Available online at www.sciencedirect.com
ScienceDirect
Journal of Magnetic Resonance 189 (2007) 90–103
JMR Journal of Magnetic Resonance
www.elsevier.com/locate/jmr

Periodicity, planarity, and pixel (3P): A program using the intrinsic residual dipolar coupling periodicity-to-peptide plane correlation and phi/psi angles to derive protein backbone structures

Jinbu Wang [a], Joseph D. Walsh [a], John Kuszewski [b], Yun-Xing Wang [a,*]

[a] *Protein Nucleic Acid Interaction Section, Structural Biophysics Laboratory, NCI-Frederick, NIH, Frederick, MD 21702, USA*
[b] *Imaging Sciences Laboratory, Division of Computational Biology, CIT, NIH, Bethesda, MD 20892, USA*

Received 19 June 2007; revised 21 August 2007
Available online 6 September 2007

Abstract

We present a detailed description of a theory and a program called 3P. "3P" stands for periodicity, planarity, and pixel. The 3P program is based on the intrinsic periodic correlations between residual dipolar couplings (RDCs) and in-plane internuclear vectors, and between RDCs and the orientation of peptide planes relative to an alignment tensor. The program extracts accurate rhombic, axial components of the alignment tensor without explicit coordinates, and discrete peptide plane orientations, which are utilized in combination with readily available phi/psi angles to determine the three-dimensional backbone structures of proteins. The 3P program uses one alignment tensor. We demonstrate the utility and robustness of the program, using both experimental and synthetic data sets, which were added with different levels of noise or were incomplete. The program is interfaced to Xplor-NIH via a "3P" module and is available to the public. The limitations and differences between our program and existing methods are also discussed.
Published by Elsevier Inc.

Keywords: Residual dipolar coupling; RDC wave; Peptide plane; RDC–PP correlation

1. Introduction

Since the end of the last decade, residual dipolar couplings (RDCs) have proved their utility in a number of different areas, including the validation of NMR and X-ray structures [1,2]; direct refinement of structures using distance, torsion angle and RDC restraints [3–7]; direct structure determination with two alignment tensors [8]; and protein folding [8–12]. Since RDCs contain information about the orientation of internuclear bond vectors in relation to the alignment tensor, the RDC constraints are independent from one another.

A caveat in the use of RDCs in the direct refinement protocol is the degeneracy problem [1]. In general, any orientations of a bond vector along a cone about the principle alignment tensor axis and its inversion give rise to the same RDC value [7]. Consequently, the continuum of possible orientations represented by the cones leads to an extremely large number of local minima. These multiple minima make global structure determination by a simulated annealing protocol in which only RDCs are the main source of restraints impractical.

Using peptide planes as structural building units has been reported [13–15]. The solid-state NMR method developed mainly by Opella et al. yields two polar angles, which are not sufficient to specify the orientation of a plane [14]. To overcome this insufficiency, they utilized the Ramachandran energy of phi and psi angles to discern the plane orientations [14]. The approach of Quine and Cross applied an elegant scheme that "connects the dots" of backbone atom positions and used dipolar and ^{15}N chemical shift data to reduce the number of possible orientations of a biplane to four [16]. Mueller et al. used peptide planes as

* Corresponding author. Fax: +1 301 846 6231.
E-mail address: wangyu@ncifcrf.gov (Y.-X. Wang).

1090-7807/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jmr.2007.08.018

a unit to interpret RDC data and to refine an existing NOE-based structure [17,18]. Hus et al. employed the concept of a chiral motif (a peptide plane followed by a tetrahedral center) and two alignment tensors to derive the backbone structure of ubiquitin [8].

Opella's group reported observations of a "dipolar wave," in which dipolar couplings measured in both solid- and solution-state NMR show sinusoidal oscillations when plotted against residue numbers in α-helical peptides [19–22]. Dipolar waves have been used to derive the orientation information of alignment tensor coordinates in relationship to the molecular frame using an empirical fitting function [20–22] or an exact analytical expression (RDC–periodicity correlation) [23,24]. The information that can be obtained from dipolar waves is intriguing. The dipolar waves provide a direct correlation between structures and experimentally measured dipolar couplings or RDCs from respective fully/partially aligned samples. In addition to the RDC–periodicity correlation, the RDCs and bond vectors of a peptide plane are also periodically correlated (RDC–planarity correlation) [25]. Both the RDC periodicity and planarity correlations (RDC–PP correlations) provide restraints to define peptide plane orientations with three angles (Fig. 1): the tilt angle δ; the phase angle ρ of the peptide plane normal vector; and the pitch angle γ of an in-plane bond vector in regular secondary-structure regions [25].

The intuitive periodicity correlation between a regular secondary structure and an RDC vanishes in non-regular secondary-structure regions. Nevertheless, the correlation between specific peptide plane orientations and RDCs can be established with the RDC–PP correlations complemented by readily available phi/psi dihedral torsion angles. The combined use of the RDC–PP correlations and phi/psi angle predictions leads to determination of peptide plane orientations. These oriented peptide planes are subjected to constraints of the covalent peptide bond lineage. In essence, peptide plane orientations $O(\delta,\rho,\gamma)$ are used as the protein's structural "pixels," analogous to the computer graphic pixels (contrast, brightness, and hue), and form the basis for determining backbone structures of proteins.

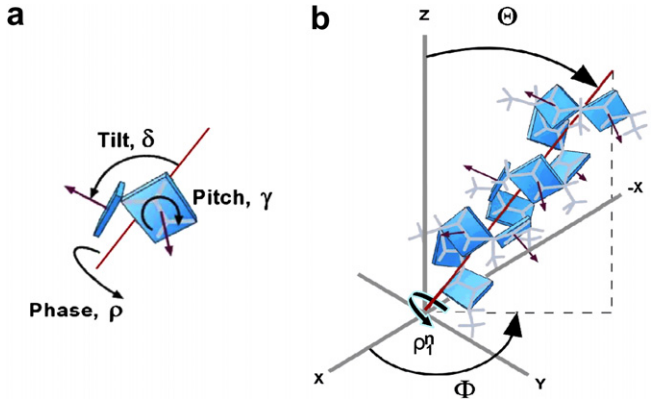

Fig. 1. (a) The definition of the peptide plane orientation tilt (δ^n), phase (ρ^n), and pitch (γ^n). The orientation of the peptide plane normal vector $\hat{n}$ is determined by the tilt and phase in the usual spherical coordinate sense according to $\hat{n} = (\sin\delta^n\cos\rho^n, \sin\delta^n\sin\rho^n, \cos\delta^n)$. The pitch is the clockwise rotation of the bond vector $\hat{r}^{AB}$ about $\hat{n}$, which determines the pitch of a helix at that plane. (b) An α-helix backbone structure defined by consecutive peptide plane orientations. The plane normal vectors, which define the plane orientations, are indicated with red arrows.

2. Theory

2.1. Periodicity and plane orientation in periodical regular secondary structures

Peptide plane normal vectors can be treated as pseudo-bond vectors, whose orientation varies periodically, similar to the peptide plane orientation in regular secondary structures. A correlation exists between the periodic behavior of these normal vectors and the periodicity in the RDC wave. The RDC–PP correlations encompass the intricate correlations among the RDC periodicity, secondary-structure periodicity, and periodicity of the bond vectors within the peptide plane [25]. These correlations can be exploited once the peptide plane normal vectors are expressed in terms of three angles, namely the phase ρ, tilt δ, and pitch γ angles (Fig. 1a). The peptide plane orientation defined with the RDC–PP correlations represents a unifying principle that allows the unambiguous interpretation of RDC data [13–16,26].

Bond vector orientations directly correlate with their RDCs when the orientation of the alignment frame in relationship to the molecular frame is known [23,24]. The extraction of orientational information from the RDCs is made possible by the explicit analytical equation (Eq. (1)) that expresses the RDC, D_{AB}, in relation to the bond vector in the alignment tensor coordinates [24]. When considering structural elements of known types, such as an α-helix, or a duplex in nucleic acids, D_{AB} can be expressed in terms of the bond vector orientation (δ_i, ρ_i) in relationship to a secondary-structure axis that is oriented at angles (Θ,Φ) with respect to the alignment frame [23,24].

$$\begin{aligned} D_{AB,i} = {} & C_1(\Theta,\Phi,\delta_i)\cos 2\rho_i + C_2(\Theta,\Phi,\delta_i)\sin 2\rho_i \\ & + C_3(\Theta,\Phi,\delta_i)\cos\rho_i + C_4(\Theta,\Phi,\delta_i)\sin\rho_i \\ & + C_5(\Theta,\Phi,\delta_i) \end{aligned} \quad (1)$$

Eq. (1) is universally applicable to any periodic structural element. In this equation, $\rho_i = (\rho_1 + 2\pi(i-1)/T)$ is the phase of the bond vector of the ith residue, which is related to the phase of the bond vector of the first residue, ρ_1, and the period $T \approx 3.6$ residues/turn for the α-helix (Fig. 1b) and $T \approx 2$ residues/turn for a β-strand. The slant angle δ_i is the angle the bond vector AB makes with the secondary-structure element axis, and the coefficients $C_k = C_k\ (\Theta,\Phi,\delta_i)$ are functions of both the helical axis orientation and δ_I [24]. We use subscript index i to indicate

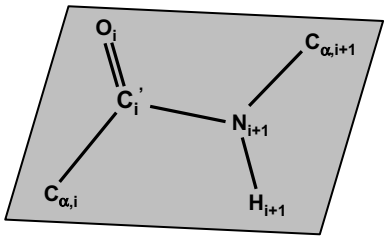

Fig. 2. The relationship between the peptide plane and residue numbering i, $i+1$. All atoms in the figure lie in a plane by the first order approximation, which we identify by the two residue numbers whose backbone atoms form the peptide bond. The residue plane i refers the peptide plane consisting of amino acid motifs, carbonyl group and C_α of residue i, and the amino group and C_α of the residue $i+1$.

the amino acid number and subscript index j to indicate the peptide plane number throughout. The index j on C_α, as in expression $\hat{r}_j^{C_\alpha H_\alpha}$, refers to the bond vector between peptide planes j and $j+1$. The residue plane (rp) number is defined as shown in Fig. 2.

