dynamical information. The major bottleneck in utilizing RDC data in recent years has been attributed to a lack of RDC analysis tools capable of extracting the pertinent information embedded within this complex data source.

Here we present a comprehensive strategy for structure calculation and reconstruction of discrete state dynamics from RDC data based on the SVD method of order tensor estimation. In addition to structure determination, we provide a mechanism of producing an ensemble of conformations for the dynamical regions of a protein from RDC data. The developed methodology has been tested on simulated RDC data with ±1Hz of error from an 83 residue α protein (PDB ID 1A1Z). In nearly all instances, our method reproduced the protein structure, including the conformational ensemble, within less than 2Å. Based on our investigations, arc motions with more than 30° of rotation are recognized as internal dynamics and are reconstructed with sufficient accuracy. Furthermore, states with relative occupancies above 20% are consistently recognized and reconstructed successfully. Arc motions with a magnitude of 15° or relative occupancy of less than 10% are consistently

unrecognizable as dynamical regions within the context of ±1Hz of error.

We also introduce a computational approach named REDCRAFT that allows for uncompromised and concurrent characterization of protein structure and dynamics. We have subjected DHFR (PDB-ID 1RX2), a 159-residue protein, to a fictitious but plausible, mixed-mode internal dynamics model. In this simulation, DHFR was segmented into seven regions. The two dynamical and rigid-body segments experienced an average orientational modification of 7˚ and 12˚, respectively. Observable RDC data for backbone C'-N, N-H, and C'-H were generated from 102 frames that described the molecular trajectory. The Dynamic Profile generated by REDCRAFT allowed for the recovery of individual fragments with bb-rmsd of less than 1Å and the identification of different dynamical regions of the protein. Following the recovery of fragments, structural assembly correctly assembled the four rigid fragments with respect to each other, categorized the two domains that underwent rigid-body dynamics, and identified one dynamical region for which no conserved structure can be defined. In conclusion, our approach successfully identified dynamical domains, recovery of structure where it is meaningful, and relative assembly of the domains when possible.

Table of Contents

List of Tables

List of Figures

Chapter 1

Proteins

Proteins are large, complex molecules that make up over 50% of the dry weight of cells. Proteins have diverse biological functions ranging from DNA replication[1], forming cytoskeletal structures[2], transporting oxygen around the bodies of multicellular organisms[3] to cell signaling[4] and ligand binding[5].

1.1 Protein Structure-Function Relationships

Despite the functional diversity, all proteins consist of a linear arrangement of amino acids assembled into a polypeptide chain. Proteins differ primarily in their sequence of amino acids, which is dictated by the nucleotide sequence of their genes. However, this two-dimensional representation of the polypeptide chain does not give information about the actual three-dimensional structure that defines its characteristic functional properties[6,7]. Understanding the arrangement of atoms within proteins and how these topologies are uniquely suited to their biological roles allows us to probe the structure-function relationship of the protein; for example, the mechanism of oxygen binding in hemoglobin[8], or understanding substrate and ligand binding[9].

1.2 Protein Structure Hierarchy

To understand the properties of proteins, we must first describe the "building blocks" of proteins and their properties. This section describes Amino acids as well as the structure of proteins. Note that the term structure, when used with proteins, takes on a much more complex meaning since proteins have four different levels of structure: primary, secondary, tertiary, and quaternary.

1.2.1 Amino Acids: The Building Blocks of Proteins

Amino acids are organic compounds that consist of two charged functional groups; (amine *(+NH_2)* group and carboxylic acid (-*COOH*) group), as well as a sidechain (*R*) group and a single Hydrogen atom, all of which are connected to an α-carbon. There are 20 different types of standard amino acids, each with a unique sidechain (*R*). An individual amino acid is encoded using a three-letter code; for example, glycine is typically referred to as GLY. The side chain (*R)* is responsible for the different properties of individual amino acids, and it contributes considerably to the physical properties of proteins. The number of amino acids in a protein can differ significantly from one protein to another, ranging from fifty to thousands.

1.2.2 Primary Sequence

The primary structure is the linear order of amino acids along the polypeptide chain. Proteins are unique in the composition and order of amino acids along the polypeptide chain that represents them. Amino acids are called residues when they are part of a peptide bond. Peptide bonds are formed when the amino group of one amino acid reacts with the carboxyl group of another amino acid. This reaction results in the elimination of water and the formation of a dipeptide. When three residues are joined together by two peptide bonds, they form a tripeptide, and so on. The amino acid sequence of a protein is read from left to right; (from the amino (N-terminal) to the carboxyl (C-terminal)). Residues in a protein are divided into the main chain, the backbone, and the side chain. The backbone of a protein consists of the amide *N*, the α-carbon, and the carbonyl © linked together via peptide bonds. The peptide bond has a double-bond character between the carbon and nitrogen atoms, which prevents rotation about this bond, thus providing stability and planarity of the peptide plane[10]. Therefore, the peptide plane consists of six atoms: the α-carbon atom, the *CO* group

from the first amino acid, the *NH* group, and the α-carbon atom from the second amino acid, as seen in *Figure 1.1*.

1.2.3 Secondary Structure

The secondary structure is a local representation of the spatial relationship of residues closest together in the primary sequence. The secondary structures are defined based on the pattern of the backbone torsion/dihedral angles. Protein's backbone is defined using three dihedral angles Phi (φ), Psi (ψ), and Omega (ω)[10]. Due to the planarity of the peptide bond, the Omega (ω) torsion angle is restricted to be either 180° or 0°. So, the critical factor that decides the basic conformation of the secondary structure is the values adopted by the other two dihedral angles (φ and ψ) and their effect on the hydrogen bonding patterns. The Phi (φ) torsion angle represents the rotation of the backbone chain around the bonds between *N*-C_α. In contrast, the Psi (ψ) torsion angle represents the rotation of the backbone around the bonds between Cα-C. The Omega (ω) torsion angle represents the rotation of the backbone around the bonds between *C-N*, as shown in *Figure 1.1*. Protein's secondary structure can be divided into three basic conformations: the Helix, the β-strand, and Coils.

The Helix

Helices make up almost 30% of secondary structures in globular proteins. The helix is formed when values adopted for the torsion/dihedral angles Phi(φ) and Psi (ψ) allow some of the backbone atoms to form hydrogen bonds that result in a spiral conformation[11]. The hydrogen bonds occur between the carbonyl oxygen of one residue of the backbone known as acceptor and the amide hydrogen of the fourth residue ahead in the polypeptide chain known as a donor. One of the essential features

of these hydrogen bonds is that they are linear and parallel to the helical axis, thus stabilizing the helical structure (see *Figure 1.2)*.

The *B*-strand

In the β-strand, the backbone atoms are elongated, so the hydrogen bonding occurs between strands (inter-strand) instead of within a strand (intra-strand). A single β-strand is not stable by itself. However, when two or more β-strands form additional hydrogen bonding, a stable β-sheet arrangement is created, see *Figure 1.2*. In β-sheets, adjacent strands can align in parallel or antiparallel configurations with the orientation established by determining the direction of the polypeptide chain from the N to the C-terminal.

The Turn/Coil

Coils are flexible loop regions in a protein that link other secondary structure elements together. Their primary role is to enable the polypeptide chain to change direction and, in some cases, to reverse back on itself[12]. Turns are classified according to the number of residues they contain. A γ-turn contains three residues and frequently links adjacent strands of anti-parallel β-sheet; on the other hand, β-turns has four residue turns. *Figure 1.2* shows an example of a single protein with multiple secondary structures.

Ramachandran Plot

Not all combinations of the two torsion angles φ and ψ, also known as Ramachandran angles, are possible due to steric collisions between atoms. Ramachandran plot[13] is a two-dimensional plot used to visualize the energetically allowed distribution of torsion angles Phi and Psi in a protein structure. Each type of secondary structure has a specific range of torsion angle values mapping to different regions observed on the Ramachandran plot. As shown in the diagram below

(*Figure 1.3*), the white areas correspond to sterically disallowed regions for all amino acids except glycine, which is unique because it has a small side chain that allows flexibility. The blue-colored areas correspond to β-sheet conformations with no steric clashes, and the green-colored areas correspond to allowed alpha-helical regions.

1.2.4 Tertiary Structure

The tertiary structure stands for the spatial arrangement of residues in the primary sequence of protein[14,15]. In other words, the protein's geometric shape results from linking together one or more secondary structures[10] or/and protein domains, as can be seen in *Figure 1.2*. A protein domain is a sub-unit used to organize the tertiary structure of proteins with a large residue number[16]. For a tertiary structure to be stable, proteins must form favorable interactions between secondary structure elements rather than repulsive ones. The process that determines the most stable interaction is known as protein folding. Elements of secondary structure interact via hydrogen bonds; and depend on disulfide bridges[17], electrostatic interactions[18,19], van der Waals interactions[20], hydrophobic contacts, and hydrogen bonds between non-backbone groups.

The protein tertiary (3-D) structure is defined by its atomic coordinates. Each atom's X Y Z coordinates in the protein are usually stored in a PDB file. PDB files contain other information besides the atomic coordinates of atoms; some of this information includes header information describing the primary sequence, the method used to determine the structure, the secondary structure elements, etc. All PDB files are deposited in a database called Protein Data Bank (PDB)[21]. The PDB database is currently maintained at Rutgers University (http://www.rcsb.org/pdb). Numerous software packages have been developed for viewing PDB files, most of which are public domain software; our research employs two of them: Molmol[22] and VMD[23].

1.2.5 Quaternary Structure

Many proteins contain more than one polypeptide chain. The quaternary structure best describes the interaction between these chains[24]. These interactions include all the ones responsible for tertiary structure stability, except they occur between two or more polypeptide chains. In the quaternary structure, polypeptide chains are called subunits. This research focuses on tertiary structures and doesn't examine any quaternary structures.

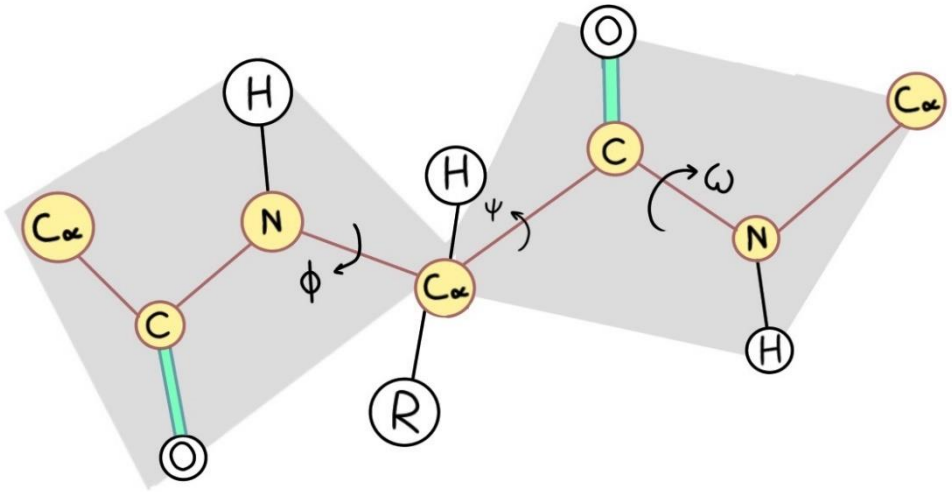

Figure 1.1: This peptide plane shows all six atoms that contribute to it in yellow color and the double bond between C and O atoms. Also, all three torsion angles are labeled.

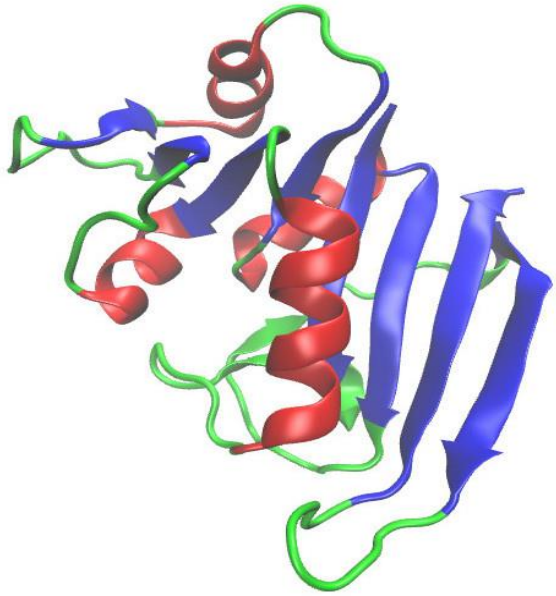

Figure 1.2: DHFR protein with all secondary structures present; the helix structure is colored red, the β-strand/sheet is in blue, and the coils are green.

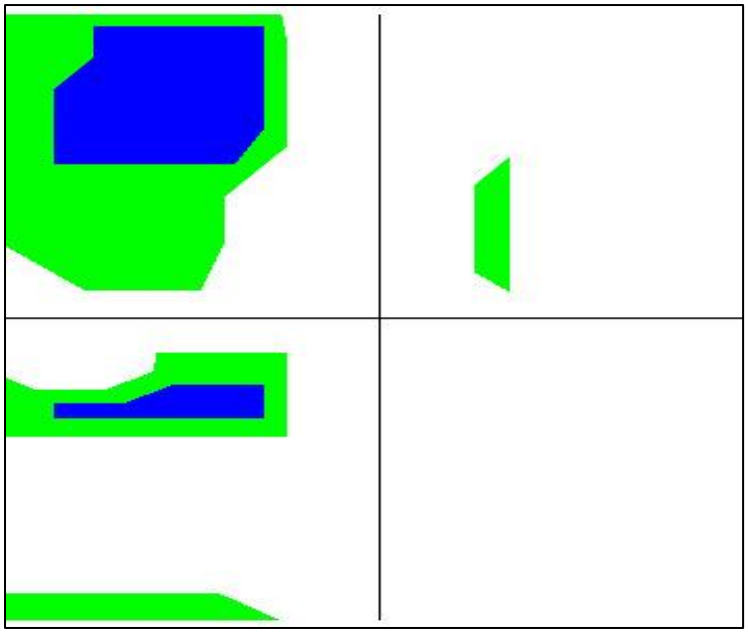

Figure 1.3: Ramachandran plot with all the allowed phi/psi combination in colored regions; the blue-colored regions represent β-sheets combinations while the green-colored regions represent the helical ones.

Chapter 2

Structure Determination and Dynamics

Protein Structure determination calculates the secondary, tertiary, and quaternary structure from the primary sequence and potential inclusion of empirical restraints. Understanding the arrangement of atoms within a protein and how its topology uniquely suits its biological role gives deeper insight into its function. However, despite all the advances in protein structure determination methods, the protein folding problem (without any empirical restraints) remains one of the most fundamental unsolved problems in computational molecular biology today. As of February 2022, the number of protein sequences reported in the latest release of UniProtKB databases[25] (http://www.uniprot.org); UniProtKB/Swiss-Prot and UniProtKB/TrEMB are 565,928 and 225,013,025 respectively. Meanwhile, the number of identified protein structures according to the current Protein Data Bank[21] (http://www.rcsb.org) holding list are only 186,934. With less than 2% of structures identified of the overall known proteins, it's clear that much more effort is still needed to narrow the gap.

2.1 Structure Calculation Methods

There are two main methods for protein structure determination: experimental and computational methods. But for the scope of this research, we will focus more on the experimental methods.

Experimental methods in structure determination are typically time-consuming, highly labor-intensive, and relatively expensive. Over 90% of all experimentally derived structures deposited in the Protein Data Bank result from Crystallographic studies, while the remaining structures are solved using Nuclear Magnetic Resonance (NMR) spectroscopy.

2.1.1 X-ray Crystallography

X-ray crystallography[26–28] is the prominent technique for protein structure determination. This method applies X-ray diffraction principles to determine the arrangement of atoms of a crystalline solid in three-dimensional space. To have successful results of X-ray crystallography, a crystal form is needed where the arrangement of the atoms needs to be in an ordered, periodic structure for them to diffract the x-ray beams. Then a series of mathematical calculations (Bragg's Law[29]) are used to produce a diffraction pattern characteristic to the particular arrangement of atoms in that crystal.

One of the significant drawbacks of X-ray crystallography is its dependence on whether a crystal of a protein can be obtained or not. The requirements for protein crystallization are challenging to satisfy, laborious, time-consuming, and expensive[30,31]. Furthermore, X-ray crystallography proved accurate with small molecules with less than 100 atoms in their crystal form. In contrast, macromolecular crystallography[32] often involves tens of thousands of particles in the unit cell; hence the atomic-level picture provided by X-ray crystallography becomes less well-resolved for a given number of observed reflections.

2.1.2 NMR Spectroscopy

Nuclear magnetic resonance spectroscopy[33,34] is a technique based on the NMR phenomenon[35] of nuclei to study the molecules' physical, chemical, and biological properties by utilizing electromagnetic radiations. NMR spectroscopy is used for various applications; one-dimensional NMR techniques typically provide detailed chemical structures. On the other hand, more complicated two-dimensional methods are used for structure determination. Time-domain NMR spectroscopic techniques[36] are used to probe molecular dynamics in solutions. On the other hand, solid-state NMR spectroscopy is used for solids structure determination.

The main advantages of NMR spectroscopy include the study of molecules in their aqueous and potentially native environments. Also, NMR spectra are unique, well-resolved, analytically tractable, and often highly predictable for some molecules. Moreover, it's the only way to study partially or wholly intrinsically unstructured proteins[37,38] and is typically used for determining conformation-activity relationships[39]. A disadvantage is that NMR spectrometers are relatively expensive, and the timescale of NMR is rather long, and thus it results in an averaged spectrum when observing fast phenomena.

To determine the 3D structure of proteins using NMR spectroscopy, some measurable parameters or restraints are computed. The three widely used restraints are distance, angle, and orientation restraints. Distance restraints, commonly known as the Nuclear Overhauser effect (NOE)[40,41] connect resonances from nuclei that are spatially close to each other; less than 6^0 A. Angle restraints are restraints on the torsion angles (Phi and Psi) of the peptide bonds, and they can be generated using either chemical shifts[42] or coupling constants[43]. In this research, we employ the

orientation restraints measured from NMR spectroscopy; hence they will be discussed in detail in the following chapter.

In theory, it's possible to define a complete three-dimensional structure of a protein if sufficient angles and distances between atoms are defined. However, in reality, all NMR-derived restraints are within a range of possible values. As a result, it's most likely that many conformations for a single protein are consistent with the measurements. Although it's challenging to define a unique structure from NMR spectroscopy, it still can yield a family of closely related structures.

2.1.3 Computational Folding Methods

Although experimental methods are still the primary source for determining protein structures, they still failed in determining the structure for some categories of proteins (protein complexes) and considering the significantly large number of protein sequences that have been identified due to the development of sequencing methods[44]; it is simply neither possible nor functional to determine all the protein structures by experimental methods. As a result, there is a pressing need for devising efficient computational methods for structure prediction.

Since it was discovered that proteins are capable of folding into their unique functional 3D structures based on their primary sequence alone, and the compelling evidence that some diseases like cystic fibrosis are a result of a misfolded protein because of a change in its primary sequence, decades of research has increased our ability to predict the 3D protein structure from sequences only[45]. All proteins fold into their most stable form, called the native state. While covalent bonds contribute equally to the stability of the folded and unfolded proteins, the non-covalent interactions cumulatively contribute to the increased stability of the native state. Non-covalent interactions include hydrogen bonds, hydrophobic forces, and interactions

between charged groups. The native form of a protein is the conformation in which the magnitude of favorable interactions outweighs the sum of the unfavorable ones. At its core, all computational methods aim to derive that conformation.

The computational approaches to protein structure prediction can be broken down into three major categories: comparative modeling[46,47], fold recognition[48,49], and ab initio prediction[50]. The main difference between those categories is how each category utilizes the available information from the known structure databases.

2.2 The Study of Protein Dynamics

The relation between protein structure and function isn't straightforward enough for reliable function predictions. Recent studies have shown that different structures can have the same function; for example, the enzyme proteases[51] occur in many branches of the classification trees of CATH[52] and SCOP[53]. Moreover, similar structures can perform different functions, such as the TIM-barrel fold[54]. As a result, recent approaches in protein function predictions are built on the assumption that a critical element of the protein function is determined by the conformational dynamics of a protein encoded in their structures, see *Figure 2.1*.

Recent advances in structure identification methods rapidly increased the number of identified proteins with conformational dynamics[55,56]. Nevertheless, a detailed understanding of how dynamics leads to function is still limited. Therefore, the development of methods leading to elucidation of the structure and dynamics of proteins is an active research area.

The traditional methods applied to study protein dynamics can be separated into two main areas: Experimental or Computational methods. While experiments are used to determine what is moving and how fast it is, MD simulations, on the other hand, define the underlying forces and corresponding energies behind that movement.

2.2.1 Experimental Approaches

Protein dynamics have been extensively studied using a variety of experimental approaches, including diffraction methods, solution-state spectroscopy, and solid-state NMR spectroscopy. Regardless of the enormous impact, experimental methods have on the protein structure determination field, they proved inadequate for investigating the dynamics-function linkage. The Study of dynamical proteins with X-ray crystallography is fundamentally complex under the desiccated and restrictive crystalline environment that may interfere with the native state dynamics of aqueous proteins. Nuclear Magnetic Resonance (NMR) spectroscopy, on the other hand, is sensitive to local structure and dynamics in many distinct time windows, each of which may have different functional implications. To describe molecular dynamics as an exchange between multiple states, it is necessary to obtain a metric by which these states can be distinguished. Fortunately, NMR provides many observables which may suit this task, including chemical shift δ[42], relaxation rates T1 and T2[57–59], Liparo-Szabo order parameters[60], and scalar coupling J[43]. However, such traditional approaches fail to provide an atomic-level description of the conformational states. This is partly due to the insensitivity of the commonly employed, short-range Nuclear Overhauser Effect (NOE) restraints to conformational changes[40,41]. Hence, if there is no observable difference in a particular metric between the exchanging states, it does not necessarily mean there are no dynamics. Instead, it means there is no detectable difference in that observable between the multiple states under those experimental conditions. Furthermore, some dynamic processes may not be visible within a particular experimental window. The traditional Liparo-Szabo approaches are sensitive to time scales faster than the overall correlation time (τc)[61–63], while functionally relevant events often take place on time scales (sec-msec).

2.2.2 Molecular Dynamic Simulations (MD)

Molecular dynamics (MD) simulation is often used for investigating the dynamics of proteins[64,65]. MD techniques can obtain details concerning individual particle motions as a function of time and the global molecular motions of proteins at spatial and temporal scales that are difficult to access experimentally[65]. The basic idea behind molecular dynamics is to iteratively solve Newton's equation of motion for each atom in a molecule. A pre-defined force field calculates the potential energy; some frequently used force fields are Amber[66], CHARM[67,68], and GROMOS[69]. Peptide geometries such as bonds, angles, and dihedral are all included in the bonded terms of the potential energy calculations used in those force fields. Some examples of modern Molecular Dynamics Simulation (MD) software are CHARMM[70], NAMD[71], GROMACS[72], AMBER[73,74], and XPLOR-NIH[75,76].

The combination of increased computer power and improved potential functions has resulted in an ability to generate simulations that approach the point at which they can survive critical examination by the experimentalists who determine the structures of the simulated proteins. However, there are several drawbacks for MD simulations: First, the calculation of potential energy function has to be performed for each atom in the protein once every femtosecond or so, which increases the computational cost of the simulations. Second, the conformational changes suitable for MD simulations are more high speed compared to conformational changes observed experimentally. Third, at best, the potential energy function at the heart of these simulations is only an approximation. And finally, currently, there is no molecular model of water to be included in these simulations that describe all water properties correctly.

Simulations are most effective when analyzed in close conjunction with experimental intermediates, essential in validating and improving the simulations. The main idea of this combined approach is to obtain the underlying distribution that gives the average value obtained in the NMR experiment from the MD simulations. Unfortunately, this is a non-trivial matter due to the non-linear and multiple-valued relationships between NMR observables and protein structure. Also, it is conceivable that many different distributions can result in the same average value. This ambiguity in results depending on the criteria used and software applied; casts a shadow on the accuracy of such approach.

Many NMR observables are typically used as restraints in MD simulations. Still, for the scope of this research, we will only focus on the use of residual dipolar coupling (RDC) in conjunction with MD simulation (Chapter 5).

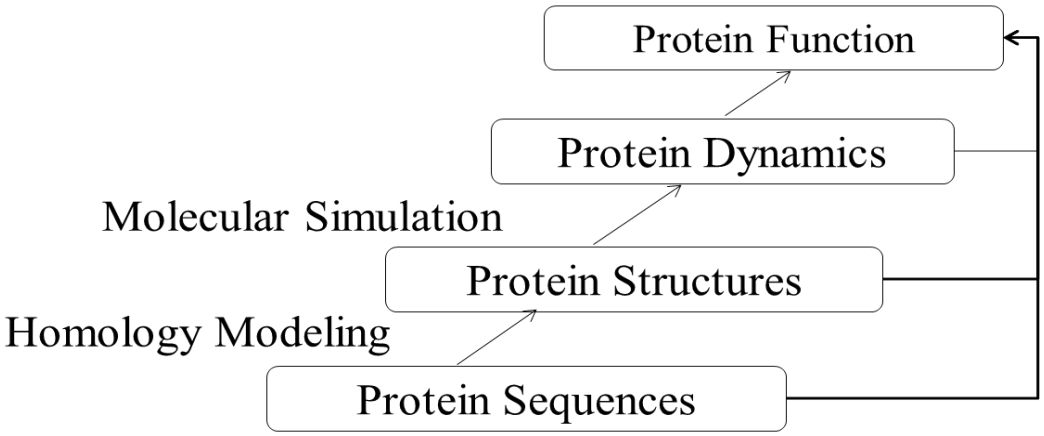

Figure 2.1: Hierarchy of the study of Protein and Function

Chapter 3

Residual Dipolar Couplings

Residual Dipolar Couplings (RDCs) have been observed as early as 1963[77]. However, it was not until the recent reintroduction of Residual Dipolar Couplings acquired by NMR spectroscopy that new opportunities for structure determination and study of internal dynamics were presented. Availability of RDCs has expanded the macromolecular investigations beyond structural characterization and into the probing of internal dynamics and molecular interaction[78]. RDCs hold the promise to report on an extensive, comprehensive range of motional timescales spanning both sub-τc and supra- τc windows[79,80]. This research mainly focuses on the implementation of RDC observables in combination with our approach to define and characterize dynamics in proteins.

3.1 RDC Foundations

The fundamental physical principle behind RDCs is the dipole-dipole (DD) interaction. Dipolar coupling measures the interaction between two magnetic nuclei in the presence of an external magnetic field. For the scope of this research, we limit our discussion to nuclei with a spin quantum number of 1⁄2. Let B_0 be the external magnetic field, *i* and *j* represent the two magnetic nuclei. The coupling magnitude D_{ij} is given by the formula in *Equations 3.1, 3.2,* from which all mathematical derivation of the RDC interactions (for a pair of spin 1⁄2 nuclei) begin.

$$D_{ij} = \left(\frac{D_{max}}{r_{ij}^3}\right)\left\langle\frac{3cos(\Theta)^2 - 1}{2}\right\rangle \tag{3.1}$$

$$D_{max} = \frac{-\mu_0 h \gamma_i \gamma_j}{8\pi^3} \tag{3.2}$$

In these Equations, D_{max} represents the magnitude of the dipolar coupling at its strongest, when the vector connecting nuclei *i* and *j* is parallel to the magnetic field B_0; where μ_0 is the magnetic permeability of free space, *h* is Plank's constant, γ_i and γ_j are the gyromagnetic ratios of nuclei *i* and *j*, respectively. r_{ij} is the distance between the nuclei *i* and *j*, Θ is the angle between the bond vector and B_0. It is important to note that the RDC value D_{ij} (reported in units of Hz) is a function of the time-dependent angle Θ(t) averaged over time *t*, as represented by the angular brackets in *Equation 3.1*.

