Algorithms and Analytic Solutions using Sparse Residual Dipolar Couplings for High-Resolution Automated Protein Backbone Structure Determination by NMR

Anna Yershova[†], Chittaranjan Tripathy[†], Pei Zhou, Bruce Randall Donald*

Abstract Developing robust and automated protein structure determination algorithms using nuclear magnetic resonance (NMR) data is an important goal in computational structural biology. Algorithms based on global orientational restraints from residual dipolar couplings (RDCs) promise to be quicker and more accurate than approaches that use only distance restraints. Recent development of analytic expressions for the roots of RDC equations together with protein kinematics has enabled exact, linear-time algorithms, highly desirable over earlier stochastic methods. In addition to providing guarantees on the number and quality of solutions, exact algorithms require a minimal amount of NMR data, thereby reducing the number of NMR experiments. Implementations of these methods determine the solution structures by explicitly computing the intersections of algebraic curves representing discrete RDC values. However, if additional RDC data can be measured, the algebraic curves no longer generically intersect. We address this situation in the paper and show that globally optimal structures can still be computed analytically as points closest to all of the algebraic curves representing the RDCs. We present new algorithms that expand the types and number of RDCs from which analytic solutions are computed. We evaluate the performance of our algorithms on NMR data for four proteins: human ubiquitin, DNA-damage-inducible protein I (DinI), the Z domain of staphylococcal protein A (SpA), and the third IgG-binding domain of Protein G

[†] These authors contributed equally.

[†]Anna Yershova
Department of Computer Science, Duke University, Durham, NC 27707, USA

[†]Chittaranjan Tripathy
Department of Computer Science, Duke University, Durham, NC 27707, USA

Pei Zhou
Department of Biochemistry, Duke University Medical Center, Durham, NC 27707, USA

*Bruce Randall Donald (corresponding author)
Department of Computer Science, Duke University, and Department of Biochemistry, Duke University Medical Center, Durham, NC 27707, USA, e-mail: brd+wafr2010@cs.duke.edu

(GB3). The results show that our algorithms are able to determine high-resolution backbone structures from a limited amount of NMR data.

1 Introduction

Understanding the structures of biologically important proteins is one of the long-term goals in biochemistry. While automation has advanced many other aspects of biology, three-dimensional (3D) protein structure determination remains a slower, harder, and more expensive task. The speed at which protein structures are being discovered today is, for example, several orders of magnitude slower than that of gene sequencing.

Two established experimental approaches for protein structure determination are X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. While crystallography can provide high resolution structures when good quality crystals can be grown, NMR spectroscopy is the only experimental technique currently capable of measuring atomic-resolution geometric restraints and dynamics of proteins in physiologically-relevant solution state conditions. This sets a challenging goal in computational structural biology: design efficient automated structure determination techniques that exploit geometric restraints collected using NMR.

Even with recent advances in reducing the acquisition time for NMR data [1] it still remains advantageous to be able to determine protein structure from fewer experiments [2]. Therefore, to design successful computational approaches the interplay between all of the components of the structure determination process should be taken into account: the time and cost of NMR data acquisition, the types of NMR data and their information content, the algorithmic complexity of the problem of computing the 3D structure, and the accuracy of the obtained solution.

Previous algorithms for structure determination problem by NMR can be divided into three categories: *stochastic search, systematic search*, and *exact* algorithms. The first category includes many widely-used approaches [3, 4, 5, 6, 7, 8], which perform stochastic search over possible structures, scored according to their agreement with experimental data. These methods suffer from well-known pitfalls, such as local minima, undersampling, non-convergence, and missed solutions. While stochastic methods may perform adequately in data-rich settings, collecting a large number of NMR spectra increases the time and cost of the experiments, and still provides no guarantee on the quality of solutions.

The second category involves systematic grid search over possible structures [9, 10, 11, 12] and scoring according to the experimental data fit. Excessive computational cost and undersampling due to insufficient resolution are the limitations of these methods.

The development of sampling methodology was influenced by the computational complexity of the structure determination problem using the *nuclear Overhauser effect* (NOE) data. The recent introduction of *residual dipolar couplings* (RDCs) [13, 14] has enabled novel attacks on the problem. In contrast to NOEs, which represent local distance restraints, RDCs measure the global orientation of internuclear vectors. While many RDC-based methods still use stochastic search

[15, 16, 9, 17, 18, 12, 19, 10] despite its pitfalls, exact algorithms [20, 21, 2] have recently emerged. These methods explicitly represent the RDC equations as well as protein kinematics in algebraic form to compute structures that optimize the fit to the RDC data. The exact algorithms not only guarantee completeness and polynomial running time, they also use a sparse set of RDC measurements (e.g. only two or three RDCs per residue), which reduces the time and cost of collecting experimental data.

Implementations of these methods determine the solution structures by explicitly computing the intersections of algebraic curves representing discrete RDC values and protein kinematics [20, 21, 2]. However, if additional RDC data is measured, the algebraic curves representing the RDCs no longer generically intersect. Even though collecting additional experimental data may improve convergence of stochastic structure determination methods, development of efficient new exact methods is needed to handle this scenario. In this paper, we present algorithms that expand the types and number of RDC data from which analytic solutions are computed. When additional orientational restraints can be measured, the globally optimal structures are calculated as points closest to all of the algebraic curves representing the RDC constraints. Therefore, the structures that optimally agree with the collected experimental data are determined analytically.

Moreover, with additional RDC data our algorithms can compute structures of *loops*, more challenging regions in a protein for structure determination compared to *secondary structure elements (SSEs)*, such as α-helices and β-sheets. Loops increase computational complexity of the exact algorithms, because there is less physical constraints on the structures that are possible for loops to assume, as is reflected in their Ramachandran statistics [22]. Therefore, we expand the domain of applicability of exact algorithms to structure determination problems which in practice only stochastic methods could handle before. This extends the benefits of analytic solutions into a new and important domain.

In summary, the following contributions are made in this paper:

- We present a general framework for computing protein backbone structures from sparse RDC measurements.
- We describe new algorithms that handle two particular types of RDC data: two RDCs per residue in one or two alignment media, and multiple RDCs per residue in multiple media.
- We present the analysis of our methods in terms of the upper bound on the number of optimal structures the algorithm can generate.
- We evaluate the performance of our methods on NMR data sets for four proteins: human ubiquitin, DNA-damage-inducible protein I (DinI), the Z domain of staphylococcal protein A (SpA), and the third IgG-binding domain of Protein G (GB3).

