PROTEINS

VOLUME 80, NUMBER 2, FEBRUARY 2012

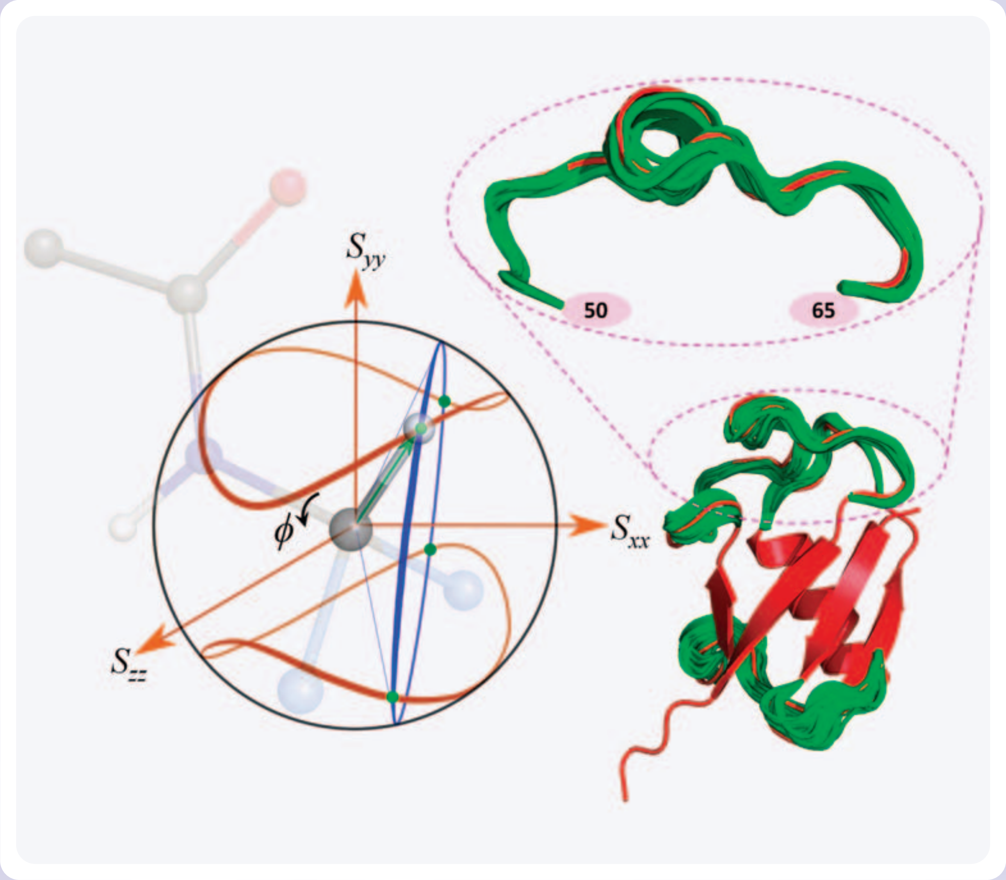
PAGES 335-676

WILEY-BLACKWELL

PROTEINS

STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

VOLUME 80, NUMBER 2, FEBRUARY 2012



Protein Loop Structures from RDCs

WILEY-BLACKWELL

ISSN 0887-3585

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Protein loop closure using orientational restraints from NMR data

Chittaranjan Tripathy,[1] Jianyang Zeng,[1] Pei Zhou,[2] and Bruce Randall Donald[1,2,*]

[1] Department of Computer Science, Duke University, Durham, North Carolina 27708
[2] Department of Biochemistry, Duke University Medical Center, Durham, North Carolina 27710

ABSTRACT

Protein loops often play important roles in biological functions. Modeling loops accurately is crucial to determining the functional specificity of a protein. Despite the recent progress in loop prediction approaches, which led to a number of algorithms over the past decade, few rigorous algorithmic approaches exist to model protein loops using global orientational restraints, such as those obtained from residual dipolar coupling (RDC) data in solution nuclear magnetic resonance (NMR) spectroscopy. In this article, we present a novel, sparse data, RDC-based algorithm, which exploits the mathematical interplay between RDC-derived sphero-conics and protein kinematics, and formulates the loop structure determination problem as a system of low-degree polynomial equations that can be solved exactly, in closed-form. The polynomial roots, which encode the candidate conformations, are searched systematically, using provable pruning strategies that triage the vast majority of conformations, to enumerate or prune all possible loop conformations consistent with the data; therefore, completeness is ensured. Results on experimental RDC datasets for four proteins, including human ubiquitin, FF2, DinI, and GB3, demonstrate that our algorithm can compute loops with higher accuracy, a three- to six-fold improvement in backbone RMSD, versus those obtained by traditional structure determination protocols on the same data. Excellent results were also obtained on synthetic RDC datasets for protein loops of length 4, 8, and 12 used in previous studies. These results suggest that our algorithm can be successfully applied to determine protein loop conformations, and hence, will be useful in high-resolution protein backbone structure determination, including loops, from sparse NMR data.

Proteins 2012; 80:433–453.

Key words: protein loops; loop closure; nuclear magnetic resonance; residual dipolar couplings; spheroconic; inverse kinematics; structural biology; algorithms.

INTRODUCTION

Protein loops are the segments of polypeptide chain that connect two relatively fixed segments of protein backbone. Although loops do not contain any regular units of secondary structure elements (SSEs), they often connect two SSEs such as α-helices or β-strands. In addition to serving as linkers between SSEs, loops often play crucial roles in protein stability and folding pathways, and in many other important biological functions such as binding, recognition, catalysis, and allosteric regulation.[1–7] Often, the structural difference in the loops within a fold family provides a basis to ratiocinate and describe the variability in the functional specificity.

Although the *global fold*, that is, the conformations and orientations of the SSEs of a protein, can often be determined with high accuracy via traditional experimental techniques such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, modeling loops that seamlessly close the gap between two consecutive SSEs by satisfying the geometric, biophysical, and data constraints remains a difficult problem. In X-ray crystallography, for instance, the disorder in a protein crystal can render interpretation of the resulting electron density for loops difficult. As a result, protein structures found in the protein data bank (PDB)[8] often have missing loops or disordered loops. The problem of computing loops that are biophysically reasonable and geometrically valid is called the *loop closure problem* or the *loop modeling problem*. Since its introduction four decades ago in the classic paper by Gō and Scheraga,[9] the loop closure problem has been an active area of research. In fact, modeling of loops can be regarded as an ab initio protein folding problem at a smaller scale.[10] It is also an important problem in de novo protein structure prediction.[11–13] Therefore, solutions and algorithms for accurate modeling of loops are highly desirable for understanding of the physical–chemical principles that determine protein structure and function.