This time-averaging phenomenon may account for molecular motions caused by natural bond vibrations, internal dynamics, or overall molecule tumbling in the solution state.

3.2 Saupe Order Tensor Matrix Formulation

Given an arbitrary molecular frame, the mathematical transformation of *Equation 3.1* can produce a computationally amiable formulation of the RDC phenomenon, as shown in *Equations 3.3, 3.4*[81,82]. In this representation of the RDC interaction, *v* signifies the normalized orientation of the interacting vector, S_{ij} denotes the ij^{th} element of the Saupe order tensor matrix[83], which is the averaged projection of axes of the molecular frame onto the direction of B_0 with $\beta_{x,y,z}$ specifying axis, δ_{ij} is Kronecker delta and the remaining constants are subsumed into a single constant, D_{max}.

$$D = D_{max} \cdot \bar{v^T} \cdot \begin{pmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{yx} & S_{yy} & S_{yz} \\ S_{zx} & S_{zy} & S_{zz} \end{pmatrix} \cdot \bar{v} \tag{3.3}$$

$$S_{ij} = \left\langle \frac{3cos(\beta_i)cos(\beta_j) - \delta_{ij}}{2} \right\rangle, ij = \{X, Y, Z\} \tag{3.4}$$

3.3 Information Content of Saupe Order Tensor Matrix

The Saupe Order Tensor matrix, also known as alignment tensor or order tensor matrix, is a symmetric, traceless 3 ×3 matrix; therefore, according to the spectral theorem of Linear Algebra, it can be decomposed in the form of $S=R\cdot \acute{S}\cdot R^T$ such that $\acute{S}$ is the diagonal matrix of eigenvalues of S known as the principal order matrix. R is the rotation matrix whose columns are the eigenvectors of S[84]. Now, *Equation 3.3* can be rewritten in the form of *Equation 3.5*. This Equation makes use of the principal order matrix $\acute{S}$ and Euler rotation matrices $R(\alpha,\beta,\gamma)$ and $R^T(\alpha,\beta,\gamma)$ where α, β, and γ are Euler angles, and T indicates conjugate transpose. The Euler rotation of Cartesian coordinates in the molecular frame *XYZ* of *Equation 3.3* results in new coordinates *XYZ* for the bond vector within the principal order frame as expressed in *Equation 3.7*.

$$D = \frac{3}{2}\cdot D_{max}\cdot \bar{v}^T\cdot R^T(\alpha,\beta,\gamma)\cdot \acute{S}\cdot R(\alpha,\beta,\gamma)\cdot \bar{v} \tag{3.5}$$

$$\acute{S} = \begin{pmatrix} A_{xx} & 0 & 0 \\ 0 & A_{yy} & 0 \\ 0 & 0 & A_{zz} \end{pmatrix} \tag{3.6}$$

$$\begin{pmatrix} \bar{X} \\ \bar{Y} \\ \bar{Z} \end{pmatrix} = R(\alpha,\beta,\gamma)\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{3.7}$$

3.4 RDC Degeneracies

RDC degeneracies arise from the mathematical form of the magnetic dipole-dipole interaction and its properties under anisotropic averaging *Equation 3.2*. Due to possible symmetries such as planarity of the protein and the alignment tensor symmetry, there exist multiple solutions for the overall orientation of a protein given a set of residual dipolar couplings (RDCs). The exact number of solutions depends on the relative orientations of the bond vectors considered and on the internal symmetry of the protein. When the rhombicity is zero, rotation about the symmetry axis of the

alignment tensor does not affect the RDCs. Therefore, the number of orientations that fulfill a given experimental RDC set belonging to a single alignment is infinite. On the other hand, for nonzero rhombicity, it was initially thought that for a rigid fragment with three or more co-planar dipolar vectors, the RDC equation resulted in 8-fold degeneracy for peptide plane orientations[85], until recently when it was shown that the analytical solution of the RDC equation contains a 16-fold degeneracy[86]. The RDC degeneracy can be reduced to four if the regular patterns of the dipolar couplings for secondary structure domains are considered[87,88]. Furthermore, it was proven that this inherent degeneracy could be resolved using two or more alignment media[89]. Nonetheless, RDC degeneracies play a vital role in ill-defined structure determination combined with conformational dynamics.

3.5 RDC Alignment

In isotropic solution, inter-nuclear magnetic dipole couplings average to zero due to rotational diffusion. Partial molecular alignment is necessary to induce an incomplete averaging of dipole couplings (RDCs). There are multiple approaches for inducing weak alignment conditions. The most common method is to depend on a medium that can be mechanically manipulated to create an anisotropic matrix that can be aligned under the presence of an external magnetic field[82]. In this case, the weak alignment of proteins is generated from the interaction between proteins and the media. Several alignment media are designed to induce partial alignments, such as bicelles[90,91], filamentous phage[92,93], and polyacrylamide gel[94,95].

3.6 Practical Aspects of RDC Acquisition in Proteins

For this research, we will focus our discussion here on a particular set of RDC measurements in proteins. RDC measurements are bond dependent; hence, in total, we consider six RDCs that are measured from a single protein residue; {*N-H, C_α-H_α,*

C-C, *N-C*, H_α-*H*, and *H*-H_α dipolar coupling}. The most frequently used RDCs are *N-H* and C_α-H_α. Due to the short bond length and the sizeable gyromagnetic ratio of H, the dipolar couplings of both RDCs are large and can be measured more accurately than other backbone one-bond RDCs. Moreover, the *N-H* RDC correlated spectra are generally well resolved, which makes the measuring process more manageable than that of C_α-H_α RDCs.

On the other hand, the C_α-H_α RDC values can be as large as 50 Hz under normal conditions due to the larger bond[96], which makes the orientational information extracted from C_α-H_α RDCs non-redundant to the one provided by *N-H* RDCs and thus valuable for structure determination and refinement.

For the *C-C* RDCs, two measurements can be computed: one for the $C_\alpha - C$ bond and the other for the $C_\alpha - C_\beta$ bond. The intrinsic RDC values of a $C_\alpha - C$ bond are about one-fifth of that of N-H RDCs, still larger than both $C - N$ and $N - C_\alpha$ RDCs. The RDCs of $C_\alpha - C_\beta$ bonds provide non-redundant orientational constraints concerning the RDCs of bond vectors located in a peptide plane. This information can be critical for the accurate determination of alignment tensors[97,98]. In addition, $C_\alpha - C_\beta$ RDCs provide information on side-chain orientation and backbone dihedral angles[99,100]. Unfortunately, measurement of $C_\alpha - C_\beta$ RDCs is complicated for reasons like poor separation of the chemical shift ranges between C_α and C_β groups, fast T2 relaxation of C_α, and medium-sized $C_\alpha C_\beta$ J-couplings.

RDCs between *H-H* bonds can be observed and applied as long-range conformational constraints for structure determination[91]. *H-H* RDCs can potentially provide distance constraints longer than 5^0A, which is the limit of traditional NOE distance constraints. Nevertheless, the neighboring protons of the bond vector significantly reduce the values of *H-H* RDCs.

3.7 Applications of RDC

The measurement of Residual Dipolar Couplings in weakly aligned proteins can potentially provide unique information on their structure and dynamics in the solution state[100,101]. Additionally, RDCs have proven to be an invaluable NMR parameter to orient multi-domain proteins and protein complexes, where the detection of inter-domain or inter-protein nuclear Overhauser effects (NOEs) are very challenging[102,103]. Moreover, recent advances in alignment media and the availability of RDCs have expanded the application of RDCs such that their use spans from automated backbone resonance assignment, structure determination, protein folding to ligand-protein and protein-protein interactions.

However, for this research, we will limit our discussion to the applications of RDCs in proteins to structure determination and the study of dynamics.

3.7.1 Protein Structure Determination

Chapter 2 highlighted the importance of determining the 3D structure of proteins. This section focuses on the role that RDCs play in structure determination. This role can be classified into two main categories: structure validation and refinement, or de novo structure calculations, performed using RDC data alone or combined with other NMR data.

Structure Validation and Refinement

Structure validation and refinement is the most common application of RDCs. For structure validation purposes, the RDC data are numerically fitted to an already existing 3D structure, determined by either X-ray crystallography or NMR spectroscopy, using software packages such as PALES[104,105] and REDCAT[106]. The fitting will result in the optimized alignment tensor that best matches the measured RDCs. *Equation 3.8* is the quality factor *Q* used to quantify the agreement between a

structure and measured RDCs[107]. In this Equation, D_m, D_c represents the measured and calculated dipolar coupling, respectively, and *rms* is the root mean square.

$$Q = \frac{rms(D_m - D_c)}{rms(D_m)} \tag{3.8}$$

On the other hand, in structure refinement, RDCs are not included in the initial structure calculations phase, which is determined mainly using NOEs distance restraints[108]. However, RDCs restraints are used in Molecular Dynamic (MD) simulation combined with simulated annealing (SA) protocols to further refine the initial structure through software packages such as XPLOR-NIH[75,76].

DE Novo Structure calculation

In the last decade, the focus of the community shifted to develop approaches that depend solely on RDC data to determine proteins backbone structures or folds. Some approaches are based on heuristic methods such as simulated annealing and complemented by Molecular Dynamics or Monte- Carlo simulation to find a solution[109–111]. Other approaches employed molecular fragment replacement (MFR) methods[90,91] instead.

3.7.2 Study of Dynamics in Protein

The potential use of RDCs to study dynamics in proteins has been recognized after the renaissance of RDCs in liquid-state protein NMR spectroscopy[82,91]. RDCs hold the promise to report on an extensive, comprehensive range of motional timescales spanning both sub- and supra- τc windows[80] as they are time-averaged from femtoseconds up to milliseconds[112,113].

Several procedures to employ RDCs for the characterization of the structure and dynamics of proteins have been proposed, including analytical deconvolution[60,101,114], the Gaussian axial fluctuations method[115], restrained molecular

dynamics simulations in which the alignment tensor is either fitted to the experimental RDCs[116,117] or calculated directly from the structure[36,118] and direct comparison with molecular dynamics simulations[119,120].

The major bottleneck in utilizing RDC data in recent years has been attributed to a lack of RDC analysis tools capable of extracting the pertinent information embedded within this complex data source. Nearly all legacy NMR data analysis software packages (i.e. XPLOR-NIH[75,76], CNS[121], Cyana[122]) have been modified to accommodate RDC restraints. In recent years, other software packages have been developed specifically for structure calculation of macromolecules from RDC data, such as REDCRAFT[123,124].

In general, all approaches fall into two categories: model-based and model-free approaches. The model-based approaches constitute some of the earliest approaches in investigating internal dynamics. These methods utilize an existing protein structure (obtained by NMR spectroscopy or X-ray crystallography) and proceed by either assuming a fixed model of dynamics[125](typically a conical motion) or a presumed stochastic model[126]. While these methods do not provide an atomic-level description of the conformational states, they can be used for quantitative analysis in the amplitude of the internal dynamics.

Alternatively, The model-free approach takes advantage of the advanced Molecular Dynamics Simulation (MD) software to simulate the averaged observable RDC data over the course of conformational changes[127–129]. These approaches can provide atomic-resolution conformational states, but at the same time, rely on an existing protein structure as the starting point of the MD simulation. MD simulation has been previously discussed in *section 2.2.2*.

Independently, both approaches (model-based and model-free) proceed in two successive steps beginning with protein structures determined under the assumption of rigidity, followed by characterization of dynamics. Although structure determination protocols based on the premise of molecular rigidity may conveniently yield a structure, the degree of similarity between a static model of a protein structure and its many conformations remains poorly understood. Recent work highlighted the possibility of obtaining erroneous structures for a protein undergoing internal dynamics[129,130]. Consequently, mapping dynamics onto a false static structure may lead to a compromised motional model. This can be attributed to the fact that it is conceptually difficult to separate structure from dynamics because the two are intimately related. Thus, any attempt in structure elucidation that disregards protein dynamics (or vice versa), may run the risk of producing faulty results.

Furthermore, the strategy of structure-first followed by dynamics next imposes a collection of superfluous data, which may include: the traditional distance-based restraints and relaxation data to establish the existence of internal dynamics. Acquisition of the additional data inflates these studies' cost and time requirements.

As an alternative, the method we present in this research provide a concurrent characterization of structure and dynamics in protein using RDC data alone, which will be explained in detail in Chapter 4

Chapter 4

Simultaneous Characterization of Structure and Dynamics

A complete understanding of the structure-function relationship of proteins requires an analysis of their dynamic behaviors in addition to the static structure. However, as mentioned in Chapter 2, all current approaches to studying dynamics in proteins have their shortcomings. A conceptually attractive and alternative approach simultaneously characterizes a protein's structure and its intrinsic dynamics[114,131,132]. Ideally, such an approach could solely rely on RDC data-carrying both structural and dynamical information. However, as previously mentioned, the major bottleneck in the utilization of RDC data in recent years has been attributed to a lack of RDC analysis tools capable of extracting the pertinent information embedded within this complex source of data. Here we present an alternative approach for concurrent characterization of structure and dynamics of proteins from RDC data.

4.1 Declaration of Terms

To facilitate further discussion, here we present a declaration of terms absent in the field and will be beneficial in understanding this approach.

4.1.1 Protein Dynamic Classifications

The traditional dynamics-based protein classification scheme is the regional scheme. Under regional mode, proteins are classified according to the area where dynamical properties are confined within the protein. If the dynamical region in the protein is confined to a single domain, most likely the loop region that connects other secondary structures, then the dynamics are classified as intra-domain.

Multiple-domains, on the other hand, refer to the cases where the dynamical region of a protein is the hinge region between multiple domains; a famous example is the 'swiveling' mechanism in pyruvate phosphate dikinase[133]. This regional classification of dynamics does not capture the dynamics-function relationship in proteins. To better facilitate the discussion of dynamics, we enumerate three distinct dynamics dimensions, namely: Temporal, Structural, and Alignment, as shown in *Table 4.1*. This comprehensive classification scheme for protein dynamics extends the characterization of the dynamics of proteins to include the time scale of the fluctuations as well as the amplitude and directionality of the fluctuations. Each of the classifications is based on specific criteria and can be subdivided into further sub-categories.

The temporal dimension of dynamics can be defined by two categories: Discrete-state and Continuous-state dynamics. The distinction between the two is solely based on the temporal occupancy of conformational states that are visited during the trajectory of the dynamics.

For the Structural mode of dynamics, we define two categories: Rigid-body and Uncorrelated modes. Many proteins need to adopt a well-defined 3D structure to carry out their function; as a result, they maintain a constant internal structure as a function of time which is defined as rigid-body dynamics. On the other hand, intrinsically disordered proteins (IDPs)[37,38] and some proteins with intrinsically disordered regions (IDRs), exhibit uncorrelated dynamics where the structure is altered as a function of time.

Finally, the Alignment mode of dynamics can be described by homogeneous and heterogeneous modes of alignment. The homogeneous mode of alignment assumes fixed alignment of the protein (within the same alignment medium) as a

function of conformational changes. In contrast, in the heterogeneous mode of dynamics, the alignment of a protein is altered as a function of the conformational changes.

In principle, all eight combined modes of dynamics should be possible with examples of all four combinations of Structural and Temporal modes of dynamics having already been identified and presented in the literature[133–136][166]. In this chapter, we investigate the combination of Rigid-body, Discrete-state dynamics with the explanation that it represents the biologically most likely event. The remaining three modes (combinations of Structural and Temporal modes) can be approximated as Rigid-body and Discrete-state dynamics in some favorable instances. The discussion related to the alignment mode of dynamics needs to be deferred for future work as it is extensive and therefore distracting. Therefore, in this work, we assume a homogeneous alignment of the protein.

4.1.2 REDCRAFT Dynamic-Profile

The first step in investigating the internal dynamics of a protein is to identify the hinge regions that give rise to the internal movement. It is also essential to establish the structural mode of dynamics (Rigid-body versus Uncorrelated) after the discovery of the onset of dynamics. The dynamic-profile[123,137] that REDCRAFT produces during structure calculation sessions can assist in discovery of the onset of dynamics and structural mode of dynamics through identification of the inconsistencies during the averaging of order tensors due to internal dynamics which results in differences between the observed order tensors of the static and rigid components of a molecule. These differences of the order tensors result in an inherent inability to produce a structure that will consistently satisfy the orientational constraints between the static and dynamical regions.

An example of a typical dynamic-profile for a static protein is shown in *Figure 4.1*. Under typical and non-anomalous conditions, a dynamic-profile will start with a very low RDC-rmsd score (due to initial lack of RDC data), then it will monotonically increase until arriving at a maximum value, followed by a final phase that is characterized by a plateauing of the RDC-rmsd score that is in agreement with the data acquisition error. Any significant departure from this typical profile is indicative of some anomalous conditions. The anomalous conditions may consist of non-standard amino acid geometries (e.g., cis-Pro, impermissible dihedrals, non-standard bond lengths, etc.), the existence of internal dynamics, or miss-assignment of the RDCs, to name a few. Of particular interest to the discussion presented here, we will observe alternations of dynamic-profile as the means to identify the onset of dynamics and distinguish different structural modes of dynamics. Dynamic profiles can be generated for forward (N-terminus to C-terminus) or backward (C-terminus to N-terminus) analysis of a given protein. The forward and backward dynamic-profiles can help to corroborate the same anomalous regions with different degrees of certainty.

Analysis of REDCRAFT's dynamic-profile takes place in two steps. The first step serves to identify any form of structural anomalies by observing any deviation from a typical profile. The second step utilizes the ability of REDRAFT to perform a fragmented structure determination of a protein (discussed in *section 4.1.3*). Once the point of anomaly is established, a new session of structure determination can be initiated a few residues in advance of the point of the anomaly.

The behavior of the dynamic-profile will be indicative of the structural mode of the dynamics (Rigid-body versus Uncorrelated), which can be established by the use of the fragmented study of a protein structure in REDCRAFT. In this context,

structure calculation can be terminated prior to the onset of dynamics, and structure calculation of a new fragment can be initiated a few residues past the onset of dynamics. Analysis of the dynamic-profile of the new fragment can help establish the structural mode of dynamics. The dynamic-profile of the new fragment undergoing Rigid-body dynamics will exhibit a typical pattern (similar to *Figure 4.1*) since it is internally rigid and consist of an internally static structure as a function of time. On the other hand, the uncorrelated dynamics will exhibit a monotonically increasing score that indicates the lack of any consistent structure as a function of time. *Figure 4.2* and *Figure 4.3* illustrate examples of these two modes of dynamics: rigid-body and Uncorrelated dynamics. In the case of Rigid-body dynamics shown in Figure 4.2, the dynamic-profile of the second fragment exhibits a normal behavior indicating successful reconstruction of a coherent structure. Dynamic-profile for the case of uncorrelated dynamics is shown in *Figure 4.3*, and it exhibits a monotonically increasing behavior that indicates the absence of a coherent structure. In the case of Rigid-body dynamics, upon recovering the structure of each domain, a measure of relative dynamics between the two domains can be established based on a comparison of their corresponding order tensors.

4.1.3 REDCRAFT Fragmented Construction

The second feature of REDCRAFT that further enables the study of the structure and dynamics of proteins emanates from its ability to conduct a fragmented reconstruction of a protein. In general, the structure of a given protein can be created in numerous fragments because of data availability, biological importance, or the study of dynamical regions that undergo Rigid-body dynamics. The Study of dynamic-profile allows for the identification of hinge regions, which can then be used to establish different dynamical domains of a protein for fragmented calculation of

structures. Relevant to the current discussion, fragmented structure calculation that can be initiated based on analysis of a dynamic-profile allows structure reconstruction of all rigid components of a protein. However, they may be dynamical with respect to each other. Once the individual structure of the rigid fragments within a protein is reconstructed, they can be assembled under a dynamics scheme that reconciles the differences in the observed order tensors across all alignment media.

4.1.4 Number of States and Rates of Occupancy

As previously stated in *section 4.1.1*, this method is designed for the Rigid-body/Discrete-state dynamics combination. In order to reflect those features in our simulated models of dynamic, we define two terms. The first term: Number of states, describes the Rigid-body portion of the dynamic model by indicating the number of states where the model maintains a constant internal structure during a period of the dynamic. For example, for a 2-state model, this model of dynamics has two states where it maintains a fixed internal structure during the interval of dynamics.

The second term: Rates of occupancy, describes the Discrete-state portion of the dynamic model by indicating the percentage of time during the interval of dynamics that each of the defined states for that model resides in. Note that the total percentage of time combined for all states for a single model should sum up to 100%.

4.2 Theoretical Treatment of Dynamics

The proposed approach permits structure calculation of proteins from a relatively sparse set of RDCs[114,123,137] in the absence of dynamics using the software package REDCRAFT[123]. Furthermore, it includes the identification and characterization of different modes of dynamics based on the dynamic-profile analysis as implemented in REDCRAFT. The basic concept behind this approach is that the discrete-state dynamical regions of a protein (when present) can be reconstructed

based on perturbation of order tensors calculated from Singular Value Decomposition (SVD)-based[106,123] mechanisms.

The presented methodology proceeds in four conceptual steps: structure determination, identification of the onset of dynamics, classification of the mode of dynamics, and reconstruction of different conformational states. The following sections detail our methodology and approach in the treatment of dynamics.

The foundation of the presented work is based on reconstructing the trajectory of dynamics using discrepancies of order tensors reported from the static and dynamic domains of a protein. Therefore, the first step in the study of dynamics is the mathematical formulation of the effects of dynamics on order tensors. *Equation 4.1* formulates changes in the observable order tensor (denoted as $\hat{S}$) as a function of time (or dynamics). In this equation, the variable j denotes the j^{th} alignment medium, and integration is performed over the entire life of the dynamics. It can be argued that biological systems perform cyclical motions (returning to some original state) therefore, the lifetime of a dynamic event can be treated as finite and periodical. A discrete approximation of the continuous function shown in *Equation 4.1* can be developed as shown in *Equation 4.2.* In this formulation, δt serves as the observation's discrete-time interval, which, if selected appropriately, can provide an accurate approximation of a temporally continuous motion. This equation can be further simplified based on relative occupancies in different states of the dynamics. This simplification occurs if the conformational continuum temporal occupancy is primarily in a small number of stable states (transient states are negligible). Under these conditions, *Equation 4.3* can be formulated and adopted to recover the primary conformational states of discrete-state dynamics. In this equation, the entity S^i_j denotes the order tensor reported from the i^{th} conformational state within the j^{th}

alignment medium, where ρ_i is the relative occupancy of the i^{th} state. The second constraint shown in this equation enforces that the sum of all relative occupancies should equate to 1 (or 100%).

$$\hat{S_j} = \int_{t=0}^{\infty} S_j(t)\, dt \tag{4.1}$$

$$\hat{S_j} = \sum_{k=1}^{n} S_j(k \cdot \delta t) \tag{4.2}$$

$$\begin{cases} \hat{S_j} = \sum_{i=1}^{n} \rho_i S_j^i \\ \text{Subject to: } \sum_{i=1}^{n} \rho_i = 1 \end{cases} \tag{4.3}$$

The following steps describe our overall strategy in the calculation of structure and characterization of dynamics (*Figure 4.4*):

1) Proceed in structure calculation with REDCRAFT[123,137] under the assumption of structural rigidity.

2) Upon identifying internal dynamics from dynamic-profile, embark on a fragmented study of dynamics for each region that exhibits internal structural rigidity.

3) After successful completion of fragmented structure calculation, establish the rigid and dynamical fragments through comparison of observed order tensors in all alignment media. Comparing order tensors across different domains can establish static and dynamic domains. Fragments can be collected into relative rigid domains based on the similarity of their order tensors.

4) Construct models of dynamics that successfully explain the differences of the observed order tensors between the static and dynamic domains in all alignment media.

The scientific basis, technical requirements, and procedures to establish steps 1-3 have been previously described and can easily be accomplished using REDCRAFT[123,137] and REDCAT[138] software packages. However, additional theoretical formulations and procedural analyses are required for step 4. To facilitate the development of procedures to achieve the objectives in step 4, we first submit that in the case of two-domain dynamics, it is possible to designate one of the domains as the static domain and the other as the dynamic domain. Although, at first, the principle of relative motion may appear to introduce some ambiguity to this designation, the presence of a third entity (the external magnetic field) against which all tumbling, vibrational motions, and internal dynamics are observed disambiguates the designation. It is, therefore, possible to uniquely designate one domain as the dynamical and the other as the static domain by simply observing the General Degree of Order (GDO)[139] for each domain.

Furthermore, *Equation 4.3* can be used as the basis of expansion to reconstruct the individual discrete states, as shown in *Equation 4.4*. In this equation, the term S^a_j denotes the anchor order tensor in alignment medium *j*. It signifies the order tensor that would have been observed if the dynamical domain was fixed and void of dynamics. The anchor order tensor can be obtained from the static domain of the protein (domain with the highest GDO). The term $\check{\zeta}i$ represents the Eulerian transformation (with its three corresponding angular arguments) that maps the Rigid-body structure of the dynamical domain from any arbitrary molecular frame to the frame that defines the *ith* state of dynamics. The average observable order tensor on the left-hand side of the equation can be obtained within REDCAT[106] by analyzing the structure of the dynamical domain using the experimentally acquired RDCs. *Equation 4.4* can be used to formulate the objective function shown in *Equation 4.5*,

which can be used to obtain solutions for four unknowns (relative occupancy and three Euler angles) that define each state of a given discrete dynamics. In this equation, the symbol $\|.\|$ denotes the magnitude of the difference matrix by summing the square of its elements. This equation can be repeated for each alignment medium, contributing five additional independent equations to the overall system of equations. In total, defining *n* discrete dynamical states will require *4n-1* (relative occupancy of the last state can always be computed by one minus the sum of all the other occupancies) degrees of freedom. At the same time, *m* alignment media will provide *5m* number of equations. Therefore, a viable solution can be obtained so long as the criterion is shown in *Equation 4.6* is satisfied.

Note that an essential fact in combining information across all alignment media is that relative occupancies and orientation of the dynamical domains with respect to the static domain remain unchanged across all alignment media. We have used the least-square minimization[140,141] routine available in Maple 14 software package to obtain the solution to *Equation 4.5*.

$$\begin{cases} \hat{S_J} = \sum_{i=1}^{n} \rho_i S_j^i = \sum_{i=1}^{n} \rho_i \cdot \xi(\alpha_i, \beta_i, \gamma_i) \cdot S_j^a \cdot \xi'(\alpha_i, \beta_i, \gamma_i) \\ \text{Subject to: } \sum_{i=1}^{n} \rho_i = 1 \end{cases} \tag{4.4}$$

$$f(\rho_{1..n}, \alpha_{1..n}, \beta_{1..n}, \gamma_{1..n}) = \begin{cases} \sum_{j=1}^{m} \left\| \hat{S_J} - \sum_{i=1}^{n} \rho_i \cdot \xi(\alpha_i, \beta_i, \gamma_i) \cdot S_j^a \cdot \xi'(\alpha_i, \beta_i, \gamma_i) \right\| \\ Subject\ to: \sum_{i=1}^{n} \rho_i = 1 \end{cases} \tag{4.5}$$

$$5m \geqslant 4n - 1 \tag{4.6}$$

4.3 Materials and Methods

We rely on synthetic data from simulated dynamics to ensure proper evaluation of the existing methods and our proposed work. The use of simulated data has several advantages during our development's early stages. Simulated data will avail ground truth for evaluation purposes and a controlled measure of the sensitivity of any method to the percentage of missing data and quality of data (as a function of signal/noise).

We will utilize synthetic data from an 83 residues FADD protein (PDB-ID 1A1Z). The FADD protein is an example of a helical protein. The helical nature of this protein presents unique challenges when studied by RDC data due to the parallel orientation of their *N-H* bonds.

4.3.1 Simulated Models of Dynamics

Our different dynamics models for the 2-state and 3-state are implemented on the same FADD protein mentioned above.