We start with background on RDCs in Section 2. We formally define the protein structure determination problem in Section 3. Section 4 presents the general framework for our exact algorithms, as well as detailed explanations of the methods that

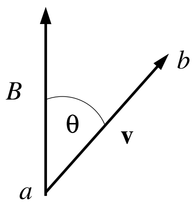

Fig. 1: For an internuclear vector **v** between two NMR-active nuclear spins *a* and *b*, and a vector representing the gradient of the magnetic field *B*, an RDC experiment detects the interaction between the nuclear spins. This interaction is dependent on the internuclear distance $r^3_{a,b}$ as well as the angle θ between vectors **v** and *B*.

handle two different types of RDC data (Sections 4.1 and 4.2). We present a study on the performance of our methods in Section 6 and conclusions in Section 7.

2 Background

The Physics of RDCs. During an NMR experiment, a protein in solution is placed in a static magnetic field. The magnetic field interacts with the nuclear spins of atoms of the protein, which allows collecting various NMR measurements.

Consider a vector **v** between two NMR-active nuclear spins, *a* and *b*, of two atoms, and a vector representing the gradient of the magnetic field *B* in Figure 1. An RDC experiment detects the interaction between the nuclear spins. This interaction is measured in units of Hertz and can be expressed as

$$D = \frac{\mu_0 \gamma_a \gamma_b \hbar}{4\pi^2 \left\langle r^3_{a,b} \right\rangle} \left\langle \frac{3\cos^2\theta - 1}{2} \right\rangle, \tag{1}$$

in which D is the residual dipolar coupling measurement, μ_0 is the magnetic permeability of vacuum, $\hbar$ is the reduced Planck's constant, γ is the gyromagnetic ratio, $r_{a,b}$ is the distance between the two spins, θ is the angle between the internuclear vector, **v**, and B. The angle brackets represent an average over time and an ensemble of proteins in solution.

Intuitively, as the protein tumbles in solution, the internuclear vector tumbles with the protein. The RDC measurement represents the time average over such movements. This measurement is of little use if the protein tumbles freely (isotropically), since the average is zero and hence no RDC value is detected. An anisotropic solution is used to constrain the motions of the protein by adding an *alignment medium* to the solution. The resulting RDC measurement represents the relationship between the internuclear vector and the set of parameters associated with the alignment medium, known as the *Saupe matrix* [23], or *alignment tensor*.

The Tensor Formulation of the RDC Equation. A more convenient form of the RDC equation (1) for computing the geometric structure of a protein can be obtained after a series of algebraic manipulations [2]:

$$D = D_{\max} \mathbf{v}^{\mathrm{T}} \mathbf{S} \mathbf{v}, \tag{2}$$

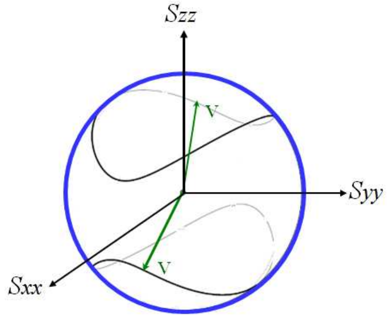

Fig. 2: The RDC sphero-quartic curves in the principal order frame for the alignment tensor (S_{xx}, S_{yy}, S_{zz}). The unit internuclear vector **v** that satisfies the RDC equation (4) is constrained to the sphero-quartic curves on the unit sphere.

in which D_{max} is the dipolar interaction constant, **v** is the unit internuclear vector, and **S** is the Saupe matrix [23] corresponding to the alignment medium used in the NMR experiment.

To derive Equation (2) from (1), a transformation to the *molecular coordinate frame* associated with the protein internuclear vector **v** is made [20, 21]. In such a coordinate frame, **v** is static, and the magnetic field vector, B, tumbles around **v**. The time and ensemble average over such tumbling of B is represented by the Saupe matrix, which we discuss later in this section.

Note that Equation (2) directly relates an RDC value to the orientation of **v** in the molecular coordinate frame. A sufficient number of such equations allows computation of all internuclear vector orientations in a protein with respect to the molecular coordinate frame. In the rest of the paper, by denoting $r = D/D_{max}$, we will work with the *normalized* RDC equation:

$$r = \mathbf{v}^{\mathrm{T}}\mathbf{S}\mathbf{v}. \tag{3}$$

Alignment Tensors. The Saupe matrix in Equations (2) and (3) is a 3×3 symmetric and traceless matrix. It contains 5 degrees of freedom, 3 of which correspond to a 3D rotation, and 2 are eigenvalues.

Each alignment tensor can be diagonalized as $\mathbf{S} = \mathbf{R}^{\mathrm{T}}\hat{\mathbf{S}}\mathbf{R}$, in which $\mathbf{R}$ is a 3D rotation matrix, and the traceless diagonal matrix $\hat{\mathbf{S}}$ has eigenvalues $(S_{xx}, S_{yy}, S_{zz}), S_{zz} = -S_{xx} - S_{yy}$. Then, the RDC equation (3) becomes

$$r = S_{xx}x^2 + S_{yy}y^2 + S_{zz}z^2, \tag{4}$$

in which $\mathbf{R}\mathbf{v} = (x, y, z)$. The coordinate frame that diagonalizes the Saupe matrix is called the *principal order frame (POF)* of the alignment medium.

To design exact algorithms, we consider algebraic representations of solutions. Figure 2 shows the solutions to the RDC equation for (x, y, z) as *sphero-quartic* curves on a sphere in the principal order frame. These curves are the intersection of the unit sphere $x^2 + y^2 + z^2 = 1$ and a hyperboloid representing the RDC equation (4).

Protein Geometry and Assumptions on Dynamics. A protein is modeled as a collection of peptide planes. Each peptide plane contains the bond vectors: C^{α}-C',

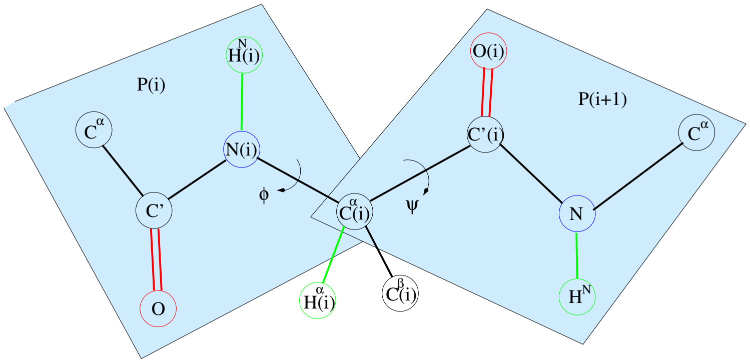

Fig. 3: Protein geometry. Two peptide planes P(i) and P(i+1) and the atoms of the corresponding residue *i* are shown. Two torsional degrees of freedom of the backbone are represented by dihedral angles ϕ and ψ.