Additional Supporting Information may be found in online version of this article
Grant sponsor: National Institutes of Health; Grant numbers: R01 GM-65982, R01 GM-079376

*Correspondence to: Bruce Randall Donald, Department of Computer Science, Levine Science Research Center, Duke University, Duham, NC 27708.
E-mail: brd+proteins11@cs.duke.edu.

Received 24 March 2011; Revised 23 August 2011; Accepted 6 September 2011
Published online 26 September 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23207

Exploring the conformation space of a protein loop to identify low energy loop conformations is a difficult computational problem. Methods to identify such loops include database search and homology modeling,[14–17] ab initio methods based on the minimization of empirical molecular mechanics energy functions,[10,18–21] and robotics-inspired inverse kinematics and optimization-based methods.[22–32] These techniques work in two phases: first, the protein conformation space is explored to find a set of candidate loop conformations, which are then evaluated in the second phase using an appropriate empirical energy function to select the most promising set of loops.

Database methods[14–17,33,34] identify a set of candidate loops from a library of fragments derived from a protein structure database such as the PDB[8] that fit the anchor residues on either end of a loop. These loops are further ranked using criteria such as the sequence homology and conformational energy. The accuracy of loop prediction by these methods heavily relies on the statistical diversity of the database, and the representation of the loops in it, for example, as in antibody hypervariable loops.[35,36] However, in general, database methods suffer from limited sampling of the loop conformations by the fragments in the database.

Ab initio loop modeling methods sample the conformation space randomly or use robotics-based sampling algorithms to generate a large number of loop conformations. Loop closure and energy minimization are done by using methods such as random tweak,[20,37] direct tweak,[10,18] analytical loop closure techniques,[25–27] molecular dynamics (MD) simulation,[38,39] Markov Chain Monte Carlo (MCMC) simulated annealing (SA),[19,40] bond-scaling-relaxation,[41] and other optimization techniques.[21,42] The accuracy of loop prediction by these methods depends on the efficacy of the conformational space exploration techniques used, and on the quality and parameterization of the force field employed to evaluate the conformational energy. These algorithms are computationally expensive, since they require a large number of random moves accompanied by repeated energy computations.

The protein loop closure problem is an *inverse kinematics* (IK) problem in computational biology. Given the poses of terminal anchor residues, it asks to find all possible values of the degrees of freedom (DOFs), that is, the values of the dihedrals ϕ and ψ, for which the fragment connects both the anchor residues. This problem has been studied widely in robotics and biology.[22–28,31,43] Tri-peptide loop closure, for which the number of DOFs is six and exactly six geometric constraints are stipulated due to the closure criterion, can be solved analytically[25–27,44,45] using exact IK solvers to give at most 16 possible solutions. For longer loops, the loop closure problem is underconstrained, so a continuous family of solutions are possible in the absence of additional constraints. Optimization-based IK solvers such as random tweak,[20,37] and the cyclic coordinate descent (CCD) algorithm[28] have been successful in dealing with a large number of DOFs, and have found many applications.[11,12,46,47] These methods iteratively solve for the DOFs until the loop closure constraints are satisfied. However, the problem of loop closure subjected to orientational restraints (e.g., from NMR data) has not been studied rigorously in the robotics or computational biology literature, and no practical deterministic algorithm exists to our knowledge.

Protein structure determination using nuclear Overhauser effect (NOE) distance restraints is NP-hard.[48] Traditional protein structure determination from solution NMR data starts with an elongated polypeptide backbone chain, and uses NOEs and dihedral angle restraints in a simulated annealing/simplified molecular dynamics (SA/MD) protocol[49–53] to compute the protein structure. Residual dipolar coupling (RDC) restraints are only incorporated in the final stages of the structure computation to refine the structures.[53,54] NOE-based structure determination protocols are known to be prone to be trapped in local minima or lead to wrong convergence. To overcome the shortcomings of NOE-based methods, approaches in Refs. 55–60 have been proposed that primarily use RDC data, which provides precise global orientational restraints on internuclear vector orientations, to determine protein backbone structure. However, most of these approaches use stochastic search, and therefore, lack any algorithmic guarantee on the quality of the solution or running time. In recent work from our laboratory,[61–64] polynomial-time algorithms have been proposed for high-resolution backbone global fold determination from a minimal amount of RDC data. These algorithms represent the RDC equations and protein kinematics in algebraic form, and use exact methods in a divide-and-conquer framework to compute the global fold. In addition, these algorithms use a sparse set of RDC measurements (e.g., only two RDCs per residue), with the goal of minimizing the number of NMR experiments, hence the time and cost to perform them.

A high-resolution protein backbone is often a starting point for structure-based protein design,[65–68] and assembly of symmetric protein homo-oligomers.[69] An accurate backbone structure facilitates the assignment of side-chain resonances (i.e., the *side-chain assignment problem*),[70] and nuclear Overhauser effect spectroscopy (NOESY) spectra (i.e., the *NOE assignment problem*),[50,71] which are prerequisites for high-resolution structure determination protocols, including side-chain conformations. For example, the algorithms in Refs. 61–64 have been used in Refs. 71–73 to develop new algorithms for NOE assignment. These algorithms led to the development of a new framework[71] for high-resolution protein structure determination, which was used prospectively to solve the solution structure of the FF Domain 2 of human transcription elongation factor CA150 (FF2; PDB id: 2kiq). The global folds obtained by Refs. 61–64 have all the loops missing which requires a new algorithm that can compute the missing loops from RDCs. A preliminary approach in computing the missing loops in Ref. 71 used a heuristic local minimization protocol.[53]

In this article, we give a solution to the loop closure problem. We present an efficient deterministic algorithm, POOL, that computes the missing loops from RDC data. Our algorithm exploits the interplay between protein backbone kinematics and the global orientational restraints derived from RDC data to naturally discretize the conformation space by polynomial-root solutions, and represents the candidate conformations using a tree. A systematic depth-first search of the conformation tree is used to enumerate all possible loop conformations that are consistent with the data. POOL uses efficient pruning strategies capable of pruning the majority of the conformations that are provably not part of a valid loop, thereby achieving a huge reduction in the search space. Unlike other algorithms, for example Ref. 57, that attempt to compute backbone structure using as many as 15 RDCs per residue recorded in two alignment media, which in general can be difficult to measure due to experimental reasons,[74] our algorithm uses as few as two or three RDCs per residue in one alignment medium, which is often experimentally feasible. As we will show in the Results and Discussion section, when given the same data, our algorithm performs better than traditional SA/MD-based approaches,[53] and also better than previous sparse-data protocols.[75] In addition, our algorithm can compute ensembles of near-native loop conformations in the presence of modest levels of protein internal dynamics. Additional RDCs, and other data that provide constraints in torsion-angle space (e.g., TALOS[76,77] dihedral restraints) or in Euclidean space (e.g., sparse NOEs), whenever available, can directly be incorporated into our algorithm. In summary, we make the following contributions in this article:

1. Derivation of quartic equations for backbone dihedrals ϕ and ψ from experimentally recorded RDC sphero-conics and backbone kinematics, which can be solved exactly and in closed form;
2. Systematic search of the roots of the polynomial equations that encode the conformations, using efficient pruning methods to eliminate the vast majority of conformations;
3. Design and implementation of an efficient algorithm, POOL, to determine protein loop conformations from a limited amount of experimental RDC data;
4. Promising results from the application of our algorithm both on experimental NMR datasets for four proteins, and on synthetic datasets for protein loops studied previously in Refs. 13,26,28,78.