In order to extract a peptide plane orientation in a structural segment from RDCs, one first needs to consider the geometric relationships among the bond vectors within the plane. For example, if the jth peptide plane orientation is expressed by its normal vector $\hat{n}_j$, then the normalized HN bond vector $\hat{r}_j^{HN}$ (defined $H \to N$) is related to the normalized $C_\alpha C'$ bond vector $\hat{r}_j^{C_\alpha C'}$ (defined $C_\alpha \to C'$) by a rotation of $\eta = 56.5°$ about $\hat{n}_j$, using the right-hand rule. The angular relation η between these two bond vectors (as well as the analogous relationships between $\hat{r}_j^{C_\alpha C'}$, $\hat{r}_j^{NC'}$ (116.9°), and $\hat{r}_j^{NH}$, $\hat{r}_j^{C_\alpha N}$ (119.5°)) is a property of covalent peptide geometry [27]. When the alignment tensor Z-axis coincides with the peptide plane normal vector, the transformation between the two bond vectors $\hat{r}_j^{C_\alpha C'}$ and $\hat{r}_j^{HN}$ is given by simply:

$$\hat{r}_j^{HN} = R_{\hat{n}}(\eta)\hat{r}_j^{C_\alpha C'} = R_{\widehat{Z}}(\eta)\hat{r}_j^{C_\alpha C'} \tag{2}$$

When the peptide plane normal does not coincide with the Z-axis, the rotation about the peptide plane normal is defined by Eq. (3):

$$R_{\hat{n}_i}(\eta) = R^{-1}_{\widehat{X}}(\alpha)R^{-1}_{\widehat{-Y}}(\beta)R_{\widehat{Z}}(\eta)R_{\widehat{-Y}}(\beta)R_{\widehat{X}}(\alpha) \tag{3}$$

The rotation matrices $R_{\widehat{X}}(\alpha)$, $R_{\widehat{-Y}}(\beta)$, and $R_{\widehat{Z}}(\eta)$ needed to express $R_{\hat{n}_i}(\eta)$ and illustration figures are shown in Supplementary material.

The geometric relationship between bond vectors lying in the peptide plane, together with Eq. (3), can be used to express the RDCs as a function of the peptide plane normal vector. Any peptide plane bond vector AB in a structure element aligned along the Z-axis can be expressed in terms of the tilt angle δ^n and phase angle ρ^n of the peptide plane normal vector, and the pitch angle γ^n of the bond vector in the peptide plane as shown in Fig. 1a (for clarity, the peptide plane subscript j has been omitted from δ^n, ρ^n, and γ^n in the equations that follow).

$$\begin{aligned}\hat{r}_j^{AB,sec.str.} &= R_Z(\rho^n)R_n^{-1}(\gamma^n)R_Y(\delta^n - \pi/2)\widehat{Z} \\ &= R_Z(\rho^n)[R_Y(\delta^n)R_Z(\gamma^n)R_{-Y}(\delta^n)]^{-1} \\ &\quad \times R_Y(\delta^n - \pi/2)\widehat{Z}\end{aligned} \tag{4}$$

In Eq. (4), we have used Eq. (3) to express $R_n(\eta)$ in terms of Cartesian rotations. The initial rotation about $\widehat{X}$ from Eq. (3) is equal to the identity matrix, since $R_Y(\delta^n - \pi/2)\widehat{Z}$ is already in the XZ-plane. Multiplying out the matrices in Eq. (4) yields an expression for any bond vector AB in terms of the peptide plane orientation angles δ^n, ρ^n, and γ^n:

$$\hat{r}_j^{AB,sec.str.} = \begin{pmatrix} -\cos\delta^n\cos\rho^n\cos\gamma^n - \sin\rho^n\sin\gamma^n \\ -\cos\delta^n\sin\rho^n\cos\gamma^n + \cos\rho^n\sin\gamma^n \\ \sin\delta^n\cos\gamma^n \end{pmatrix} \tag{5}$$

For a structure element referenced relative to (Θ, Φ), which can be an arbitrary axis, or a principle alignment axis, the bond vector orientation $\hat{r}_j^{AB}$ with respect to the alignment frame is given by:

$$\hat{r}_j^{AB} = R_Z(\Phi)R_Y(\Theta)\hat{r}_j^{AB,sec.str.} \tag{6}$$

By combining this expression for the bond vector with the general RDC Eq. (1), we obtain an expression for the RDCs in terms of the orientation of the peptide plane, tilt δ^n, phase ρ^n, and pitch γ^n:

$$\begin{aligned}&D_j^{AB}(\Theta,\Phi;\delta^n,\rho^n,\gamma^n) \\ &= D_a\{(-1+3R/2)[\sin\Phi(\cos\delta^n\sin\rho^n\cos\gamma^n - \cos\rho^n\sin\gamma^n) \\ &\quad - \cos\Theta\cos\Phi(\cos\delta^n\cos\rho^n\cos\gamma^n + \sin\rho^n\sin\gamma^n) \\ &\quad + \sin\Theta\cos\Phi\sin\delta^n\cos\gamma^n]^2 - (1+3R/2) \\ &\quad \times [-\cos\Phi(\cos\delta^n\sin\rho^n\cos\gamma^n - \cos\rho^n\sin\gamma^n) \\ &\quad - \cos\Theta\sin\Phi(\cos\delta^n\cos\rho^n\cos\gamma^n + \sin\rho^n\sin\gamma^n) \\ &\quad + \sin\Theta\sin\Phi\sin\delta^n\cos\gamma^n]^2 + 2[\cos\Theta\sin\delta^n\cos\gamma^n \\ &\quad + \sin\Theta(\cos\delta^n\cos\rho^n\cos\gamma^n + \sin\rho^n\sin\gamma^n)]^2\}\end{aligned} \tag{7}$$

Eq. (7) correlates the RDCs of bond vectors in a given peptide plane, as well as sequential peptide plane RDCs, to their plane orientations. Therefore, this equation makes it possible to extract plane orientations in the alignment tensor axis system from RDCs of coupled nuclei in a peptide plane.

2.2. Tetrahedral centers

The tetrahedral configuration around C_α represents well-defined chemical geometry. The geometry of the tetrahedral center, and thus its angular relationship to the peptide planes, is well conserved throughout a protein structure. In the case of a 0.78 Å resolution X-ray crystal structure of the 26.7 kDa subtilisin of *Bacillus lentus* (Accession code: 1GCI) [28], the mean angle $T_{NC_\alpha C'}$ between N, C_α, and C' yields a narrow distribution, with standard deviation (STD) of only 2.4°. Thus, we used this well-conserved geometry to correlate the RDCs associated with the tetrahedral center C_α to the flanking peptide planes, and applied it to resolve the ambiguity in plane

orientations due to the possible multiple minima from the grid search, and to verify and refine the plane orientations. In the actual implementation of the 3P method, the angle of the tetrahedral geometry is set to be $109 \pm 8°$ to accommodating possible larger deviation form the norm and experimental errors in RDC measurements.

2.3. Fourfold degeneracy

A peptide plane followed by a tetrahedral center constitutes a chiral structure element [8,24]. Interpreting RDCs in the frame of a chiral structural element greatly reduces the degeneracy to the four subsets of orientations for each peptide plane. These four possible orientations in terms of (δ, ρ, γ) can readily be computed by searching the minima in the full (δ, ρ, γ) space using the following equation:

$$\mathrm{RMSD} = \sum_{j=1}^{n} \frac{\sqrt{\sum_{i=1}^{m} f_i (D_i^j(\mathrm{exp}) - D_i^j(\mathrm{calc}))^2 / m}}{n} \tag{8}$$

where m and n are the numbers of different types of one-bond RDCs and the number of residue planes, respectively; f_i is the normalizing factor with respect to the NH RDC; $D_i^j(\mathrm{exp})$ and $D_i^j(\mathrm{calc})$ are the experimental RDC and the RDC calculated using Eq. (7), respectively. These four possible orientations are (Θ, Φ), $(\pi - \Theta, 2\pi - \sigma)$, $(\pi - \Theta, \pi - \Phi)$, and $(\Theta, \pi + \Phi)$ [24]. Fig. 3 illustrates the four possible orientations of residue plan 29 of ubiquitin in the (δ, ρ, γ) space using experimental data. Furthermore, the orientations of the proceeding and following planes can be selected based on the tetrahedral angle centered at C_α. Therefore, theoretically, a peptide backbone structure can be calculated, given a complete set of RDCs, three in-plane and one tetrahedral, from only one alignment tensor. The detailed correlation expression in the vector space is given in Supplementary material.

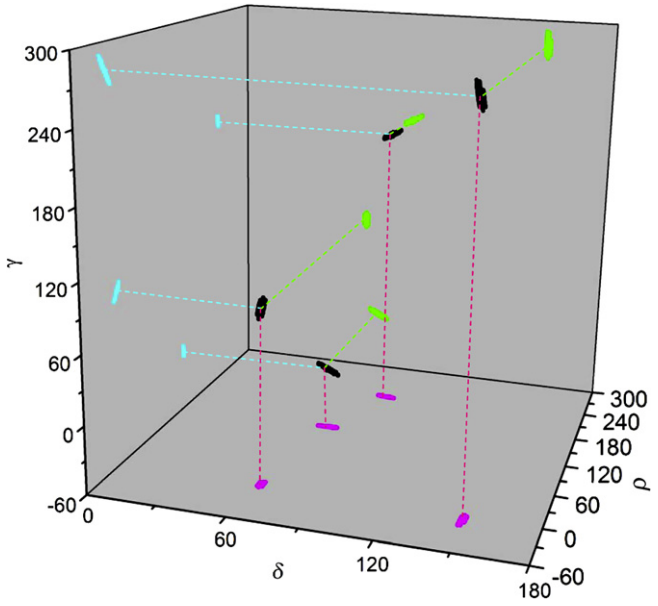

Fig. 3. An illustration of four possible rp orientations that satisfy both the in-plane (NH, $C_\alpha C'$, NC) and the tetrahedral ($C_\alpha H_\alpha$) RDCs of the four possible orientations of rp 29 of ubiquitin are drawn in (δ, ρ, γ). These possibilities were computed using Eq. (8) and experimental RDCs in the full (δ, ρ, γ). For any one of the four orientations, the orientations of the proceeding and the following peptide plane can be easily identified by the tetrahedral angle. For clarity, the four possible orientations are plotted in black and their projections on three planes are plotted in light green, cyan, and magenta.