C′-N, C′-O, N-H$^{\mathrm{N}}$ and N-C$^{\alpha}$. Peptide planes are joined through the bond vectors N-C$^{\alpha}$ and C$^{\alpha}$-C′, with two torsional degrees of freedom. *Dihedral angles*, ϕ and ψ, parametrize these joints (see Figure 3). The bond vectors C$^{\alpha}$-H$^{\alpha}$ and C$^{\alpha}$-C$^{\beta}$ do not belong to peptide planes and form a fixed tetrahedral geometry with both N-C$^{\alpha}$ and C$^{\alpha}$-C′. Often, it is convenient to divide a protein into *residues*. Most of the residues (excluding glycine and proline with different geometry) contain consecutive atoms H, N, C$^{\alpha}$, H$^{\alpha}$, C$^{\beta}$, C′ and O, as shown in Figure 3. The *backbone* of a protein is a sequence of repeating C$^{\alpha}$, C′, and N atoms.

There exist NMR techniques to measure RDCs on the C$^{\alpha}$-H$^{\alpha}$, C$^{\alpha}$-C′, C$^{\alpha}$-C$^{\beta}$, N-H$^{\mathrm{N}}$, and C′-N bond vectors as well as on C′-H$^{\mathrm{N}}$ internuclear vector.

Crucial assumptions about the *dynamics* of the protein are often necessary to solve for the large number of unknown variables in RDC equations. Initially, neither internuclear vectors $\mathbf{v}$ nor Saupe matrices $\mathbf{S}$ are known in the structure determination process. The assumption that the protein fluctuates in a small ensemble about one principal mode, that is, the protein is more or less rigid, leads to a simplified dynamics model in which the gradient vector B of the magnetic field is assumed to tumble similarly with respect to each internuclear vector. We work with this model, because it is applicable to many proteins. However, when the protein is known to be more flexible in solution, additional parameters representing this flexibility must be integrated into the model.

3 The Protein Structure Determination Problem

The protein structure determination problem considered in this paper is: given sparse RDC data for a protein, determine the alignment tensor(s) and dihedral angles of the backbone that produce the best fit to the experimental data as well as represent a valid protein model.

Sparse RDC data. We consider the case of experimental data, containing as few as two RDCs per residue in one or two media. To ease the presentation of our methods, we differentiate between RDC measurements based on their location

ϕ-defining	C^{α}-H^{α}, C^{α}-C', C^{α}-C^{β}
ψ-defining	N-H^{N}, C'-N, C'-H^{N}

Table 1: A ϕ-defining RDC is used to compute the backbone dihedral ϕ, and a ψ-defining RDC is used to compute the backbone dihedral ψ.

within a residue. Denote RDC measurements on internuclear vectors C^{α}-H^{α}, C^{α}-C', and C^{α}-C^{β} as *ϕ-defining* RDCs, and N-H^{N}, C'-N, and C'-H^{N} as *ψ-defining* RDCs (Table 1). In this paper we design exact algorithms for the following types of RDC data: (a) One ϕ-defining RDC and one ψ-defining RDC per residue in one or two media; (b) Multiple ϕ-defining RDCs and multiple ψ-defining RDCs per residue in one or more media.

In general, it may not be possible to record a complete set of such RDC data for the entire protein. When we describe our methods in Section 4, we assume that all of the RDC values are present for a *protein fragment*. We describe how we deal with some cases of incomplete data in Section 5.

Data fit. The data fit function that we use is RDC root mean square deviation (RMSD). Denote $\mathbf{S}_j, j = 1, \ldots, m$, all of the alignment tensors involved, and $\phi_i, \psi_i, i = 1, \ldots n$, the dihedral angles for residues 1 through n in the protein backbone. The RDC RMSD for given alignment tensors and dihedral angles is computed using the equation:

$$\sigma(\{\mathbf{S}_j\}_{j=1}^{m}, \{\phi_i, \psi_i\}_{i=1}^{n}) = \sqrt{\frac{1}{l}\sum_{k=1}^{l}(r_k^b - r_k^e)^2}, \tag{5}$$

in which l is the total number of RDCs for all residues, r_k^e is the experimental RDC, and r_k^b is the RDC value back-computed from the alignment tensors and the structure defined by the dihedral angles computed using Equation (3).

Validation. To ensure that the solution structure is biologically meaningful, we validate it according to two criteria: Ramachandran regions [22], and van der Waals packing [24].

Next we proceed to the methods which solve the protein structure determination problem defined in this section.

4 Methods

The key to our methods is the use of explicit representation of the protein kinematics incorporated into the RDC equations.

At this point, it is useful to introduce several notations. Without loss of generality, we choose the principal order frame of $\mathbf{S}_1$ (POF_1) as the global coordinate frame. Within this coordinate frame, $\mathbf{S}_1$ is diagonal, with eigenvalues $S_{1,xx}$, $S_{1,yy}$, and $(-S_{1,xx} - S_{1,yy})$. We denote the diagonal components of any other alignment tensor, $\mathbf{S}_j$, as $S_{j,xx}$ and $S_{j,yy}$. $\mathbf{R}_{j,O}$ denotes the orientation of the principle order frame of alignment tensor $\mathbf{S}_j$ in POF_1.

Within a protein fragment, there will be several coordinate frames associated with different internuclear vectors in the portion, and related to each other through the protein kinematics. Our algorithms keep representations of these frames in POF_1.

```
BRANCH-AND-BOUND ($\mathbf{S}_1,\ldots,\mathbf{S}_m$, $\mathbf{R}_{i,P}$)
Input: Global orientation of the coordinate frame for the current internuclear vector,
the RDC data for the next internuclear vector, and the alignment tensor(s)
Output: Those dihedral angles that represent valid protein structures and
best fit to the experimental data
Branch:   Use methods from Sections 4.1 and 4.2 to solve RDC equations.
          Each solution is a dihedral that represents a child node.
Bound:    Prune invalid children nodes based on:
          RDC RMSD, Ramachandran regions, and van der Waals packing
Recurse:  Compute global orientation of the coordinate frame for each valid
          child node, $\mathbf{R}_{i+1,P}$, call BRANCH-AND-BOUND ($\mathbf{S}_1,\ldots\mathbf{S}_m$, $\mathbf{R}_{i+1,P}$)
```

Fig. 4: The outline of the branch-and-bound tree search algorithm. The tree encodes all the solutions to the system of RDC equations together with the kinematic equations that relate consecutive internuclear vectors in the protein backbone. The algorithm systematically searches through these solutions and prunes those that do not satisfy the bound conditions. Different types of RDC data require different branch-and-bound criteria, which we cover in Sections 4.1 and 4.2.