METHODS

Overview

POOL solves the following loop closure problem. Let the residues of the protein be numbered from 1 to n (from N- to C-terminus). Suppose the global fold of the protein has been determined from RDCs in a *principal order frame* (POF) of RDCs (see subsection RDC sphero-conics), as we showed was feasible in Refs. 61–64,71. In principle, the global fold of proteins could also be computed using protein structure prediction,[79] or homology modeling[80,81]; alternatively, X-ray structures (with missing loops) can be used. Given two consecutive SSEs with n_1 and n_2 being the last residue of the first SSE and first residue of the second SSE, respectively, the missing loop $[n_1, n_2]$ is defined as the fragment between residues n_1 and n_2 with both end residues included. The residues n_1 and n_2 that are part of the SSEs will be called the *stationary anchors*, and those of a candidate loop will be called the *mobile anchors*. We assume that the n_1 mobile anchor of the loop is attached to the n_1 stationary anchor of the first SSE. Then the loop closure problem is stated as follows: in the POF, given the poses of the stationary anchors n_1 and n_2 [points in $\mathbb{R}^3 \times SO(3)$], compute a complete set of conformations of fragments $[n_1, n_2]$ so that n_2 mobile anchor of each fragment in the set assumes the pose of the stationary anchor n_2, while satisfying the RDC data and standard protein geometry.

Table I
ϕ-Defining and ψ-Defining RDCs

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

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 exactly and in closed form.

Our algorithm builds upon the initial work from our laboratory,[62–64,71] where the authors developed polynomial time algorithms to compute high-resolution backbone global fold de novo from N–H^N and C^{α}–H^{α} RDCs in one alignment medium. These sparse-data algorithms have been extended to incorporate combinations of different types of RDCs (see Table I) in one or two alignment media. The new generalized framework is called RDC-ANALYTIC.[64,71] POOL implements a novel algorithm to determine protein loop backbone structures from a minimal amount of RDC data, and is a crucial addition to the RDC-ANALYTIC suite, which did not compute loops before.

Table I describes the RDC types that POOL uses to compute the backbone dihedrals exactly and in closed form. 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. The input data to POOL include: (1) the global fold of the protein computed by Refs. 62,63,71; (2) the alignment tensor, which generally can be computed from the global fold using Refs. 61,82; (3) at least one ϕ-defining and one ψ-defining RDCs per residue, and optionally other data, for example, additional RDCs, TALOS[76,77] dihedral restraints and sparse NOEs; and (4) the primary sequence of the protein.

Solving a system of equations from RDCs, protein kinematics and loop closure constraints simultaneously is a difficult computational problem since it leads to solving a high-degree polynomial system. However, since RDCs are very precise measurements, an algorithm which is able to compute protein fragments by inductively solving low-degree polynomial equations derived from RDCs and backbone kinematics, and drives the computation to satisfy the loop closure criterion, will achieve the desired objective. Our algorithm POOL is based on this key insight. Starting from a stationary anchor, it solves each DOF sequentially using the equations derived in the following subsections. The discrete values of the DOFs computed from the polynomial roots, are represented by a conformation tree grown recursively as we solve for the DOFs progressively. An internal (i.e., non-leaf) node in the tree represents the conformation of a part of a candidate loop, and a leaf node represents a candidate loop conformation computed from RDCs. Figure 1 illustrates a conformation tree for a loop. As each node is visited in a depth-first traversal of the tree, if the conformation represented by that node fails the conformation filters (see subsection Pruning with conformation filters), it is called a *dead-end* node, and the subtree rooted at that node is pruned. Dead-end nodes identified at lower levels (i.e., closer to the root) of the conformation tree prune more conformations than those identified at higher levels. Finally, all remaining unpruned conformations (leaf nodes) already close to the stationary anchor (since they satisfy the reachability criterion as described in subsection Pruning with conformation filters), are evaluated for loop closure. At this stage, minimization techniques can be applied to improve the closure. Conformations satisfying the closure criterion are added to the final ensemble of loops. POOL enumerates all loop conformations that satisfy the RDC data and pass the conformation filters; therefore, it guarantees completeness.

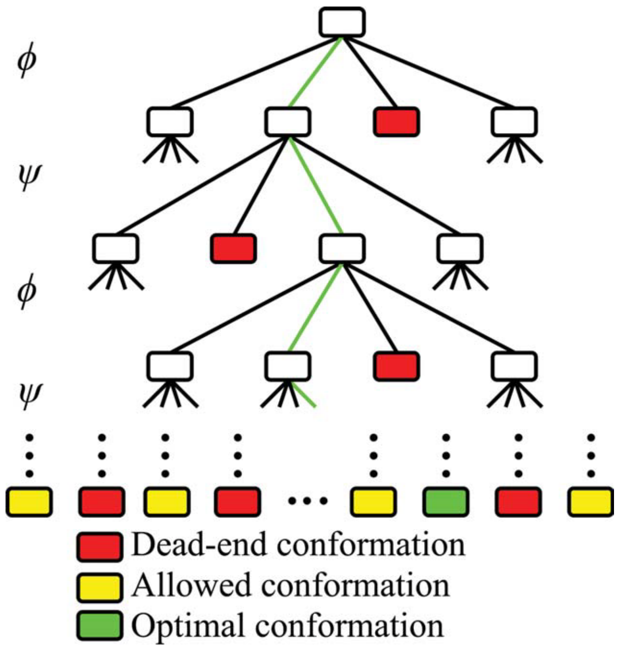


Figure 1

An example conformation tree. A non-leaf node represents a part of a candidate loop, and a leaf node represents a candidate loop conformation. Dead-end conformations detected by the conformation filters are pruned. Allowed conformations are subject to the test for loop closure. An optimal conformation passes all the tests; therefore, belongs to the ensemble of computed loops. Shown in green is an accepting computation path for an optimal conformation. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