2.4. phi/psi angle supplements

Although in theory the in-plane and tetrahedral RDCs are sufficient to define four discrete orientations of a peptide plane, in practice, the four subsets of peptide plane orientations may not be so distinctly defined due to shallow minimum and noisy data. Instead of using the second set of RDCs from a non-correlated tensor, we resort to predicted phi/psi angles, which are readily available once backbone assignments are completed [29]. Combining tensors from solid-state NMR measurements with phi/psi restraints from databases was pioneered by Opella's group [14]. Restrained by phi/psi angles, a consistent orientation between the two can usually be identified. Using phi/psi torsion-angle restraints as an aid for determining backbone structures has an obvious advantage over other methods that may require using of multi-alignment tensors.

2.5. Determining D_a and R values

The magnitude D_a and rhombicity R values of an alignment tensor are prerequisites for interpreting RDC data. In addition, the accuracy of D_a and R values in interpreting and utilizing RDCs is critical in a structure determination when RDC data is the main source of restraints. Several methods for deriving D_a and R values from RDC data have been reported. Trial-and-error approaches that employ simulated-annealing [30] and the singular value decomposition (SVD) [31] methods were the first two methods used to extract D_a and R values from RDC data, but both methods required pre-existing structural coordinates. The extended histogram method (EHM) [32] and the maximum likelihood method (MLM) [33] are two alternatives that do not require pre-existing, three-dimensional coordinates, but the former is less accurate and the latter has limited accuracy when dealing with RDCs of anisotropically distributed spin pairs. The 3P method uses the RDC–PP correlations in the regions with high information content (IC), usually regular secondary-structure elements, to derive D_a and R values for a protein under the rigid-body assumption. This method becomes feasible because, in the region with both high RDC and phi/psi IC, or in a periodical structure element like an α-helix, the correlation between RDC and the orientation of spin pairs in the peptide planes is defined explicitly, eliminating a large number of false minima that would otherwise be encountered when the

correlation between spin-pair orientation and RDC is considered in a non-concerted fashion.

3. Methods

3.1. Procedures

The implementation and overall outline of the 3P program are illustrated in Fig. 4. The 3P program was coded in Python and C/C++, and run either in a single CPU mode or on a multiple-CPU Linux cluster (Dell Poweredge 2650), using a parallel computing mechanism (Message Passing Interface (MPI)). The running script, written in Python, is less than 100 lines long, and a sample script is provided in Supplementary material.

To extract the peptide plane orientations, we used Levmar, Levenberg–Marquardt nonlinear least-squares algorithms with bound constraints in C/C++ (http://www.ics.forth.gr/~lourakis/levmar) to best fit RDC data to Eq. (7). Levmar is the GPL native ANSI C implementation of the Levenberg–Marquardt optimization algorithm, available also in C++. Both unconstrained and constrained (under linear equations and box constraints) Levenberg–Marquardt variants are included. The Levenberg–Marquardt (LM) algorithm is an iterative technique that finds a local minimum of a function that is expressed as the sum of squares of nonlinear functions. It has become a standard technique for nonlinear least-squares problems and can be thought of as a combination of steepest descent and the Gauss–Newton method. Moreover, Levmar is the most compatible with Python language. Overall, the calculation is accomplished in several steps, discussed in the modules that follow. For all fittings, the calculations start with randomly generated initial plane orientations.

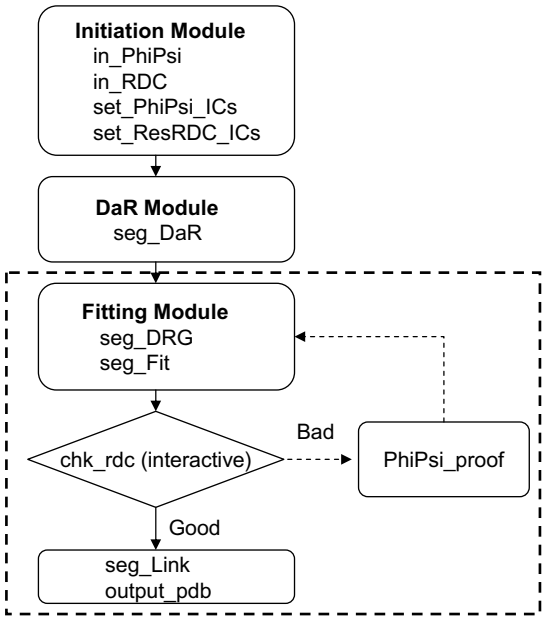

Fig. 4. The flow chart of the 3P program. 3P consists of three modules: initiation, DaR, and fitting. The initiation module mainly consists of four functions, two for handling input RDC data and phi/psi angles, in_RDC and in_Phi/Psi, respectively, and two, set_ResRDC_ICs and set_PhiPsi_ICs, for registry of RDC and Phi/Psi ICs for individual residues, respectively. The DaR module calls function seg_DaR to calculate the values of the axial and rhombic components, D_a and R, of the alignment tensor based on the segments with high ICs. The third module, the fitting module, which constitutes the core of the 3P program, has five main functions: seg_DRG estimates anchor plane orientation (D stands for δ, R for ρ, G for γ); the seg_fit function performs calculations of peptide plan orientations for segments. In the case where raw phi/psi angles from the output of TALOS are directly used for the fit, the program proceeds with the Phi/Psi_proof procedure to detect and correct erroneous torsion angle predictions based on RMSDs in RDC of fit segments. Currently, the Phi/Psi_proof is written as a standing alone module and is used interactively by a user to perform the phi/psi error proof (Supporting materials). In the case where phi/psi angles are error proofed, the 3P program takes the input RDC and phi/psi angles to generate backbones of segments, which are linked by the seq_link function and the final backbone structures are written in PDB format by the output_pdb function, and the whole calculations are fully automated.

3.2. The initiation module

This module takes RDC and phi/psi torsion angles from TALOS or other sources, if available, as inputs to calculate the Information Content IC for each peptide plane. IC is used for both dividing the peptide chain into segments and defining anchor planes and anchor segments (see next section). Five one-bond RDCs per amino acid ($^1D_{NH}$, $^1D_{C_\alpha C'}$, $^1D_{NC'}$, $^1D_{C_\alpha H_\alpha}$, and $^1D_{C_\alpha C_\beta}$) are usually measured, and each one is counted as an IC score of 0.2. An RDC IC of 1.0 is assigned if all five RDCs are available for a given plane. The phi/psi IC is assigned based on the standard deviation of the torsion angles predicted by the TALOS output. For STDs of both phi/psi that are 10°, 20°, 30°, 40°, 50°, and 60°, the phi/psi ICs of 1.0, 0.8, 0.6, 0.4, 0.2, and 0.0, respectively, are assigned. Therefore, the IC value is an indicator of how much information is available for determining the orientation of each peptide plane along a peptide chain.

Two types of segments are used in the 3P program, the anchor segment and non-anchor segment. Anchor segments are usually a stretch of covalently connected peptide planes with the highest overall ICs and the anchor plan is the plan with the highest IC values. They break off at residues with low ICs. The lengths of these anchor segments may vary from 10 to 30 peptide planes. The anchor segment is usually in a regular secondary-structure region, where RDCs are more easily measured and phi/psi are better predicted. Accurate D_a and R values of the overall protein, with a rigid-body assumption, can be derived readily from the anchor segment and used for the subsequent calculation of the protein backbone structure (see the following section). The anchor plane can also be derived readily from the anchor segment. Further, the two anchor planes that have the highest ICs in a segment, are assigned for each non-anchor segment. Since anchor planes have the highest ICs, they serve as a starting point and a checkpoint to verify chain propagation in the calculation of a backbone conformation. The orientations of the anchor planes

serve as fixed starting points from which the orientations of the preceding and following planes are calculated sequentially to simultaneously satisfy periodicity-planarity correlations expressed in Eq. (7) and the phi/psi torsion-angle restraints. The orientations of the anchor/non-anchor planes are determined using the nonlinear least-squares algorithm with bound constraints, starting from randomly generated initial orientations to best fit the RDC data. The conformations of segments are derived so that the RMSD in the RDC for anchor planes, as well as the overall RMSD differences between the experimental and the back-calculated RDC, are comparable to the experimental errors.

3.3. $D_a R$ module

The D_a, and R values are most accurately determined by using an empirical weighting of the RDC data in the alignment fit, in which the ${}^1D_{C_\alpha C'}$ data is scaled by a factor of 2.5, ${}^1D_{NC'}$ data by a factor of 4.2, and ${}^1D_{C_\alpha C_\beta}$ by a factor of 0.3 (or, alternately, ${}^1D_{C_\alpha H_\alpha}$ by a factor of 0.1), relative to ${}^1D_{NH}$. The empirical values originate from the relative gyromagnetic ratios, internuleic distances, and the relative expected measurement errors. We find the accuracy of our results comparable to those using published methods, such as EHM and MLM [32,33]. This module can be bypassed if D_a and R values are already accurately known.

To extract D_a and R values, the module randomly chooses a maximum of three segments of the peptide chain with the highest ICs that will best fit to the RDC data. For each segment, this module yields a set of D_a and R values starting from a number of randomly generated initial orientations for the anchor planes. The number of the randomly generated initial orientations chosen is a compromise between the best RMSD in the RDC and the highest computing speed, and the default number is 20. Among the 20 sets of D_a and R, only the three sets, which give the lowest RMSD in the RDC, comparable to the measurement error, are chosen. In total, nine sets of D_a and R values are extracted from the three segments, and the mean values of the three sets with the lowest RMSD in the RDC are taken as the D_a and R values.