2-state rigid body dynamics

Our exploration of 2-state dynamics consisted of two different models of dynamics. The two dynamics models consist of an arc motion and a more complex motion resulting from rotation about two axes. The 2-state models of arc motion are generated by rotating the φ angle of 1A1Z protein at the 71^{st} residue (denoted by φ_{71}) by 15˚, 30˚ and 60˚. Consequently, in the arc model of dynamics, this protein is segmented into two domains: a static domain that consists of residues 1-69 and the dynamic domain that consists of residues 73-83. An example of arc motion with 60˚ perturbation of φ_{71} is shown in *Figure 4.5*. In this figure, the protein segment shown in blue is the static region, while the red and green domains represent the two

conformations of the dynamical region. It is noteworthy that this partitioning introduces additional challenges since the dynamical region is a single helix.

The more complex motion (example shown in *Figure 4.6*) was created by performing a 30˚rotation of the φ and ψ angles at residue 58 (30˚rotation of φ_{58} followed by 30˚rotation of the ψ_{58}) of the protein 1A1Z. In this case, the two domains were defined as residues 1-56 (the static region) and residues 60-83 (dynamic region). In *Figure 4.6,* the blue portion of the structure represents the static region, while the red and orange portions represent the two alternate states of the dynamical region.

3-state rigid body dynamics

Our exploration of the 3-state dynamics consists of building on the complex model of 2-state dynamics. Here the two states from the complex 2-state motion will be used as states one and two of the complex 3-state motion. The third state will be created by rotating the ψ angle of residue 58 (only ψ_{58}) by 60˚from the original structure. As in the case of 2-state complex motion, the domains will be defined by residues 1-56 and 60-83 as the static and dynamic domains, respectively. The simulated three conformations are shown in *Figure 4.7,* where the red, green, and orange fragments illustrate states 1, 2, and 3 of the dynamical domain while the static domain is illustrated in blue.

Extended-state rigid body dynamics (4 & 5 & 6)

Theoretically, our proposed method can be extended to accommodate the rigid-body/discrete-state model of dynamics with several states up to six. Our exploration of the extended-state dynamics includes 4-state, 5-state, and 6-state models that are biologically plausible; they do not violate basic geometry or result in disallowed collisions.

4.3.2 Simulated Data

Simulation of RDC values for an arbitrary pair of nuclei requires a priori knowledge of an order tensor. In this research, we use the formulation of a Saupe order tensor[142] by providing principal order parameters S_{xx}, S_{yy}, and S_{zz} and rotational Euler angles α, β, and γ. Using the atomic coordinates, order parameters, and Euler angles, REDCAT[106] was used to produce computed RDC values. We have utilized several order tensors typically observed from similar size/shape proteins in our investigations to observe the dependency of our method on order tensors, passively. *Table 4.2-Table 4.5* summarizes the order tensors used for each of our dynamics models.

The simulated RDCs used in the 2-state and 3-state part of the research contains the following set of RDCs: {*C'-N, N-H, C'-H, Cα -Hα*}. Also, note that the RDCs used in these models are accompanied by a uniform random change in the RDC values in the range of ±1 Hz to account for simulated error or noise. To simulate different percentages of occupancies, *Equation 4.7* was used to average the sets of RDCs from different conformations, where ρ_i and RDC^i_j denote the relative occupancy and RDC values for vector j in the i^{th} conformational state, respectively. In this equation, n is the total number of discrete conformational states.

$$\begin{cases} \overline{RDC_j} = \sum_{i=1}^{n} \rho_i \cdot RDC_j^i \\ \text{Subject to: } \sum_{i=1}^{n} \rho_i = 1 \end{cases} \tag{4.7}$$

4.3.3 Software Resources

RECRAFT

REDCRAFT software package[123,124,137] is designed for structure determination purely from orientational restraints. REDCRAFT is well suited for the study of

structure and dynamics because of its key feature of calculating the optimal structure by appending one residue at a time. This elongation process is consistent with the biological synthesis of proteins and allows for progressive examination of the rigidity assumption of a protein's structure. In this research, we focus on the features of Redcraft that are relevant for the study of structure and dynamics of proteins; the Dynamic-profile and Fragmented reconstruction. Both features are explained in detail in *sections 4.1.24.1.3*. The REDCRAFT software package is available for download from (https://ifestos.cse.sc.edu).

REDCAT

REDCAT software package[106,138] is used for computing the synthesized RDC values for the given order tensors in *section 4.3.2*. REDCAT is also used to perform the order tensor analysis executed in all chapters. REDCAT is available for download from (https://ifestos.cse.sc.edu).

VMD

VMD[23] is a molecular analysis and display software. For this research, VMD was used to manipulate and generate different simulated models of dynamic, and bb-rmsd analysis.

Molmol

MOLMOL[22] is a molecular analysis and display software. For this research, MOLMOL was used to manipulate and generate the different simulated models of dynamic.

Maple

We applied many procedures using Maple 14 software package, such as the Gram−Schmidt procedure[143] and least-square minimization[141] routine.

4.4 Testing and Validation

The general testing and validation strategy relies on simulated RDC data. Using simulated data during the early stages of method development is critical. Prior knowledge of the dynamics (ground-truth) allows for meaningful comparison of the recovered results to the known model of dynamics to establish the accuracy of the recovery method. Furthermore, simulated scenarios allow for systematic exploration of the strengths and limitations of the presented methodology.

The overall process consists of generating average sets of RDC data from different dynamics models, reconstructing fragmented structures based on steps 1-3 as listed in *section 4.2*, and reconstructing the dynamical states from the recovered Euler rotations (after solving *Equation 4.5*). Following reconstruction of the discrete states, validation is based on quantifying the backbone deviation between the reconstructed and target states. Our experiments, utilized a synthetic model and data from 1A1Z protein, as mentioned in *sections 4.3.1* and *4.3.2*.

The traditional method of reporting results for the reconstructed structure of a protein is based on the measure of backbone-root-mean-square-deviation (*bb-rmsd*). In this research, the simple use of *bb-rmsd* is not sufficient to report our findings since it might generate results biased in favor of this method. Therefore, a more stringent approach is required to preserve the relative orientation of a protein's fragments. Complete validation of the recovered structures in this research will comprise three consecutive steps. The first step assembles the individual structural components, including the dynamical region’s different conformations. Assembly of different conformational states will be accomplished by utilizing the Euler angles obtained from the minimization of the objective function shown in *Equation 4.5*. These Euler angles facilitate the correct orientation of the conformational domains with respect to

the static domain. Furthermore, any existing orientational degeneracies (e.g., inversion degeneracy, etc.) are automatically resolved because the information from more than one alignment medium is used. Note that upon completing this step, while the individual components of the protein are in a correct orientational relationship with respect to each other, they may exhibit a substantial translation in space.

During the second step of the validation, the target structure (including all of its conformational states) will be rotated to a relative orientation with respect to the reconstructed structure to serve as a template for measurement of the *bb-rmsd* similarity. During this step, we will use MOLMOL[22]/VMD[23] visualization software to optimally superimpose the target protein's static domain onto the reconstructed structure's static domain through rotational and translational modifications. Completion of this step provides a measure of backbone similarity between the static domain of the target and reconstructed structures.

The third evaluation step consists of establishing the orientational accuracy of the reconstructed conformations for the dynamic domain by allowing only translational modifications (disallowing orientational modification) of the domains. Calculation of *bb-rmsd* based on optimized translation and disallowing rotational modification will be performed by the software backbone that is included within the REDCRAFT software package[123,124]. It is important to note that the reported *bb-rmsd* measures are upper-bound estimates.

Table 4.1 : Modes of Dynamics

Temporal	Structural	Alignment
Discrete-state	Rigid-body	Homogeneous
Continuous-state	Uncorrelated	Heterogeneous

Table 4.2: Order parameters used for the simulated 2-state arc motion.

	S_{xx}	S_{yy}	S_{zz}	α	β	γ
S1	3.00×10^{-4}	5.00×10^{-4}	-8.00×10^{-4}	0º	0º	0º
S2	-4.00×10^{-4}	-6.00×10^{-4}	1.00×10^{-3}	40º	50º	-60º

Table 4.3: Order parameters used for the complex 2-state model of dynamics.

	S_{xx}	S_{yy}	S_{zz}	α	β	γ
S1	-3.00×10^{-4}	-5.00×10^{-4}	8.00×10^{-4}	0º	0º	0º
S2	2.00×10^{-4}	5.00×10^{-4}	-7.00×10^{-4}	-40º	-50º	60º

Table 4.4: Order parameters used for the complex 3-state, 4-state model of dynamics.

	S_{xx}	S_{yy}	S_{zz}	α	β	γ
S1	3.00×10^{-4}	5.00×10^{-4}	-8.00×10^{-4}	0º	0º	0º
S2	2.00×10^{-4}	5.00×10^{-4}	-7.00×10^{-4}	-40º	-50º	60º
S3	-7.00×10^{-4}	-1.00×10^{-4}	8.00×10^{-4}	20º	-40º	20º

Table 4.5: Order parameters used for the 5-state, and 6-state model of dynamics.

	S_{xx}	S_{yy}	S_{zz}	α	β	γ
S1	3.00×10^{-4}	5.00×10^{-4}	-8.00×10^{-4}	0º	0º	0º
S2	2.00×10^{-4}	5.00×10^{-4}	-7.00×10^{-4}	-40º	-50º	60º
S3	-7.00×10^{-4}	-1.00×10^{-4}	8.00×10^{-4}	20º	-40º	20º
S4	-4.00×10^{-4}	3.00×10^{-4}	1.00×10^{-4}	46º	-28º	152º

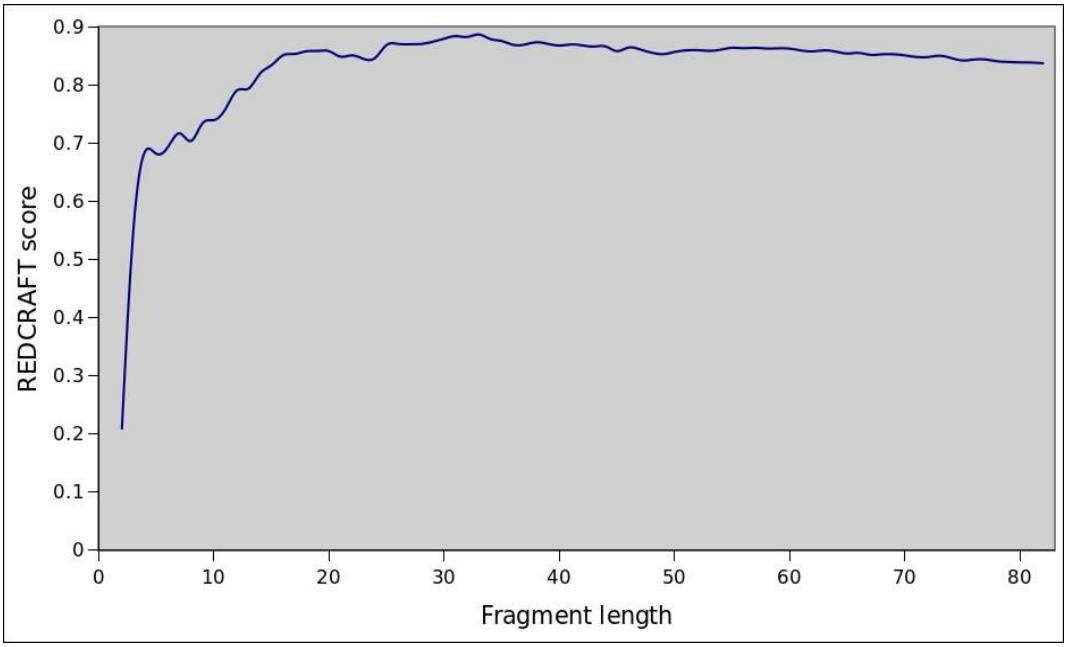

Figure 4.1: Example of a typical dynamic-profile for the protein 1A1Z in the absence of internal dynamics with simulated ±1Hz of uniformly distributed noise.

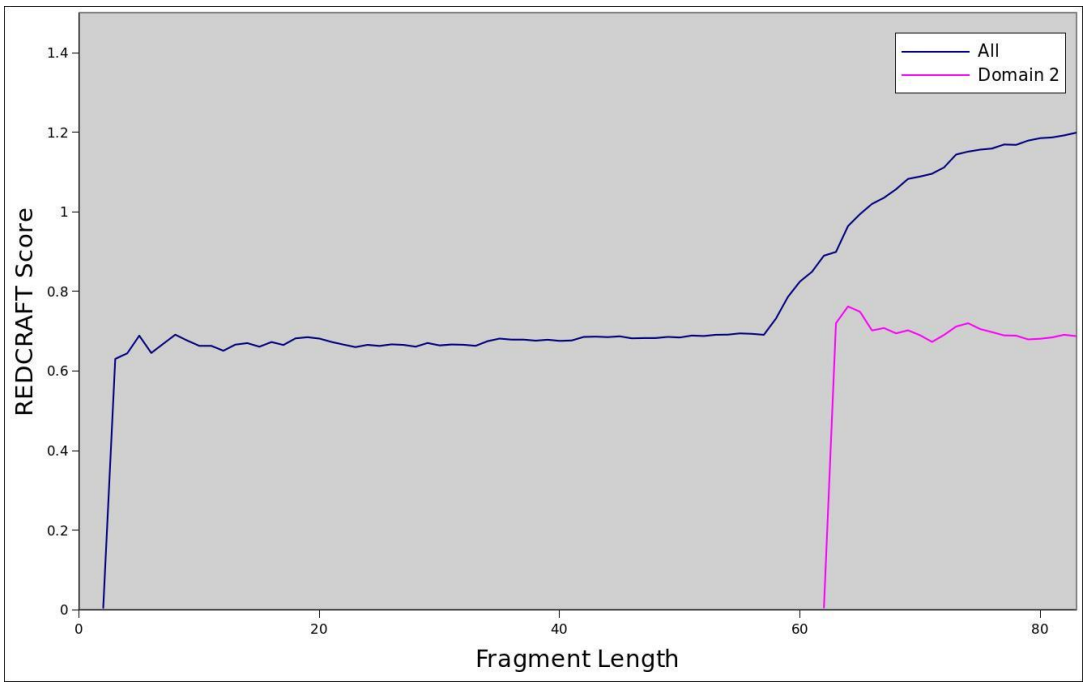

Figure 4.2: Example of a dynamic-profile of Rigid-body dynamics

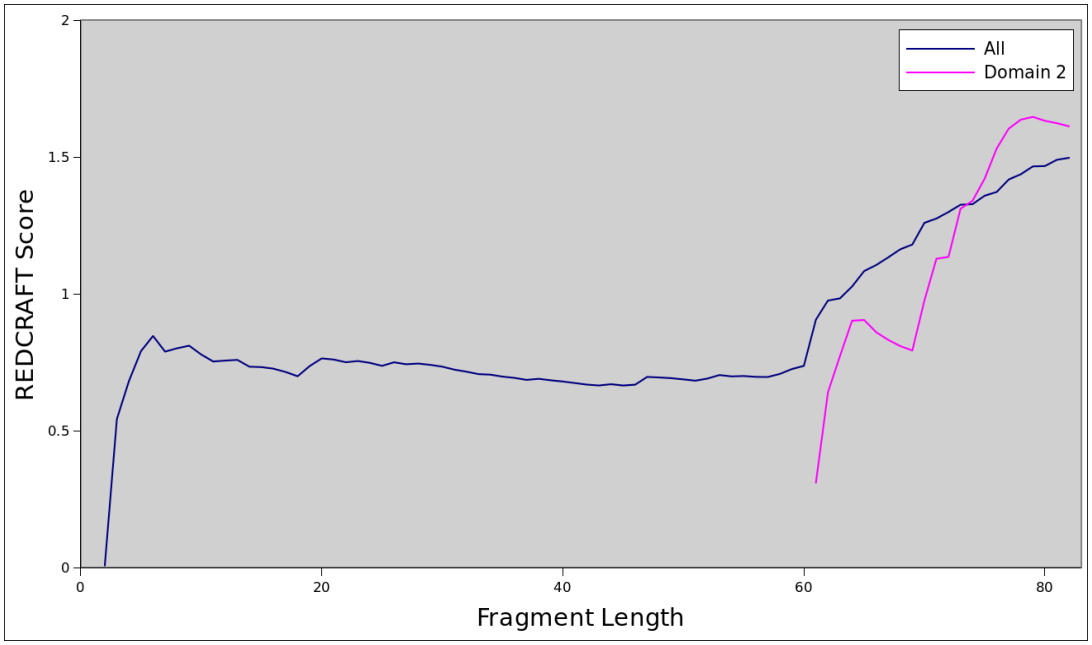

Figure 4.3: Example of a dynamic-profile of uncorrelated dynamics

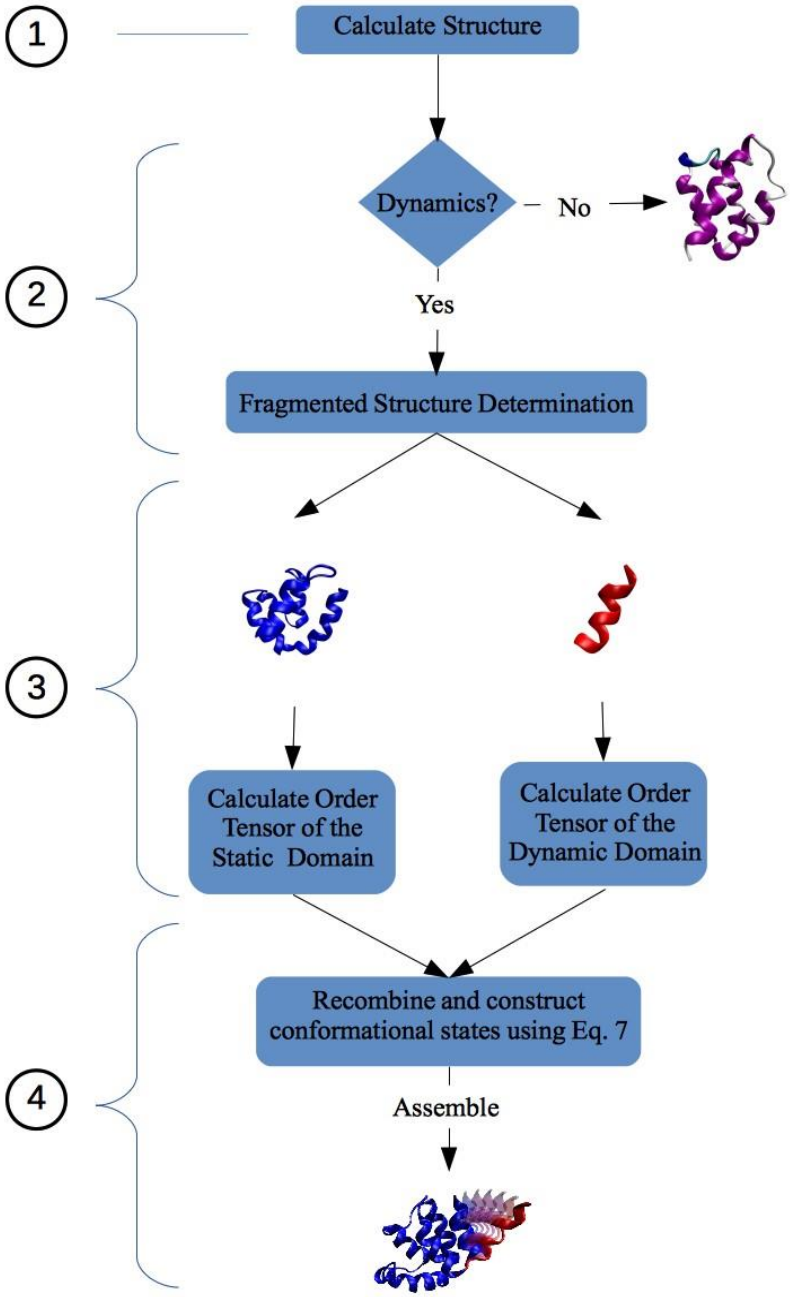

Figure 4.4: Theoretical treatment flowchart

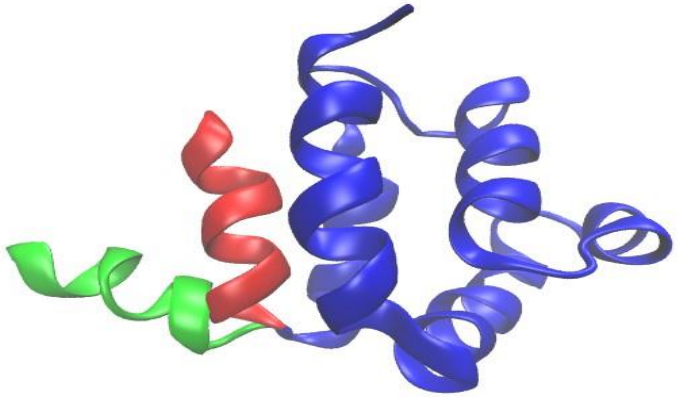

Figure 4.5: 2-state arc motion of the protein 1A1Z by 60º perturbation of the ψ71 dihedral at residue 71

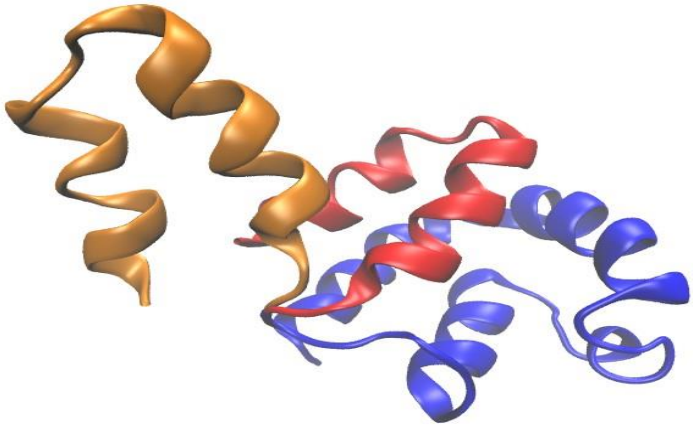

Figure 4.6: 2-state complex motion created by altering the dihedral angles of the protein 1A1Z at residue 58.

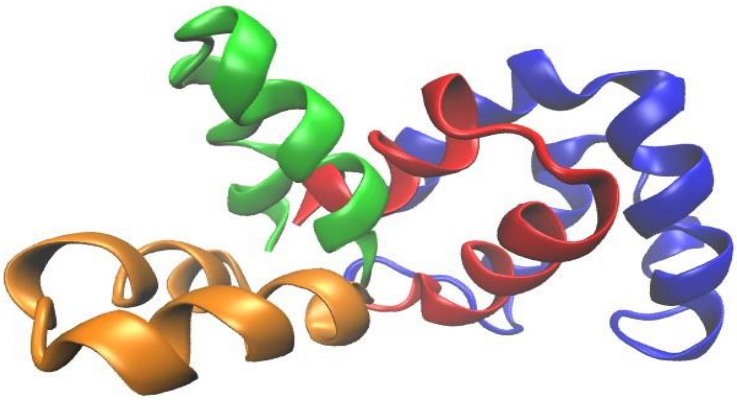

Figure 4.7: 3-state complex model of dynamics with blue representing the static domain and the dynamic domain shown in red, green and orange correspond to the conformational states 1, 2 and 3 respectively

Chapter 5

Evaluation of Molecular Dynamic Simulation Approach to Study Protein Dynamics with RDC Constraints

The previous chapter presented a method that allows for concurrent characterization of protein structure and dynamics using only RDC data. To further validate the contribution of our method, in this chapter, we evaluate the performance of one of the most popular existing approaches in the study of dynamical proteins, Molecular Dynamic Simulation, when they are executed with RDC data.

5.1 Framework

This comparative study aims to specify the requirements needed for the MDS to produce an accurate trajectory that reflects the simulated model of dynamics that the RDC data represents. The evaluation of MD simulation will start under the most pragmatic conditions and will proceed by including additional restraints to test if it is possible to produce a successful reconstruction of structure and dynamics, see *Figure 5.1*.

The first evaluation phase; will utilize only the RDC restraints from the specified model of simulated dynamics mentioned in *section 5.2.2* in the MDS.

The second evaluation phase; will utilize both RDC and partial structural constraints. The structural restraints include the dihedral angles of the domain that undergoes dynamics in the chosen simulated model.

The last evaluation phase differs from the first two by including RDCs as orientation restraints to replica-averaged MD Simulation[118,144–147].

After executing the simulations, an analysis will be performed to test the success and correctness of the simulations by confirming how compliant they are to the restraints used with the MD simulations, see *section 5.3*.

5.2 Methods and Material

Figure 5.1 states the steps performed to apply all evaluation steps of MD Simulations. The first step for all three phases is to generate a simulated dynamics model. Next, RDCs are calculated from the developed model and averaged according to the desired occupancies for the different conformations of the simulated model. The following step, the implementation, is what differs between the three evaluation steps as explained in *section 5.2.3*. The final step is the analysis of the trajectory for validation of the results.

5.2.1 Simulated Model of Dynamics

The 2-state model of dynamics from 1A1Z FADD protein described in *section 4.3.1* will be used in the evaluation phases, *Figure 4.5*. This 2-state model showcases an arc motion with φ angle rotation of 60˚at the 71st residue (denoted by φ_{71}) with a 50% occupancy rate for each state.

5.2.2 Simulated Data

Order tensor values from *Table 4.2* (*section 4.3.2*) were used to simulate RDC values for the simulated dynamics model states. *Equation 4.1* was used to average the sets of RDCs from different conformations based on a 50% occupancy rate. A uniformly distributed noise in the range of ±1 Hz was added to all RDC data. These simulated averaged RDCs are used as orientation restraint in all MD simulations performed in this chapter.

5.2.3 Molecular Dynamic Simulation

GROMACS MD software[148–150](series 2021.5) is used to perform the MD simulations executed in this chapter. The 1A1Z protein structure was minimized in terms of energy, temperature, and pressure to arrive at a more equilibrated state. In the next step, a constrained molecular dynamics simulation was performed in GROMACS (series 2021.5) software. The simulation was conducted for 2500000 steps with a step size of 0.0005 psec in a 300 K bath temperature in vacuum. A total of 1002 uniformly sampled frames were produced during the molecular trajectory to be used to calculate ensemble RDC data.

In the first and second evaluation phases, we utilized two forcefields, CHARMM27 and AMBER99 force fields, to test the effect of forcefield selection on the final result. Also, two simulations per forcefield were performed with different starting coordinates, one with state one and the other with state two.

Replica Average MD Simulations

Replica exchange simulations' primary purpose is to enhance the sampling of energy landscapes that feature many minima over accessible simulation time scales. The basic idea behind it is to simultaneously simulate multiple replicas of the same structure under the same conditions but with different temperatures and periodically exchange the coordinates of replicas between these ensembles. The probability of observing a replica in a particular ensemble depends on the potential energy and the temperature. If two states have a likelihood that they would be observed in two independent ensembles, then an exchange of the coordinates between both replicas in the ensembles occurs. A few criteria need to be taken into consideration to achieve a statistically correct sampling:

- The temperature scheme used; Temperatures should be distributed in geometric progression
- The choice of starting coordinates for each replica
- The acceptable exchange probability between ensembles; according to litreture[151,152], exchange acceptance probability around 0.2 is acceptable.
- The number of replicas needed; choosing the number to achieve the desired exchange probability depends on the temperature range and the available computational resources. Unfortunately, there is no universally correct answer; some experimentation is needed to find the right number of replicas.

For the third evaluation phase, 12 REMD simulations were performed with GROMACS software[148,149] using the RDC data generated from the simulated model discussed in *sections 5.2.1* and *5.2.2*. Each simulation was conducted for 5000000 steps with a step size of 0.002 psec in vacuum with AMBER99 forcefield. The number of replicas used was 4, 6, and 8. Multiple temperature schemes were sampled from the 300-400K to the 300- 700k.

5.3 Testing and Validation

Validation of the results is done in two steps. The first validation phase compares the bb-rmsd of the frames generated in the simulation trajectory with the bb-rmsd value between state1 and state2 from the simulated model equal to 3.7 Å to see if the MD simulation has generated enough magnitude of motion that reflects the actual dynamics.