Consider a coordinate frame defined at the peptide plane P_i with z-axis along the bond vector $\mathrm{N}(i) \rightarrow \mathrm{H}^{\mathrm{N}}(i)$ of residue i, in which the notation $a \rightarrow b$ means a vector from the nucleus a to the nucleus b. The y-axis is on the peptide plane i and the angle between the y-axis and the bond vector $\mathrm{N}(i) \rightarrow \mathrm{C}^{\alpha}(i)$ is 29.14° as described in [20]. The x-axis is defined based on right-handedness. Let $\mathbf{R}_{i,P}$ denote the orientation (rotation matrix) of P_i with respect to POF_1. Then, $\mathbf{R}_{1,P}$ denotes the relative rotation matrix between the coordinate system defined at the first peptide plane of the current protein portion and the principal order frame of the alignment tensor $\mathbf{S}_1$. We call it the *orientation of the first peptide plane*.

Our methods follow the following framework. First, the diagonal and rotational components of the alignment tensors and the orientation of the first peptide plane, $\mathbf{R}_{1,P}$, are estimated using methods described in Section 5. Next, the branch-and-bound tree search algorithm shown in Figure 4 is called, which computes the dihedral angles of the protein portion backbone. The tree encodes all the solutions to the system of RDC equations together with the kinematic equations that relate consecutive internuclear vectors in the protein backbone. Depending on the types of RDC data, the branch-and-bound search performs different computations. We cover these differences in Sections 4.1 and 4.2.

4.1 Exact Solutions for Peptide Plane Orientations from One ϕ-defining RDC and One ψ-defining RDC

In this section, consider the case of experimental RDC data that only contains one ϕ-defining and one ψ-defining RDC measurement per residue. This covers, for example, the case of having RDC measurements for C^{α}-H^{α}, and N-H^{N} in one medium. It also covers having C^{α}-H^{α} RDC in one medium and N-H^{N} RDC in second medium. Next, we describe the inductive step of the branch-and-bound tree search algorithm in Figure 4.

The algorithm uses $\mathbf{R}_{i,P}$ to derive $\mathbf{R}_{i+1,P}$ inductively after it computes the backbone dihedral angles ϕ_i and ψ_i. $\mathbf{R}_{i+1,P}$ is then used to compute the dihedrals of the $(i+1)^{st}$ peptide plane. The angles ϕ_i and ψ_i are computed using the following propositions.

Proposition 1 *Given the diagonalized alignment tensor components $S_{1,xx}$ and $S_{1,yy}$, the orientation of the i^{th} peptide plane $\mathbf{R}_{i,P}$, and a ϕ-defining RDC, r, for the corresponding internuclear vector, $\mathbf{v}$, there exist at most 4 possible values of the dihedral angle ϕ_i that satisfy the RDC. The possible values of ϕ_i can be computed exactly and in closed form by solving a quartic equation.*

Proof. Let the unit vector $\mathbf{v}_0 = (0,0,1)^T$ represent the N-H$^{\mathrm{N}}$ bond vector of residue i in the local coordinate frame defined on the peptide plane P_i. Let $\mathbf{v} = (x,y,z)^T$ denote the internuclear vector for the ϕ-defining RDC for residue i in the principal order frame. We can write the forward kinematics relation between $\mathbf{v}$ and $\mathbf{v}_0$ as

$$\mathbf{v} = \mathbf{R}_{i,P}\,\mathbf{R}_l\,\mathbf{R}_z(\phi_i)\,\mathbf{R}_r\,\mathbf{v}_0, \tag{6}$$

in which $\mathbf{R}_l$ and $\mathbf{R}_r$ are constant rotation matrices that describe the kinematic relationship between $\mathbf{v}$ and $\mathbf{v}_0$. $\mathbf{R}_z(\phi_i)$ is the rotation about the z-axis by ϕ_i.

Let c and s denote $\cos\phi_i$ and $\sin\phi_i$, respectively. Using this while expanding Equation (6) we have

$$x = A_0 + A_1c + A_2s,\; y = B_0 + B_1c + B_2s,\; z = C_0 + C_1c + C_2s, \tag{7}$$

in which A_i, B_i, C_i for $0 \le i \le 2$ are constants. Using Equation (7) in the RDC equation (4) and simplifying we have

$$K_0 + K_1c + K_2s + K_3cs + K_4c^2 + K_5s^2 = 0, \tag{8}$$

in which K_i, $0 \le i \le 5$ are constants. Using half-angle substitutions

$$u = \tan(\frac{\phi_i}{2}),\; c = \frac{1-u^2}{1+u^2},\; \text{and}\; s = \frac{2u}{1+u^2} \tag{9}$$

in Equation (8) we obtain

$$g(u) = L_0 + L_1u + L_2u^2 + L_3u^3 + L_4u^4 = 0, \tag{10}$$

in which L_i, $0 \le i \le 4$ are constants. Equation (10) is a quartic equation which can be solved exactly and in closed form. Let $\{u_1, u_2, u_3, u_4\}$ denote the set of four real solutions (at most) of Equation (10). For each u_i the corresponding dihedral angle ϕ_i can be computed using Eq. (9). □

Proposition 2 *Given the diagonalized alignment tensor components $S_{1,xx}$ and $S_{1,yy}$, the orientation of the i^{th} peptide plane $\mathbf{R}_{i,P}$, the dihedral ϕ_i, and a ψ-defining RDC, r, for the corresponding internuclear vector, $\mathbf{v}'$, on the peptide plane P_{i+1}, there exist at most 4 possible values of the dihedral angle ψ_i that satisfy the RDC. The possible values of ψ_i can be computed exactly and in closed form by solving a quartic equation.*

Proof. After representing the internuclear vector $\mathbf{v}'$ through $\mathbf{v}_0$ using protein kinematics:

$$\mathbf{v}' = \mathbf{R}_{i,P}\,\mathbf{R}_l\,\mathbf{R}_z(\phi_i)\,\mathbf{R}_r\,\mathbf{R}'_l\,\mathbf{R}_z(\psi_i)\,\mathbf{R}'_r\,\mathbf{v}_0, \tag{11}$$

the proof is similar to that in Proposition 1, since the value of ϕ_i is known. □