3.4. The fitting module

After D_a and R are obtained, this module determines the backbone conformations of segments and links the segments together to derive the overall backbone structures of the proteins. The first step in this process is to determine the orientations $(\delta^n, \rho^n, \gamma^n)$ of the anchor peptide planes, where n is the number of anchor planes, for each segment by fitting Eq. (7) to RDCs with restraints of phi/psi angles and the fixed D_a and R. Out of 20 calculations started from randomly generated initial orientations, three possible plane orientations, $O^n(\delta^n, \rho^n, \gamma^n)$, for each anchor plane are accepted based on the RMSD comparable to measurement errors. These anchor plane orientations are used simultaneously as initial planes to generate preceding/subsequent plane orientations by searching for minimums in RMSDs in RDC in $O^n(\delta^n, \rho^n, \gamma^n)$ space. There are four possible plane orientations that satisfy the both the in-plane and the tetrahedral RDCs (Fig. 3). For each segment, the second anchor plane with the higher residue number index is used as a check to verify the conformation of the segment, by comparing it with that of the same plane in the segment using criteria comparable to experimental errors. The calculation accepts three conformers for each segment that give the best RMSD in the RDC. All three conformers are then used to pair with those possible conformers of the preceding/subsequent segments in the next round of calculations.

In the next step, the segments are linked together to build the protein backbone structure. Because there are, by default, three accepted conformers for each segment, calculations have to be performed for each possible combination that may lead to one of the 3^N structures, where N is the number of fitting segments. For cutinase, which is divided into nine segments and used as a test case in the following section, the calculation would have yielded $3^9 = 19{,}683$ possible backbone structures. To speed up the calculation, the 3P program randomly pairs the accepted segments that are long enough to generate a set of the backbones with the best RMSDs in the RDC. The strict requirements for the normal covalent chemical geometry at the joint tetrahedral C_α and a global fit of possible backbone structures to the RDC–PP correlation and phi/psi restrictions eliminate the fourfold degeneracy associated with a chiral motif of a known structure [24]. Specifically, the program ensures a good tetrahedral angle, $T_{NC_\alpha C'}$, and the best fit to the experimental ${}^1D_{C_\alpha H_\alpha}$ (see Section 2) at a joint between two segments (Fig. 5). The 3P program uses the algorithm we refer to as Soft Link to link two continuous segments by imposing sets of the angles, and dihedral and tetrahedral restrictions of the linker regions on the fitting to RDC data. By default, the length of a linker region spans six residues, three on each side of a joint, but it may be adjusted by the user. The length of a linker region may be adjusted by users in the running script in the program. The default of six is chosen for the balance between the highest computing speed and the best fit. Finally, the backbone structures generated from 3P are then regularized using a "3P" module interfaced with Xplor-NIH to add side-chain atoms and remove possible van der Waals violations [34].

4. Results

4.1. Determination of D_a and R

We tested the robustness of the program for determining D_a and R values using simulated sets of data for the following scenarios: three realistic noise levels embedded in the simulated data, which are also compounded with different rhombic component values from 0.1 to 0.67. The RDC and phi/psi angle data were generated based on the

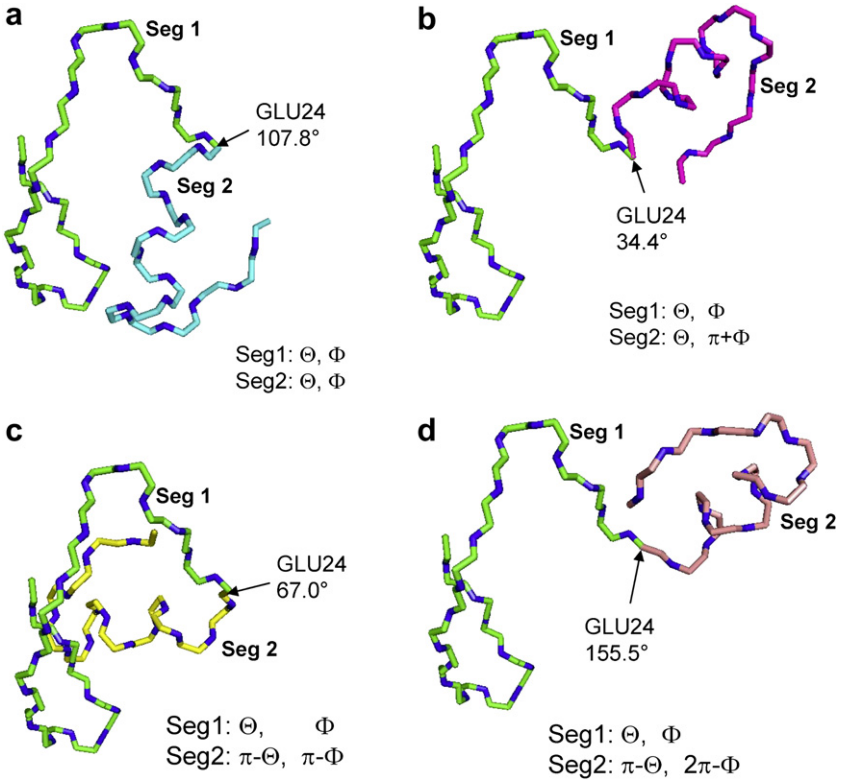

Fig. 5. An illustration of connecting two segments. There are four possible subsets of orientations for each segments, drawn in C′, N, and C_α of the backbone for clarity: (Θ,Φ), $(\Theta,\pi+\Phi)$, $(\Theta,-\Phi)$, $(\pi-\Theta,2\pi-\Phi)$, where (Θ,Φ) is an arbitrary reference in a spherical coordinate axis. When one randomly selects a segment, for example, Seg 1 with the orientation of (Θ,Φ), there are another four possible orientations for the following segment, Seg 2. The strict requirement of the co-valent tetrahedral geometry at C_α, $\angle NC_\alpha C'$, $109 \pm 8°$, eliminates other three possibilities in (b), (c), and (d) by the program.

cutinase (Accession code: 1CEX) X-ray structure. The results are tabulated in Table 1. To add noise to the data, we used both the Gaussian and random distributions with the distribution tails cut off at $3\sigma^{AB}$ and σ^{AB}, respectively. For a known α-helix element in a protein structure, accurate D_a and R values can be accurately derived with the restraint of the RDC–periodicity correlation alone [24].

4.2. Backbone structures of non-regular secondary motifs

Monte-Carlo simulations of the effect of errors in both RDC and phi/psi angles superimposed on the structure were performed by generating noisy data sets of different degrees and with various error ranges for phi/psi torsion angles for a short peptide that consists of seven peptide planes. We tested scenarios using φ/ψ angle error ranges of ±45°, $D_a = -10$, with $R = 1/3$, using what we refer to here as standard errors (HN, $C_\alpha C'$, NC′, and $C_\alpha H_\alpha$ RDCs of 0.5, 0.25, 0.25, and 1.0 Hz, respectively). Out of the 100 calculate structures using the noisy dataset, the 77 accepted structures gave a backbone RMSD of 0.18 Å to the original structure based on which the noisy dataset was generated; increasing all errors by a factor 1.5 resulted in 74 accepted structures with RMSD of 0.25 Å; increasing the errors by a factor of 2 yielded 34 accepted structures with

Table 1
Calculate D_a and R values

R (target)	D_a	R
Added error: $\sigma^{HN} = 0.50$ *Hz,* $\sigma^{C_\alpha C} = 0.25$ Hz, $\sigma^{NC} = 0.25$ *Hz,* $\sigma^{C_\alpha C_\beta} = 0.25$ Hz		
0.00	-14.982 ± 0.083	0.011 ± 0.007
0.10	-14.996 ± 0.081	0.100 ± 0.013
0.30	-15.030 ± 0.079	0.299 ± 0.009
0.50	-15.038 ± 0.103	0.498 ± 0.010
2/3	-15.054 ± 0.095	0.660 ± 0.005
Added error: $\sigma^{HN} = 0.75$ *Hz,* $\sigma^{C_\alpha C} = 0.38$ Hz, $\sigma^{NC} = 0.38$ *Hz,* $\sigma^{C_\alpha C_\beta} = 0.38$ Hz		
0.00	-14.980 ± 0.114	0.018 ± 0.011
0.10	-15.013 ± 0.134	0.102 ± 0.017
0.30	-15.037 ± 0.128	0.298 ± 0.017
0.50	-15.066 ± 0.125	0.493 ± 0.019
2/3	-15.056 ± 0.132	0.656 ± 0.014
Added error: $\sigma^{HN} = 1.0$ *Hz,* $\sigma^{C_\alpha C} = 0.5$ Hz, $\sigma^{NC} = 0.5$ *Hz,* $\sigma^{C_\alpha C_\beta} = 0.5$ Hz		
0.00	-15.001 ± 0.182	0.022 ± 0.010
0.10	-15.026 ± 0.161	0.102 ± 0.016
0.30	-15.041 ± 0.163	0.304 ± 0.018
0.50	-15.097 ± 0.168	0.497 ± 0.023
2/3	-15.080 ± 0.178	0.650 ± 0.021

Note: The cutinase protein alignment parameters from a Monte-Carlo simulation of 20 noisy RDC data sets with 20% missing RDC for each scenario, generated from a target alignment of $D_a = -15.0$ with a number of different target rhombicities. Gaussian distributions of errors with tail cutoff at $3\sigma^{AB}$ were added to the simulated data.

an RMSD of 0.28 Å. Therefore, most of the increase in the structural RMSD resulted from increasing the standard errors by a factor of 1.5, with a smaller increase occurring beyond 1.5, although the convergence rate dropped from ~3/4 to ~1/3.