The second step will use the simulated RDC generated from the simulated model of dynamic described in *sections 5.2.1* and *5.2.2* and employed as orientation restraints in the MD simulations to validate that they closely reproduce the conformational properties of the original simulated model trajectory. The RDC values

are back calculated as average over all frames in the produced trajectory. These resulting average RDC values are expected to match the simulated RDC data used as restraints.

Using the trajectory produced from the MD simulation, 1002 frames were generated uniformly to span the entire course of the dynamics. Auxiliary tools were used to separate these frames in a PDB format and generate a corresponding REDCAT file. The software package REDCAT[106] was used to calculate the RDCs values for backbone {*C'-N, N-HN, Cα -Hα, C'-HN, H-Hα, and Hα-H}* for each frame of the trajectory using the order tensors shown in *Table 4.2* in two alignment media. REDCAT's internal utility functions were used to create the observable RDCs by averaging the individual RDCs across the entire course of the dynamics (defined by 1002 frames). Finally, REDCAT's internal utility functions were used to calculate the difference between these averaged RDCs and the simulated RDCs used as restraints in the simulation to calculate the order parameter and violations of the MD trajectory.

For the replica average MD simulation, validation will follow in two steps: in the first step, the exchange probability has to be around .2 for the REMD to be deemed producible, the second step will involve the same order tensor analysis performed in the case of regular MD simulation described above.

5.4 Results and Discussion

Molecular Dynamic Simulation with Orientational Restraints

Table 5.1 contains the results of the bb-rmsd analysis of each MD simulation trajectory run in both forcefields and with each starting structure. As can be seen, the highest value of bb-rmsd during the simulation was less than the expected 3.7 Å. While the choice of forcefield used has an effect on the magnitude of motion produced, it is not as significant as the effect of the starting structure chosen. The most considerable

magnitude of motion resulted from starting the simulation from state2 in AMBER99 forcefield. This result is expected; since state1 is the most stable state while state2 is simulated and not plausible in real life, it is improbable for the protein to move and violate the parameters optimized in the MD software.

On the other hand, *Table 5.2* contains the order parameter calculated from the back calculated average RDCs frames the simulation trajectory. For each alignment media, we calculated the order parameter, the number of violations between the back calculated RDCs, and the simulated RDCs, and the highest value of violations recorded to compare with the 1Hz of error of noise added to the simulated RDCs. In *Table 5.2*, although the calculated order parameter is close to the ones used in creating the simulated RDCs, there is a slight difference. In addition, there are many violations with tremendous values, which indicates that the simulation run doesn't comply with all the RDC restraints.

Molecular Dynamic Simulation with Orientational and Dihedral Restraints

The results of the MD simulation run with RDC restraints, and the addition of dihedral restraints were very disappointing. The results of the bb-rmsd analysis of the trajectory of each MD simulation were the same as the MD run with RDC restraints only. These results prove that the dihedral restraints added no more information to the simulation.

Moreover, the order parameter analysis results shown in *Table 5.3* agree with the results from the previous section. The calculated order parameter is comparable to the ones used to create the simulated RDCs with a slight difference. In addition, there are many violations with substantial values which indicates that the addition of dihedral restraint didn't provide further improvements, hence deemed unnecessary.

Replica Averaged MDS with Orientational Restraints

The different simulations overall results were unproductive; both REMD runs executed with 8 replicas failed. The first one used a temperature scheme of 300-700K and failed due to the system's violation of the LINCS constraint algorithm applied in GROMCS (some bond rotation exceeded the maximum allowed threshold because of the high temperature). On the other hand, the second attempt of REMD with 8 replicas was executed with 300-450k. As a result, the ensembles were too similar for any exchange to take place.

The same issues were encountered with the 6 and 4 replica ensembles; high-temperature ranges, simulation halts, and low-temperature ranges yield no exchange. Only one generated an exchange rate of .22 between two of the four replicas with the few completed runs. With further analysis, the results of this successful run were not any better than the results of a single MD simulation with RDC restraint discussed above.

5.5 Conclusions

The results reported in this evaluation indicate that using MD simulation with RDC data alone will not yield positive results, which agrees with the results reported in the literature.

Although MD simulation is a valuable tool for studying protein dynamics, they are simulations, and to some degree, we get back what we put into the simulation. The choice of forcefield to use, the coordinates of the starting structure, what violations to deem acceptable, and even the option of MD software to use all affect the simulation results.

All this points to the need for a more conclusive, dependable, reproducible, and mathematical approach in studying dynamics as the one presented in Chapter 4.

Table 5.1: BB-rmsd for MD run trajectory with both cases of using RDC restraint alone or with the addition of dihedral restraint

Force Field	Starting Structure	BB-rmsd in A^0		
		Minimum	Average	Maximum
CHARMM27	State 1	0.0005	0.623	1.346
	State 2	0.0005	0.937	1.433
AMBER99	State 1	0.0005	0.622	1.118
	State 2	0.0005	1.169	2.027

Table 5.2: Order Tensor analysis of 1002 frames in trajectory of MD run with RDC restraints

Force Field	Starting Structure	Align-ment Media	Calculated Order Parameter			Number of Vio-lations	Max Value of Violation
			S_{xx}	S_{yy}	S_{zz}		
CHARMM27	State 1	M1	3.1×10^{-4}	4.7×10^{-4}	-7.8×10^{-4}	71	4.5 Hz
		M2	-3.7×10^{-4}	-5.6×10^{-4}	9.4×10^{-4}	112	9.7 Hz
	State 2	M1	3.1×10^{-4}	4.8×10^{-4}	-7.8×10^{-4}	90	4.8 Hz
		M2	-3.7×10^{-4}	-5.6×10^{-4}	9.3×10^{-4}	114	11.8 Hz
AMBER99	State 1	M1	3.0×10^{-4}	4.9×10^{-4}	-8.0×10^{-4}	131	5.3 Hz
		M2	-3.8×10^{-4}	-5.8×10^{-4}	9.6×10^{-4}	167	15.8 Hz
	State 2	M1	3.1×10^{-4}	4.7×10^{-4}	-7.9×10^{-4}	113	6.3 Hz
		M2	-3.7×10^{-4}	-6.1×10^{-4}	9.8×10^{-4}	129	4.3 Hz

Table 5.3: Order Tensor analysis of 1002 frames in trajectory of MD run with dihedral and RDC restraints

Force Field	Starting Structure	Alignment Media	Calculated Order Parameter			Number of Violations	Max Value of Violation
			S_{xx}	S_{yy}	S_{zz}		
CHARMM27	State 1	M1	3.1×10^{-4}	4.7×10^{-4}	-7.8×10^{-4}	81	3.24 Hz
		M2	-3.7×10^{-4}	-5.7×10^{-4}	9.4×10^{-4}	118	10.7 Hz
	State 2	M1	3.1×10^{-4}	4.7×10^{-4}	-7.8×10^{-4}	101	5.1 Hz
		M2	-3.8×10^{-4}	-5.6×10^{-4}	9.4×10^{-4}	104	5.8 Hz
AMBER99	State 1	M1	3.0×10^{-4}	5.0×10^{-4}	-8.0×10^{-4}	91	8.5 Hz
		M2	-3.8×10^{-4}	-5.7×10^{-4}	9.6×10^{-4}	122	15.9 Hz
	State 2	M1	3.2×10^{-4}	4.8×10^{-4}	-7.9×10^{-4}	83	16.4 Hz
		M2	-3.6×10^{-4}	-6.0×10^{-4}	9.6×10^{-4}	108	10.5 Hz

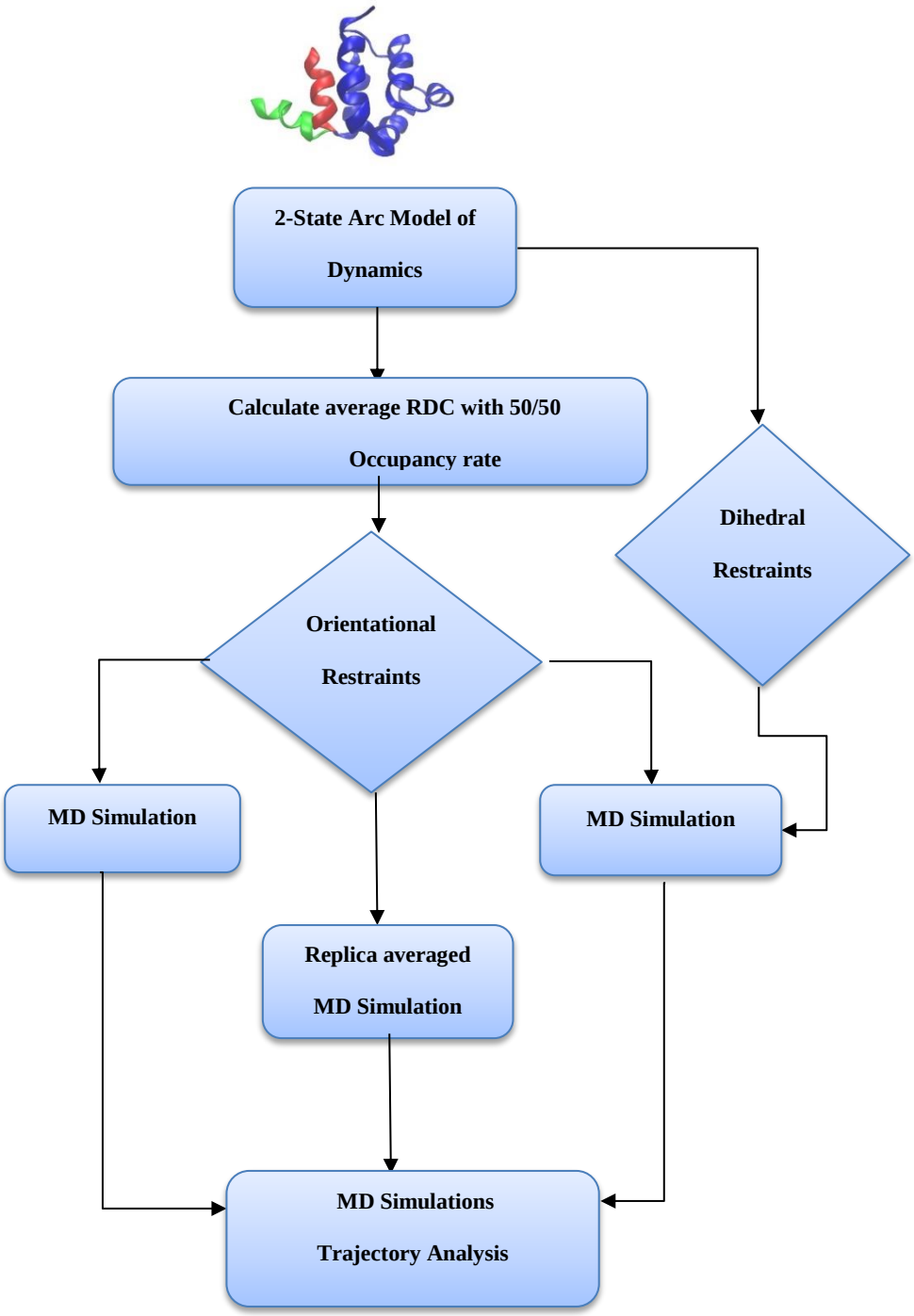

Figure 5.1: Flowchart of the MD simulation evaluation phases.

Chapter 6

Results of the Simultaneous Characterization of Structure and Dynamics Approach

In the following sections, we provide results supporting our approach in the treatment of structure and dynamics of proteins discussed in Chapter 4. Our results first focus on the ability of REDCRAFT to accurately identify the onset of dynamics and allude to the structural mode of the dynamic. Next, we present our results in the reconstruction of conformations from two, three, and four state dynamics. We conclude our results with a discussion of the limitations of the presented work and anomalies related to the study of dynamics from RDC data.

6.1 Discovery of Onset of Dynamics and Structural Modes of Dynamic by Dynamic-Profile from REDCRAFT

As the first example in the utility of the dynamic-profile, we present the case of 2-state dynamics. We utilized the dynamical model shown in *section 4.3.1* (two states generated through perturbation of φ71) and utilized the averaged RDCs to perform a forward and reverse structure calculation of the protein 1A1Z. An example of the dynamic-profile of a 2-state dynamic can be seen in *Figure 6.1*. In this figure, the blue and red profiles correspond to the forward and reverse structure calculations, respectively. In contrast to the typical profile shown in *Figure 4.1*, an anomalous increase has manifested in the vicinity of residue 71 on both forward and reverse sessions of REDCRAFT. This result is consistent with the model of dynamics that was used during this exercise. While both forward and

reverse analyses exhibit an increase in the RDC score of the dynamic-profile; this phenomenon is more prominently observed in the case of the reverse structure determination than the forward. This inequality arises because in the case of forward run, the anomalous region is discovered after 73 residues and RDC data from only 11 residues are inconsistent with respect to the remainder of the protein. This small portion will have a relatively more minor effect in perturbation of the RDC score reported by REDCRAFT.

In contrast, a much larger discrepancy is observed in the case of reverse folding of the protein because a much larger portion of the data can contribute to any observed inconsistencies. A similar exercise was conducted for the 3-state dynamics described in *section 4.3.1*. In this model geometry of the 58th residue was altered to simulate dynamics. *Figure 6.2* illustrates the dynamic-profile of this 3- state model of dynamics. Consistent with the model of dynamics, the dynamic-profile identifies the onset of the dynamics at around residue 57-58. However, unlike the previous exercise, and since a larger portion of the protein is undergoing dynamics, an approximately equal increase is observed in the dynamic-profiles of forward and reverse structure calculation by REDCRAFT.

The above two examples demonstrated the ability of REDCRAFT in identifying the onset of internal dynamics via the use of dynamic-profile analysis. The structural mode of dynamics (Rigid-body versus Uncorrelated) can be established by using the fragmented study of a protein structure in REDCRAFT as described in *section 4.1.2*.

6.2 2-State Rigid Body Dynamics

We begin discussion of our results with the case of 60˚ arc motion (shown in *Table 6.1*). As can be seen in each table, the relative occupancies of each exercise are

listed on the left-hand most column of the table. The “Minimum” value corresponds to the lowest value (in units of Hz scaled to N-H vectors) obtained from minimizing the objective function shown in *Equation 4.5*. This value helps to establish the success of the general approach; minimum values in the vicinity of the experimental noise indicate successful reconstruction of the states. The “Conformation #” denotes the conformation number, the BB-rmsd, and relative occupancies correspond to each state's structural/orientational similarity and relative occupancy, respectively.

In general, all states of the dynamics were reconstructed very accurately, including orientation of the states and relative occupancies. In some instances (such as 50/50), the relative occupancies were in error by as much as 13%. The only exercise that exhibited an anomalous outcome was the case of 90/10. Here the first conformation was reconstructed with a high degree of accuracy (0.37Å with respect to the target protein), while the second state was created with bb-rmsd of 9.4Å with respect to its corresponding state. Our explanation for this behavior is the low relative occupancy of this particular scenario marginalizes the perturbation of RDCs due to dynamics. The small perturbation of RDCs (compared to the noise) has rendered its effect moot. This phenomenon is observed in other instances discussed in the future sections. Since the effect of the second state is negligible, it was reconstructed in nearly an irrelevant orientation giving rise to its high bb-rmsd to the target conformation.

Next, to further investigate the sensitivity of our method with respect to the magnitude of motion, we reduced the change of φ71 to 30°. The results of these experiments are shown in *Table 6.2* and are very similar to that of the 60° dynamics except for the case of 80/20. In this case, the second state could not be reconstructed with much accuracy. We suspect the reason for this inconsistency is the combination

of a smaller angle of rotation and lower relative occupancy. It appears that at only 30° rotation, a state with less than 20% occupancy gets subsumed into the original state, as occurred in the case of 90/10 of the previous example.

To further investigate the effect and behavior of our approach on small and negligible motions, we investigated the case of 15˚ arc motion. Although internal variation exists in the second domain, the REDCRAFT dynamic profile does not identify internal dynamics indicating the absence of any anomalous behavior. Despite this finding, we proceeded to reconstruct the two orientations, and the results are shown in *Table 6.3*. The overarching observation that can be concluded from the results in this table consists of accurate reconstruction of the first state and nearly complete failure to reconstruct the second state (both orientation and relative occupancy) despite achieving a low value for the objective function. This behavior is consistent with the results for 60˚ arc motion and 90/10 occupancy exercise. Both of these exercises help establish the boundaries of the information content of the RDC data. In summary, the particular instance of 15° motion did not provide sufficient alteration of RDCs (and therefore order tensors) to indicate the existence of internal dynamics at any relative occupancies.

The results from the complex 2-state model are shown in *Table 6.4* and portray an outcome consistent with the case of arc motion. Both conformations were reconstructed with a high degree of accuracy despite the complexity of the dynamics. However, it can be seen that as the relative occupancy of the second state (the ending state) decreases, the predicted orientation's bb-rmsd to the target structure increases as well. This deterioration in performance is observable in the case of 80/20 and clearly so in the case of 90/10. In both cases, the first state was reconstructed with reasonable accuracy, while the reconstructed second state deteriorated as a function of

occupancies. A relative occupancy of 10% can be seen as almost negligible in the course of a dynamic movement when using RDCs with ±1Hz of error.

6.3 3-State Rigid Body Dynamics

Results of the 3-state complex dynamics are shown in *Table 6.5* for several different relative state occupancies. Similar to the case of 2-state, the relative occupancies of each exercise are listed on the left-hand most column of this table. The "Minimum" value corresponds to the lowest value (in units of Hz scaled to N-H vectors) obtained from minimizing the objective function shown in *Equation 4.5*. This value helps to establish the success of the general approach; minimum values in the vicinity of the experimental noise indicate successful reconstruction of the states. The "Conformation #" denotes the conformation number, the BB-rmsd, and Relative Occupancies correspond to the structural/orientational similarity and relative occupancy of each state, respectively.

As seen in *Table 6.5*, our presented method has successfully reconstructed the conformational states and rates of occupancies with less than 2Å in structural resolution. We note a higher variability in the recovered measure of relative occupancies, variations as much as 0.19%.

6.4 Extended-State Rigid-Body Dynamics

To create an extended-state model of dynamics, large-size proteins are needed to develop a biologically plausible model that does not violate basic geometry or result in disallowed collisions. However, large proteins exhibit complicated structures of sub-domains. In addition, as mentioned in *section 6.2* and discussed in detail in *section 6.6*, our system failed to detect dynamics with a small magnitude of rotation or/and occupancy rate less than 20%. Hence, it becomes increasingly difficult to find four independent states that satisfy our method conditions and basic geometry while

still staying with occupancy rates larger than 20% to represent a 4-state model. Moreover, it is impossible to develop a 5-state or 6-state model with an occupancy rate larger than 20%. Also, note that while an extended-state model of dynamics could exist in theory, in actuality, the highest model of dynamics that has been described in the literature is a 4-state model of dynamics. For these reasons, the extended-state models are left for future work.

6.5 Modeling of 2-State Dynamics as 3-State Dynamics or a 3-State as 2-State

All the dynamic models discussed previously assume a priori knowledge in the number of dynamical states. It is reasonable to consider the cases where the number of stable conformations is unknown prior to analysis. In this case, a parsimonious approach can be employed to assist in the discovery of the appropriate jump states. More specifically, 2-state dynamics can serve as a starting point of any investigation. The total number of conformational states can be explored incrementally until a satisfactory result is achieved. A satisfactory result is quantified by the fitness of experimental data to the computed ones to within the data acquisition error.

To demonstrate this approach, analysis of 3-state dynamics described in *Section 4.3.1* was utilized. Based on this parsimonious approach, the reconstruction of conformations will proceed based on the assumption of 2-state dynamics. Results of the 2-state recovery of the 3-state dynamics are shown in *Table 6.6*. In principle, and in agreement with the results shown in this table, the incomplete modeling should be problematic and manifest itself in an unacceptably high objective function value. In *Table 6.6*, the left-most column indicates the true relative occupancies of each state during the simulation of dynamics. The information marked as “Minimum” denotes fitness of the objective function (*Equation 4.5*) scaled to the units of Hz for *N-H*

vectors. Increasing the number of states to 3, produces the results shown in *Table 6.5* with minimum values of the objective function that indicate successful recovery of states. The cases of 60/30/10 and 50/30/20 exhibited potentially acceptable objective functions because they can be treated as a two-state dynamics by disregarding the state with relatively low occupancies (10% or 20%). This serves as another affirmation that relative occupancies of less than 20% are potentially negligible within the framework of ±1Hz of experimental error.

Conversely, 2-state dynamics can be forced to be modeled as 3-state. In theory, a 2-state dynamics should be classified as a 3-state dynamic where two of the recovered states correspond to the two conformations and a third phantom state with a relative occupancy of 0%. To illustrate this point, two experiments in which 2-state models of dynamics were forced into a 3-state recovery. Recovery of 3-state dynamics requires RDC data from at least three alignment media. The three alignment media shown in *Table 4.4*, along with the two-state arc motion and two-state complex motion described in *Sections 4.3.2, 4.3.1* was utilized in this exercise. In both cases, equal 50% relative occupancies were used to simulate the RDC data.

As shown in *Table 6.7*, Conformation 3 in both the arc and the complex motions have occupancy rates of 0.03 and 0.01, respectively. These conformations correspond to neither state 1 nor state 2 of their respective model of dynamics. An occupancy rate of 1-3% is, in practice, negligible, making the corresponding state clearly inconsequential. *Figure 6.3* shows the results from the two-state arc motion with the extraneous conformation shown in yellow and the two conformations (1 and 2 in *Table 6.7*) that align well with the original model of dynamics.

The results of these experiments are essential because they reveal exciting insights into the presented method. They show that the inclusion of more data

improves the preciseness in the reconstruction of domains, as evidenced by the low bb-rmsd of the reconstructed states in *Table 6.7*. In addition, it can be reasonably argued that in application to natural examples of dynamics where the actual number of discrete states are not known, our method appears to successfully identify the correct number of states that describe a model of dynamics.

6.6 Limitations in Recovery of Discrete State Dynamics

In *section 6.2,* we demonstrated the inability of the present work to reconstruct conformations with relative occupancies less than 10% in the cyclical life of a dynamical event. In addition, we also demonstrated the limitation in reconstructing conformational changes imposed by as small as 15 ° of arc motion. These limitations are due to the overall contribution of dynamics (either relative occupancy or small motion) relative to the experimental precision of data acquisition. Therefore, in the absence of any other information, these types of limitations are universal, and no approach will be able to recover useful information related to the internal dynamics.

Another category of limitations can be described as inherent to any approach that relies on an analysis of order tensors to recover conformational information. More specifically, these limitations arise from the fact that order tensors span a five-dimensional space (degrees of freedom of an order tensor). Therefore, regardless of the number of alignment media explored, no more than five independent alignment tensors can be obtained. Considering the relationship shown in *Equation 4.6*, this imposes a limitation on our approach of recovering a maximum of six conformations.

Table 6.1: Results for 60º arc motion.

50/50	Minimum	9.13×10^{-11} (0.23 Hz)	
	Conformation	1	2
	BB-RMSD	0.93Å	1.02Å
	Relative Occupancy	0.63	0.37
60/40	Minimum	1.25×10^{-10} (0.27 Hz)	
	Conformation	1	2
	BB-RMSD	0.38Å	0.42Å
	Relative Occupancy	0.61	0.39
70/30	Minimum	1.31×10^{-10} (0.28Hz)	
	Conformation	1	2
	BB-RMSD	0.44Å	0.45Å
	Relative Occupancy	0.72	0.38
80/20	Minimum	2.37×10^{-10} (0.37 Hz)	
	Conformation	1	2
	BB-RMSD	0.59Å	1.52Å
	Relative Occupancy	0.85	0.15
90/10	Minimum	2.29×10^{-10} (0.37 Hz)	
	Conformation	1	2
	BB-RMSD	0.37Å	9.4Å
	Relative Occupancy	0.896	0.103

Table 6.2: Results for 30° arc motion

50/50	Minimum	5.12×10-11 (0.17 Hz)	
	Conformation	1	2
	BB-RMSD	0.46Å	0.44Å
	Relative Occupancy	0.45	0.55
60/40	Minimum	7.27×10-11 (0.2 Hz)	
	Conformation	1	2
	BB-RMSD	0.5Å	0.58Å
	Relative Occupancy	0.66	0.34
70/30	Minimum	1.12×10-10 (0.26Hz)	
	Conformation	1	2
	BB-RMSD	0.65Å	2.00Å
	Relative Occupancy	0.88	0.12
80/20	Minimum	9.4×10-11 (0.23 Hz)	
	Conformation	1	2
	BB-RMSD	0.66Å	5.2Å
	Relative Occupancy	0.95	0.05
90/10	Minimum	4.49×10-11 (0.16 Hz)	
	Conformation	1	2
	BB-RMSD	0.5Å	6.9Å
	Relative Occupancy	0.98	0.02

Table 6.3: Results for 15º arc motion.

50/50	Minimum	$7.\times10^{-10}$ (0.65 Hz)	
	Conformation	1	2
	BB-RMSD	0.78Å	7.8Å
	Rate of Occupancy	0.96	0.04
60/40	Minimum	1.9×10^{-10} (0.34 Hz)	
	Conformation	1	2
	BB-RMSD	0.73Å	9.6Å
	Rate of Occupancy	0.85	0.15
70/30	Minimum	3.48×10^{-10} (0.46Hz)	
	Conformation	1	2
	BB-RMSD	0.68Å	9.3Å
	Rate of Occupancy	0.88	0.12
80/20	Minimum	7.13×10^{-10} (0.65 Hz)	
	Conformation	1	2
	BB-RMSD	0.66Å	9.1Å
	Rate of Occupancy	0.88	0.12
90/10	Minimum	1.14×10^{-9} (0.82Hz)	
	Conformation	1	2
	BB-RMSD	0.58Å	5.2Å
	Rate of Occupancy	0.95	0.05

Table 6.4: Results for 2-state complex dynamics experiments.

50/50	Minimum	2.27×10^{-10} (0.36 Hz)	
	Conformation	1	2
	BB-RMSD	0.76Å	0.83Å
	Rate of Occupancy	0.42	0.58
60/40	Minimum	1.6×10^{-10} (0.31 Hz)	
	Conformation	1	2
	BB-RMSD	1.1Å	1.4Å
	Rate of Occupancy	0.47	0.53
70/30	Minimum	1.4×10^{-10} (0.29 Hz)	
	Conformation	1	2
	BB-RMSD	1.2Å	1.6Å
	Rate of Occupancy	0.53	0.47
80/20	Minimum	6.04×10^{-11} (0.19 Hz)	
	Conformation	1	2
	BB-RMSD	0.69Å	2.3Å
	Rate of Occupancy	0.66	0.34
90/10	Minimum	1.7×10^{-10} (0.32 Hz)	
	Conformation	1	2
	BB-RMSD	0.83Å	6.33Å
	Rate of Occupancy	0.95	0.05

Table 6.5: Results for 3-state dynamics experiments.

50/25/25	Minimum	2.9×10^{-11} (0.13 Hz)		
	Conformation #	1	2	3
	BB-RMSD	0.95Å	1.9Å	0.67Å
	Rate of Occupancy	0.42	0.32	0.26
34/33/33	Minimum	2.6×10^{-11} (0.12 Hz)		
	Conformation #	1	2	3
	BB-RMSD	1.4Å	0.38Å	1.3Å
	Rate of Occupancy	0.25	0.41	0.33
50/30/20	Minimum	3.4×10^{-11} (0.14 Hz)		
	Conformation #	1	2	3
	BB-RMSD	1.08Å	1.5Å	0.4Å
	Rate of Occupancy	0.32	0.34	0.34
60/30/10	Minimum	7.8×10^{-11} (0.21 Hz)		
	Conformation #	1	2	3
	BB-RMSD	0.64Å	1.3Å	1.3Å
	Rate of Occupancy	0.52	0.35	0.1

Table 6.6: Results for modeling of a 3-state dynamic as a 2-state.