Proposition 3 *Given the diagonalized alignment tensor components, the orientation of the i^{th} peptide plane $\mathbf{R}_{i,P}$, a ϕ-defining RDC and a ψ-defining RDC for ϕ_i and ψ_i, respectively, there exist at most 16 orientations,$\mathbf{R}_{i+1,P}$, of the peptide plane P_{i+1} that satisfy the RDCs.*

Proof. This follows from the direct application of Propositions 1 and 2. □

Proposition 4 *Given the diagonalized alignment tensor components $S_{1,xx}$ and $S_{1,yy}$ for medium 1, $S_{2,xx}$ and $S_{2,yy}$ for medium 2, a relative rotation matrix $\mathbf{R}_{2,O}$, the orientation of the i^{th} peptide plane $\mathbf{R}_{i,P}$, a ϕ-defining RDC in medium 1 and a ψ-defining RDC in medium 2 for ϕ_i and ψ_i, respectively, there exist at most 16 orientations, $\mathbf{R}_{i+1,P}$, of the peptide plane P_{i+1} that satisfy the RDCs, which can be computed exactly and in closed form by solving two quartic equations.*

Proof. It follows the proof of Proposition 3, once the transformation to the principal order frame of medium 2 is made by $v' = \mathbf{R}_{2,O} v$ to compute the value of ψ_i. □

4.2 Exact Minima for Peptide Plane Orientations from Multiple ϕ-defining RDCs and Multiple ψ-defining RDCs.

Now, consider the case when additional RDC data has been collected, and more than one ϕ and ψ-defining RDC measurements are available in one or more media. This covers, for example, the case of having RDCs for C^{α}-H^{α}, C^{α}-C', N-H^{N}, and C'-N in one medium. It also covers the case of having C^{α}-H^{α} and N-H^{N} RDCs in two media. The inductive step of the tree search in Figure 4 is performed using the following propositions.

Proposition 5 *Given the diagonalized alignment tensor components $S_{j,xx}$ and $S_{j,yy}$, the rotations between principal order frames, $\mathbf{R}_{j,O}$, the orientation of the i^{th} peptide plane $\mathbf{R}_{i,P}$, and multiple ϕ-defining RDC for the corresponding internuclear vector* $\mathbf{v}$ *of residue i, the global minimum of the RDC RMSD function for* $\mathbf{v}$ *can be computed exactly. There exist at most 4 possible values of the dihedral angle ϕ_i that minimize the RDC RMSD function, and such values of ϕ_i can be computed exactly.*

Proof. Let l be the number of RDC equations available for the internuclear vector $\mathbf{v}$. The RDC RMSD function for $\mathbf{v}$ is a univariate function of ϕ:

$$\sigma(\phi) = \sqrt{\frac{1}{l}\sum_{k=1}^{l}(r_k^b - r_k^e)^2}, \tag{12}$$

in which r_k^b is the back computed RDC value, $r_k^b = \mathbf{v}^T \mathbf{S}_j \mathbf{v}$, for the appropriate alignment medium $\mathbf{S}_j$. Similarly to the proof of Proposition 1, $\mathbf{v}$ can be represented as

$$\mathbf{v} = \mathbf{R}_{i,P}\, \mathbf{R}_l\, \mathbf{R}_z(\phi_i)\, \mathbf{R}_r\, \mathbf{v}_0. \tag{13}$$

After denoting $\cos\phi_i$ and $\sin\phi_i$ as c and s, respectively, Equation (13) becomes

$$x = A_0 + A_1 c + A_2 s,\; y = B_0 + B_1 c + B_2 s,\; z = C_0 + C_1 c + C_2 s, \tag{14}$$

in which A_i, B_i, C_i for $0 \le i \le 2$ are constants. Substituting x, y, and z into each RDC term of Equation (12) and using half-angle substitutions we obtain:

$$\sigma(u) = \sqrt{\frac{1}{2}\sum_{k=1}^{2}(g_k(u))^2}, \tag{15}$$

in which $g_k(u)$ are quartic polynomials for each medium k as in Equation 10.

Equation (15) defines a univariate function of u that can be minimized exactly, by finding zeroes of its derivative function. Let $\{u_1, u_2, u_3, u_4\}$ denote the set of four minima (at most) of Equation (15). For each u_i the corresponding dihedral angle ϕ_i can be computed using Eq. (9). □

Proposition 6 *Given the diagonalized alignment tensor components $S_{j,xx}$ and $S_{j,yy}$, the rotations between principal order frames, $\mathbf{R}_{j,O}$, the orientation of the i^{th} peptide plane $\mathbf{R}_{i,P}$, the dihedral ϕ_i, and multiple ψ-defining RDCs for the corresponding internuclear vector $\mathbf{v}'$ on peptide plane P_{i+1}, the global minima of the RDC RMSD function for $\mathbf{v}'$ can be computed exactly. There exist at most 4 possible values of the dihedral angle ψ_i that minimize the RDC RMSD function, and such values of ψ_i can be computed exactly.*

Proof. The proof is similar to that in Proposition 5, after the transformation as in Proposition 2 is used. □

Proposition 7 *Given the alignment tensors $\{\mathbf{S}_j\}_{j=1}^{m}$, the orientation of the peptide plane P_i, multiple ϕ-defining RDC and multiple ψ-defining RDC for ϕ_i and ψ_i, respectively, there exist at most 16 orientations of the peptide plane P_{i+1} with respect to P_i that minimize the RDC RMSD functions for each of the internuclear vectors.*

Proof. This follows from the direct application of Propositions 5 and 6. □

Note that the case of data described in this section allows comparing RDC RMSD for the branches of the search tree in Figure 4. This enables reducing the size of the search tree by pruning the branches based on RDC RMSD, which was not possible for the data described in Section 4.1 since RDC RMSD was always 0. Pruning based on Ramachandran regions and steric clashes alone is not always enough to compute the structures of protein loops. In Section 6 we show that if RDCs in second medium are measured, pruning based on RDC RMSD allows computation of loops.

5 Alignment Tensors, Orientation of the First Peptide Plane, and Packing

In our current implementation we estimate alignment tensor(s) similarly to [21, 20], by using singular value decomposition (SVD) [25] method to fit experimental RDC data to the corresponding vectors of an α-helix with ideal geometry. After that, the alignment tensor(s) are iteratively refined by using the computed helix structures by our exact algorithms. Once the values of the alignment tensor(s) are estimated, other fragments of the protein are computed using these values.