Monte-Carlo simulations of the effect of the rhombicity on the accuracy of the fit non-periodic backbone structure were performed by generating 100 noisy data sets (using the standard errors) allowing phi/psi angle error ranges of ±45° in the fit, $D_a = -10$, with $R = 2/3$, 1/3, and 0. These simulations were repeated for a number of different orientations in the alignment tensor, and the results are shown in Table 2. The effect of the rhombicity on the RMSD of the calculated structures to the original structure is small, with RMSDs = 0.16 Å and 0.17 Å for $R = 2/3$ and $R = 1/3$, respectively, but the effect on convergence is noticeable. Only in the axially symmetrical case with $R = 0$, where the angular IC of the RDCs is less than those in axially non-symmetrical cases, does the RMSD increase appreciably, to 0.25 Å. These calculations illustrate that, with a complete set of RDC data measured in one medium and phi/psi torsion angle restraints, non-regular secondary backbone structures of small fragments can be determined with reasonable accuracy.

4.3. Effect of errors in RDC and phi/psi on the accuracy of backbone structures of proteins

We have tested the 3P program using simulated RDC data based on whole proteins. To test the robustness of the program, we chose to generate simulated sets of data, using various scenarios and the cutinase X-ray crystal structure. We tested the cases with various levels of errors in both RDC and phi/psi. To the RDC data we added the Gaussian and flat random errors of 5% and 10%, with the distribution tails cut off at $3\sigma^{AB}$ and σ^{AB}, respectively. The noisy RDC data were further compounded with loosely defined phi/psi angle ranges, which were set to ±15°, ±30°, ±45°, and ±60°. In each case, out of 200 structures, we chose 10 with the best RMSD in RDC for tallying the statistics. We tested the utility of the program with four Gaussian and random noisy levels, 5%, 10%, 20%, and 30% of the RDC data sets. The backbone RMSDs ranged from 0.93 Å with 5% randomly distributed errors (Figs. 6 and 7) to about 3.0 Å with 30% Gaussian errors, while phi/psi error ranges remain ±30° (Fig. 7), and these RMSDs increased to 1.85 Å and 3.70 Å, respectively, when phi/psi error ranges were set to ±60° (Fig. 7). It appears that the phi/psi angles can be loosely restrained up to ±45° of the error range without dramatically increasing the RMSDs of the calculated structures relative to the original ones (Fig. 6).

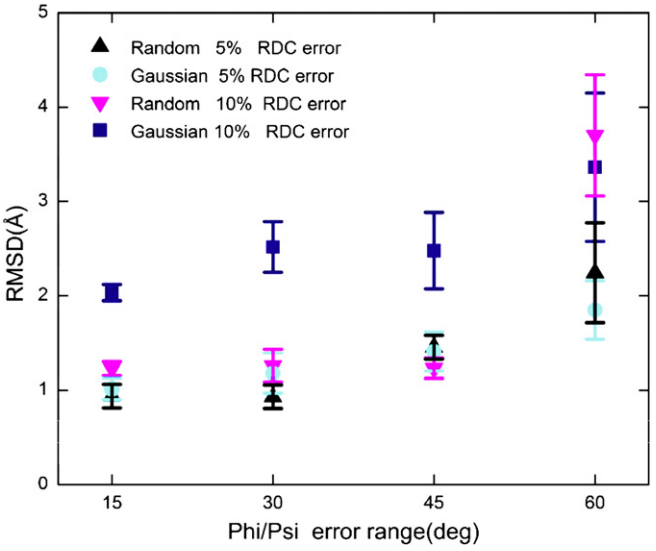

Fig. 6. The effect of the phi/psi error ranges with different noisy RDC data sets on the RMSDs of the cutinase protein backbone structures relative to that of the original X-ray crystal structure (Accession code: 1CEX). The Gaussian error cutoffs tail at $3\sigma^{AB}$, whereas the randomly error cutoffs tail at $1\sigma^{AB}$. The error bars represent the standard deviation of RMSDs relative to the original structure among the 10 best structures (see the text).

Table 2
Backbone accuracy and convergence[a]

R	Error (HN) (Hz)	RDCs used	#Accepted/#total	RMSD[b]
1/3	0.5	HN, $C_\alpha C'$, NC′, $C_\alpha H_\alpha$	77/100	0.16 Å
1/3	0.75	HN, $C_\alpha C'$, NC′, $C_\alpha H_\alpha$	74/100	0.25 Å
1/3	1.0	HN, $C_\alpha C'$, NC′, $C_\alpha H_\alpha$	34/100	0.28 Å
0	0.5	HN, $C_\alpha C'$, NC′, $C_\alpha H_\alpha$	72/100	0.25 Å
1/3	0.5	HN, $C_\alpha C'$, NC′, $C_\alpha H_\alpha$	77/100	0.16 Å
2/3	0.5	HN, $C_\alpha C'$, NC′, $C_\alpha H_\alpha$	94/100	0.17 Å
1/3	0.5	HN, $C_\alpha C'$, $C_\alpha H_\alpha$	36/50	0.19 Å (0.16 Å)[c]
1/3	0.5	$C_\alpha C'$, NC′, $C_\alpha H_\alpha$	48/50	0.24 Å (0.23 Å)
1/3	0.5	HN, NC′, $C_\alpha H_\alpha$	37/50	0.29 Å (0.27 Å)
1/3	0.5	HN, $C_\alpha C'$, NC′	44/50	0.36 Å (0.30 Å)

[a] The table show the number of accepted of fit structures from Monte-Carlo simulations of a 7-peptide plane, non-periodic backbone segment (see text). The magnitude of superimposed RDC errors used is indicated by the HN RDC error, and errors in the other three types of RDCs were scaled accordingly. We tested the effects of rhombicity, errors, and types of available RDCs on the convergence and RMSDs relative to the target structure.

[b] RMSD of accepted structures (RMSD of 10 best RDC-fit structures).

[c] RMSD after removing outlier from the bundle. With outlier, RMSD = 0.24 Å.

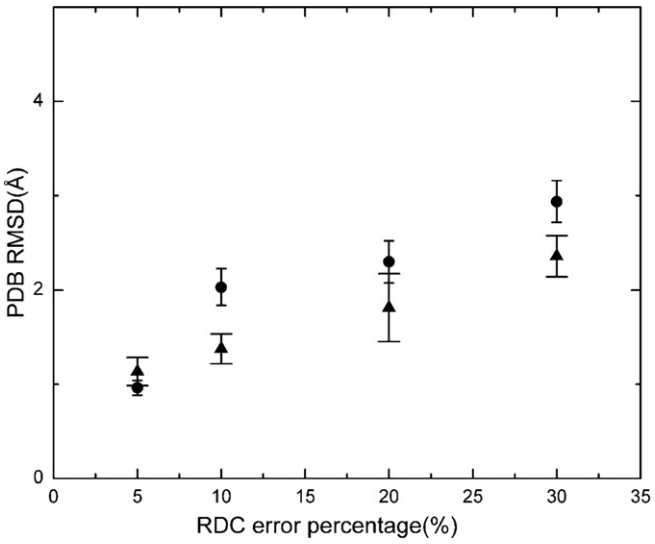

Fig. 7. The effect of superimposed errors in the RDC on the RMSDs of the cutinase protein backbone structures relative to that of the orignial X-ray crystal structure. The simulated RDC data were added following different levels of noise, simulated RDC data superimposed with a Gaussian RDC error with the cutoff at $3\sigma^{AB}$; 7 simulated RDC data superimposed with a random RDC error with the cutoff at σ^{AB}. In all calculations, the error ranges for phi/psi angles were kept at ±30.

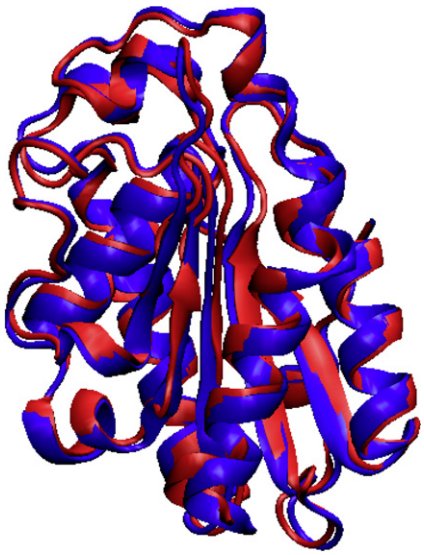

Fig. 8. Structure comparison between the 3P-calculated non-regularized structure (red) and the X-ray crystal structure (blue, Accession code: 1CEX) of cutinase with the backbone RMSD = 0.92 Å. The simulated RDC data (HN, NC, $C'C_\alpha$, $C_\alpha H_\alpha$, $C_\alpha C_\beta$) were superimposed with 0.5 Hz Gaussian error with cutoff tailed at $3\sigma^{AB} = 1.5$ Hz, and 20% of the RDC were randomly removed.

When the error ranges for phi/psi were set to ±90°, the program yields structures with the backbone RMSD greater than 19 Å, and the backbone structure determination becomes impractical. In practice, the standard deviations of phi/psi predictions from TALOS in well-defined regions are well under 30°, whereas the STDs of those in regions labeled as "new" by TALOS are well below 60°. The superimposed ribbon diagrams of the cutinase structures are shown in Fig. 8. It takes about 300 s to calculate 200 backbone structures of cutinase on a Dell Precision 670, a dual CPU Linux computer.

4.4. Calculating backbone structures with random incomplete RDC data sets

We first tested the effect of incomplete RDC data sets using fragments. Using Monte-Carlo simulation, we estimated the effect of incomplete RDC data sets by randomly removing 4 of the 28 RDCs (14%) and superimposing errors to generate 50 noisy data sets. This process was then repeated 10 times, using $R = 1/3$ and phi/psi angle ranges of ±45°. The 10 best fits to the RDCs from each run showed RMSDs of 0.15–0.27 Å to the original structure. Similarly, the removal of four RDCs, all belonging to one particular peptide plane, yielded different RMSDs, depending on the location of the peptide plane in the sequence: If it was at the beginning of the sequence, the RMSD was 0.16 Å; if it was at the middle, the RMSD was 0.27 Å.