True Occupancies	Minimum
34/33/33	2.99×10^{-8} (4.21 Hz)
50/25/25	4.097×10^{-7} (15.56 Hz)
50/30/20	5.8×10^{-9} (1.85 Hz)
60/30/10	6.9×10^{-9} (2.02 Hz)

Table 6.7: Results for simulating 2-state dynamics in our 3-state dynamic equation.

Arc Motion (50/50/0)	Minimum	3.15×10^{-13} (0.013 Hz)		
	Conformation	1	2	3
	BB-RMSD	0.7Å	0.63Å	4-7Å
	Rate of Occupancy	0.47	0.50	0.03
Complex Motion (50/50/0)	Minimum	1.6×10^{-10} (0.31 Hz)		
	Conformation	1	2	3
	BB-RMSD	0.66Å	0.6Å	9-10Å
	Rate of Occupancy	0.44	0.55	0.01

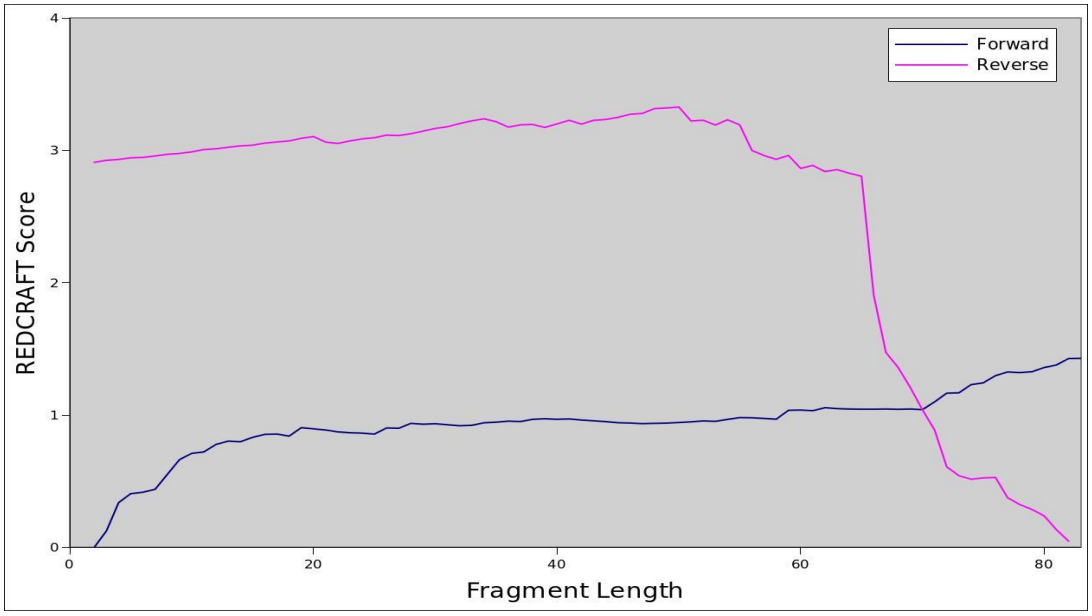

Figure 6.1: An example of the dynamic profile for a 2-state model of dynamics. The blue line represents the RDC-RMSD score from REDCRAFT in forward configuration and the red line denotes REDCRAFT in reverse configuration. In this particular model of dynamics, the phi angle of the 71st residue of the protein was rotated 60 degrees. The dynamic profile indicates an anomaly around that same area.

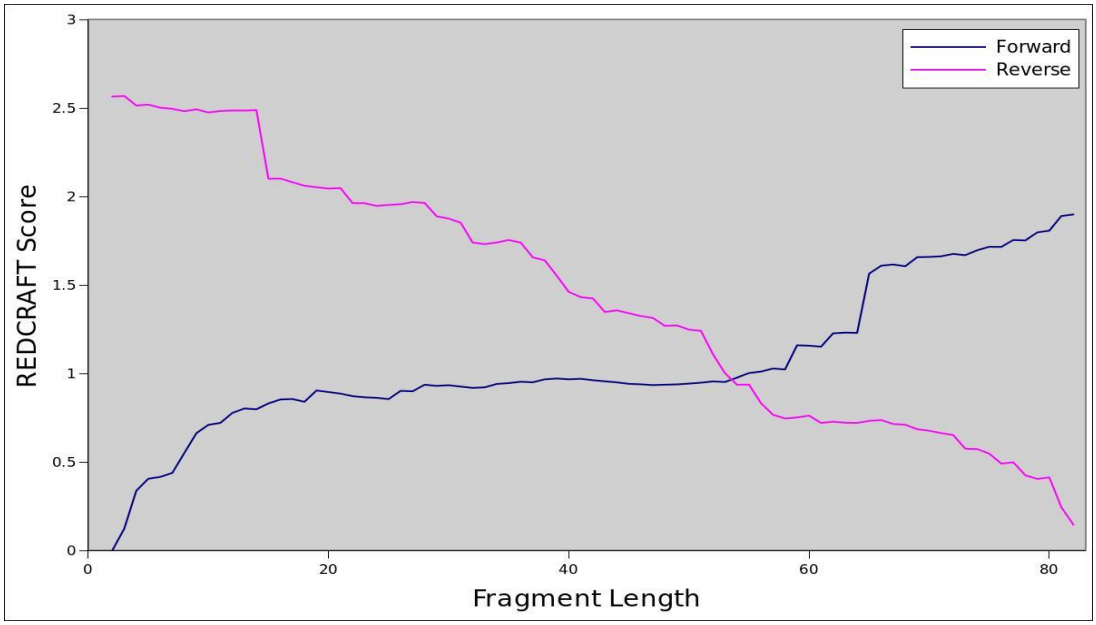

Figure 6.2: An example of the dynamic profile for a 3-state model of dynamics. The blue line represents the RDC-RMSD score from REDCRAFT in forward configuration and the red line denotes REDCRAFT in reverse configuration. In this particular model of dynamics, the 58th was mutated to simulate dynamics. The dynamic profile indicates an anomaly around that same area.

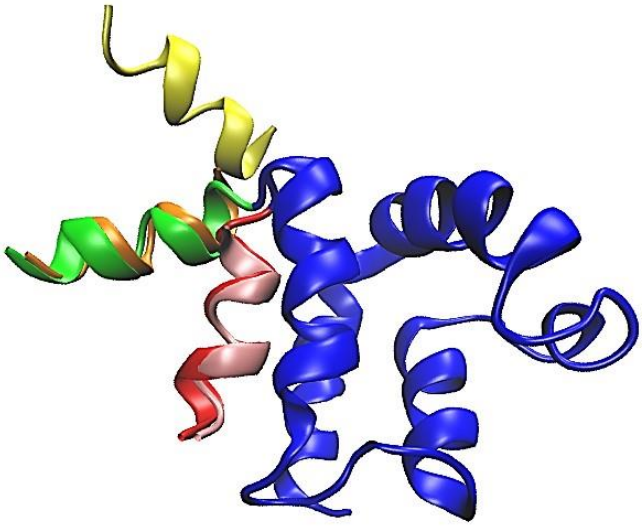

Figure 6.3: The resulting conformations from forced modeling of a 2-state dynamic as a 3-state are shown here. Fragments shown in red and green correspond to the two actual conformational states while yellow depicting the phantom irrelevant conformation with 1% relative occupancy.

Chapter 7

Concurrent Identification and Characterization of Protein Structure and Continuous Internal Dynamics with REDCRAFT[1]

Hanin Omar, Aaron Hein, Casey A. Cole, Homayoun Valafar*

Department of Computer Science and Engineering

University of South Carolina

Columbia, SC 29201

*Corresponding author: homayoun@cse.sc.edu

[1] Concurrent Identification and Characterization of Protein Structure and Continuous Internal Dynamics with REDCRAFT, Hanin Omar, Aaron Hein, Casey A. Cole, and Homayoun Valafar, 2020. Frontiers in Molecular Bioscience. 10.3389/fmolb.2022.806584. Reprinted here with permission of the publisher.

7.1 Abstract

Internal dynamics of proteins can play a critical role in the biological function of some proteins. Several well documented instances have been reported, such as MBP, DHFR, hTS, DGCR8, and NSP1 of the SARS-CoV family of viruses. Despite the importance of internal dynamics of proteins, there currently are very few approaches that allow for meaningful separation of internal dynamics from structural aspects using experimental data. Here we present a computational approach named REDCRAFT that allows for concurrent characterization of protein structure and dynamics. Here, we have subjected DHFR (PDB-ID 1RX2), a 159- residue protein, to a fictitious, mixed mode model of internal dynamics. In this simulation, DHFR was segmented into 7 regions where 4 of the fragments were fixed with respect to each other, two regions underwent rigid-body dynamics, and one region experienced uncorrelated and melting event. The two dynamical and rigid-body segments experienced an average orientational modification of 7° and 12°, respectively. Observable RDC data for backbone C′-N, N-HN, and C′-HN were generated from 102 uniformly sampled frames that described the molecular trajectory. The structure calculation of DHFR with REDCRAFT by using traditional Ramachandran restraint produced a structure with 29 Å of structural difference measured over the backbone atoms (bb-rmsd) over the entire length of the protein and an average bb-rmsd of more than 4.7 Å over each of the dynamical fragments. The same exercise repeated with context-specific dihedral restraints generated by PDBMine produced a structure with bb-rmsd of 21 Å over the entire length of the protein but with bb-rmsd of less than 3 Å over each of the fragments. Finally, utilization of the Dynamic Profile generated by REDCRAFT allowed for the identification of different dynamical regions of the protein and the recovery of individual fragments with bb-rmsd of less than 1 Å.

Following the recovery of the fragments, our assembly procedure of domains (larger segments consisting of multiple fragments with a common dynamical profile) correctly assembled the four fragments that are rigid with respect to each other, categorized the two domains that underwent rigid-body dynamics, and identified one dynamical region for which no conserved structure could be defined. In conclusion, our approach was successful in identifying the dynamical domains, recovery of structure where it is meaningful, and relative assembly of the domains when possible.

Keywords

REDCRAFT, RDC, protein, dynamics, computational, REDCAT, order tensor, PDBMine

7.2 Introduction

Mounting evidence demonstrates the importance of internal dynamics of biomolecules, including proteins, in their enzymatic and biological functions. A number of biologically important proteins have been the subjects of dynamic investigations, confirming the importance of internal dynamics in their function. The breathing motion of myoglobin[153–156] can be cited as a historical instance of this property. Studies of other biologically important proteins such as lipases and hydrolases[157], dihydrofolate reductase (DHFR)[158,159], maltose binding protein (MBP)[160–163], and others[164–167] have revealed the importance of internal dynamics in their function.

Computational approaches such as CHARMM[67,70], AMBER[73,74], GROMACS[72], or NAMD[71] provide simulations of molecular dynamics (MD) from first principles. These platforms incorporate nearly all of the understood biophysical forces at the atomic level, and while the accuracy of the underlying potentials is not perfect, MD methods have the potential to generate reliable models of protein

dynamics if given reasonably accurate starting points. X-ray crystallography is also used to study conformational sampling of some proteins (e.g., DHFR [159], MBP[168,169]). Although studies of dynamics by X-ray crystallography can provide high-resolution descriptions of the multiple conformational states of proteins, these structures and/or their temporal occupancies may be perturbed by the crystal lattice. In fact, it is entirely plausible that functionally unimportant transient states are selected by a crystal lattice. In addition, the timescales of the dynamical events and occupancy of the conformational states are not recoverable by crystallography. Nuclear Magnetic Resonance (NMR) spectroscopy, including measurements of T1 and T2 relaxation rates[57,59,170], and relaxation-dispersion experiments[60], also provide powerful methods for investigating internal dynamics of macromolecules. However, there are few robust NMR studies of the equilibrium distributions of conformations that define the conformational landscape of the "native" protein structure.

Conceptually, from the experimental perspective it is difficult to separate the contribution of structure from dynamics since the two are intimately related. The existing approaches for characterization of protein dynamics from NMR measurements are typically performed in two separate steps—with the protein's structure determined first, followed by an assessment of its motion using the calculated structure. Our recent work[131,171] has demonstrated the potential for obtaining erroneous structures when dynamically-averaged NMR data is best-fit to a single static structure. Subsequent mapping of dynamic information onto such an erroneous structure will likely lead to compromised models of motion. Therefore any attempt in structure elucidation that disregards the dynamics of a protein (or vice versa) can produce erroneous results[172,173]. In this work, we demonstrate a more practical and rigorous approach to characterize a protein's structure and its dynamics

simultaneously through the use of Residual Dipolar Couplings (RDCs)[123,124,131,137,171,174], which are sensitive reporters of both structure and dynamics[63]. The reported results will constitute the first instance of studying structure and dynamics of a protein from RDCs under a continuous and mixed-mode dynamics.

7.3 Theoretical Background

7.3.1 Residual Dipolar Couplings Data

Numerous reviews[101,113,174–178] highlight the utility of RDC data in a broad spectrum of applications to biological macromolecules. RDCs have been used in studies of carbohydrates[143,179–181], nucleic acids[103,125,177,182,183] and proteins[143,156,184–188]. Until recently, the role of RDCs in structure determination has generally been to provide supplemental restraints to a large number of distance-based NOE restraints. Recent developments[123,189–192] have demonstrated the success of structure determination of macromolecules by using primarily or exclusively RDC data. The use of RDCs can lead to a significant reduction in data collection and analysis[131,137,193–195] while providing simultaneous resonance assignment, structure determination, and identification of dynamical regions[126,143,191,192,196,197].

RDCs arise from the interaction of two magnetically active nuclei in the presence of the external magnetic field of an NMR instrument[82,175,198,199]. This interaction is normally reduced to zero, due to the isotropic tumbling of molecules in their aqueous environment. The introduction of partial order to the molecular alignment reintroduces dipolar interactions by minutely limiting isotropic tumbling. This partial order can be introduced in numerous ways[200], including inherent magnetic anisotropy susceptibility of molecules[101], incorporation of artificial tags (such as lanthanides) that exhibit magnetic anisotropy[201], or in a liquid crystal aqueous solution[200]. The RDC interaction phenomenon can be formulated in different

ways[82,90]. In our work we utilize the matrix formulation of this interaction as shown in *Eq. 7.1.* The entity S shown in *Eqs 7.1, 7.2* represents the Saupe order tensor matrix[101,106,142] (the 'order tensor') that can be described as a 3 × 3 symmetric and traceless matrix. D_{max} in *Eq. 7.1* is a nucleus-specific collection of constants, r_{ij} is the separation distance between the two interacting nuclei (in units of Å), and v_{ij} is the corresponding normalized internuclear vector. The order tensor formulation of the RDC interaction provides a convenient mechanism of probing internal dynamics of proteins. Decomposition of the alignment tensor[106,202] can reveal information regarding the level of order[106,198,203] and the preferred direction of alignment[106,203] . A careful comparison of order tensors obtained from different regions of a macromolecule can provide a diagnostic tool in identifying relative orientations between structural elements and/ or the presence of internal dynamics[106,124,203].

$$D_{ij} = \left(\frac{D_{max}}{r_{ij}^3}\right) v_{ij} * S * v_{ij}^T \tag{7.1}$$

$$S = \begin{bmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{xy} & S_{yy} & S_{yz} \\ S_{xz} & S_{yz} & S_{zz} \end{bmatrix}, v_{ij} = \begin{pmatrix} \cos(\theta_x) \\ \cos(\theta_y) \\ \cos(\theta_z) \end{pmatrix} \tag{7.2}$$

The collection of RDC data imposes additional steps in sample preparation and data acquisition when compared to the requisites of the traditional data acquisition by NMR spectroscopy. Despite the additional requirements, the use of RDCs may be justified based on several of their unique features. Our most recent work[204] illustrated the sensitivity of NOEs and RDCs as reporters of protein structures. Based on this work, NOEs tend to lose sensitivity as the search approaches the native structure, while RDCs become more sensitive. Therefore, the addition of RDCs has the potential of improving the structural resolution of proteins studies by NMR spectroscopy. RDCs can also report molecular motions on time-scales ranging

from picoseconds to microseconds[63,204,205], during which many functionally important events occur. Indeed, in the 10 ns–1 s timescale window, RDCs are the most sensitive of NMR parameters[203]. Therefore, in instances of investigating internal dynamics of macromolecules, the use of RDCs can be very beneficial if not necessary. In summary, RDCs have the unique property of simultaneously reporting structural and dynamics information, which has not been fully explored. In this work, we extend our previous work by presenting the first instance of simultaneous characterization of structure and dynamics that include continuous and mixed-mode internal dynamics.

7.3.2 The Effect of Motion on Saupe Order Tensor

Previous works have described the theoretical aspects of the Suape Order Tensors (OTM)[63,196]. Here we provide a more applied summary of this topic as it pertains to this report. Under purely theoretical and hypothetical conditions, a molecule that is absolutely devoid of any motion (internal or external tumbling) will achieve the highest level of order that is represented by the order tensor described in *Eq. 7.3*. Under realistic and unperturbed conditions, the isotropic tumbling of a macromolecule results in an order tensor that has been averaged to zero due to a uniform sampling of all possible molecular orientations. After inducing a tumbling anisotropy, a nonzero order tensor will be reintroduced based on the preferred orientation of the molecular tumbling, which is the origin of observing finite RDC data. In the absence of internal dynamics, the tumbling anisotropy is equally experienced by all portions of the molecule, and therefore OTMs reported by any portion of the molecule are equal to within the experimental error. The presence of internal dynamics will result in an OTM that is different than an OTM obtained from any other portion of the macromolecule. This is due to the fact that OTM from the dynamical region will consist of the effect of anisotropic molecular tumbling

combined with the perturbation of internal dynamics. This is the primary principle that we employ in the development of our analysis. A systematic departure in OTMs reported from different portions of the protein are due to internal dynamics and can be used to identify dynamical regions, internally orchestrated motions, and be used in some instances to reconstruct the trajectory of motion[197].

$$S = \begin{bmatrix} -1/2 & 0 & 0 \\ 0 & -1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{7.3}$$

7.4 Materials and Methods

7.4.1 Target Proteins

In this study we utilized dihydrofolate reductase enzyme (DHFR) that has been selected based on the substantial existing literature in support of major conformational changes when performing their enzymatic function[158,159,168,169].

Dihydrofolate reductase enzyme (DHFR)[136] is a 159-residue long protein that has long been recognized for its central role in regulating tetrahydrofolate level in the cell, which directly aids in the synthesis of nucleic acid precursors. DHFR has been extensively studied and paramount evidence has confirmed its conformational changes as it binds to different intermediates[160,206–208]. DHFR is a single-domain, monomeric molecule; the structure of which is divided into two subdomains: the adenosine binding subdomain and the loop subdomain. The gap separating the two subdomains is occupied by a nicotinamide ring, and the pteridine ring is located in the cleft between helices B and C. Four known states have been identified for this protein: open, closed, and occluded states depending on whether the active site is open, closed, or occluded by the loop. Due to internal dynamic, sometimes it becomes crystallographically unclear or invisible, hence the last state, known as the disordered state[158]. Although there exists ample evidence of the existence of internal dynamics,

little is known regarding the exact nature of the structural rearrangement of this protein.

In this study we use DHFR to test the ability of our approach in concurrent characterization of structure and dynamics of proteins. To that end, we perform a fictitious, mixed-mode molecular dynamics simulation on DHFR (PDB-ID: 1RX2) in order to simulate RDC data and explore the possibility of identifying different dynamical regions of this protein by REDCRAFT, while providing atomic resolution structures for each dynamical domain. It is important to note that the imposed MDS is for illustration purposes only and it servers no useful information in recovering the actual dynamics of this protein in its native form.

7.4.2 Molecular Dynamic Simulation

A fictitious, molecular dynamics simulation was implemented for DHFR based on some of the information available in the literature. More specifically, the structure PDB-ID 1RX2 was fractionated and subjected to various models of internal dynamics to better test our approach. The overall model of dynamics consisted of four fixed regions, two segments that underwent rigid-body dynamics, and one unstructured region. These segments were connected by hinge regions as shown in *Figure 7.1* and *Figure 7.2*. As the first step in our MD simulation, the protein structure was minimized in order to arrive at a more equilibrated state. In the next step, A mixed-mode constrained molecular dynamics simulation was performed in XPLOR-NIH[75,76] (version 3.3) by keeping segments 1 (residue 1–11), 3 (residue42-60), 5 (residue 92–115), and 7 (residue 137–159) fixed in space. Segment 2 (residue 15–28) and segment 4 (residue 64–88) were constrained to experience rigid body dynamics by permitting the hinge regions (regions connecting each segment) to fluctuate freely in space. Segment 6 (residue 116–136) was allowed to freely move in

space without any additional constraints and therefore experienced a melting of that domain. The simulation was conducted for 100,000 steps with step size of 0.0001 psec in a 2,000 K bath temperature. A total of 102 uniformly sampled frames were produced during the course of the molecular trajectory to be used during the calculation of ensemble RDC data.

7.4.3 Calculation of RDC Data

Using the trajectory produced from the MD simulation, 102 frames were generated uniformly to span the entire course of the dynamics. Auxiliary tools were used to separate each of these frames in a PDB format and to generate a corresponding REDCAT file. The software package REDCAT[106] was used to calculate the RDCs values for backbone *C'-N*, *N-HN*, and *C'-HN* for each frame of the trajectory using the order tensors shown in *Table 7.1* in two alignment media. REDCAT's internal utility functions were used to create the observable RDCs by averaging the individual RDCs (for the three vectors) across the entire course of the dynamics (defined by 102 frames). To simulate a more realistic set of data, uniformly distributed noise in the range of ±0.5 Hz was added to all RDC data. These averaged RDCs were used for reconstruction of structure and study of the internal dynamics by REDCRAFT in a procedure highlighted in the following sections. It is important to comment on our choice of RDC data. Although a variety of highly informative RDC data (e.g., *Cα-Hα, Hα-HN*, etc.) can be collected from smaller proteins, we have not used them in our studies since they may not be available in larger systems. To extend the applicable range of NMR spectroscopy to larger proteins, protons are exchanged with deuterons to improve spectral quality. Therefore, in our study, we have confined the use of RDC data to what can be obtained from small or large and perdeuterated proteins. Finally, due to the existence of prolines, in general, the average number of

RDCs is usually less than three per residue since only backbone *N-HN* RDCs can be acquired. In the case of DHFR, the effective and average number of RDCs pre residue was reduced to 2.5 in each alignment medium.

7.4.4 Context Specific Dihedral Constraints with PDBMine

PDBMine[209] is a newly developed tool (https:// ifestos.cse.sc.edu/PDBMine/) that performs an exhaustive search of the dihedral angles for a protein in the Protein Data Bank[21]. As the first step, PDBMine creates a number of subsequences from the primary sequence of the query protein using a rolling window of size W. Therefore, for a protein of size *N* and a rolling window of size *W*, PDBMine creates *N-W+1* subsequences. In the case of DHFR (159 residue protein) and a window size of 7, a total of 153 subsequences (residues 1–7, 2–8, 3–9 ... 153–159) are created. As a second step, PDBMine gathers and aggregates an exhaustive list of all the observed dihedral angles for every residue in every subsequence present in the PDB. During the final step of its analysis, all the returned dihedral angles for all the subsequences are assembled into a final dihedral restraints for each residue of the query protein. In theory, a window size of one will reproduce the known Ramachandran dihedral space. Selection of a larger window size can be viewed as a context-sensitive Ramachandran space. Previous work[210] has illustrated the differences between the dihedral spaces for a proline that precedes a glycine, versus a proline that succeeds a glycine. Therefore, having context specific estimations of dihedrals can be very useful in accelerating the task of structure determination. Another unique feature of PDBMine is its responsiveness; an exhaustive search of the PDB for a 159-residue protein will be completed in less than 10 min.

Under pragmatic conditions, use of the largest window size that produces a set of dihedrals is recommended. However, under testing conditions, it is important to

exercise the necessary precautions to remove biases in the creation of the dihedral restraints. To that end, the primary objective is to avoid creation of the dihedral sets that are heavily populated with instance of 1RX2 or other homologous proteins. Therefore, any process that ensure diverse representation of dihedral angles will test the ability of REDCRAFT in identifying the correct dihedral angles among a large list of decoys. In this exercise, we explored window sizes of 3, 5, 7, and 9 after removing all instances of 1RX2 dihedrals. The window sizes of 3 and 5 produced an intractable number of hits, while the window size of 9 produced results that converged to the dihedrals of 1RX2 for some residues. The window size of 7 produced manageable results with at least 100 dihedrals that were separated from the actual dihedral of 1RX2 by more than 10° (some examples shown in the results section). REDCRAFT incorporates the results of PDBMine to improve its computation time by using the confined dihedral search space of the protein under investigation (in this case 1RX2). It is important to note that REDCRAFT can proceed in successful determination of protein structures in the absence of any dihedral constraints as demonstrated previously[123,211].

7.4.5 Concurrent Study of Structure and Dynamics with REDCRAFT

During the past decade, several approaches and programs for structure determination from RDC data have been described[61,126,142,184,187,188,196,201]. Each of these programs has different advantages and disadvantages. REDCRAFT[61,62,131,191,192], sets itself apart from other existing software packages by deploying a more efficient and effective search mechanism. As a result, REDCRAFT can achieve the same structure determination outcome as other methods with less data[211]. REDCRAFT also allows simultaneous study of structure and dynamics of proteins[123,124,197]. Applications of REDCRAFT in structure calculation have been demonstrated using

aqueous[123,124,212] and membrane[131] proteins with as little as two RDCs per residue[131,213,214] (in two alignment media).

REDCRAFT has introduced a novel approach to structure determination of proteins from RDC data[211]. Aside from an unorthodox search method that is robust and fast[211], REDCRAFT employs an incremental strategy to structure determination in contrast to the all-at-once approach that is adopted by other existing methods. REDCRAFT's incremental structure determination strategy has certain advantages and starts with a search for the optimal torsion angles that join two neighboring peptide planes. This seed dipeptide plane is recursively extended by one residue at a time by exploring a directed and extensive combinatorial search of the dihedral angles that extend the seed structure by one peptide plane (or amino acid) that optimally satisfies the RDC constraints. This process can start from the N-terminus of the protein and continue until the C-terminal end or traverse the structure of the protein in the reverse order (C to N-terminus).

The structural fitness that is produced by REDCRAFT during the course of fragment extension (from dipeptide to the entire protein) is termed the "Dynamic-Profile" (or DP), which plays an instrumental role in a number of analyses including assessing the quality of the final structure or elucidation of internal dynamics. Using the Dynamic-Profile, we have defined a process that allows for simultaneous identification and characterization of structure and internal dynamics. This process consists of three functional steps: standard structure determination, identification of internal dynamics (hinge regions), a grouping of the structural domains (coordinated dynamics), followed by reconstruction of the atomic resolution dynamics when possible. While the last step in the reconstruction of atomic-resolution of dynamics has been discussed in our previous work[124,131,211,212], the former steps have not been

fully described in the literature. In addition, our previous work has been applied to the cases of finite and discrete state dynamics. In this work, we will define and test a more rigorous method of studying continuous and mixed mode dynamics. The four comprehensive steps are as follows:

Standard Structure Determination

Structure calculation of static proteins with REDCRAFT using RDC data has been well described[211]. The DP of a static protein (or a static segment of a protein) generally starts with a low RDC fitness value due to the lack of experimental constraints. The underdetermined system generally produces a RDC fitness value of 0 and gradually increases during the elongation of the dipeptide seed. As the system becomes overdetermined, the RDC fitness reported by DP will increase to approximately the value of experimental error in data acquisition. Structural error defined by the actual deviation of peptide geometries from an ideal geometry (e.g., perfect planarity of the peptide planes, bond lengths, bond angles, etc.) is another source of error. Previous work has empirically determined this error to consist of 20% of the experimental data acquisition error (± 0.2 Hz in this case)[211]. Supplementary *Figure A.1* presents an example of a typical DP for a static protein with the experimental error of ±1.0 Hz.