We can use uniform samples over rotation matrices to obtain the orientation of the first peptide plane in a protein fragment that result in structures with the best fit to the RDC data. For certain types of RDC data, however, the orientation of the first peptide plane in a fragment can be computed analytically.

The resulting running time complexity of our algorithms is linear, and the analysis is similar to [21]. As described in [21], we use a divide-and-conquer approach in which the protein is first divided into $O(n)$ fragments of constant length (typically 5-14 residues) based on their secondary structure type (α-helix, β-sheet, loop), and then our algorithms from Section 4 are applied to determine the orientations and conformations of these fragments. In contrast to [21], in which an algebraic geometry approach for finding the structure that minimizes the RDC fit for various RDC data was described, in this paper we have presented algorithms that achieve the same goal, but are simple and practical to implement.

In some cases reliable computation of the structure of certain fragments in a protein is not possible from sparse RDC alone. This may happen due to missing RDCs for certain residues. It may also happen due to the large size of the set of possible solutions returned by methods above. To overcome this problem, we use a sparse set of distance restraints (NOEs) to assemble the fragments. Our packing method [30] considers all possible discrete translations of the fragments over a three dimensional grid (within a parametrized resolution) that satisfy these sparse NOEs.

We also incorporate sparse unambiguous NOEs to pack β-sheets. We use rotamers from the Richardsons' Rotamer Library [49], and model the side-chain NOEs to pack the strands while they are being computed using the methods we described in the previous sections. A composite scoring scheme is used as a bound criterion in the tree search is based on a combination of (1) RDC RMSD, (2) RMS dihedral deviation from an ideal strand, (2) hydrogen bond score, i.e., a combination of RMS deviation of proton-acceptor distance and RMS deviation of donor-proton-acceptor angle violation, (3) score from the steric checker, and (4) NOE RMSD.

However, when a sufficient number of RDCs is measured for a protein, the divide-and-conquer approach is applied to the neighboring fragments sequentially, and, therefore, packing of the fragments does not require NOEs.

6 Results

We implemented our algorithms in a software package called RDC-ANALYTIC. Table 2 shows the results of the application of RDC-ANALYTIC on datasets for human ubiquitin (PDB id: 1ubq [26], DNA-damage-inducible protein I (DinI, PDB id: 1ghh [27]), Z-Domain of Staphylococcal Protein A (SpA, PDB id: 1q2n [28]), and the third IgG-binding domain of Protein G (GB3, PDB id: 1p7e [29]). For these proteins, the experimental NMR data has been taken from the Biological Magnetic Resonance Data Bank (BMRB). For ubiquitin, our program estimates the alignment tensors for different sets of RDCs and computes the helix (Ile23-Lys33) conformations. The results show that the corresponding alignment tensor components computed from different sets of RDCs in one medium or two media agree fairly well with those computed from ubiquitin NMR structure (PDB id: 1d3z). As shown in Table 2, the backbone RMSDs of the helices compared to the X-ray structure (PDB id: 1ubq) and NMR reference structure (PDB id: 1d3z) are small. The global folds of ubiquitin, DinI and SpA computed from different sets of RDCs and sparse sets of NOEs are shown in Figure 5. The results are summarized in Table 2. For ubiquitin

Protein	RDCs^a used & RMSD (Hz)	Alignment Tensor(s) (S_{yy}, S_{zz})	Backbone RMSD (Å) vs. X-ray/NMR structure
Ubiquitin^{c,d} α:23-33	C$^\alpha$-H$^\alpha$: 1.11, N-H^N: 0.740	15.230, 24.657	1.276
β:2-7, 12-17, 41-45, 65-70	C$^\alpha$-C$'$: 0.129, N-H^N: 0.603	14.219, 25.490	1.172
DinI^{c,d}	C$^\alpha$-C$'$: 0.483, N-H^N: 1.203	10.347, 33.459	1.111
α:18-32,58-72			
β:2-8, 39-44, 49-53			
SpA^d	(run1) C$^\alpha$-H$^\alpha$: 0.458,N-H^N: 2.11	8.008, 23.063	1.063
α:8-17, 24-36, 41-54	(run2) C$^\alpha$-H$^\alpha$: 0.678,N-H^N: 0.543	8.146, 24.261	1.577
	^{b}(run3)C$^\alpha$-C$'$: 1.237,N-H^N: 1.049	7.676, 22.961	0.834
Ubiquitin α:25-31	2xC$^\alpha$-H$^\alpha$: 0.93, 2xN-H^N: 0.32	16.9, 23.2; 7.0, 52.4	0.403
loop:54-58	2xC$^\alpha$-H$^\alpha$: 2.2, 2xN-H^N: 0.7	16.9, 23.2; 7.0, 52.4	0.409
loop:59-64	2xC$^\alpha$-H$^\alpha$: 1.9, 2xN-H^N: 1.2	16.9, 23.2; 7.0, 52.4	0.652
loop/β:64-70	2xC$^\alpha$-H$^\alpha$: 3.1, 2xN-H^N: 1.2	16.9, 23.2; 7.0, 52.4	0.49
β:2-7	2xC$^\alpha$-H$^\alpha$: 2.6, 2xN-H^N: 1.4	16.9, 23.2; 7.0, 52.4	0.64
β:11-17	2xC$^\alpha$-H$^\alpha$: 2.6, 2xN-H^N: 1.5	16.9, 23.2; 7.0, 52.4	0.50
β:41-55	2xC$^\alpha$-H$^\alpha$: 2.2, 2xN-H^N: 0.8	16.9, 23.2; 7.0, 52.4	0.44
GB3$^{(e)}$ α/loop:23-39	2xC$^\alpha$-H$^\alpha$: 1.7, 2xN-H^N: 1.6	47.0, 19.2; 23.8, 12.6	0.35
loop/β:39-51	2xC$^\alpha$-H$^\alpha$: 0.9, 2xN-H^N: 1.3	16.9, 23.2; 7.0, 52.4	0.49
loop/β:51-55	2xC$^\alpha$-H$^\alpha$: 0.7, 2xN-H^N: 0.5	16.9, 23.2; 7.0, 52.4	0.54