RDC sphero-conics

The RDC r between two spin-$\frac{1}{2}$ nuclei a and b is given by

$$r = D_{\max}\mathbf{v}^T\mathbf{S}\mathbf{v}, \qquad (1)$$

where $\mathbf{v}$ is the unit internuclear vector between a and b, $D_{\max}$ is the dipolar interaction constant, and $\mathbf{S}$ is the *Saupe order matrix*,[83] or *alignment tensor*, that specifies the ensemble-averaged anisotropic orientation of the protein in the laboratory frame. $\mathbf{S}$ is a 3 × 3 symmetric, traceless, rank 2 tensor with five independent elements.[63,84–86] The dipolar interaction constant $D_{\max}$ is given by

$$D_{\max} = \frac{\mu_0 \hbar \gamma_a \gamma_b}{4\pi^2}\langle r_{ab}^{-3}\rangle, \qquad (2)$$

where μ_0 is the magnetic permeability of vacuum, $\hbar$ *is Planck's constant*, γ_a and γ_b are the gyromagnetic ratios of the nuclei a and b, respectively, and $\langle r_{ab}^{-3}\rangle$ represents the vibrational ensemble-averaged inverse cube of the distance between the two nuclei. Letting $D_{\max} = 1$ (i.e., scaling the RDCs appropriately), and considering a global coordinate frame that diagonalizes the alignment tensor $\mathbf{S}$, often called the POF, Eq. (1) can be written as

$$r = S_{xx}x^2 + S_{yy}y^2 + S_{zz}z^2, \qquad (3)$$

where S_{xx}, S_{yy}, and S_{zz} are the three diagonal elements of a diagonalized alignment tensor $\mathbf{S}$, and x, y, and z are, respectively, the x, y, and z components of the unit vector $\mathbf{v}$ in a POF that diagonalizes $\mathbf{S}$. Since $\mathbf{v}$ is a unit vector, that is,

$$x^2 + y^2 + z^2 = 1, \qquad (4)$$

an RDC constrains the corresponding internuclear vector $\mathbf{v}$ to lie on the intersection of a concentric unit sphere [Eq. (4)] and a quadric [Eq. (3)].[87] This gives a pair of closed curves inscribed on the unit sphere that are diametrically opposite to each other (see Fig. 2). These curves are known as *sphero-conics* or *sphero-quartics*.[88–90]

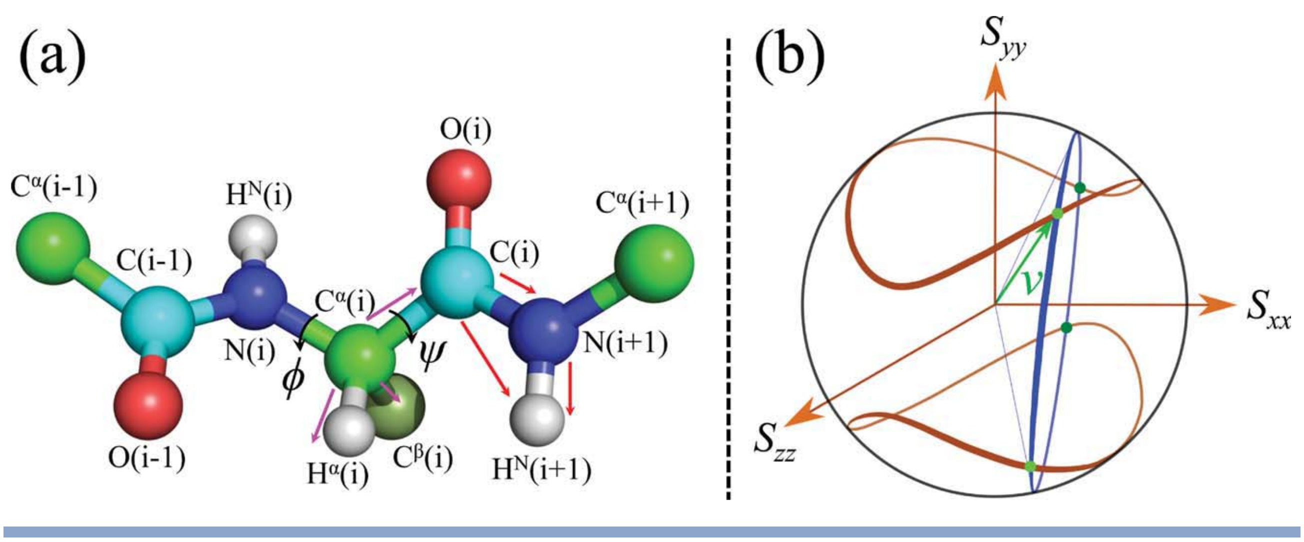


Figure 2

(**a**) The internuclear vectors (shown using arrows) for which RDCs are possible to measure. The magenta and red arrows represent ϕ-defining and ψ-defining RDCs, respectively. (**b**) The brown pringle-shaped RDC sphero-conic curves inscribed on a unit sphere constrain the internuclear vector **v** (green arrow) to lie on one of them. The kinematic circle (shown in blue almost edge-on) of **v** intersects the sphero-conic curves in at most four points (green dots) leading to a maximum of four possible orientations for the internuclear vector **v**. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Using Eq. (4) in Eq. (3), we can rewrite Eq. (3) in the following form:

$$ax^2 + by^2 = c, \tag{5}$$

where $a = S_{xx} - S_{zz}$, $b = S_{yy} - S_{zz}$, and $c = r - S_{zz}$. Henceforth, we refer to Eq. (5) as the *reduced RDC equation*.

For further background on RDCs and RDC-based structure determination, the reader is referred to Refs. 43,63,84–86. We now derive analytic solutions for peptide plane orientations using protein kinematics and the RDC sphero-conics, which are used inductively in our algorithm POOL to build the conformation tree.

Analytic solutions for peptide plane orientations from ϕ-defining and ψ-defining RDCs in one alignment medium

The derivation below assumes standard protein geometry, which is exploited in the kinematics.[61] We choose to work in an orthogonal coordinate system defined at the peptide plane P_i with z-axis along the bond vector N(i) → H^N(i), where 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 y-axis and the bond vector N(i) → C$^\alpha$(i) is fixed. The x-axis is defined based on the right-handedness. Let $\mathbf{R}_{i,\mathrm{POF}}$ denote the orientation (rotation matrix) of P_i with respect to the POF. Then, $\mathbf{R}_{1,\mathrm{POF}}$ denotes the relative rotation matrix between the coordinate system defined at the first residue of the current SSE and the POF. $\mathbf{R}_{i,\mathrm{POF}}$ is used to derive $\mathbf{R}_{i+1,\mathrm{POF}}$ inductively after we compute the dihedral angles ϕ_i and ψ_i. $\mathbf{R}_{i+1,\mathrm{POF}}$, in turn, is used to compute the $(i+1)^{st}$ peptide plane.

We derive closed-form solutions for the dihedral ϕ, and hence the corresponding internuclear vector orientations, using a ϕ-defining RDC as shown in the following proposition.