Identification of Hinge Regions and the Mode of Dynamics

The order tensor obtained from a dynamical portion of a protein will incorporate the effect of overall molecular tumbling and the effect of internal dynamics of that region. Therefore, order tensors reported from two domains of the same protein that undergo different regiments of dynamics will be incongruent. This difference in order tensors will be manifested as a sudden increase in the DP as REDCRAFT will be unable to identify a single order tensor and a static structure that

will satisfy all the RDC constraints. Therefore, a sudden rise in the DP (as illustrated in the Supplementary *Figure A.2*) that clearly exceed the expected error should be interpreted as the hinge region and signifies a transitional region between two distinctly different domains of the same protein. In such instances, the structure of the protein up to the onset of dynamics can be considered as an acceptable structure produced by REDCRAFT.

To investigate the structure of the proceeding portion of the protein, a new structural fragment can be initiated a few residues past the hinge region. In our experiments, we use a skip region of 5 residues and repeat the step 1 above. If the new fragment exhibits a well-behaved DP, then the structure will be accepted as a rigid-body, otherwise, repeat the skip-ahead-region until a rigid-body is discovered. In this process any contiguous region that does not produce a well-behaved DP can be considered undergoing dynamics without any preserved structure, which we term uncorrelated dynamics. Our choice of the term "uncorrelated" is to denote any existing correction between the individual peptide planes of a fragment. Although in practice a gap size of one residue can be used to more accurately establish the hinge regions, a larger gap size is recommended in order to reduce the number of iterations that are needed to pass the hinge region. A more precise exploration of the hinge regions can be conducted at the later stages once the fragments are fully identified. At that point, each fragment can be extended on the *C* and *N* termini to more accurately identify the hinge regions.

Grouping of the Structural Domains

The next step in the process consists of assembling the individual fragments into larger domains based on their orchestrated internal dynamics. This process will allow the integration of fragments that are separated in the primary sequence but

undergo a coordinated motion. The process of identifying the fragments that exhibit no relative internal motion with respect to each other will also complete the proper spatial orientation of the fragments with respect to each other. This process will also identify different regions of the protein that are experiencing different internal dynamics regiments. The assembly of fragments in space is previously described[103] and consists of first expressing all the fragments in a common frame (referred to as the Principal Alignment Frame, PAF) of the first alignment medium. RDC data are insensitive to inversion about each of the PAF and therefore four orientations of fragments with respect to each other are indistinguishable from each other. To eliminate the inversion degeneracy of structure assembly in one alignment medium[103], four alternative orientations of each fragment need to be explored from the perspective of the second alignment medium. The four orientations consist of each fragment as it appears and rotated by the 180° about each of the principal axes of the PAF (*x*, *y*, and *z*) for medium one. These four alternative orientations will be evaluated for fitness to the RDCs in the second alignment medium and the correct structure should exhibit the lowest score. In this exercise we use *Q*-factor[107] as the measure of fitness that normalizes for the strength of alignment. After the completion of this step, all the fragments that belong to the same regiment of internal dynamics will be assembled with a low *Q*-score. The remaining fragments with clearly defined structure can be considered domains that undergo their unique rigid-body dynamics. Finally, any fragment with an incoherent structure is a domain that undergoes uncorrelated dynamics.

Reconstruction of Atomic-Resolution Trajectory of Dynamics

Presence of any form of internal dynamics will perturb the order tensor reported by that region of a molecule. In principle, perturbation of the order tensor can

be used to recover an atomic-resolution trajectory of dynamics in some instances such as the case of discrete state dynamics. Our strategy in reconstruction of atomic resolution trajectory of dynamics has been previously discussed and therefore not presented in this report[197].

7.5 Results and Discussion

7.5.1 Dihedral Constraints for DHFR Using PDBMine

PDBMine was used as the first step to structure determination of DHFR by performing a search with a window size of 7. *Figure 7.3* illustrates the number of hits that were identified by PDBMine with window size of 7 for each residue of DHFR. In average each residue received 5,923 possible dihedral angles with residues 37 and 57 receiving the least and the most (525 and 6,813 respectively) number of dihedral angles.

Figure 7.4 illustrates the aggregated dihedral angles for residues G14 (panel A) and G85 (panel B). In this figure all the dihedral angles reported by PDBMine are illustrated in blue and the corresponding dihedral angles obtained from the PDB (1RX2) is illustrated in red. Several noteworthy observations can be stated. First, the results of PDBMine in principle converge to a Ramachandran space as a reducing window size. However, due to the context-specific nature of the search, a more restricted dihedral space is reported by PDBMine. The second notable observation further expands on the context specific nature of the PDBMine search and is illustrated in *Figure 7.4*. Both of the results correspond to a glycine, but they differ substantially due to the context in which the two glycine's appear in the primary sequence. The third important point is to confirm the proper precautions that we have deployed to remove any unintended biases in our evaluations. It is clear from these

figures that there are significant number of decoy dihedrals among which, REDCRAFT successfully selects the correct dihedral angle.

7.5.2 Summary of MD Simulation

It is important to quantify two aspects of internal dynamics. The first relates to capturing the magnitude of dynamics, and the second relates to the duration of time that was spent in different states. We first report the magnitude of dynamics for the rigid-body domains as an orientational departure from frame0 as the point of reference. *Figure 7.5* illustrates the descriptive statistics regarding the movement of two Rigid-Body domains. Panel (A) of this figure displays the angular departure of each domain (F2 and F4) with respect to the fixed domains (F1, F3, F5, F7) measured between frame i and frame0. Based on this information, Fragment 4 undergoes orientational rearrangement of as high as 32°, while Fragment 2 exhibits a much smaller motion of less than 15°. In addition to the magnitude of motion, it is important to assess the amount of time (or the number of frames) that each fragment spends in each orientational state during its trajectory. The frequency (or likelihood) of existing in a continuum of the orientational repositioning is illustrated in panel (B) of *Figure 7.5*. Based on this information, Fragment 2 spends a very small portion of its trajectory away from frame0, while spending most of the trajectory in the vicinity of the original state (less than 5°). Fragment 4 on the other hand, spends more than 50% of the time in an orientation more than 10° away from the original state. The general summary is that Fragment 2 undergoes small amount of structural rearrangement, while Fragment 4 exhibits a larger motion with respect to the fixed domains of the protein. It is important to state that the MD simulation of DHFR is purely engineered with the primary intention of exploring the sensitivity of our approach in detection of motion.

7.5.3 Structure Determination of DHFR

As the first logical step, the structure of DHFR was determined in its entirety using REDCRAFT using Ramachandran dihedral restraints. As expected, this attempt at structure determination produced unsatisfactory results as indicated by the unacceptable fitness to the RDC data (1.14 Hz), and therefore are succinctly summarized here. The additional details are provided in Supplementary material in Supplementary Table A.1 and Supplementary *Figure A.3*. In summary, the overall structure exhibited 29 Å of bb-rmsd with respect to 1RX2 over the entire length of the protein with a fitness score of 1.14 Hz to the RDC data. The bb-rmsd computed over each of the fragments exhibited an average of 4.8 Å with localized similarities ranging from 0.8 to 9.7 Å.

As a more interesting case, the structure of DHFR was computed by REDCRAFT using the context specific dihedral restraints produced by PDBMine. The examination of the REDCRAFT's DP will be crucial in assessing its success in the structure determination of this protein. The DP generated by REDCRAFT (shown in *Figure 7.6*) exhibits two indicators of the internal dynamics and therefore, a poor structure determination session. First, the final value of the fitness to the RDC data (1.2 Hz) compared to the expected value of 0.6 Hz (corresponding to the simulated error) indicates a failed attempt at structure determination. Second, the existence of sudden and anomalous increases in the DP in various places (e.g., at residues 12–14) is a potential indicator of internal dynamics that requires further examination. It is important to note the close correlation between the sudden increases in the DP and the location of hinge regions of our simulation (denoted by red markers in *Figure 7.6*).

Figure 7.7 illustrates the superimposed structure of DHFR (1RX2 shown in red) and the REDCRAFT recovered structure (shown in blue) by disregarding the existence of internal dynamics.

Table 7.2 highlights the detailed results of comparing the structure of REDCRAFT to 1RX2. As a summary, the two structures exhibit a bb-rmsd of 21 Å and the comparison of fragments exhibit structural similarity in the range of 0.7 to 3 Å. Based on this information, in addition to the divergence in the overall structure, the structural error is also manifested in local fragments. It is important to note that the improved localized structural similarity is due to the effective restraining of the dihedrals accomplished by PDBMine. It is also important to note while the inclusion of PDBMine constraints improved the structural quality of our analysis, there is still substantial room for improvement.

7.5.4 Fragmented Structure Characterization

Fragment 1: Residue 1–11—In consideration of the results shown in the previous section, fragmented study of the protein was conducted. The results of REDCRAFT for the region consisting of residues 1–11 exhibits an acceptable fitness score (around 0.5 Hz) and is devoid of any sudden increase. Therefore, the structure is deemed acceptable as the first fragment of this protein. Implementing steps 1 and 2 listed in the Methods section, the fragmented study continues from residue 16 (after skipping ahead 5 residues).

Fragment 2: Residue 17–38—Structure calculation of DHFR can proceed by investigating a new fragment. The start of the new fragment is based on skipping a fixed number of residues (i.e., 5 residues) from the onset of dynamics to pass the hinge region. The start of a new fragment essentially resets the calculation of an order tensor and therefore removes any inconsistency in the reported order tensors from two

dynamically distinct domains of the protein. Therefore, structure calculation can proceed if a well-behaved DP is exhibited. *Figure 7.8* illustrates the DP of the REDCRAFT for the new fragments starting at residue 17 and as expected, the REDCRAFT score increases at the beginning of the run due to lack of RDC data. Once stabilized, the general pattern is conserved until residue 38, at which point, the DP exhibits a distinct and anomalous increase in the REDCRAFT score. Indeed, residue 39 marks the beginning of the hinge regions and adjoins fragments 2 and 3 of this protein. Hence, we group residues 17–38 as the second Fragment in our investigation.

Fragments 3, 4, 5, 6, and 7— After completion of Fragment 2, a new structure calculation session was started from residue 44. As it can be observed in the DP for this segment (shown in *Figure 7.8*), the same general pattern as the previous two fragments is observed with an anomalous and notable increase in the REDCRAFT score at residue 61. This concluded the analysis of the third fragment that consisted of residues 44–60. The process of fragmented analysis was continued with the corresponding DP illustrated in *Figure 7.8*. The final completion of this process yielded four additional fragments F3 (44–60), F4 (65–88), F5 (97–116), and F7 (138–159). The range of the recovered fragments remarkably agree with the simulated MD. The DP of the only aberrant fragment, Fragment 6, is shown in *Figure 7.8* as multiple attempts in structure recovery. Our first attempt at structure determination of this fragments started from residue 120 after skipping 5 residues from the end of the previous fragment. This attempt at structure determination was unsuccessful since the DP exhibited monotonically increasing score that exceeded the acceptable threshold of 0.6 Hz. The process of skipping forward by 5 residues was repeated with the objective of arriving at a well-behaved region of the protein. Each attempt at structure

determination after skipping 5 residues is shown in *Figure 7.8*. This portion of the protein, unlike all other portions, never resulted in a well behaving DP due to the nature of its internal dynamics. Since the structure of this fragment was consistently modified in each frame, there is no conserved structure to recover, explaining the failure of structure calculation by REDCRAFT. This example also serves as a demonstration of cases where a gap region is larger than 5 residues.

The complete assessment of REDCRAFT's results should consist of two parts. First, to evaluate the success of REDCRAFT in delineating different dynamical regions of the protein as described above. The second portion consist of assessing the structural accuracy of the recovered regions by REDCRAFT.

Table 7.3 shows the results of the fragmented structure determination of DHFR by REDCRAFT while *Figure 7.9* provides an illustration of the fragments (shown in blue) superposed on the corresponding regions of DHFR (shown in green). In *Figure 7.9*, we have omitted the REDCRAFT calculated structure of F6 due to the absence of a meaningful structure to compare. REDCRAFT was able to accurately recover the fragments of DHFR from three RDC data with an accuracy of less than 1 Å. It is important to note that these results are based on unrefined structures in order to expose and exhibit the raw capabilities of REDCRAFT. In practice however, these structure will benefit from refinement in platforms such as Xplor-NIH[21,210], CNS[121], or CYANA[215] to name a few.

Fragment Assembly— Following the structure determination of the individual fragments, the assembly process can proceed based on the procedure described in the Methods section. We start the assembly process by transforming all the fragments into their Principal Alignment Frame (denoted at PAF1) of the first medium and perform an initial investigation of their order tensor (OTM1). The OTM for each fragment in

the second alignment medium is also established using the PAF1 as the common frame of comparison. Once the order tensors from all both alignment media have been canonicalized properly, a simple comparison of the order tensors will be sufficient to establish the relatively large motions between two fragments. In this case, F6 clearly was excluded based on the dissimilarity of its order tensors from the OTMs of any other fragment (due to one order of magnitude difference). However, since F2 and F4 were subjected to relatively small magnitudes of motion, the simple comparison of OTMs was inconclusive. A more sensitive discrimination of internal dynamics can be performed by assembling the fragments after examining all the inversion possibilities of each fragment. *Table 7.4* provides a summary of the progressive fragment assembly using *Q*-Factor as a metric of fitness computed by REDCAT. The first column in this table indicates the progressively growing fragment during the course of the assembly. The nomenclature used in this column consists of the fragment number followed by subscript indicator of the fragment inversion examined in each evaluation. The second column indicates the fitness of the assembly to the combined RDC data in the first alignment medium. The following four columns signify the fitness of the assembly to the combined RDC data from the second alignment medium, after applying the indicated inversion to the last addition to the sequence. In these columns, I, R_x, R_y, and R_z indicate no rotation (Identity or as is), rotation about x, y, and z axes respectively. The fragment assembly starts with the first fragment and as noted in the first row of this table. Note that there is no effect in the rotation of this fragment from the perspective of the second alignment medium. Using the first fragment in its original orientation, fragment 3 has been appended and *Q*-Factors have been computed for all of 4 possible orientations of F3 (not F1). Since the rotation about y yielded an acceptable score, its extension by the fragment 5 will be based on

the *y*-rotated fragment 3. As an empirically accepted practice in the community, *Q*-Factor scores with values less than 0.2 reflect a high-quality structure and are deemed acceptable[107,211]. Using this practice of evaluation, it is clear that fragments 1, 3, 5, and 7 can successfully be assembled as one unit (the fixed core), while fragments 2 and 4 cannot be successfully accepted as part of the fixed domain of the protein.

7.6 Conclusions

Residual Dipolar Coupling are sensitive reporters of structure and dynamics covering a broad range of biologically relevant timescales. However, improper use of RDCs can lead to erroneous results, which may manifest as a faulty structure or an inaccurate model of dynamics. In fact, disregarding dynamics during the course of structure determination can be very detrimental as reported previously[137]. To fully extract the information reported by RDCs, it is imperative to utilize the appropriate analytic approach, in the appropriate manner. Here we have demonstrated that the use of REDCRAFT allows for clear identification of onset of internal dynamics in a protein. In the case of our simulated DHFR, each of the hinge regions was identified very accurately to within one or two residues. Proper isolation of fragments that exhibit a consistent internal dynamics regiment allows for the recovery of structural information after removing the influence of dynamics. In this study we have demonstrated the accurate recovery of structural fragments to within 1 Å of accuracy using only three RDC data acquired in two alignment media.

In addition to accurate structure determination, we demonstrated REDCRAFT's ability to decipher between rigid-body and uncorrelated modes of dynamics as demonstrated with fragments 2, 4, and 6 of DHFR. Although the three domains underwent internal dynamics, REDCRAFT successfully recovered the structure of fragments 2 and 4, where structure was conserved during the course of the

dynamics. On the other hand, the uncorrelated mode of dynamics does not present the conservation of structural coherence throughout the course of dynamics, which renders the exercise of structure determination moot. The nature of internal dynamics of different fragments was established during the course of the fragment assembly. In this step, fragments 1, 3, 5, and 7 were successfully assembled, affirming the fixed relationship between these fragments. The inability to assemble fragments 2 and 4 with the fixed core (fragments 1, 3, 5, and 7) of the protein, when combined with confidently computed structures concludes that the two domains undergo internal dynamics with respect to the core. In regard to the magnitude of dynamics, our previous work[211] related to discrete-state dynamics concluded the inability to identify dynamics with magnitude of less than 15° of movement. This observation was reconfirmed in this study as the distortion of DP in transition from the first fragment to the second was not as notable as the distortion of DP due to the larger dynamics of Fragment 4.

Finally, in our interpretation of DP distortions, we disregarded some anomalous increases in some instances. Except for Fragment 6, all other fragments exhibited such instances with the most notable ones appearing at residue 50 in Fragment 3 or residue 74 in Fragment 4. In such instances we have accepted the results since the net RDC-fitness remained within the experimental error. The origin of these subtle distortions is due to localized departure of peptide geometries from ideal geometries such as non-ideal omega angles, slightly modified bond angles, or bond lengths. These types of structural noise[211] are the basis of expanding the threshold of acceptable RDC-fitness by 20% of the experimental error and are easily rectified during the refinement process when peptide geometries are relaxed and allowed to deviate within an acceptable range[211].

Table 7.1: Order tensors used for RDC simulations.

	S_{xx}	S_{yy}	S_{zz}	α	β	γ
M_1	$3x10^{-4}$	$5x10^{-4}$	$-8x10^{-4}$	0°	0°	0°
M_2	$-4x10^{-4}$	$-6x10^{-4}$	$10x10^{-4}$	40°	50°	-60°

Table 7.2: The BBRMSD of the different fragments generated through the complete run of REDCRAFT from residue 1 until residue159 of DHFR.

Fragment number	Residue Range	BBRMSD with 1RX2
Whole protein	1 - 159	21 Å
Fragment 1	1 - 11	0.7 A
Fragment 2	16 - 38	0.73 A
Fragment 3	44 - 60	0.9 A
Fragment 4	64 - 88	2.2 A
Fragment 5	97 - 115	2.4 A
Fragment 6	116-137	??
Fragment 7	138 - 159	0.7 A

Table 7.3: The BBRMSD of the different fragments generated through the fragmented run of REDCRAFT.

Fragment #	Actual Range	REDCRAFT Range	BBRMSD with 1RX2
Fragment 1	1 - 11	1 - 11	0.5 Å
Fragment 2	15 - 38	16 - 38	0.65 Å
Fragment 3	42 - 60	44 - 60	0.71 Å
Fragment 4	64 - 88	64 - 88	1.2 Å
Fragment 5	92 - 115	97 - 115	.75 Å
Fragment 6	116 - 137	116 - 137	N/A
Fragment 7	138 - 159	138 - 159	0.93 Å

Table 7.4: Results of progressive fragment assembly as investigation all inversion degeneracies. The reported scores are Q-Factors determined by REDCAT.

Fragment #	*M1, I*	*I*	*R_x(180˚)*	*R_y(180˚)*	*R_z(180˚)*
1	0.05	0.05	0.05	0.05	0.05
1_i 3	0.07	0.56	0.62	0.11	0.28
1 3_y 5	0.07	0.71	0.94	0.14	0.71
1 3_y 5_y 7	0.07	0.93	0.72	0.62	0.16
1 3_y 5_y 7_z 2	0.06	0.94	0.88	0.79	0.64
1 3_y 5_y 7_z 4	0.067	0.91	0.77	0.92	0.79

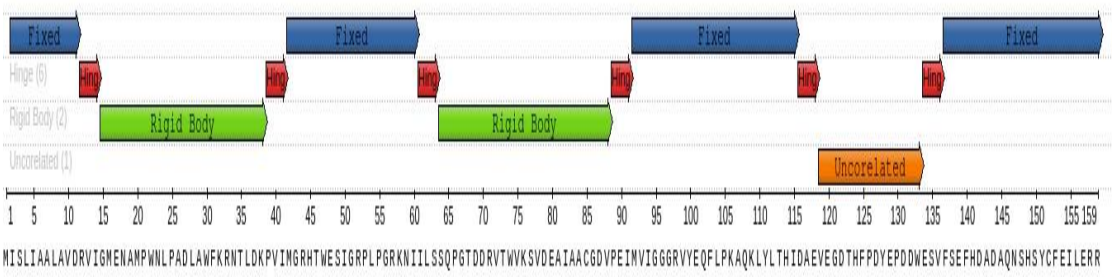

Figure 7.1: The regions of DHFR that were subjected to MD simulation.

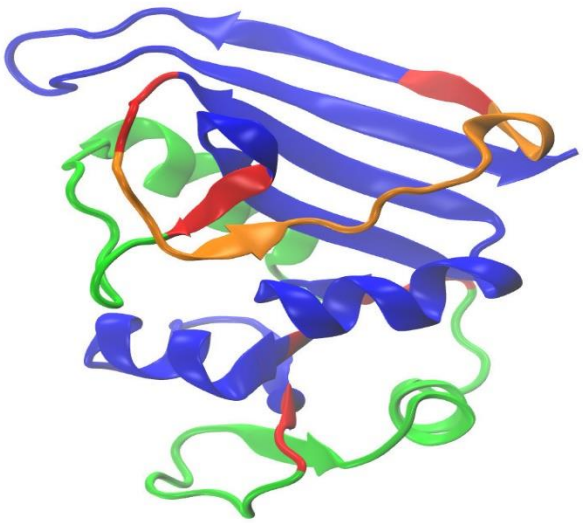

Figure 7.2: Structure of DHFR (PDB-ID 1RX2) that was used in this study with color annotation based on the simulated dynamics. The blue sections correspond to the fixed region while the green sections correspond to the rigid-body dynamics. The section illustrated in red section was subjected to no constraints and was subject to free motion (uncorrelated movement).

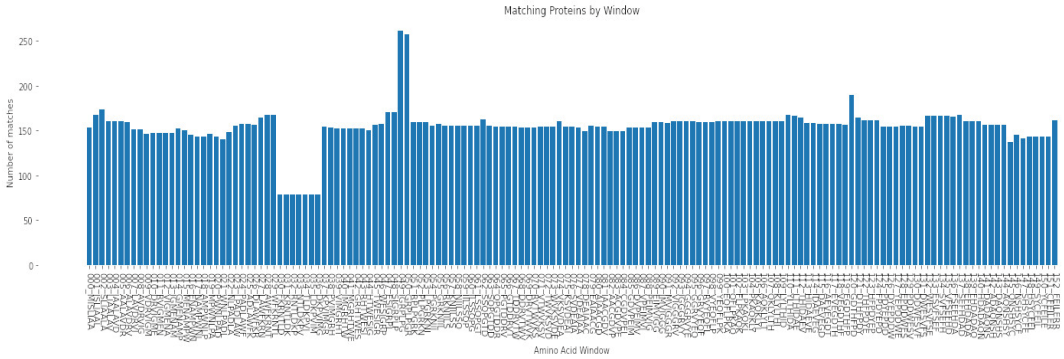

Figure 7.3: The number of dihedral angles returned by PDBMine using a window size of 7 for the DHFR protein (PDB-ID 1RX2).

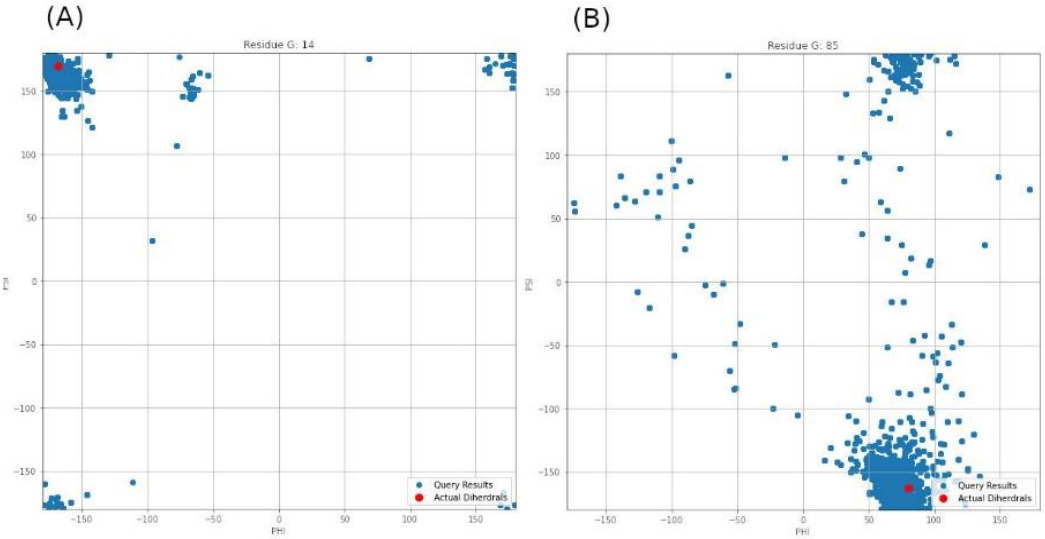

Figure 7.4: Dihedral angles produced by PDBMine using a window size of 7 for residues (A) G14 and (B) G85 of DHFR protein.

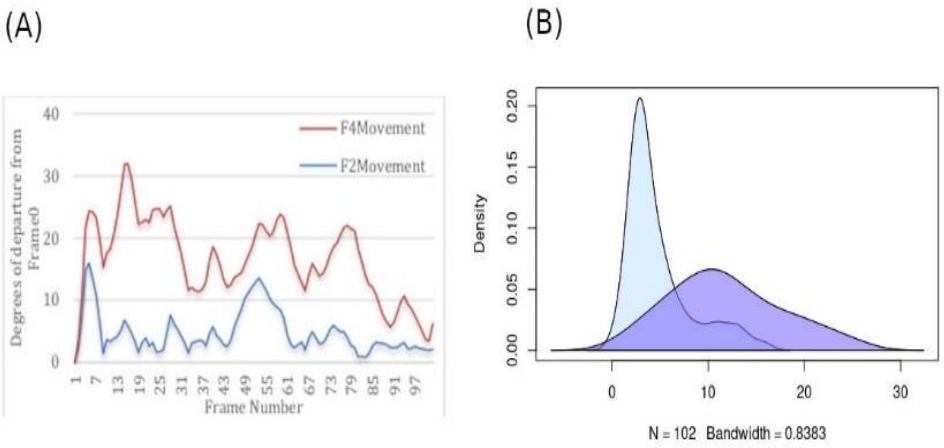

Figure 7.5 Descriptive statistics describing (A) the angular departure from the initial state (Frame0) for both Rigid-Body domains, and (B) the distribution of angular departure to assess the amount of time spent in each state.

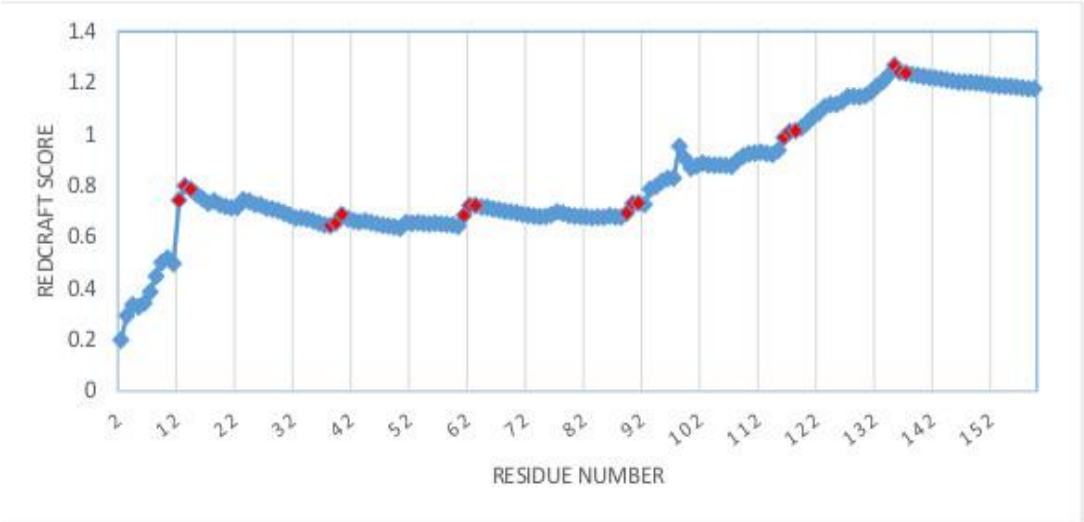

Figure 7.6: Dynamic profile of REDCRAFT for DHFR from residue 1 to 159. Hinge regions from the implemented MD simulation and marked in red to illustrate the correlation between the anomalous increases in DP and the transition between fragments with different internal dynamics.