We tested the relative importance of tetrahedral versus in-plane RDC data. Using $R = 1/3$, phi/psi angle ranges of ±45°, and the standard errors, a Monte-Carlo simulation was performed on 50 runs with all tetrahedral RDCs removed. The 10 best fits to the RDCs had an RMSD to the original structure of 0.30 Å, which is larger than the RMSDs of 0.27, 0.24, and 0.16 Å for the cases in which all HN, $C_\alpha C'$, and NC′ RDCs, respectively, were removed individually. The increase in the RMSD is largest and similar for $C_\alpha H_\alpha$ and HN; therefore, no a priori distinction appears to exist between in-plane and tetrahedral RDCs for this fitting program since these two RDCs possess almost identical relative errors (ratio of random error to D_a). The effect of omitting the $C_\alpha C'$ and NC′ RDCs is progressively weaker, which were weighed according to the scaling factors and practical errors, and therefore lesser angular IC. Finally, in the case in which seven RDCs are removed randomly from among all the RDCs, the RMSDs ranged from 0.12 to 0.31 Å, for the 10 random selections made.

It is noteworthy that the RDC data measured in a single medium for three well-measured proteins, ubiquitin, GB3, and DinI, are about 14–18% incomplete. We then tested incomplete RDC data sets for protein cutinase using three scenarios in which 10%, 20%, and 30% of RDC data are randomly missing. These RDC data sets were also superimposed with 5% and 10% random errors. For each scenario, we repeated the calculations using five different sets of randomly incomplete data. The RMSD results are the average of the top 10 structures for each scenario of calculations (Fig. 9). The RMSDs are 1.26 ± 0.14 Å, 1.87 ± 0.30 Å, and 2.41 ± 0.17 Å, respectively, for 10%, 20%, and 30% missing RDC data superimposed with 5% random errors.

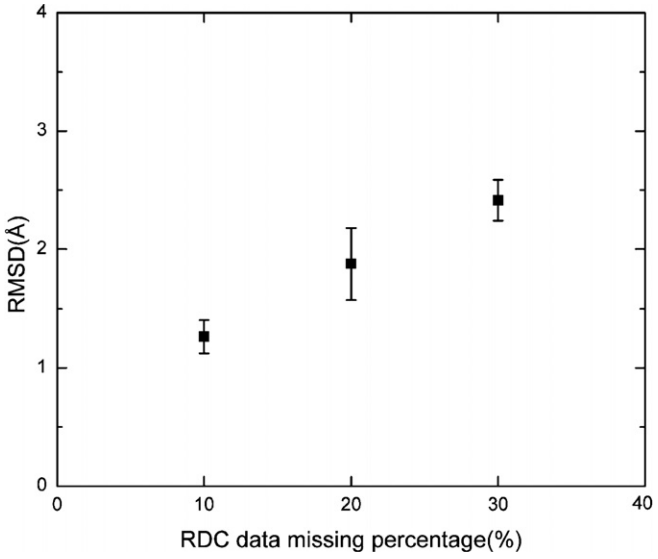

Fig. 9. The effect of incomplete RDC data on the RMSDs of the cutinase protein backbone structures calculated by the 3P method from the simulated RDC data superimposed with 5% random errors and different levels (10%, 20%, and 30%) of missing RDC data. For each level of missing data, five cutinase backbone structures are calculated with five randomly different missing RDC data superimposed with a 5% random error. The phi/psi error range is ±30°.

The RMSDs increased to 1.60 ± 0.18 Å, 2.04 ± 0.28 Å, and 2.65 ± 0.16 Å when the RDC data were superimposed with 10% random errors in each case. In all of these calculations, phi/psi angle error ranges were set to be ±30°. In the calculations using the incomplete RDC data sets, we excluded data sets where all missing data are concentrated in small regions.

4.5. Structure determination of ubiquitin, GB3 and DinI

We have tested the program using the experimental RDC data of ubiquitin, GB3, and DinI [20]. The STDs for phi/psi angles from the TALOS output were used as the error range for the phi/psi restraints. One of the practical problems when using the phi/psi torsion angles from TALOS is erroneous predictions. Roughly, 2–5% of TALOS predictions are outside of the STDs of the true values. For this reason, we also added a module that detects and corrects the phi/psi angles. This module checks the consistency between possible discrete peptide plane orientations derived from experimental RDC and the TALOS phi/psi angles. When phi/psi torsion angles from TALOS fall outside the range of possible correct phi/psi torsion angles (as defined by the discrete possible peptide plane orientations derived from the RDC–planarity correlation and the tetrahedral geometry at C_α), the RMSD between the experimental and back-calculated RDC increases dramatically. Fig. 10 shows examples of a jump in RMSD in RDC due to incorrect phi/psi torsion angles, residues 53 (phi) and 60 (phi/psi). In each case, the module then calculates possible correct torsion angles that give the RMSDs in RDC comparable to the average of the segment. Tests on the module show it is capable of detecting/correcting two consecutive incorrect phi/psi torsion angles in a restraint file, given a sufficiently large RDC IC. A detailed description for the phi/psi proof is provided in Supplementary material.

The backbone RMSD of the 3P-calculated structure relative to that of the X-ray crystal structures (1UBQ) of the ubiquitin is about 0.8 Å. We also replaced the STDs with an error range of ±30° in the phi/psi throughout the

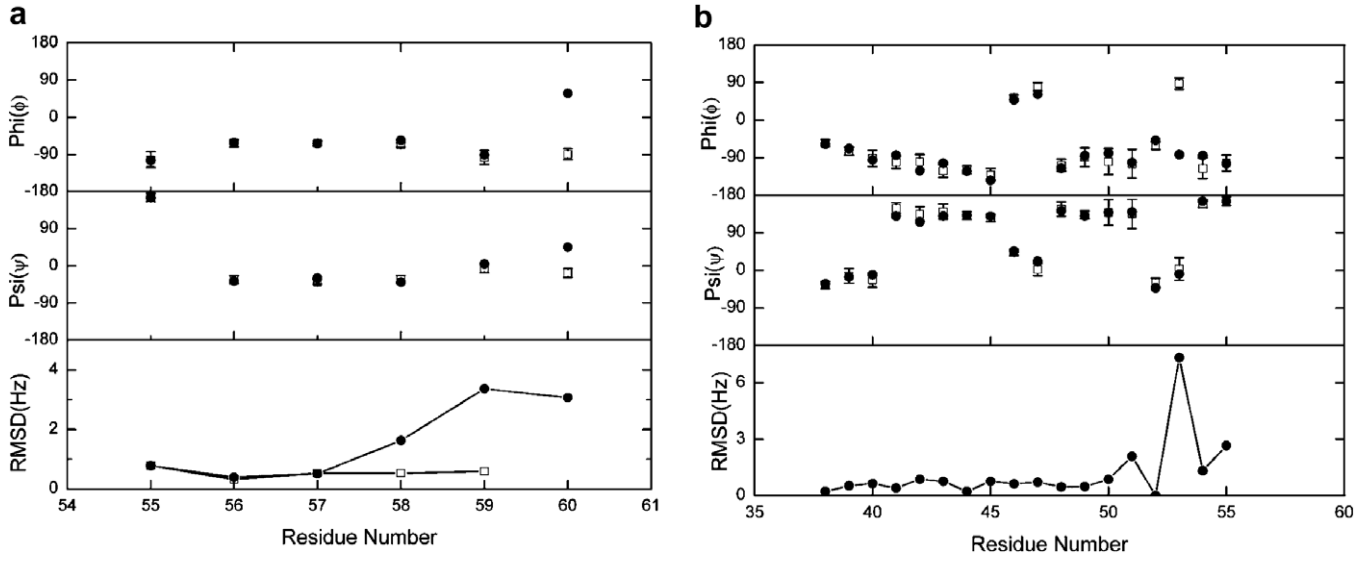

Fig. 10. phi/psi error proof of the ubiquitin TALOS phi/psi angles. The RMSDs in RDC start to jump as a result of erroneous phi/psi angles that contradict the experimental RDC data. (a) The phi/psi angles at residue 60 were erroneous, as indicated by a jump in the RMSD in RDC between the experimental data and the back-calculated data. The phi and psi angles in the crystal structure are 57.9° and 45.5°, respectively, whereas the TALOS-predicted values are $-89.5° \pm 13.0°$ and $-17.4° \pm 10.7°$, respectively. (b) For residue 53, the phi dihedral angle of the crystal structure is −82.9°, whereas TALOS predicted this angle to be 87.6 ± 13.6.

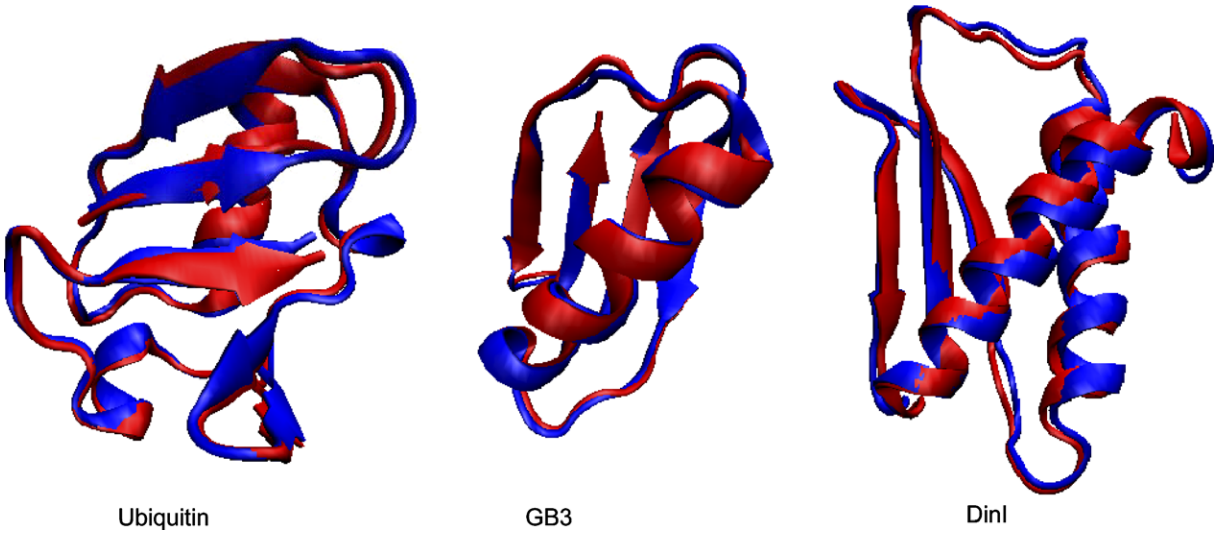

Fig. 11. Ribbon diagrams of the structure comparisons between the 3P-derived ubiquitin, GB3, and DinI (red), and their respective original X-ray crystal structures (blue). The following residues were not included in the calculations: residues (1–2) and the disordered C-terminus (70–76) of ubiquitin; residue 1 of GB3; residues 79–81 of DinI (no data). The RMSDs in pdb of the backbone structures are 0.92 Å (residues 3–69), 0.72 Å (residues 2–55), 0.89 Å (residues 2–78) for ubiquitin, GB3, and DinI, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

peptide chain, and the calculation yielded backbone structure with similar RMSDs, about 0.8 Å. In addition, we calculated the backbone structure of GB3 using one set of RDC data measured from one alignment medium, and the backbone RMSD relative to the X-ray crystal structure was about 0.7 Å (Fig. 11).