Proposition 1. *Given the diagonalized alignment tensor components S_{xx} and S_{yy}, the peptide plane P_i, and a ϕ-defining RDC r for the corresponding internuclear vector of residue i, there exist at most four possible values of the dihedral angle ϕ_i that satisfy the RDC r. 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^N bond vector of residue i in the local coordinate frame defined on the peptide plane P_i. Let $\mathbf{v}_1 = (x, y, z)^T$ denote the internuclear vector for the ϕ-defining RDC for residue i in the POF. We can write the forward kinematics relation between $\mathbf{v}_0$ and $\mathbf{v}_1$ as follows:

$$\mathbf{v}_1 = \mathbf{R}_{i,\mathrm{POF}}\, \mathbf{R}_l\, \mathbf{R}_z(\phi_i)\, \mathbf{R}_r\, \mathbf{v}_0. \tag{6}$$

Here, $\mathbf{R}_l$ and $\mathbf{R}_r$ are constant rotation matrices that describe the kinematic relationship between $\mathbf{v}_0$ and $\mathbf{v}_1$. $\mathbf{R}_z(\phi_i)$ 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 Eq. (6) we have

$$x = A_0 + A_1 c + A_2 s, \qquad y = B_0 + B_1 c + B_2 s, \qquad z = C_0 + C_1 c + C_2 s, \tag{7}$$

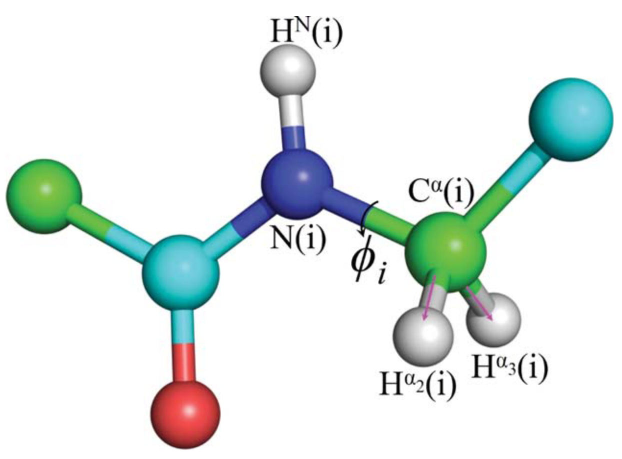


Figure 3

The amino acid residue glycine. The two H^α atoms are denoted by H^{α_2} and H^{α_3}, respectively. The C^α–H^α RDC is the sum of the RDCs measured for the bond vectors C^α–H^{α_2} and C^α–H^{α_3}. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

in which A_i, B_i, C_i for $0 \leq i \leq 2$ are constants. Using Eq. (7) in the reduced RDC equation [Eq. (5)], and simplifying we obtain

$$K_0 + K_1 c + K_2 s + K_3 cs + K_4 c^2 + K_5 s^2 = 0, \tag{8}$$

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

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

in Eq. (8) we have

$$L_0 + L_1 u + L_2 u^2 + L_3 u^3 + L_4 u^4 = 0, \tag{10}$$

in which L_i, $0 \leq i \leq 4$ 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 (at most) four real solutions of Eq. (10). For each u_i, the corresponding ϕ_i value can be computed using Eq. (9). □

The amino acid residue glycine (Gly), shown in Figure 3, has two H^α atoms which we denote by H^{α_2} and H^{α_3}. The C^α–H^α RDC, measured for Gly is the sum of the RDCs for the bond vectors $C^\alpha-H^{\alpha_2}$ and $C^\alpha-H^{\alpha_3}$. Here we show that given C^α–H^α RDC for a Gly residue, we can compute all possible solutions for the dihedral ϕ.

Proposition 2. *Given the diagonalized alignment tensor components S_{xx} and S_{yy}, the peptide plane P_i, and the C^α–H^α RDC r for residue i which is a glycine, there exist at most four possible values of the dihedral angle ϕ_i that satisfy the C^α–H^α RDC r. 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^N bond vector of residue i in the local coordinate frame defined on the peptide plane P_i. Let $\mathbf{v}_1 = (x_1, y_1, z_1)^T$ and $\mathbf{v}_2 = (x_2, y_2, z_2)^T$ be the unit vectors defined in the POF to represent $C^\alpha-H^{\alpha_2}$ and $C^\alpha-H^{\alpha_3}$, respectively. We can write the forward kinematics relations between $\mathbf{v}_0$ and $\mathbf{v}_1$, and between $\mathbf{v}_0$ and $\mathbf{v}_2$ as follows:

$$\mathbf{v}_1 = \mathbf{R}_{i,\mathrm{POF}}\, \mathbf{R}_l\, \mathbf{R}_z(\phi_i)\, \mathbf{R}_r\, \mathbf{v}_0 \tag{11}$$

$$\mathbf{v}_2 = \mathbf{R}_{i,\mathrm{POF}}\, \mathbf{R}_l\, \mathbf{R}_z(\phi_i)\, \mathbf{R}_r'\, \mathbf{v}_0. \tag{12}$$

Here $\mathbf{R}_l$, $\mathbf{R}_r$, and $\mathbf{R}_r'$ are constant rotation matrices. $\mathbf{R}_z(\phi_i)$ 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 Eqs. (11) and (12) we have

$$x_1 = A_{10} + A_{11}c + A_{12}s, \qquad y_1 = B_{10} + B_{11}c + B_{12}s, \\ z_1 = C_{10} + C_{11}c + C_{12}s \tag{13}$$

$$x_2 = A_{20} + A_{21}c + A_{22}s, \qquad y_2 = B_{20} + B_{21}c + B_{22}s, \\ z_2 = C_{20} + C_{21}c + C_{22}s, \tag{14}$$

where A_{ij}, B_{ij}, C_{ij} for $1 \leq i \leq 2$ and $0 \leq j \leq 2$ are constants.

For glycine, since C^α–H^α RDC is the sum of the RDCs for the bond vectors $C^\alpha-H^{\alpha_2}$ and $C^\alpha-H^{\alpha_3}$, we can write the RDC equation as

$$r = D_{\max}\left[\mathbf{v}_1^T \mathbf{S} \mathbf{v}_1 + \mathbf{v}_2^T \mathbf{S} \mathbf{v}_2\right]. \tag{15}$$

Without loss of generality, we let $D_{\max} = 1$, which is done by scaling the RDCs appropriately. Now, since $\mathbf{v}_1$ and $\mathbf{v}_2$ are unit vectors,

$$x_1^2 + y_1^2 + z_1^2 = 1 \tag{16}$$

$$x_2^2 + y_2^2 + z_2^2 = 1. \tag{17}$$

Using Eqs. (16) and (17) we can expand Eq. (15), and rewrite it in the following form:

$$a'(x_1^2 + x_2^2) + b'(y_1^2 + y_2^2) = c', \tag{18}$$

where

$$a' = S_{xx} - S_{zz}, \qquad b' = S_{yy} - S_{zz}, \qquad c' = r - 2S_{zz}.$$

Using Eqs. (13) and (14) in Eq. (18), and simplifying we obtain

$$K_0 + K_1 c + K_2 s + K_3 cs + K_4 c^2 + K_5 s^2 = 0, \qquad (19)$$

where K_i, $0 \leq i \leq 5$ are constants.