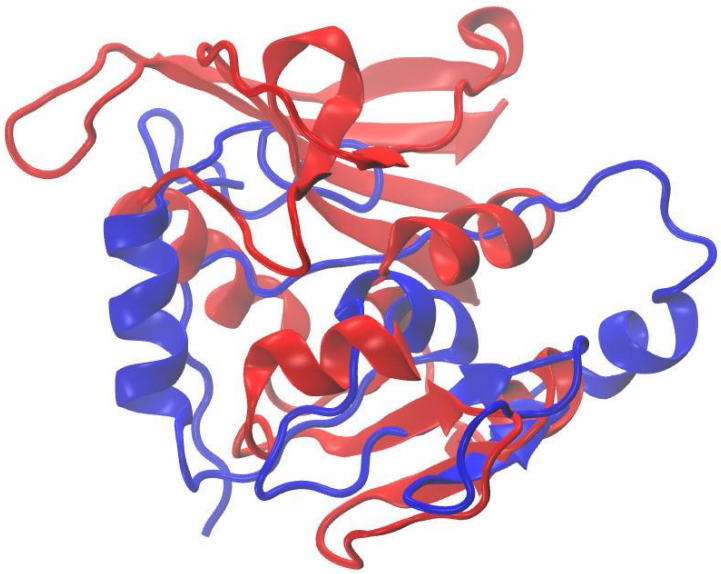

Figure 7.7: Superposition of the structure of 1RX2 (red) over the structure determined by REDCRAFT (blue). The two structures exhibit 21Å of bb-rmsd.

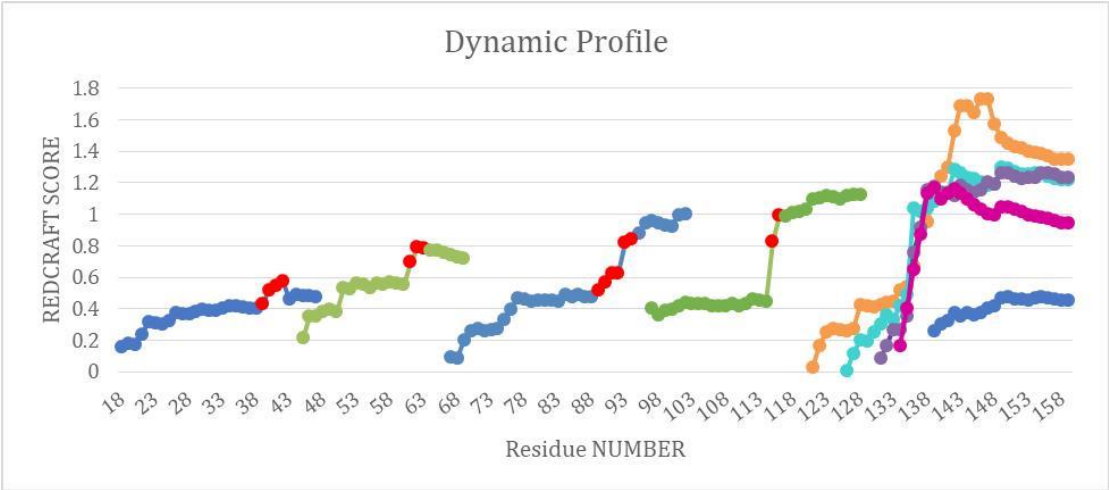

Figure 7.8: The combined dynamic profile for all REDCRAFT runs. The blue segments represent the dynamic profile of the fixed regions in DHFR, the green segments represent the dynamic profile for the rigid body dynamic parts of DHFR, different runs for the uncorrelated dynamics fragment are represented in orange, cyan, purple and pink. Last, the red points indicate the start of increase in scores in the specific dynamic profile for that run.

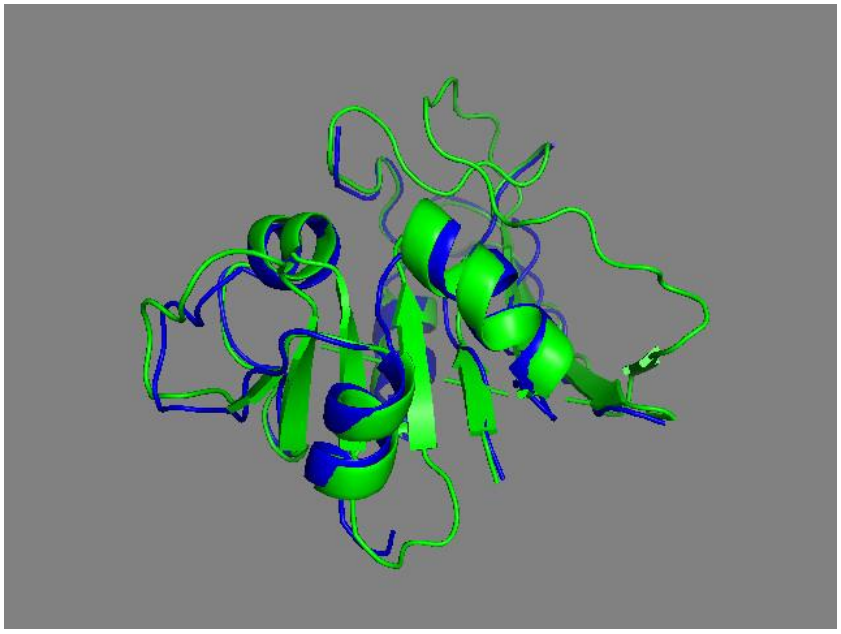

Figure 7.9: Superposition of the calculated fragments by REDCRAFT (blue) and the X-ray structure of DHFR (green).

Future Work

We conclude our research with a discussion of some limitations of the presented work in both Chapter 4 and Chapter 7 and some anomalies that we encountered during the process of developing the presented work. All cases discussed in this chapter required further analysis and research to explain them hence they are recommended for future work.

Anomaly Case Investigation

During the process of generating simulated models of dynamic from the FADD protein (PDB-ID 1A1Z) to validate our approach in *section 4.3.1*, we encountered an anomalous case where our proposed approach for concurrent characterization of structure and dynamics failed at successfully completing the fragmented structure calculation. The anomalous model was a 2-state arc motion with a rotation of 90° on the φ angle of the protein at the 71st residue (denoted by φ_{71}). The model is segmented into two domains: a static domain that consists of residues 1-69 and a dynamic domain that consists of residues 73-83. In this case, Redcraft[124,210] was unable to reconstruct the helical structural elements of the second dynamic domain based on the four RDC vectors from two alignment media in *Table 4.2*.

For future work, an investigation of the source behind such an anomaly is recommended. The foundation of such investigation depends on the assumption that there exists a combination of order tensors, a rotation angle, and chosen occupancies for the different conformational models that render our proposed method for fragment

reconstruction using REDCRAFT useless. We base the hypothesis on two possible scenarios: The first one is based on RDC degeneracies, discussed in detail in *section 3.4*. In this scenario, we assume that the combination of the order tensors, rotation angle, and specific occupancies causes the RDC data from more than one alignment medium to collapse into one, making it impossible to find a single optimal solution or structure that satisfies the RDC data since there is an infinite number of solutions. The second scenario suggests that the combination of the order tensors, rotation angle, and specific occupancies will result in an RDC vectors that are parallel to the S_{ZZ} vector, as a result, regardless what rotation you perform on the S_{yy} *and* S_{xx} vectors , only one solution can be attained.

Limitation of Presented Methodologies

The results detailed in Chapter 6 of the proposed approach for concurrent characterization of structure and dynamics from RDC data explained in Chapter 4 indicate that the approach does not produce the expected results with a small magnitude of dynamics and/or low occupancy rates of less than 20%. The results in *Table 6.3* indicate that at just 15˚ degrees of movement, the approach can reconstruct one of the states (State 1) with reasonable accuracy but fails to reconstruct the second state. However, it can be observed that when the motion is extended to a 60˚ or 30˚ movement (results shown in *Table 6.1* and *Table 6.2* respectively), both states can be reconstructed with reasonable accuracy as long as the relative occupancies exceed 20%. The general explanation for both cases is that the contribution of dynamics is less than the experimental noise, and therefore meaningful calculations are moot.

Furthermore, the computational approach presented in Chapter 7 reconfirmed these results as the distortion of DP in transition from the fragments that undergo a

small amount of structural rearrangement was not as notable as the distortion of DP in the fragment that exhibits a more significant motion with respect to the fixed domain.

References

1. Hartsock, A. & Nelson, W. J. Adherens and tight junctions: Structure, function and connections to the actin cytoskeleton. *Biochimica et Biophysica Acta - Biomembranes* vol. 1778 660–669 (2008).
2. Pittman, R. N. *Regulation of Tissue Oxygenation*. *Colloquium Series on Integrated Systems Physiology: From Molecule to Function* vol. 3 (2011).
3. Cooper, G. M. & Hausman, R. E. *The Cell: A Molecular Approach 2nd Edition*. *Sinauer Associates* (2007).
4. Kristiansen, K. Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: Molecular modeling and mutagenesis approaches to receptor structure and function. *Pharmacology and Therapeutics* vol. 103 21–80 (2004).
5. Alberts, B. *et al. Molecular Biology of the Cell, Fourth Edition*. *Molecular Biology* (2002). doi:citeulike-article-id:691434.
6. Sekhar, A. & Kay, L. E. NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. *Proc. Natl. Acad. Sci.* **110**, 12867 LP – 12874 (2013).
7. Hsia, C. C. W. Respiratory Function of Hemoglobin. *N. Engl. J. Med.* 239–247 (1998) doi:10.1056/NEJM199801223380407.
8. Teif, V. B. & Rippe, K. Statistical-mechanical lattice models for protein-DNA binding in chromatin. *J. Phys. Condens. Matter* **22**, 414105 (2010).
9. Bakhle, Y. S. Structure of COX-1 and COX-2 enzymes and their interaction with inhibitors. in *Drugs of Today* vol. 35 237–250 (1999).

10. Richardson, J. S. The Anatomy and Taxonomy of Protein Structure. in *Advances in Protein Chemistry* vol. 34 167–339 (1981).

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

12. Rose, G. D., Gierasch, L. M. & Smith, J. A. *Turns in peptides and proteins. Advances in protein chemistry* vol. 37 (1985).

13. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99 (1963).

14. Nič, M. IUPAC Compendium of Chemical Terminology 2nd Edition (1997). *IUPAC Compend. Chem. Terminol. Gold B.* **2**, 1997–1997 (1997).

15. Branden, C. & Jhon, T. *Introduction to Protein Structure*. (Garland Publishing).

16. Bork, P. Shuffled domains in extracellular proteins. *FEBS Lett.* **286**, 47–54 (1991).

17. He, H. T. *et al.* Synthesis and chemical stability of a disulfide bond in a model cyclic pentapeptide: Cyclo(1,4)-Cys-Gly-Phe-Cys-Gly-OH. *J. Pharm. Sci.* **95**, 2222–2234 (2006).

18. Stewart, J. *Intermediate Electromagnetic Theory*. (World Scientific Publishing Co., 2001).

19. Robert S., E. *Electromagnetics: History, Theory, and Applications*. (Wiley-IEEE Press, 1999).

20. Van Oss, C. J., Absolom, D. R. & Neumann, A. W. Applications of net repulsive van der Waals forces between different particles, macromolecules, or biological cells in liquids. *Colloids and Surfaces* **1**, 45–56 (1980).

21. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242

(2000).

22. Koradi, R., Billeter, M. & W??thrich, K. MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55 (1996).

23. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

24. Chou, K. C. & Cai, Y. D. Predicting protein quaternary structure by pseudo amino acid composition. *Proteins* **53**, 282–289 (2003).

25. Apweiler, R. The universal protein resource (UniProt) in 2010. *Nucleic Acids Res.* **38**, (2009).

26. Bragg, W. L. The Analysis of Crystals by the X-ray Spectrometer. *Math. Phys. Character* **8924087**, 468–489 (1914).

27. BRAGG, W. The diffraction of short electromagnetic waves by a crystal. *Proc. Camb. Philol. Soc.* **17**, 43–57 (1913).

28. Drenth, J. *Principles of Protein X-ray crystallography*. (Springer science and buisness media,LLC, 2007).

29. Bragg, W. The Specular Reflection of X-rays. *Nature* vol. 90 410–410 (1912).

30. Geerlof, A. *et al.* The impact of protein characterization in structural proteomics. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **62**, 1125–1136 (2006).

31. Rupp, B. & Wang, J. Predictive models for protein crystallization. *Methods* vol. 34 390–407 (2004).

32. Huang, Y.-F. Study of Minning Protein Structural properties and it's application. (National Taiwan University, 2007).

33. Wüthrich, K. The way to NMR structures of proteins. *Nat. Struct. Biol.* **8**, 923–925 (2001).

34. Wüthrich, K. Protein structure determination in solution by NMR spectroscopy.

J. Biol. Chem. **265**, 22059–22062 (1990).

35. Rabi, I., Zacharias, J., Millman, S. & Kusch, P. A New Method of Measuring Nuclear magnetic Moment. *Phys. Rev.* **53**, 318 (1938).

36. Salmon, L. *et al.* Multi-timescale conformational dynamics of the SH3 domain of CD2-associated protein using NMR spectroscopy and accelerated molecular dynamics. *Angew. Chem. Int. Ed. Engl.* **51**, 6103–6 (2012).

37. Dunker, A. K. *et al.* Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59 (2001).

38. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).

39. Taylor, R. E., Chen, Y., Galvin, G. M. & Pabba, P. K. Conformation-activity relationships in polyketide natural products. Towards the biologically active conformation of epothilone. *Org. Biomol. Chem.* **2**, 127–32 (2004).

40. Overhauser, A. W. Polarization of nuclei in metals. *Phys. Rev.* **92**, 411–415 (1953).

41. Kaiser, R. Use of the Nuclear Overhauser Effect in the Analysis of High-Resolution Nuclear Magnetic Resonance Spectra. *J. Chem. Phys.* **39**, 2435 (1963).

42. Silverstein, R. M., Bassler, G. C. & Morrill, T. C. *Spectrometric Identification of organic compounds*. (1981).

43. Hahn, E. L. & Maxwell, D. E. Spin echo measurements of nuclear spin coupling in molecules. *Phys. Rev.* **88**, 1070–1084 (1952).

44. Pettersson, E., Lundeberg, J. & Ahmadian, A. Generations of sequencing technologies. *Genomics* vol. 93 105–111 (2009).

45. Nelson, D. L. & Cox, M. *Lehninger Principles of Biochemistry*.

(M.W.H.Freeman, 2008).

46. Martí-Renom, M. a *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).

47. Ginalski, K. Comparative modeling for protein structure prediction. *Current Opinion in Structural Biology* vol. 16 172–177 (2006).

48. Bowie, J. U., Lüthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–170 (1991).

49. Jones, D. T., Taylor, W. R. & Thornton, J. M. A new approach to protein fold recognition. *Nature* vol. 358 86–89 (1992).

50. Samudrala, R., Xia, Y., Huang, E. & Levitt, M. Ab initio protein structure prediction using a combined hierarchical approach. *Proteins* **Suppl 3**, 194–8 (1999).

51. Rawlings, N. D. & Barrett, A. J. MEROPS: The peptidase database. *Nucleic Acids Research* vol. 27 325–331 (1999).

52. Sillitoe, I. *et al.* CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.* **43**, D376–D381 (2015).

53. Hubbard, T. J. P., Ailey, B., Brenner, S. E., Murzin, A. G. & Chothia, C. SCOP: A structural classification of proteins database. *Nucleic Acids Research* vol. 27 254–256 (1999).

54. Wierenga, R. K. The TIM-barrel fold: A versatile framework for efficient enzymes. *FEBS Letters* vol. 492 193–198 (2001).

55. Namanja, A. T. *et al.* Toward flexibility-activity relationships by NMR spectroscopy: Dynamics of Pin1 ligands. *J. Am. Chem. Soc.* **132**, 5607–5609 (2010).

56. Boehr, D. D., Dyson, H. J. & Wright, P. E. An NMR perspective on enzyme

dynamics. *Chem. Rev.* **106**, 3055–79 (2006).

57. Cavanagh, J., Fairbrother, W. Jr., Palmer III, A. J., Skelton, N. J. & Rance, M. *Protein NMR Spectroscopy, Principles and Practice*. (Academic Press, 2006).

58. Barbato, G., Ikura, M., Kay, L. E., Pastor, R. W. & Bax, A. Backbone Dynamics of Calmodulin Studied by 1 5 N Relaxation Using Inverse Detected Two-Dimensional NMR Spectroscopy: The Central Helix Is Flexible? *Biochemistry* **31**, 5269–5278 (1992).

59. Lorieau, J. L., Louis, J. M. & Bax, A. Whole-body rocking motion of a fusion peptide in lipid bilayers from size-dispersed 15N NMR relaxation. *J. Am. Chem. Soc.* **133**, 14184–14187 (2011).

60. Lipari, G. & Szabo, A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.* **104**, 4546–4559 (1982).

61. Bouvignies, G. *et al.* Identification of slow correlated motions in proteins using residual dipolar and hydrogen-bond scalar couplings. *Proc. Natl. Acad. Sci.* **102**, 13885–13890 (2005).

62. Clore, G. M. & Schwieters, C. D. Amplitudes of protein backbone dynamics and correlated motions in a small α/β protein: Correspondence of dipolar coupling and heteronuclear relaxation measurements. *Biochemistry* **43**, 10678–10691 (2004).

63. Tolman, J. R., Flanagan, J. M., Kennedy, M. A. & Prestegard, J. H. NMR evidence for slow collective motions in cyanometmyoglobin. *Nat. Struct. Biol.* **4**, 292–297 (1997).

64. Persson, F. & Halle, B. Transient access to the protein interior: Simulation versus NMR. *J. Am. Chem. Soc.* **135**, 8735–8748 (2013).

65. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).

66. Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).

67. Brooks, B. R. *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217 (1983).

68. MacKerel Jr., A. D. *et al.* CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. in *The Encyclopedia of Computational Chemistry* vol. 1 271–277 (1998).

69. Scott, W. R. P. *et al.* The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A* **103**, 3596–3607 (1999).

70. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614 (2009).

71. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).

72. Hess, B., Kutzner, C., Van Der Spoel, D. & Lindahl, E. GRGMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **4**, 435–447 (2008).

73. Case, D. A. *et al.* The Amber biomolecular simulation programs. *Journal of Computational Chemistry* vol. 26 1668–1688 (2005).

74. Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the Amber biomolecular simulation package. *WIREs Comput. Mol. Sci.* **3**, 198–210 (2013).

75. Schwieters, C. D., Kuszewski, J. J. & Marius Clore, G. Using Xplor-NIH for NMR molecular structure determination. *Progress in Nuclear Magnetic*

Resonance Spectroscopy vol. 48 47–62 (2006).

76. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160**, 65–73 (2003).

77. Saupe, A. Kernresonanzen in kristallinen fl??ssigkeiten und in kristallinfl??ssigen l??sungen. teil I. *Zeitschrift fur Naturforsch. - Sect. A J. Phys. Sci.* **19**, 161–171 (1964).

78. Tolman, J. R. & Ruan, K. NMR Residual Dipolar Couplings as Probes of Biomolecular Dynamics. *Chem. Rev.* **106**, 1720–1736 (2006).

79. Lakomek, N. A., Carlomagno, T., Becker, S., Griesinger, C. & Meiler, J. A thorough dynamic interpretation of residual dipolar couplings in ubiquitin. *J. Biomol. NMR* **34**, 101–115 (2006).

80. Lakomek, N. A. *et al.* Self-consistent residual dipolar coupling based model-free analysis for the robust determination of nanosecond to microsecond protein dynamics. *J. Biomol. NMR* **41**, 139–155 (2008).

81. Griffiths, D. *Introduction to Quantum Mechanics*. (2005).

82. Tolman, J. R., Flanagan, J. M., Kennedy, M. A. & Prestegard, J. H. Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 9279–9283 (1995).

83. Saupe, A. & Englert, G. High resolution nuclear magnetic resonance spectra of various polyisoprenes. *Phys. Rev. Lett.* **11**, (1962).

84. Strang, G. *Introduction to Linear Algebra*. (Pearson, 2009).

85. Mueller, G. A. *et al.* Global folds of proteins with low densities of NOEs using residual dipolar couplings: application to the 370-residue maltodextrin-binding protein. *J. Mol. Biol.* **300**, 197–212 (2000).

86. Hus, J.-C. *et al.* 16-fold degeneracy of peptide plane orientations from residual dipolar couplings: analytical treatment and implications for protein structure determination. *J. Am. Chem. Soc.* **130**, 15927–37 (2008).

87. Mesleh, M. F., Veglia, G., DeSilva, T. M., Marassi, F. M. & Opella, S. J. Dipolar waves as NMR maps of protein structure. *J. Am. Chem. Soc.* **124**, 4206–4207 (2002).

88. Mascioni, A. & Veglia, G. Theoretical analysis of residual dipolar coupling patterns in regular secondary structures of proteins. *J. Am. Chem. Soc.* **125**, 12520–12526 (2003).

89. Al-Hashimi, H. M. *et al.* Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. *J. Magn. Reson.* **143**, 402–6 (2000).

90. Bax, A. & Tjandra, N. High-resolution heteronuclear NMR of human ubiquitin in an aqueous liquid crystalline medium. *J. Biomol. NMR* **10**, 289–292 (1997).

91. Tjandra, N. & Bax, A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* **278**, 1111–4 (1997).

92. Ottiger, M. & Bax, A. Characterization of magnetically oriented phospholipid micelles for measurement of dipolar couplings in macromolecules. *J. Biomol. NMR* **12**, 361–372 (1998).

93. Hansen, M. R., Mueller, L. & Pardi, A. Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nat. Struct. Biol.* **5**, 1065–1074 (1998).

94. Clore, G. M., Starich, M. R. & Gronenborn, A. M. Measurement of Residual Dipolar Couplings of Macromolecules Aligned in the Nematic Phase of a

Colloidal Suspension of Rod-Shaped Viruses. *J. Am. Chem. Soc.* **120**, 10571–10572 (1998).

95. Sass, H. J., Musco, G., Stahl, S. J., Wingfield, P. T. & Grzesiek, S. Solution NMR of proteins within polyacrylamide gels: diffusional properties and residual alignment by mechanical stress or embedding of oriented purple membranes. *J. Biomol. NMR* **18**, 303–9 (2000).

96. Meiler, J. & Baker, D. The fumarate sensor DcuS: progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy. *J. Magn. Reson.* **173**, 310–6 (2005).

97. Clore, G. M., Gronenborn, a M. & Bax, a. A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information. *J. Magn. Reson.* **133**, 216–221 (1998).

98. Evenäs, J., Mittermaier, A., Yang, D. & Kay, L. E. Measurement of (13)C(alpha)-(13)C(beta) dipolar couplings in (15)N,(13)C,(2)H-labeled proteins: application to domain orientation in maltose binding protein. *J. Am. Chem. Soc.* **123**, 2858–2864 (2001).

99. Mittermaier, A. & Kay, L. E. χ1 torsion angle dynamics in proteins, from dipolar couplings. *J. Am. Chem. Soc.* **123**, 6892–6903 (2001).

100. Prestegard, J. H. New techniques in structural NMR--anisotropic interactions. *Nat. Struct. Biol.* **5 Suppl**, 517–522 (1998).

101. Prestegard, J. H., Al-Hashimi, H. M. & Tolman, J. R. NMR structures of biomolecules using field oriented media and residual dipolar couplings. *Q. Rev. Biophys.* **33**, S0033583500003656 (2000).

102. Fischer, M. W. F., Losonczi, J. A., Weaver, J. L. & Prestegard, J. H. Domain orientation and dynamics in multidomain proteins from residual dipolar

couplings. *Biochemistry* **38**, 9013–9022 (1999).

103. Al-Hashimi, H. ., Bolon, P. . & Prestegard, J. . Molecular Symmetry as an Aid to Geometry Determination in Ligand Protein Complexes. *J. Magn. Reson.* **142**, 153–158 (2000).

104. Meiler, J. & Baker, D. Rapid protein fold determination using unassigned NMR data. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15404–9 (2003).

105. Zweckstetter, M. NMR: prediction of molecular alignment from structure using the PALES software. *Nat. Protoc.* **3**, 679–690 (2008).

106. Valafar, H. & Prestegard, J. H. REDCAT: a residual dipolar coupling analysis tool. *J. Magn. Reson.* **167**, 228–241 (2004).

107. Cornilescu, G., Marquardt, J. L., Ottiger, M. & Bax, A. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *J. Am. Chem. Soc.* **120**, 6836–6837 (1998).

108. Anet, F. A. L. & Bourn, A. J. R. Nuclear Magnetic Resonance Spectral Assignments from Nuclear Overhauser Effects 1. *J. Am. Chem. Soc.* **87**, 5250–5251 (1965).

109. Tao, Y., Rao, Z.-H. & Liu, S.-Q. Insight derived from molecular dynamics simulation into substrate-induced changes in protein motions of proteinase K. *J. Biomol. Struct. Dyn.* **28**, 143–158 (2010).

110. Clore, G. M. & Iwahara, J. Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. *Chem. Rev.* **109**, 4108–39 (2009).

111. Thiele, C. M. Residual dipolar couplings (RDCs) in organic structure determination. *European Journal of Organic Chemistry* 5673–5685 (2008)

doi:10.1002/ejoc.200800686.

112. Boehr, D. D., McElheny, D., Dyson, H. J. & Wright, P. E. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* **313**, 1638–42 (2006).

113. Blackledge, M. Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. *Prog. Nucl. Magn. Reson. Spectrosc.* **46**, 23–61 (2005).

114. Bouvignies, G., Markwick, P. R. L. & Blackledge, M. Simultaneous definition of high resolution protein structure and backbone conformational dynamics using NMR residual dipolar couplings. *ChemPhysChem* vol. 8 1901–1909 (2007).

115. Salmon, L. *et al.* Protein conformational flexibility from structure-free analysis of NMR dipolar couplings: Quantitative and absolute determination of backbone motion in ubiquitin. *Angew. Chemie - Int. Ed.* **48**, 4154–4157 (2009).

116. Hess, B. & Scheek, R. M. Orientation restraints in molecular dynamics simulations using time and ensemble averaging. *J. Magn. Reson.* **164**, 19–27 (2003).

117. De Simone, A., Richter, B., Salvatella, X. & Vendruscolo, M. Toward an Accurate Determination of Free Energy Landscapes in Solution States of Proteins. *J. Am. Chem. Soc.* **131**, 3810–3811 (2009).

118. Huang, J. R. & Grzesiek, S. Ensemble calculations of unstructured proteins constrained by RDC and PRE data: A case study of urea-denatured ubiquitin. *J. Am. Chem. Soc.* **132**, 694–705 (2010).

119. Fenwick, R. B., Esteban-Martín, S. & Salvatella, X. Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *Eur. Biophys. J.* **40**, 1339–1355 (2011).

120. Stelzer, A. C., Frank, A. T., Bailor, M. H., Andricioaei, I. & Al-Hashimi, H. M. Constructing atomic-resolution RNA structural ensembles using MD and motionally decoupled NMR RDCs. *Methods* **49**, 167–73 (2009).

121. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D. Biol. Crystallogr.* **54**, 905–921 (1998).

122. Herrmann, T., Guntert, P. & Wuthrich, K. Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J Biomol NMR* **24**, 171–189 (2002).

123. Simin, M., Irausquin, S., Cole, C. A. & Valafar, H. Improvements to REDCRAFT: A software tool for simultaneous characterization of protein backbone structure and dynamics from residual dipolar couplings. *J. Biomol. NMR* **60**, (2014).