Protein DinI is relatively difficult case in which numerous errors are found in the phi/psi angles predicted by TALOS. In addition, a number of planes (planes 11, 13, 33, 45, and 53) have no RDC. For example, the region spanning residues 11 to 16 is missing about 50% of its RDC data, and has three errors in the phi/psi angle predictions from TALOS. Therefore, the phi/psi proof and the backbone structure calculation of this region are challenging but still manageable. The backbone RMSD of the 3P structures relative to that of 1GHH is 0.9–1.1 Å in the top 10% structures accepted based on the RMSDs in RDCs after regularization (Fig. 11). The RMSD for the segment of residues 2–53 is 0.61 Å, comparable to 0.79 Å obtained using MFR+ using two alignment tensors [11].

5. Discussion

Periodicity and planarity, and the intricate relation between these two and RDCs, form the basis of the current work and led to the derivation of the explicit analytical expression of RDCs in terms of the plane orientations in an equation that allows the alignment tensor axis system to be determined relative to the molecular frame, using a set of RDC data measured from a single alignment medium prior to a 3D structure determination. In the coordinate system of $(\delta^n, \rho^n, \gamma^n)$, the peptide plane orientations are fully defined by searching through these three-angle spaces and locating the RMSD minima between experimental and expected RDCs, aided by the restriction of phi/psi angles.

Both our method and the MECCANO [8] method utilize the peptide plane orientation, but the approaches are conceptually different. We use the RDC–PP correlations, which define the direct relationship between the peptide plane orientation and the RDCs, to calculate the structure. This relationship originates from the intrinsic correlation between the periodic secondary structure and the RDCs. At the end of the data analysis, we extract explicit peptide plane orientations relative to the alignment tensor axis system, using the equation and the RDC data. Consequently, our method requires only one alignment medium to solve the structure. In contrast, MECCANO requires two non-correlated sets of RDCs measured from two alignment media and uses a least-squares-based search algorithm to find the best solution. Obtaining more than one non-correlated set of RDCs from different alignment media is achievable in a favorable case, such as ubiquitin [6,7,35], but it can be rather challenging for a less well-behaved protein. This difficulty is exacerbated in cases where the structure is altered by differential interaction between the protein and various alignment media [36].

Uncertainty in the determined plane orientations may arise from several sources, including incomplete RDC measurements, deviations from planarity of the peptide plane, measurement errors in the RDC data, or departure from the rigid-body assumption. The interpretation of RDCs for dynamic systems is a subject under investigation [35,37–43] and the effect of errors in the measurements has been illustrated in the previous section. We will focus our discussion here on the first two sources of errors.

Incomplete RDC data sets may result in ambiguity in the peptide plane orientation. In many favorable cases, this

ambiguity can be resolved by the neighboring tetrahedral RDC ${}^1D_{C_\alpha H_\alpha}$ (or ${}^1D_{C_\alpha C_\beta}$) and available phi/psi angles. On the other hand, since the 3P program relies on the RDC data as the main source of restraints for structure determination, missing a significant number of RDCs results in an overall low IC and greater uncertainty in defining peptide plane orientations. In our test cases, the program appears able to converge with as much as 30% incomplete RDC data sets (Fig. 8). In addition, because the program interprets RDC in a concerted way in the context of chiral structure elements, it is able to handle cases where neither RDCs nor phi/psi angles are available for a residue in the middle of a structure because of the restraints that the orientation of the chiral structure elements imposes on both the sides and the chemical bond lineage of the residue.

The 3P program relys on mainly RDC data, supplemented by phi/psi angles, to determine protein backbone structures. The extensive absence of both RDC and phi/psi angles, such as in totally flexible loops that parse structured regions, will lead to a large RMSD or may yield no structure. In such a case, a few distance restraints among structural segments will minimize the translational errors [44].

One of the key assumptions of the 3P program is that peptide groups are planar to the first order of approximation (Eq. (7)). In practice, the ω angle deviates from 180° in protein structures. We used two methods to estimate the effect of non-planarity on the accuracy of the individual peptide plane orientations. First, the fit to the synthesized RDCs of an α-helix (residues 134–143) of subtilisin without a superimposed error was used as a sample of peptide ω angles. The mean angular RMSD between the extracted peptide plane normal vectors and the crystal structure average peptide plane normal vectors was 2.5°. A deviation of 5.9° was the largest observed for the plane of residues 138–139 ($\Delta\omega = 6.4°$, where $\Delta\omega$ is the deviation from the perfect planarity), and the smallest deviation was 0.3° for plane of residues 142–143 ($\Delta\omega = 0.5°$). The second method was running a Monte-Carlo simulation to estimate the error in the fit peptide plane orientation as a function of the deviation from planarity. The error in the fit plane orientation was found to increase approximately linearly with $\Delta\omega$, depending on the orientation of the plane in the alignment frame. These two results indicate that most peptide planes in proteins, which are within ~6° of planarity, can be well fit by the method. In some cases of alignment, non-planarity on one plane results in the selection of an incorrect (degenerate) orientation of the subsequent peptide plane because it agrees better with the nominal assumed tetrahedral geometry. A significant deviation from planarity is reflected in the residual errors in $T_{NC_\alpha C'}$ and ${}^1D_{C_\alpha H_\alpha}$. Non-planar peptide bonds result in deflections in $\hat{r}_j^{C_\alpha N}$ and $\hat{r}_{j+1}^{C_\alpha C}$ bond vectors from their respective (best-fit) peptide planes, which in turn result in a structure with deviation in planarity, causing discernable errors in calculating the tetrahedral angle $T_{NC_\alpha C'}$. For example, a moderate deviation from planarity in sutilisin resulted in an average deviation of the calculated tetrahedral angles $T_{NC_\alpha C'}$ related to the plane of residues (138–139) of 8.1° from the nominal value (Figs. S3 and S4, Supplementary material). The plane of residues 133–134 has the next-largest non-planarity ($\Delta\omega = 3°$) and yields a calculated mean tetrahedral angle deviation of 4.5°.

A survey of ω angles of the subtilisin structure indicates that deviations, greater than 7.5°, from planarity occur almost exclusively in the tight turns at either the beginnings or ends of α-helices or β-strands, or in non-regular secondary structure regions most likely at Gly residues. One remedy for alleviating the problem due to the large deviations is to restrain structures at these positions with RDC data directly during the regularization procedure.

Our approach differs in both philosophy and implementation from approaches such as the molecular fragment replacement+ (MFR+) [11], even though both methods use similar information. For example, instead of determining the tensors (by SVD) of a large group of peptide fragments in a database to find fits to the experimental RDCs of the target protein segment, we determine the tensor (by RDC–PP) [25] of a protein directly, without needing an explicit coordinate. The search for the best fit to a given set of RDC data within the restricted phi/psi range using the in-plane RDC periodicity is concerted in terms of plane orientation and is therefore more efficient than minimizing the sum of the difference between the measured and searched individual RDCs [11]. When a sufficient number of in-plane RDCs is available, simultaneously satisfying the consecutive in-plane periodicity correlations along a peptide backbone, which is also constrained by phi/psi, reduces the possibilities to a small group of discrete ensemble of orientations, whose RMSD is limited only by the quality of RDC measurements and the accuracy of the phi/psi.

The 3P program differs from programs REDCAT [45] and REDCRAFT [46] in a number of aspects. While REDCAT is essentially a graphics implementation of the SVD method, which extracts alignment tensors from RDC data and pre-existing coordinates [31] and REDCRAFT determines protein backbone structures using on RDCs measured in multiple alignment media and numbers of non-RDC restraints, such as NOEs, *J*-couplings, etc. for verification [46]. The 3P program extracts an alignment tensor from RDCs without a need for pre-knowledge of protein coordinates and calculates backbone structures using RDCs measured from a single medium aided with phi/psi angles.

6. Conclusion

The 3P program uses RDCs measured from one alignment medium, together with readily available predictions of phi/psi angles, to determine protein backbone structures. The program appears to be robust and can handle common, difficult scenarios such as those with incomplete or noisy RDC data, inaccurate phi/psi predictions, or non-regular covalent geometry.

7. Software availability

The 3P program package, including testing examples, tutorials, a tool box, program codes and installation instructions, can be downloaded from the web site http://sblweb.ncifcrf.gov/PNAI/files. The 3P tool box includes the following Python scripts: sim_bRDC.py and sim_pRDC.py for generating simulated RDC data based on either bond or peptide plane orientations extracted from a pdb file, respectively; sim_gauss_err_RDC.py and sim_rand_err_RDC.py for adding noises to simulated RDC data using either a Gaussian or a random distribution, respectively; sim_incomp_RDC.py for generating a set of given percentage incomplete RDC data set; PPP_xplor.py for generating peptide plane orientation restraints from a 3P backbone structure for regularization by Xplor-NIH. We also include Xplor-nih input scripts for regularization in directories wherever necessary.

Acknowledgment

This research was supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmr.2007.08.018.