Using half-angle substitutions

$$u = \tan\left(\frac{\phi_i}{2}\right), \qquad c = \frac{1-u^2}{1+u^2}, \quad \text{and} \quad s = \frac{2u}{1+u^2} \qquad (20)$$

in Eq. (19) we have

$$L_0 + L_1 u + L_2 u^2 + L_3 u^3 + L_4 u^4 = 0, \qquad (21)$$

where L_i, $0 \leq i \leq 4$ are constants.

Equation (21) 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 (at most) four real solutions of Eq. (21). For each u_i, the corresponding ϕ_i value can be computed using Eq. (20). □

We next derive closed-form solutions for the dihedral ψ, using a ψ-defining RDC as shown in the following proposition.

Proposition 3. *Given the diagonalized alignment tensor components S_{xx} and S_{yy}, the peptide plane P_i, the dihedral ϕ_i, and a ψ-defining RDC r for the corresponding internuclear vector on peptide plane P_{i+1}, there exist at most four possible values of the dihedral angle ψ_i that satisfy the RDC r. The possible values of ψ_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^N bond vector of residue i in the local coordinate frame defined on the peptide plane P_i. Let $\mathbf{v}_1 = (x, y, z)^T$ denote the internuclear vector for the ψ-defining RDC for residue i in the POF. Note that the internuclear vector for a ψ-defining RDC has at least one nucleus that belongs to residue $i + 1$. The forward kinematics relation between $\mathbf{v}_0$ and $\mathbf{v}_1$ can be written as follows:

$$\mathbf{v}_1 = \mathbf{R}_{i,\mathrm{POF}}\, \mathbf{R}_l\, \mathbf{R}_z(\phi_i)\, \mathbf{R}_m\, \mathbf{R}_z(\psi_i)\, \mathbf{R}_r\, \mathbf{v}_0. \qquad (22)$$

Here, $\mathbf{R}_l$, $\mathbf{R}_m$, and $\mathbf{R}_r$ are constant rotation matrices. $\mathbf{R}_z(\phi_i)$ is the rotation about the z-axis by ϕ_i, and is a constant rotation matrix since ϕ_i is known (already computed before computing ψ_i by using Proposition 1). $\mathbf{R}_z(\psi_i)$ is the rotation about the z-axis by ψ_i.

Let c and s denote $\cos\psi_i$ and $\sin\psi_i$, respectively. Using this and expanding Eq. (22) we have

$$\begin{aligned} x &= A_0 + A_1 c + A_2 s, \qquad y = B_0 + B_1 c + B_2 s, \\ z &= C_0 + C_1 c + C_2 s, \end{aligned} \qquad (23)$$

in which A_i, B_i, C_i for $0 \leq i \leq 2$ are constants. Using Eq. (23) in the reduced RDC equation [Eq. (5)], and simplifying we obtain

$$K_0 + K_1 c + K_2 s + K_3 cs + K_4 c^2 + K_5 s^2 = 0, \qquad (24)$$

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

$$u = \tan\left(\frac{\psi_i}{2}\right), \qquad c = \frac{1-u^2}{1+u^2}, \quad \text{and} \quad s = \frac{2u}{1+u^2} \qquad (25)$$

in Eq. (24) we have

$$L_0 + L_1 u + L_2 u^2 + L_3 u^3 + L_4 u^4 = 0, \qquad (26)$$

in which L_i, $0 \leq i \leq 4$ are constants.

Equation (26) 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 (at most) four real solutions of Eq. (26). For each u_i, the corresponding ψ_i value can be computed by using Eq. (25). □

Putting the preceding propositions together, we obtain the following result for the number of peptide plane orientations: Given the diagonalized alignment tensor components S_{xx} and S_{yy}, the peptide plane P_i, a ϕ-defining RDC and a 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 satisfy the RDCs.

Sampling the DOFs when RDCs are missing

Protein loops can be modeled as kinematic chains.[24,31,43,47] In a kinematic chain, a *redundant* DOF is defined as a DOF for which no kinematic constraint is available.[47,91,92] Mathematically, with no restraints from experimental measurements, for a loop with n (>6) DOFs, three translational and three orientational constraints are stipulated due to loop closure; therefore, the remaining $n - 6$ DOFs are redundant. Hence, $n - 6$ equality constraints are necessary to solve for the loop conformations so that the number of conformations is discrete and finite. When RDCs are missing, we sample the corresponding DOFs. We systematically sample, at 5° resolution, the dihedrals from the Ramachandran map (and TALOS dihedral restraints if available) for the DOFs for which RDCs are missing, and use analytic equations derived above to solve for the other dihedrals for which RDCs are available, to compute an ensemble of loops complete to the resolution of sampling. If RDCs can be recorded for the missing ones in a second alignment medium, POOL can use them as shown in Supporting Information Appendix A. Table II shows the number of missing RDCs, and when as many as five RDCs are missing in a loop, POOL still could compute the loops accurately.