124. Bryson, M., Tian, F., Prestegard, J. H. & Valafar, H. REDCRAFT: a tool for simultaneous characterization of protein backbone structure and motion from RDC data. *J. Magn. Reson.* **191**, 322–334 (2008).

125. Al-Hashimi, H. M. *et al.* Concerted motions in HIV-1 TAR RNA may allow access to bound state conformations: RNA dynamics from NMR residual dipolar couplings. *J. Mol. Biol.* **315**, 95–102 (2002).

126. Bernadó, P. & Blackledge, M. Local dynamic amplitudes on the protein backbone from dipolar couplings: toward the elucidation of slower motions in biomolecules. *J. Am. Chem. Soc.* **126**, 7760–1 (2004).

127. Yang, S. & Al-Hashimi, H. M. Unveiling Inherent Degeneracies in Determining Population-Weighted Ensembles of Interdomain Orientational Distributions Using NMR Residual Dipolar Couplings: Application to RNA Helix Junction

Helix Motifs. *J. Phys. Chem. B* **119**, 9614–26 (2015).

128. Simone, A. De, Montalvao, R. W. & Vendruscolo, M. Determination of Conformational Equilibria in Proteins Using Residual Dipolar Couplings. 4189–4195 (2011).

129. Ulmer, T. S., Ramirez, B. E., Delaglio, F. & Bax, A. Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. *J. Am. Chem. Soc.* **125**, 9179–9191 (2003).

130. Sang, H. P., Woo, S. S., Mukhopadhyay, R., Valafar, H. & Opella, S. J. Phage-induced alignment of membrane proteins enables the measurement and structural analysis of residual dipolar couplings with dipolar waves and ??-maps. *J. Am. Chem. Soc.* **131**, 14140–14141 (2009).

131. Shealy, P., Simin, M., Park, S. H., Opella, S. J. & Valafar, H. Simultaneous structure and dynamics of a membrane protein using REDCRAFT: membrane-bound form of Pf1 coat protein. *J. Magn. Reson.* **207**, 8–16 (2010).

132. Yao, L., Vögeli, B., Torchia, D. a & Bax, A. Simultaneous NMR study of protein structure and dynamics using conservative mutagenesis. *J. Phys. Chem. B* **112**, 6045–6056 (2008).

133. Loria, J. P., Berlow, R. B. & Watt, E. D. Characterization of enzyme motions by solution NMR relaxation dispersion. *Acc. Chem. Res.* **41**, 214–221 (2008).

134. Gibson, L. M., Lovelace, L. L. & Lebioda, L. The R163K mutant of human thymidylate synthase is stabilized in an active conformation: structural asymmetry and reactivity of cysteine 195. *Biochemistry* **47**, 4636–43 (2008).

135. Schnell, J. R., Dyson, H. J. & Wright, P. E. Structure, dynamics, and catalytic function of dihydrofolate reductase. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 119–140 (2004).

136. Sawaya, M. R. & Kraut, J. Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: Crystallographic evidence. *Biochemistry* **36**, 586–603 (1997).

137. Valafar, H., Simin, M. & Irausquin, S. A Review of REDCRAFT. Simultaneous Investigation of Structure and Dynamics of Proteins from RDC Restraints. *Annu. Reports NMR Spectrosc.* **76**, 23–66 (2012).

138. Schmidt, C., Irausquin, S. J. & Valafar, H. Advances in the REDCAT software package. *BMC Bioinformatics* **14**, 302 (2013).

139. Tolman, J. R., Al-Hashimi, H. M., Kay, L. E. & Prestegard, J. H. Structural and dynamic analysis of residual dipolar coupling data for proteins. *J. Am. Chem. Soc.* **123**, 1416–1424 (2001).

140. Greshenfeld, N. . *The Nature of Mathmatical Modeling*. (Cambridge University Press, 1998).

141. Levenberg, K. *A METHOD FOR THE SOLUTION OF CERTAIN NON-LINEAR PROBLEMS IN LEAST SQUARES*. (Brown University, 1944).

142. Saupe, A. & Englert, G. High-Resolution Nuclear Magnetic Resonance Spectra of Orientated Molecules. *Phys. Rev. Lett.* **11**, 462–464 (1963).

143. Tian, F., Al-Hashimi, H. M., Craighead, J. L. & Prestegard, J. H. Conformational analysis of a flexible oligosaccharide using residual dipolar couplings. *J. Am. Chem. Soc.* **123**, 485–492 (2001).

144. De Simone, A., Richter, B., Salvatella, X. & Vendruscolo, M. Toward an accurate determination of free energy landscapes in solution states of proteins. *J. Am. Chem. Soc.* **131**, 3810–3811 (2009).

145. De Simone, A., Montalvao, R. W. & Vendruscolo, M. Determination of conformational equilibria in proteins using residual dipolar couplings. *J. Chem.*

Theory Comput. **7**, 4189–4195 (2011).

146. De Simone, A. *et al.* Structures of the excited states of phospholamban and shifts in their populations upon phosphorylation. *Biochemistry* **52**, 6684–6694 (2013).

147. De Simone, A., Montalvao, R. W., Dobson, C. M. & Vendruscolo, M. Characterization of the interdomain motions in hen lysozyme using residual dipolar couplings as replica-averaged structural restraints in molecular dynamics simulations. *Biochemistry* **52**, 6480–6486 (2013).

148. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–56 (1995).

149. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**–**2**, 19–25 (2015).

150. Pronk, S. *et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).

151. Abraham, M. & Gready, J. Ensuring Mixing Efficiency of Replica-Exchange Molecular Dynamics Simulations. *J. Chem. Theory Comput. - J CHEM THEORY Comput* **4**, (2008).

152. Bernhardt, N. A., Xi, W., Wang, W. & Hansmann, U. H. E. Simulating Protein Fold Switching by Replica Exchange with Tunneling. *J. Chem. Theory Comput.* **12**, 5656–5666 (2016).

153. Shimada, H. & Caughey, W. Dynamic protein structures. Effects of pH on conformer stabilities at the ligand-binding site of bovine heart myoglobin carbonyl. *J. Biol. Chem.* **257**, 11893–11900 (1982).

154. Cupane, A., Leone, M., Vitrano, E. & Cordone, L. Structural and Dynamic

Properties of the Heme Pocket in Myoglobin Probed by Optical Spectroscopy. *Biopolymers* **27**, 1977–1998 (1988).

155. Emerson, S. D., Lecomte, J. T. J. & La Mar, G. N. Proton NMR resonance assignment and dynamic analysis of phenylalanine CD1 in a low-spin ferric complex of sperm whale myoglobin. *J. Am. Chem. Soc.* **110**, 4176–4182 (1988).
156. Bertini I, Luchinat C, Turano P, Battaini G & Casella L. The magnetic properties of myoglobin as studied by NMR spectroscopy. Chemistry. *PMID 12772306* (2003) doi:10.1002/chem.200204562.
157. Yu, X.-W., Xu, Y. & Xiao, R. Lipases from the genus Rhizopus: Characteristics, expression, protein engineering and application. *Prog. Lipid Res.* **64**, 57–68 (2016).
158. Bystroff, C. & Kraut, J. Crystal structure of unliganded Escherichia coli dihydrofolate reductase. Ligand-induced conformational changes and cooperativity in binding. *Biochemistry* **30**, 2227–2239 (1991).
159. Osborne, M. J., Schnell, J., Benkovic, S. J., Dyson, H. J. & Wright, P. E. Backbone Dynamics in Dihydrofolate Reductase Complexes: Role of Loop Flexibility in the Catalytic Mechanism. *Biochemistry* **40**, 9846–9859 (2001).
160. Evenäs, J. *et al.* Ligand-induced structural changes to maltodextrin-binding protein as studied by solution NMR spectroscopy11Edited by P. E. Wright. *J. Mol. Biol.* **309**, 961–974 (2001).
161. Hwang, P. M., Skrynnikov, N. R. & Kay, L. E. Domain orientation in β-cyclodextrin-loaded maltose binding protein: Diffusion anisotropy measurements confirm the results of a dipolar coupling study. *J. Biomol. NMR* **20**, 83–88 (2001).
162. Tang, C., Schwieters, C. D. & Clore, G. M. Open-to-closed transition in apo

maltose-binding protein observed by paramagnetic NMR. *Nature* **449**, 1078–1082 (2007).

163. Millet, O., Hudson, R. P. & Kay, L. E. The energetic cost of domain reorientation in maltose-binding protein as studied by NMR and fluorescence spectroscopy. *Proc. Natl. Acad. Sci.* **100**, 12700 (2003).

164. Aramini, J. M. *et al.* The RAS-Binding Domain of Human BRAF Protein Serine/Threonine Kinase Exhibits Allosteric Conformational Changes upon Binding HRAS. *Structure* **23**, 1382–1393 (2015).

165. Kerns, S. J. *et al.* The energy landscape of adenylate kinase during catalysis. *Nat. Struct. Mol. Biol.* **22**, 124–131 (2015).

166. Palmer, A. G. Enzyme Dynamics from NMR Spectroscopy. *Acc. Chem. Res.* **48**, 457–465 (2015).

167. Wilson, C. *et al.* Kinase dynamics. Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science* **347**, 882–886 (2015).

168. Diez, J. *et al.* The crystal structure of a liganded trehalose/maltose-binding protein from the hyperthermophilic Archaeon Thermococcus litoralis at 1.85 Å. *J. Mol. Biol.* **305**, 905–915 (2001).

169. Duan, X., Hall, J. A., Nikaido, H. & Quiocho, F. A. Crystal structures of the maltodextrin/maltose-binding protein complexed with reduced oligosaccharides: flexibility of tertiary structure and ligand binding1 1Edited by I. A. Wilson. *J. Mol. Biol.* **306**, 1115–1126 (2001).

170. Barbato, G., Ikura, M., Kay, L. E., Pastor, R. W. & Bax, A. Backbone dynamics of calmodulin studied by nitrogen-15 relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible. *Biochemistry* **31**, 5269–5278 (1992).

171. Park, S. H., Son, W. S., Mukhopadhyay, R., Valafar, H. & Opella, S. J. Phage-induced alignment of membrane proteins enables the measurement and structural analysis of residual dipolar couplings with dipolar waves and lambda-maps. *J. Am. Chem. Soc.* **131**, 14140–14141 (2009).

172. Tejero, R., Bassolino-Klimas, D., Bruccoleri, R. E. & Montelione, G. T. Simulated annealing with restrained molecular dynamics using CONGEN: Energy refinement of the NMR solution structures of epidermal and type-α transforming growth factors. *Protein Sci.* **5**, 578–592 (1996).

173. Montelione, G. T. *et al.* Recommendations of the wwPDB NMR Validation Task Force. *Structure* **21**, 1563–1570 (2013).

174. Tolman, J. R. Dipolar couplings as a probe of molecular dynamics and structure in solution. *Current Opinion in Structural Biology* vol. 11 (2001).

175. Clore, G. M., Gronenborn, a M. & Tjandra, N. Direct structure refinement against residual dipolar couplings in the presence of rhombicity of unknown magnitude. *J. Magn. Reson.* **131**, 159–162 (1998).

176. Zhou, H., Vermeulen, A., Jucker, F. M. & Pardi, A. Incorporating residual dipolar couplings into the NMR solution structure determination of nucleic acids. *Biopolymers* **52**, (1999).

177. Al-Hashimi, H. M., Gorin, A., Majumdar, A., Gosser, Y. & Patel, D. J. Towards structural genomics of RNA: Rapid NMR resonance assignment and simultaneous RNA tertiary structure determination using residual dipolar couplings. *J. Mol. Biol.* **318**, (2002).

178. De Alba, E. & Tjandra, N. NMR dipolar couplings for the structure determination of biopolymers in solution. *Progress in Nuclear Magnetic Resonance Spectroscopy* vol. 40 (2002).

179. Azurmendi, H. F., Martin-Pastor, M. & Bush, C. A. Conformational studies of Lewis X and Lewis A trisaccharides using NMR residual dipolar couplings. *Biopolymers* **63**, (2002).

180. Azurmendi, H. F. & Bush, C. A. Conformational studies of blood group A and blood group B oligosaccharides using NMR residual dipolar couplings. *Carbohydr. Res.* **337**, (2002).

181. Adeyeye, J. *et al.* Conformation of the hexasaccharide repeating subunit from the Vibrio cholerae O139 capsular polysaccharide. *Biochemistry* **42**, (2003).

182. Tjandra, N., Tate, S. I., Ono, A., Kainosho, M. & Bax, A. The NMR structure of a DNA dodecamer in an aqueous dilute liquid crystalline phase. *J. Am. Chem. Soc.* **122**, (2000).

183. Vermeulen, A., Zhou, H. & Pardi, A. Determining DNA global structure and DNA bending by application of NMR residual dipolar couplings. *J. Am. Chem. Soc.* **122**, (2000).

184. Cornilescu, G., Delaglio, F. & Bax, A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13**, (1999).

185. Fowler, C. A., Tian, F., Al-Hashimi, H. M. & Prestegard, J. H. Rapid determination of protein folds using residual dipolar couplings. *J. Mol. Biol.* **304**, (2000).

186. Andrec, M., Du, P. & Levy, R. M. Protein backbone structure determination using only residual dipolar couplings from one ordering medium. *J. Biomol. NMR* **21**, 335–347 (2001).

187. Clore, G. M. & Bewley, C. A. Using conjoined rigid body/torsion angle simulated annealing to determine the relative orientation of covalently linked

protein domains from dipolar couplings. *Journal of Magnetic Resonance* vol. 154 (2002).

188. Assfalg, M. *et al.* 15N-1H residual dipolar coupling analysis of native and alkaline-K79A Saccharomyces cerevisiae cytochrome c. *Biophys. J.* **84**, (2003).

189. Tian, F., Valafar, H. & Prestegard, J. H. A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. *J. Am. Chem. Soc.* **123**, 11791–11796 (2001).

190. Dosset, P., Hus, J. C., Marion, D. & Blackledge, M. A novel interactive tool for rigid-body modeling of multi-domain macromolecules using residual dipolar couplings. *J. Biomol. NMR* **20**, 223–231 (2001).

191. Prestegard, J. H., Mayer, K. L., Valafar, H. & Benison, G. C. Determination of protein backbone structures from residual dipolar couplings. *Methods Enzymol.* **394**, (2005).

192. Valafar, H. *et al.* Backbone solution structures of proteins using residual dipolar couplings: Application to a novel structural genomics target. *J. Struct. Funct. Genomics* **5**, (2005).

193. Raman, S. *et al.* NMR structure determination for larger proteins using backbone-only data. *Science (80-.).* **327**, (2010).

194. Lange, O. F. *et al.* Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc. Natl. Acad. Sci. U. S. A.* **109**, (2012).

195. Tang, Y. *et al.* Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat. Methods* **12**, (2015).

196. Shealy, P., Liu, Y., Simin, M. & Valafar, H. Backbone resonance assignment and order tensor estimation using residual dipolar couplings. *J. Biomol. NMR* **50**,

(2011).

197. Cole, C. A., Mukhopadhyay, R., Omar, H., Hennig, M. & Valafar, H. Structure Calculation and Reconstruction of Discrete-State Dynamics from Residual Dipolar Couplings. *J. Chem. Theory Comput.* **12**, (2016).

198. Pomeranz, S. B. & Gershenfeld, N. The Nature of Mathematical Modeling. *Am. Math. Mon.* **107**, (2000).

199. Tjandra, N., Grzesiek, S. & Bax, A. Magnetic Field Dependence of Nitrogen−Proton J Splittings in 15N-Enriched Human Ubiquitin Resulting from Relaxation Interference and Residual Dipolar Coupling. *J. Am. Chem. Soc.* **118**, 6264–6272 (1996).

200. Prestegard, J. H. & Kishore, A. I. Partial alignment of biomolecules: An aid to NMR characterization. *Current Opinion in Chemical Biology* vol. 5 (2001).

201. Nitz, M. *et al.* Structural origin of the high affinity of a chemically evolved lanthanide-binding peptide. *Angew. Chemie - Int. Ed.* **43**, (2004).

202. Losonczi, J. A., Andrec, M., Fischer, M. W. F. & Prestegard, J. H. Order Matrix Analysis of Residual Dipolar Couplings Using Singular Value Decomposition. *Journal of Magnetic Resonance* vol. 138 (1999).

203. Tolman, J. R., Al-Hashimi, H. M., Kay, L. E. & Prestegard, J. H. Structural and Dynamic Analysis of Residual Dipolar Coupling Data for Proteins. *J. Am. Chem. Soc.* **123**, 1416–1424 (2001).

204. Peti, W., Meiler, J., Brüschweiler, R. & Griesinger, C. Model-free analysis of protein backbone motion from residual dipolar couplings. *J. Am. Chem. Soc.* **124**, (2002).

205. Meiler, J., Prompers, J. J., Peti, W., Griesinger, C. & Brüschweiler, R. Model-free approach to the dynamic interpretation of residual dipolar couplings in

globular proteins. *J. Am. Chem. Soc.* **123**, 6098–6107 (2001).

206. Rod, T. H. & Brooks, C. L. How dihydrofolate reductase facilitates protonation of dihydrofolate. *J. Am. Chem. Soc.* **125**, (2003).

207. Antikainen, N. M., Smiley, R. D., Benkovic, S. J. & Hammes, G. G. Conformation coupled enzyme catalysis: Single-molecule and transient kinetics investigation of dihydrofolate reductase. *Biochemistry* **44**, (2005).

208. Mauldin, R. V & Lee, A. L. Nuclear magnetic resonance study of the role of M42 in the solution dynamics of Escherichia coli dihydrofolate reductase. *Biochemistry* **49**, 1606–1615 (2010).

209. Cole, C., Ott, C., Valdes, D. & Valafar, H. PDBMine: A reformulation of the protein data bank to facilitate structural data mining. in *Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019* (2019). doi:10.1109/CSCI49370.2019.00272.

210. Cole, C., Parks, C., Rachele, J. & Valafar, H. *Improvements of the REDCRAFT Software Package*. https://bitbucket.org/hvalafar/redcraft/src/master/.

211. Cole, C. A., Daigham, N. S., Liu, G., Montelione, G. T. & Valafar, H. REDCRAFT: A computational platform using residual dipolar coupling NMR data for determining structures of perdeuterated proteins in solution. *PLOS Comput. Biol.* **17**, e1008060 (2021).

212. Cole, C., Ishimaru, D., Hennig, M. & Valafar, H. *An Investigation of Minimum Data Requirement for Successful Structure Determination of Pf2048.1 with REDCRAFT*. (2020).

213. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR* **44**, (2009).

214. Shen, Y. & Bax, A. Protein structural information derived from nmr chemical shift with the neural network program talos-n. *Methods Mol. Biol.* **1260**, (2015).

215. Güntert, P. Automated NMR Structure Calculation With CYANA. in *Protein NMR Techniques* (ed. Downing, A. K.) 353–378 (Humana Press, 2004). doi:10.1385/1-59259-809-9:353.

Appendix A

Supplementary Material

Supplementary Tables

Table A.1: The structure computed by REDCRAFT using standard Ramachandron restraints. As expected, the structure is locally and globally compromised due to the influence of dynamics on RDC data.

Fragment number	Residue Range	BBRMSD with 1RX2
Whole protein	1 - 159	29 Å
Fragment 1	1 - 11	2.0Å
Fragment 2	16 - 38	0.8Å
Fragment 3	44 - 60	5.5Å
Fragment 4	64 - 88	6.0Å
Fragment 5	93 - 114	9.7Å
Fragment 6	115-137	8.9Å
Fragment 7	138 - 159	0.7Å

Supplementary Figures

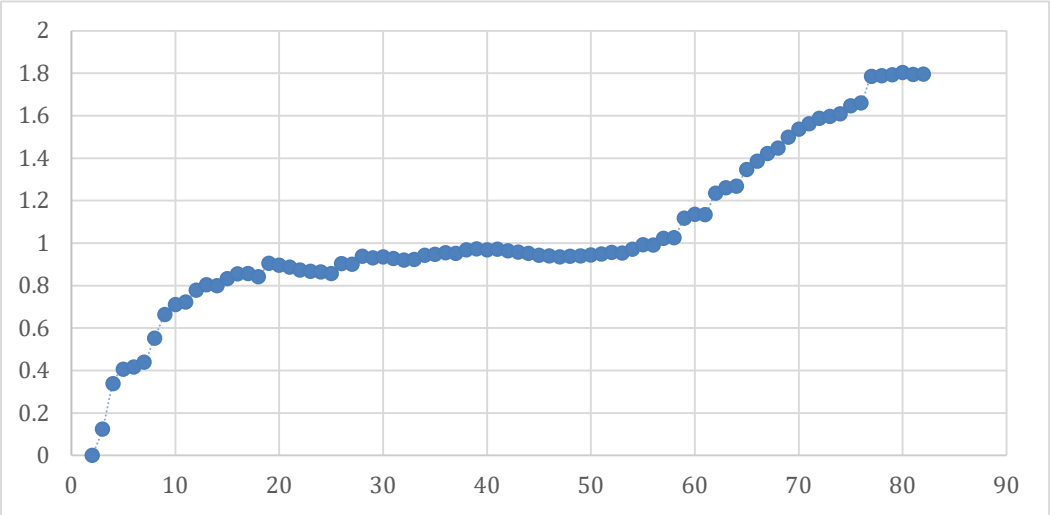

Figure A.1: Typical DP in structure with no dynamics (generated from structure).

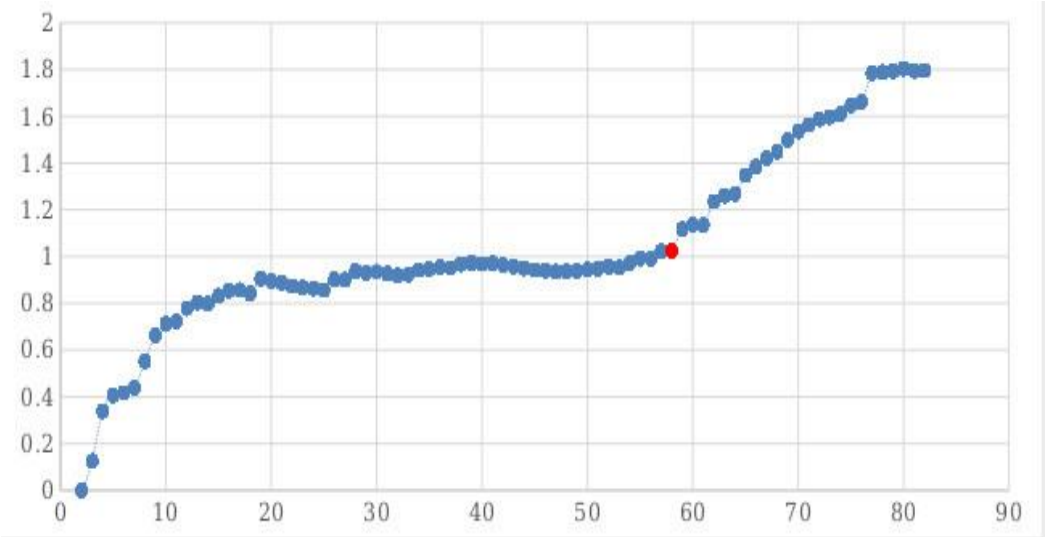

Figure A.2: DP of PDB ID 1A1Z with a simulated 2 state motion starting at residue 58 (shown in red). A uniformly distribute noise of ±1Hz was added to all RDC data.

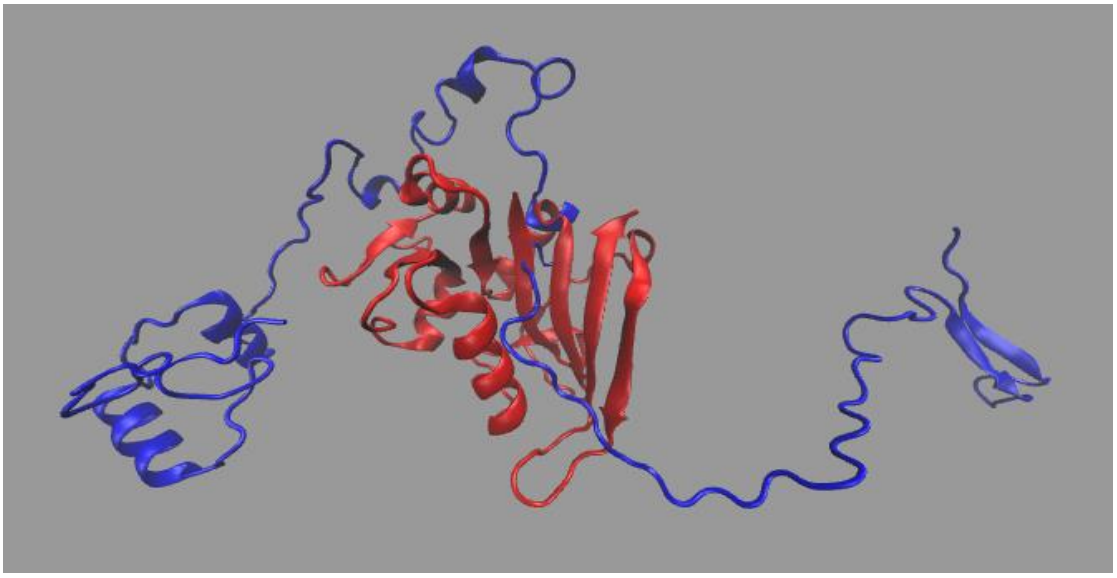

Figure A.3: Structure of DHFR determined by REDCRAFT (shown in blue) using typical Ramachandron dihedral restraints superposed on the actual X-ray structure (shown in red) with more than 35Å of bb-rmsd.

Appendix B

Information on Chapter 7

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

HV oversaw the entire project, assisted in the composition of the manuscript, development of the computational models, analysis of the data, and the design of the experiments. HO, the lead Ph.D. student, assisted with the analysis of RDC data in REDCRAFT and REDCAT software packages, assisted with the composition of the manuscript, analysis of the MD simulation, and contributed to the overall progress of this work. AH, a Ph.D. student, assisted with the data collection and analysis related to the PDBMine. He also contributed to the composition of the manuscript, software development, and data analysis. CC, assisted with composition of the manuscript, contributed to the methods development, software development, and data analytics.

FUNDING

Funding was granted to Valafar from NIGMS branch of NIH, award number 5P20GM103499-21.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.806584/full#supplementary-material , see Appendix A.

Appendix C

Permission to Reprint

The following information was taken from frontiers website on their policies (https://www.frontiersin.org/about/policies-and-publication-ethics)

Theses and Dissertations

Frontiers allows the inclusion of content which first appeared in an author's thesis so long as this is the only form in which it has appeared, is in line with the author's university policy, and can be accessed online. If the thesis is not archived online, it is considered original, unpublished data and is subject to the unpublished data restrictions of some article types. Inclusion of material from theses or dissertations should be noted in the Acknowledgements section of the manuscript AND cited accordingly in the reference list.

Preprints

Frontiers' supportive preprint policy encourages full open access at all stages of a research paper, to share and generate knowledge researchers need to support their work. Authors publishing in Frontiers journals may share their work ahead of submission to a peer-reviewed journal, as well as during the Frontiers review process, on repositories or preprint servers (such as arXiv, PeerJ Preprints, OSF, and others), provided that the server imposes no restrictions upon the author's full copyright and re-use rights. Also note that any manuscript files shared after submission to Frontiers journals, during the review process, cannot contain the Frontiers logo or branding.

Correct attribution of the original source in repositories or preprint servers must be included within the manuscript on submission or added at re-submission if the deposition is done during the review process. We ask that the preprint is both listed within the acknowledgement section and the full citation included in the reference list.

If the article is published, authors are then strongly encouraged to link from the preprint server to the Frontiers publication to enable readers to find, access, and cite the final peer-reviewed version. Please note that we cannot consider for publication content that has been previously published, or is already under review, within a scientific journal, book or similar entity.