Table 2: **Results of** RDC-ANALYTIC**.** (a) Experimental RDC data for ubiquitin (PDB id: 1d3z), DinI (PDB id: 1ghh), SpA (PDB id: 1q2n), and GB3 (PDB id: 1p7e) are taken from the Biological Magnetic Resonance Data Bank (BMRB). The SSE backbones are computed for different combinations of RDCs in one or two media. If RDC measurements in two media are collected for a bond vector, we denote it by 2x in the table (e.g. 2xC$^\alpha$-H$^\alpha$). For ubiquitin the computed SSEs are compared with both the X-ray structure (PDB id: 1ubq) and the NMR structure (PDB id: 1d3z, Model 1). For DinI, SpA and GB3, since only the NMR structures are available, we compare our SSEs with Model 1 of the respective ensemble. (b) Simulated RDCs obtainted from the reference structure are used. (c) Simultaneous structure computation and assembly of β-strands into β-sheets of ubiquitin and DinI are done using 13 and 6 NOEs, respectively, which involve only amide and methyl protons obtainable using ^{1}H-^{13}C-ILV methyl labeling. (d) For ubiquitin, DinI and SpA we used 5, 10 and 10 C$^\alpha$-C$^\alpha$ distance restraints, respectively, to pack the SSEs and obtain the maximum likelihood backbone folds. (e) Simulated RDCs from 1p7e Model 1 are used only for the missing RDC values in the experimental data.

and DinI we used sparse sets of NOEs which involve only amide and methyl protons obtainable from ^1H-^{13}C-ILV methyl labeling. We also used C$^\alpha$-C$'$ and N-H^N RDCs. For ubiquitin, the backbone RMSDs between the structures computed our algorithm and the reference structures (Model 1 of 1d3z, and 1ubq) is $<$ 1.28 Å. For, DinI we computed the backbone fold using C$^\alpha$-C$'$ and N-H^N RDCs. Compared to the reference structure (Model 1 of 1ghh) the backbone RMSD was 1.11 Å.

For SpA we performed three runs of our program. In the first two runs, we used C$^\alpha$-H$^\alpha$ and N-H^N RDCs and selected different sets of parameters. For the first run, the backbone fold computed by our algorithm is within 1.1 Å of the reference structure (Model 1 of 1q2n). For the second run, we used a narrow sampling interval for N-H^N RDCs, and as a result the N-H^N RDC RMSD of the structure computed was better than that for the structure computed in the first run. However, when we packed the SSEs computed from the second run and then compared the resulting backbone fold with the reference structure (Model 1 of 1q2n), the backbone RMSD was 1.58 Å, slightly higher than that from the first run. We found that when the first two helices (Glu24-Asp36, Ser41-Ala54) are compared with the reference structure, the backbone RMSD was 0.72 Å. The N-H^N RDCs are missing for Glu8 and Gln9 that define the first two residues of the helix Glu8-Leu17, which probably led to

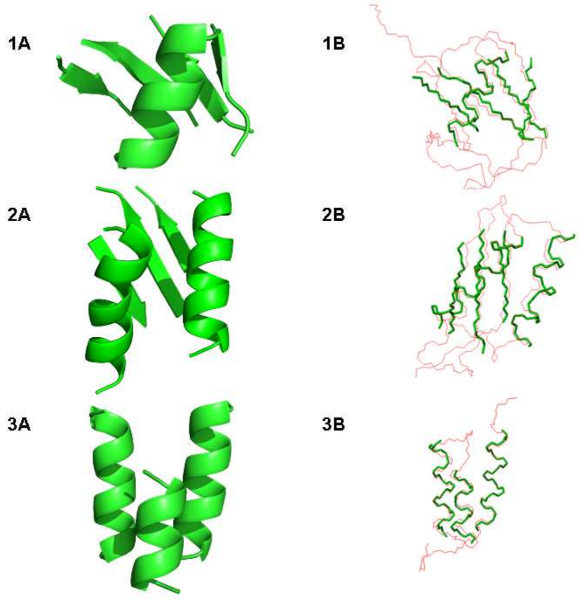

Fig. 5: The global folds of ubiquitin (1A), DinI (2A) and SpA (3A), computed by RDC-ANALYTIC, using C^{α}-C' and N-H^{N} RDCs and a sparse set of NOEs. For ubiquitin and DinI experimental C^{α}-C' and N-H^{N} RDCs are used. For SpA simulated C^{α}-C' and N-H^{N} RDCs are used. (1B) Overlay of the ubiquitin global fold (green) computed by RDC-ANALYTIC with the X-ray structure (red). The backbone RMSD is 1.17 Å. (2B) Overlay of the global fold of DinI (green) computed by RDC-ANALYTIC with the Model 1 (red) of the reference structure (PDB id: 1ghh). The backbone RMSD is 1.11 Å. (3B) The global fold of SpA computed by RDC-ANALYTIC is overlaid on the Model 1(red) of the reference structure (PDB id: 1q2n). The backbone RMSD is 0.83 Å.

somewhat poor conformation for this helix. We then simulated the C^{α}-C' and N-H^{N} RDCs using 1q2n Model 1. Using these simulated RDCs, we computed the global fold of SpA during the third run. When compared with the reference structure, the backbone RMSD was 0.83 Å.

For both ubiquitin and GB3 we applied RDC-ANALYTIC to compute portions (including helices, loops and β-strands) from C^{α}-H^{α} and N-H^{N} RDCs in two media. The results for portions of ubiquitin are reported in Table 2. The resulting overlay of the residues 23-55 of GB3 backbone (red, green, and black) computed by RDC-ANALYTIC with the NMR reference structure (PDB id:1p7e Model 1 [29]) is shown in Figure 6. Since the experimental data for GB3 is not complete (28 out of 132 C^{α}-H^{α} and N-H^{N} RDCs in two media are missing for for residues 23-55), we simulated the missing RDCs using the NMR reference structure. We then computed several consecutive portions of the protein in a divide-and-conquer fashion. The overall backbone RMSD with the reference structure was 1.47 Å.

The above tests on both real NMR data and simulated data demonstrate the capability of RDC-ANALYTIC to determine high-quality backbone fold. As part of our

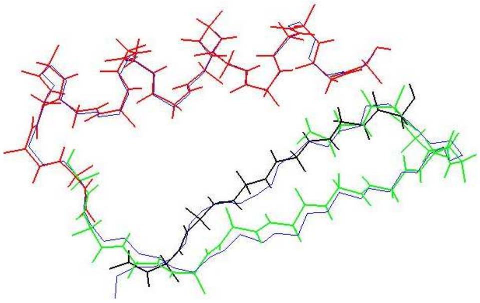

Fig. 6: Overlay of the residues 23-55 of GB3 backbone (red, green, and black) computed by RDC-ANALYTIC with the NMR reference structure PDB id:1p7e Model 1 [29] (blue). Several portions of the protein were computed in a divide-and-conquer fashion. The backbone RMSD of the portions are the following residues: 23-39(red) - 0.35 Å, 39-51(green) - 0.49 Å, 51-55(black) - 0.54 Å. The overall backbone RMSD with the reference structure is 1.47 Å.

future work, we plan to apply our algorithms to determine backbone structures from NMR data collected for larger proteins.