Table II
The Minimum RMSD (Å) from the NMR Reference Loops

Protein loop[a]	Length[b]	Types of RDCs[c]	RDCs missing[d]	RMSD[e] (Å) (POOL)	RMSD[f] (Å) (XPLOR-NIH)	RMSD[g] (Å) (CS-ROSETTA)
Ubiquitin 7–12	6	C^{α}–H^{α}, N–H^{N}	2	0.64	1.40	1.74
Ubiquitin 17–23	7	C^{α}–H^{α}, N–H^{N}	2	0.60	2.25	0.50
Ubiquitin 33–41	9	C^{α}–H^{α}, N–H^{N}	2	0.89	2.07	0.92
Ubiquitin 45–48	4	C^{α}–H^{α}, N–H^{N}	0	0.27	1.58	0.51
Ubiquitin 50–65	16	C^{α}–H^{α}, N–H^{N}	2	0.66	3.94	0.63
Ubiquitin 7–12	6	C^{α}–C′, N–H^{N}	3	0.37	0.67	1.60
Ubiquitin 17–23	7	C^{α}–C′, N–H^{N}	3	0.60	3.54	0.49
Ubiquitin 33–41	9	C^{α}–C′, N–H^{N}	5	0.58	3.11	0.66
Ubiquitin 45–48	4	C^{α}–C′, N–H^{N}	0	0.11	1.02	0.28
Ubiquitin 50–65	16	C^{α}–C′, N–H^{N}	4	1.06	4.48	0.67
FF2 18–27	10	C^{α}–H^{α}, N–H^{N}	3	1.41	3.20	2.08
FF2 33–38	6	C^{α}–H^{α}, N–H^{N}	3	0.34	1.09	0.95
FF2 42–48	7	C^{α}–H^{α}, N–H^{N}	4	1.31	2.14	1.34
DinI 8–17	10	C^{α}–H^{α}, N–H^{N}	5	1.57	4.17	2.51
DinI 32–39	8	C^{α}–H^{α}, N–H^{N}	3	0.61	3.45	0.58
DinI 45–49	5	C^{α}–H^{α}, N–H^{N}	2	0.28	2.27	2.16
DinI 53–58	6	C^{α}–H^{α}, N–H^{N}	2	0.42	2.62	0.81
GB3 8–13	6	C^{α}–H^{α}, N–H^{N}	0	0.43	1.07	2.59
GB3 19–23	5	C^{α}–H^{α}, N–H^{N}	0	0.34	0.23	0.55
GB3 36–42	7	C^{α}–H^{α}, N–H^{N}	1	0.27	1.34	1.27
GB3 46–51	6	C^{α}–H^{α}, N–H^{N}	0	0.65	3.61	1.76
Average				**0.64**	**2.35**	**1.17**

[a] The anchor residues are always included.
[b] number of residues.
[c] experimental RDCs used. The C^{α}–H^{α}, C^{α}–C′ and N–H^{N} RDC RMSDs of loops computed by POOL are less than 2.0, 0.2 and 1.0 Hz, respectively.
[d] Missing means unavailable.
[e,f,g] Backbone RMSD computed versus the NMR reference loops. The results show that the loops computed by POOL are more accurate than those computed by XPLOR-NIH[53] using the same sparse data. In most cases, the loops computed by POOL are more accurate than those computed by CS-ROSETTA. CS-ROSETTA used the same RDCs as POOL, plus the backbone chemical shifts.

Pruning with conformation filters

Loop conformations are generated by traversing a conformation tree in a depth-first search order (see subsection Overview). At each node, conformation filters are applied as *predicates*. If the node passes all the filters, then the subtree rooted at that node is visited; otherwise, the subtree is pruned. Failing a predicate at lower levels (closer to the root) of the conformation tree prunes more conformations than that detected at higher levels (farther from the root). In fact, pruning at depth i eliminates $O(b^{n-i})$ conformations, where b is the average number of branches in the conformation tree, and n is the height of the conformation tree. For loops with constrained work-space, substantial pruning can be achieved resulting in significant speedup. POOL uses the following conformation filters.

Real solution filter

While solving the analytic equations derived earlier to compute the dihedrals from RDCs, all non-real roots with the imaginary parts greater than a chosen threshold are discarded.[71] In addition, multiplicities of the roots are eliminated as follows: if two roots r_1 and r_2 are such that $|r_1 - r_2| < \delta$ for a chosen small number δ, then one of the roots is eliminated in favor of the other. This prunes the entire subtree rooted at the eliminated root.

Ramachandran and TALOS filters

There exist regions in the Ramachandran map that are forbidden for certain combinations of (ϕ, ψ) values for a given residue type. Therefore, any disallowed value for a dihedral suggested by the Ramachandran map, whenever it appears in the conformation tree, is pruned. We used the data from Ref. 93, and implemented a *residue-specific* Ramachandran filter. Our implementation considers four residue types: Gly, Pro, pre-Pro, and other general amino acid types (called *general*). It has been specifically optimized for $O(1)$-time queries for the *favored* or *allowed* intervals for ϕ, and ψ given ϕ. If $M_{\mathcal{T}}$ is the Ramachandran map for residue type $\mathcal{T}$, and $I_{\mathcal{T}}$ is the set of all allowed ϕ-intervals for $\mathcal{T}$, we evaluate if $\phi \in I_{\mathcal{T}}$ for a computed ϕ. Similarly, when a ψ is computed, we evaluate if $\psi \in I_{\mathcal{T}}|_{\phi}$. TALOS[76,77] dihedral information, whenever available, are used as follows. If for the dihedral ϕ_i of the residue i of type $\mathcal{T}$, $I_{\mathcal{L}}$ is the TALOS-predicted interval, then for a computed ϕ for the residue i, we evaluate if $\phi \in I_{\mathcal{T}} \cap I_{\mathcal{L}}$. Similarly, for a computed ψ, the predicate $\psi \in I_{\mathcal{T}}|_{\phi} \cap I_{\mathcal{L}}$ is evaluated. The subtree rooted at the node representing the dihedral is pruned if any of these predicates fail. Further, in the absence of RDC data for a dihedral, finite-resolution uniform sampling of the Ramachandran map is used for that dihedral.

Steric filter

We use our in-house implementation of the steric checker similar to that in Ref. 94. During the depth-first search of the conformation tree, at each node corresponding to a newly added residue, the steric check is performed for (i) self-collision, that is, if the fragment clashes with itself, and (ii) collision with the rest of the protein. If the clash score[94] is greater than a user-defined threshold, then the branch is pruned and the search backtracks.

Reachability criterion

As each node of the conformation tree is visited, we test if the rest of the fragment, if grown using the best possible kinematic chain, can ever reach the stationary anchor. The node is pruned if this test fails. For long loops, this test prunes a large fraction of conformations, especially at the tree nodes at higher levels (farther from the root).

Closure criterion

When the distance between the mobile anchor (i.e., the conformation at a leaf node) and the stationary anchor is less than a user-specified threshold (chosen to be 0.2 Å), called the *closure distance*, and defined as the root-mean-square distance between the N, C^{α}, and C' atoms of the mobile anchor and stationary anchor, the conformation is accepted and added to the ensemble of computed loops. Otherwise, the conformation is subject to a minimization over the last few dihedrals to improve the closure distance to below 0.2 Å while maintaining the user-defined RDC RMSD thresholds. If after minimization the closure is achieved, the conformation is accepted; otherwise, rejected. The RDC RMSD between back-computed and experimental RDCs is computed using the equation $\mathrm{RMSD}_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r^b_{x,i} - r^e_{x,i})^2}$, where x is either a ϕ-defining or a ψ-defining RDC type, n is the number of RDCs, $r^e_{x,i}$ is the experimental RDC, and $r^b_{x,i}$ is the corresponding back-computed RDC.

Pruning using unambiguous NOEs

When unambiguous backbone NOEs are available, they can be used as predicates to prune unsatisfying conformations.