References

[1] A. Bax, G. Kontaxis, N. Tjandra, Dipolar couplings in macromolecular structure determination, Methods Enzymol. 339 (2001) 127–174.
[2] N.R. Skrynnikov, L.E. Kay, Assessment of molecular structure using frame-independent orientational restraints derived from residual dipolar couplings, J. Biomol. NMR 18 (3) (2000) 239–252.
[3] N. Tjandra, J.G. Omichinski, A.M. Gronenborn, G.M. Clore, A. Bax, Use of dipolar H-1-N-15 and H-1-C-13 couplings in the structure determination of magnetically oriented macromolecules in solution, Nat. Struct. Biol. 4 (9) (1997) 732–738.
[4] J.R. Tolman, H.M. Al-Hashimi, L.E. Kay, J.H. Prestegard, Structural and dynamic analysis of residual dipolar coupling data for proteins, J. Am. Chem. Soc. 123 (7) (2001) 1416–1424.
[5] H. Zhou, A. Vermeulen, F.M. Jucker, A. Pardi, Incorporating residual dipolar couplings into the NMR solution structure determination of nucleic acids, Biopolymers 52 (4) (1999) 168–180.
[6] B.E. Ramirez, O.N. Voloshin, R.D. Camerini-Otero, A. Bax, Solution structure of DinI provides insight into its mode of RecA inactivation, Protein Sci. 9 (11) (2000) 2161–2169.
[7] B.E. Ramirez, A. Bax, Modulation of the alignment tensor of macromolecules dissolved in a dilute liquid crystalline medium, J. Am. Chem. Soc. 120 (35) (1998) 9106–9107.
[8] J.C. Hus, D. Marion, M. Blackledge, Determination of protein backbone structure using only residual dipolar couplings, J. Am. Chem. Soc. 123 (7) (2001) 1541–1542.
[9] F. Delaglio, G. Kontaxis, A. Bax, Protein structure determination using molecular fragment replacement and NMR dipolar couplings, J. Am. Chem. Soc. 122 (9) (2000) 2142–2143.
[10] C.A. Fowler, F. Tian, J.H. Prestegard, An NMR method for the rapid determination of protein folds using dipolar couplings, Biophys. J. 78 (1) (2000) 2827Pos.
[11] G. Kontaxis, F. Delaglio, A. Bax, Molecular fragment replacement approach to protein structure determination by chemical shift and dipolar homology database mining, Methods Enzymol. 394 (2005) 42–78.
[12] H. Valafar, J.H. Prestegard, Rapid classification of a protein fold family using a statistical analysis of dipolar couplings, Bioinformatics 19 (12) (2003) 1549–1555.
[13] R.R. Ketchem, W. Hu, T.A. Cross, high-resolution conformation of gramicidin-a in a lipid bilayer by solid-state NMR, Science 261 (5127) (1993) 1457–1460.
[14] S.J. Opella, P.L. Stewart, Solid-state nuclear magnetic-resonance structural studies of proteins, Method Enzymol. 176 (1989) 242–275.
[15] R. Tycko, P.L. Stewart, S.J. Opella, Peptide plane orientations determined by fundamental and overtone N-14 NMR, J. Am. Chem. Soc. 108 (18) (1986) 5419–5425.
[16] J.R. Quine, T.A. Cross, Protein structure in anisotropic environments: unique structural fold from orientational constraints, Concepts Magn. Reson. 12 (2) (2000) 71–82.
[17] G.A. Mueller, W.Y. Choy, D.W. Yang, J.D. Forman-Kay, R.A. Venters, L.E. Kay, 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 (1) (2000) 197–212.
[18] G.A. Mueller, W.Y. Choy, N.R. Skrynnikov, L.E. Kay, A method for incorporating dipolar couplings into structure calculations in cases of (near) axial symmetry of alignment, J. Biomol. NMR 18 (3) (2000) 183–188.
[19] S. Lee, M.F. Mesleh, S.J. Opella, Structure and dynamics of a membrane protein in micelles from three solution NMR experiments, J. Biomol. NMR 26 (4) (2003) 327–334.
[20] M.F. Mesleh, S.J. Opella, Dipolar waves as NMR maps of helices in proteins, J. Magn. Reson. 163 (2) (2003) 288–299.
[21] M.F. Mesleh, S. Lee, G. Veglia, D.S. Thiriot, F.M. Marassi, S.J. Opella, Dipolar waves map the structure and topology of helices in membrane proteins, J. Am. Chem. Soc. 125 (29) (2003) 8928–8935.
[22] M.F. Mesleh, G. Veglia, T.M. DeSilva, F.M. Marassi, S.J. Opella, Dipolar waves as NMR maps of protein structure, J. Am. Chem. Soc. 124 (16) (2002) 4206–4207.
[23] A. Mascioni, G. Veglia, Theoretical analysis of residual dipolar coupling patterns in regular secondary structures of proteins, J. Am. Chem. Soc. 125 (41) (2003) 12520–12526.
[24] J. Walsh, J. Cabello-Villegas, Y.-X. Wang, Periodicity in residual dipolar couplings and nucleic acid structures, JACS 126 (7) (2004) 1938–1939.
[25] J.D. Walsh, Y.X. Wang, Periodicity, planarity, residual dipolar coupling, and structures, J. Magn. Reson. 174 (1) (2005) 152–162.
[26] L.E. Chirlian, S.J. Opella, Improvement in determining structural information from solid-state NMR spectra, New Polymers 2 (3) (1992) 239–290.
[27] R.A. Engh, R. Huber, Accurate bond and angle parameters for X-ray protein-structure refinement, Acta Crystallogr. A 47 (1991) 392–400.
[28] P. Kuhn, M. Knapp, S.M. Soltis, G. Ganshaw, M. Thoene, R. Bott, The 0.78 Å structure of a serine protease: *Bacillus lentus* subtilisin, Biochemistry 37 (39) (1998) 13446–13452.
[29] G. Cornilescu, F. Delaglio, A. Bax, Protein backbone angle restraints from searching a database for chemical shift and sequence homology, J. Biomol. NMR 13 (3) (1999) 289–302.
[30] G.M. Clore, A.M. Gronenborn, A. Bax, 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 (1) (1998) 216–221.
[31] J.A. Losonczi, M. Andrec, M.W.F. Fischer, J.H. Prestegard, Order matrix analysis of residual dipolar couplings using singular value decomposition, J. Magn. Reson. 138 (2) (1999) 334–342.
[32] D.L. Bryce, A. Bax, Application of correlated residual dipolar couplings to the determination of the molecular alignment tensor

magnitude of oriented proteins and nucleic acids, J. Biomol. NMR 28 (3) (2004) 273–287.
[33] J.J. Warren, P.B. Moore, Application of dipolar coupling data to the refinement of the solution structure of the sarcin-ricin loop RNA, J. Biomol. NMR 20 (4) (2001) 311–323.
[34] J.D. Walsh, J. Kuszweski, Y.X. Wang, J.D. Walsh, Y.X. Wang, Determining a helical protein structure using peptide pixels, periodicity, planarity, residual dipolar coupling, and structures, J. Magn. Reson. 177 (1) (2005) 155–159.
[35] K.B. Briggman, J.R. Tolman, De novo determination of bond orientations and order parameters from residual dipolar couplings with high accuracy, J. Am. Chem. Soc. 125 (34) (2003) 10164–10165.
[36] J.J. Chou, J.D. Kaufman, S.J. Stahl, P.T. Wingfield, A. Bax, Micelle-induced curvature in a water-insoluble HIV-1 Env peptide revealed by NMR dipolar coupling measurement in stretched polyacrylamide gel, J. Am. Chem. Soc. 124 (11) (2002) 2450–2451.
[37] G.M. Clore, C.D. Schwieters, How much backbone motion in ubiquitin is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation? J. Am. Chem. Soc. 126 (9) (2004) 2923–2938.
[38] J. Meiler, W. Peti, C. Griesinger, Dipolar couplings in multiple alignments suggest alpha helical motion in ubiquitin, J. Am. Chem. Soc. 125 (27) (2003) 8072–8073.
[39] J. Meiler, J.J. Prompers, W. Peti, C. Griesinger, R. Bruschweiler, Model-free approach to the dynamic interpretation of residual dipolar couplings in globular proteins, J. Am. Chem. Soc. 123 (25) (2001) 6098–6107.
[40] W. Peti, J. Meiler, R. Bruschweiler, C. Griesinger, Model-free analysis of protein backbone motion from residual dipolar couplings, J. Am. Chem. Soc. 124 (20) (2002) 5822–5833.
[41] J.R. Tolman, Dipolar couplings as a probe of molecular dynamics and structure in solution, Curr. Opin. Struct. Biol. 11 (5) (2001) 532–539.
[42] N.A. Lakomek, T. Carlomagno, S. Becker, C. Griesinger, J. Meiler, G. Bouvignies, P. Bernado, M. Blackledge, A thorough dynamic interpretation of residual dipolar couplings in ubiquitin protein backbone dynamics from N–HN dipolar couplings in partially aligned systems: a comparison of motional models in the presence of structural noise, J. Biomol. NMR 34 (2) (2006) 101–115.
[43] G. Bouvignies, P. Bernado, M. Blackledge, Protein backbone dynamics from N–HN dipolar couplings in partially aligned systems: a comparison of motional models in the presence of structural noise, J. Magn. Reson. 173 (2) (2005) 328–338.
[44] J.D. Walsh, J. Kuszweski, Y.X. Wang, Determining a helical protein structure using peptide pixels, periodicity, planarity, residual dipolar coupling, and structures, J. Magn. Reson. 177 (1) (2005) 155–159.
[45] H. Valafar, K.L. Mayer, C.M. Bougault, P.D. LeBlond, F.E. Jenney Jr., P.S. Brereton, M.W.W. Adams, J.H. Prestegard, Backbone solution structures of proteins using residual dipolar couplings: application to a novel structural genomics target, J. Struc. Funct. Genomics 5 (4) (2004) 241–254.
[46] H. Valafar, J.H. Prestegard, J. Magn. Reson. 167 (2) (2004) 228–241.