7 Conclusions

We developed algorithms for protein structure determination using residual dipolar couplings collected by solution-state NMR. Our algorithms take into account different aspects of the structure determination process, such as the time and cost of NMR data acquisition, the types of NMR data and their information content, the algorithmic complexity of extracting the 3D structure, and the accuracy of the obtained solution. By using RDCs, we reduce the algorithmic complexity of the structure determination problem to linear time. To reduce the cost of NMR data acquisition, our methods use sparse RDC data, specifically, as little as two RDC measurements per residue in a protein. We develop exact algorithms to compute analytic solutions that optimally fit the NMR data producing high quality structures.

The key to our algorithms is the explicit representations of the RDC equations together with the protein kinematics. Geometrically, this representation results in algebraic curves on a unit sphere that may or may not intersect, depending on the type and number of RDC measurements collected for a single internuclear vector in a residue. Our algorithms find the points on the unit sphere located closest to all of the algebraic curves exactly. These points correspond to the dihedral angles of the protein that optimally fit the RDC data.

We tested our algorithms on NMR data for several proteins: human ubiquitin, DinI, SpA, and GB3. Previous versions of our algorithms [30] (which were not exact for as many types of RDC data as the new algorithms presented in this paper) have been extensively tested on NMR datasets for different proteins, as well as used in a prospective study to solve the structure of the FF Domain 2 of human transcription elongation factor CA150 (PDB id: 2kiq [30]). We plan to do more extensive experimental tests on different NMR data using the algorithms proposed in this paper. We also plan to apply our algorithms to solve other new protein structures in our future work.

Acknowledgements We thank the members of Donald, Zhou and Richardson labs for their assistance and helpful comments. This work is supported by the following grants from National Institutes of Health: R01 GM-65982 to B.R.D. and R01 GM-079376 to P.Z.

References

1. Coggins, B., Venters, R., and Zhou, P. (2010) *Progr NMR Spectr, (in press)*.
2. Donald, B. R. and Martin, J. (2009) *Progr NMR Spectr* **55(2)**, 101–127.
3. Clore, G. M., Gronenborn, A. M., and Tjandra, N. (1998) *J Magnet Res* **131**, 159–162.
4. Güntert, P. (2003) *Progr NMR Spectr* **43**, 105–125.
5. Mumenthaler, C., Güntert, P., Braun, W., and Wüthrich, K. (1997) *J Biomol NMR* **10(4)**, 351–362.
6. Gronwald, W., Moussa, S., Elsner, R., Jung, A., Ganslmeier, B., Trenner, J., Kremer, W., Neidig, K.-P., and Kalbitzer, H. R. (2002) *J Biomol NMR* **23**, 271–287.
7. Kuszewski, J., Schwieters, C. D., Garrett, D. S., Byrd, R. A., Tjandra, N., and Clore, G. M. (2004) *J Am Chem Soc* **126(20)**, 6258–6273.
8. Huang, Y. J., Tejero, R., Powers, R., and Montelione, G. T. (2006) *Proteins: Structure Function and Bioinformatics* **62(3)**, 587–603.
9. Delaglio, F., Kontaxis, G., and Bax, A. (2000) *J Am Chem Soc* **122**, 2142–2143.
10. Andrec, M., Du, P., and Levy, R. M. (2001) *J Biomol NMR* **21(4)**, 335–347.
11. Rienstra, C. M., Tucker-Kellogg, L., Jaroniec, C. P., Hohwy, M., Reif, B., Mcmahon, M. T., Tidor, B., Lozano-Pérez, T., and Griffin, R. G. August 2002 *Proceedings of the National Academy of Sciences of the United States of America* **99(16)**, 10260–10265.
12. Tian, F., Valafar, H., and Prestegard, J. H. (2001) *J Am Chem Soc* **123**, 11791–11796.
13. Tolman, J. R., Flanagan, J. M., Kennedy, M. A., and Prestegard, J. H. (1995) *Proceedings of the National Academy of Sciences USA* **92**, 9279–9283.
14. Tjandra, N. and Bax, A. (1997) *Science* **278**, 1111–1114.
15. Brünger, A. T. (1992) X-PLOR, version 3.1. A system for X-ray crystallography and NMR, Yale University Press, New Haven, CT, .
16. Schwieters, C. D., Kuszewski, J. J., Tjandr, N., and Clore, G. M. (2003) *J Magnet Res* **160**, 65–73.
17. Rohl, C. A. and Baker, D. (2002) *J Am Chem Soc* **124**, 2723–2729.
18. Hus, J.-C., Marion, D., and Blackledge, M. (2001) *J Am Chem Soc* **123**, 1541–1542.
19. Giesen, A., Homans, S., and Brown, J. (2003) *J Biomol NMR* **25**, 63–71.
20. Wang, L. and Donald, B. R. (2004) *J Biomol NMR* **29(3)**, 223–242.
21. Wang, L., Mettu, R. R., and Donald, B. R. (2006) *J Comp Bio* **13(7)**, 1276–1288.
22. Lovell, S. C., Davis, I. W., Arendall, W. B., deBakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C. February 2003 *Proteins: Structure, Function, and Genetics* **50(3)**, 437–450.
23. Saupe, A. (1968) *Ang Chemie* **7(2)**, 97–112.
24. Word, Lovell, S. C., Labean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., Richardson, J. S., and Richardson, D. C. January 1999 *J Mol Bio* **285(4)**, 1711–1733.
25. Losonczi, J. A., Andrec, M., Fischer, M. W. F., and Prestegard, J. H. (1999) *J Magnet Res* **138**, 334–342.
26. Vijay-Kumar, S., Bugg, C. E., and Cook, W. J. (1987) *J Mol Bio* **194**, 531–44.
27. Ramirez, B. E., Voloshin, O. N., Camerini-Otero, R. D., and Bax, A. (2000) *Protein Science* **9**, 2161–2169.
28. Zheng, D., Aramini, J. M., and Montelione, G. T. (2004) *Protein Science* **13**, 549–554.
29. Ulmer, T., Ramirez, B., Delaglio, F., and Bax, A. (2003) *J Am Chem Soc* **125(13)**, 9179–9191.
30. Zeng, J., Boyles, J., Tripathy, C., Wang, L., Yan, A., Zhou, P., and Donald, B. R. (2009) *J Biomol NMR* **45(3)**, 265–281.