NMR experimental procedures

The NMR data for FF2 was recorded and collected using Varian 600 and 800 MHz spectrometers at Duke University. NMRPipe[95] was used to process the NMR spectra. All NMR peaks were picked by the programs NMRView[96] or Xeasy/Cara,[97] followed by manual editing. Backbone assignments were obtained from the set of triple resonance NMR experiments HNCA, HN(CO)CA, HN(CA)CB, HN(COCA)CB, and HNCO, combined with the HSQC spectra using the program Paces,[98] followed by manual checking. The C^{α}–H^{α} and N–H^{N} RDC data for FF2 was measured from a 2D ^{1}H–^{15}N IPAP experiment[99] and a modified (HACACO)NH experiment,[100] respectively. The C^{α}–C' and C'–N RDCs of FF2 were measured from a set of HNCO-based experiments.[101] The RDC data for human ubiquitin (PDB id: 1d3z),[102] the DNA damage inducible protein I (DinI; PDB id: 1ghh),[103] and the third IgG-binding domain of Protein G (GB3; PDB id: 2oed)[104] were obtained from the BioMagResBank (BMRB).[105]

RESULTS AND DISCUSSION

To study the effectiveness of our algorithm POOL, we tested it on experimental NMR datasets for four proteins. We further tested POOL on synthetic datasets for three sets of canonical loops of length 4, 8, and 12 residues that were investigated by three previous protein loop closure algorithms.[26,28,78] To further assess the added value of NMR restraints, we used the same set of twenty 12-residue long loops published in Ref. 13, for which we simulated RDCs as described in Supporting Information Appendix B. POOL was then used to compute the loop conformations using these synthetic RDCs. The results, in addition to providing a way to compare our algorithm with these loop prediction approaches, enable us to study the robustness of our algorithm to minor variations in standard peptide geometry. We further show that in the presence of a moderate level of dynamics, POOL can compute an ensemble of near-native loop conformations from sparse RDC measurements.

Tests on experimental NMR data

We applied POOL to compute the loops of four proteins: FF2 (PDB id: 2kiq),[71] human ubiquitin (PDB id: 1d3z),[102] DinI (PDB id: 1ghh),[103] and GB3 (PDB id: 2oed).[104] The experimental details of RDC data collection is provided in subsection NMR experimental procedures. For each of these proteins, we used the NMR Model 1 with loops removed as the respective test structures. RDCs were perturbed within the experimental-error window[61] to account for experimental errors. The following were input to POOL: (1) the core, that is, the SSEs of the NMR models with no loops on it; (2) the alignment tensor computed from the core of the respective NMR models and the experimental RDCs using singular value decomposition (SVD)[61,82]; (3) RDC data for the loop in one alignment medium; and (4) the primary sequence of the loop to instantiate the appropriate residue-specific Ramachandran map. POOL was then invoked to compute the loops.

Table II summarizes the results computed by POOL. For ubiquitin two different combinations of RDCs, viz.

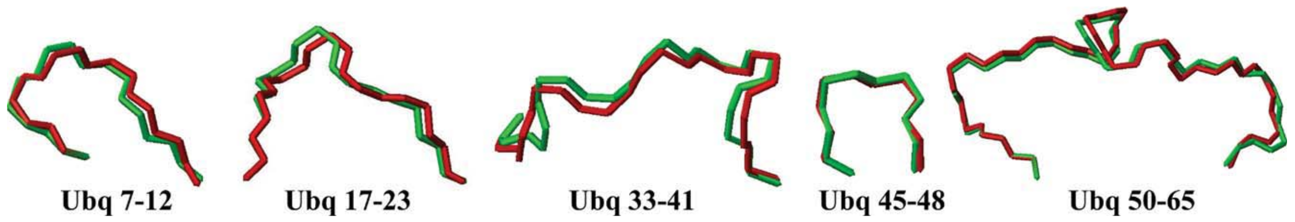


Figure 4

Overlay of the lowest RMSD loops (green) of ubiquitin computed by POOL using C^{α}–H^{α} and N–H^{N} RDCs versus the corresponding loops (red) in the NMR reference structure (1d3z Model 1) without any structural alignment. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

(C^{α}–H^{α}, N–H^{N}) and (C^{α}–C′, N–H^{N}), were used to analytically compute ensembles of satisfying loop conformations in order to test the performance of POOL on different types of RDC data. In most cases, *sub-angstrom* RMSD loops were computed by POOL. Figure 4 shows the overlay of the minimum RMSD loops computed for ubiquitin using C^{α}–H^{α} and N–H^{N} RDCs with the corresponding loops from the NMR reference structure. No structural alignment of the computed loops with the reference loops was done during the overlay, and while computing the backbone RMSD. This is because any such alignment, while likely to yield a lower backbone RMSD value, can invalidate the loop closure due to the translation and rotation of the mobile anchors away from the stationary anchors induced by the local structural alignment. In addition, a loop conformation reoriented during the local structural alignment can no longer be in the same POF as the stationary anchors and the other SSEs; and therefore, will no longer fit the RDCs. For FF2, DinI and GB3, the results show that POOL is able to compute accurate loops when as many as five RDCs are missing.

The algorithm POOL can be regarded as a conformation generator that computes an ensemble of all satisfying loop conformations from as few as two RDCs per residue. Whenever additional RDCs are present, POOL can use them as a *filter* during the conformation tree-search, to prune the conformation tree (Fig. 1) encoding the analytic solutions, as described in Ref. 71. To test the ability of POOL to compute ensembles of near-native loop conformations, we performed the following computational experiment using RDCs for ubiquitin in one alignment medium. POOL used C^{α}–H^{α} and N–H^{N} RDCs to compute the analytic solutions for the backbone dihedrals to build the conformation tree. An additional set of C^{α}–C′ RDCs was used as a filter to prune the conformation tree.[71] Figure 5(a) shows an ensemble of 48 loop conformations (green) for the loop 50–65 of ubiquitin computed by POOL. The corresponding loop from the NMR reference structure is shown in red. In Figure 5(b) the ensembles of loops (green) computed for all of the ubiquitin loop regions are shown. The NMR reference structure is shown in red. Table III summarizes the ensemble of loops computed by POOL for each ubiquitin loop, computed as described earlier. The results in Table III show that POOL computed sub-angstrom accuracy loop conformations in every case, and low-RMSD near-native loop conformations were often included in the ensemble. When additional RDCs are not available, or an all-atom energy-based refinement of the loop conformations is needed, the set of loops generated by POOL can in principle be evaluated using a molecular mechanics energy function to obtain an ensemble of low-energy loop conformations.[10,18,34]

The run-time complexity analysis of POOL is similar to that in Ref. 62. In practice, for short loops, POOL runs in minutes, and for longer loops (e.g., ubiquitin 50–65) it runs in hours on a 2.5 GHz dual-core processor Linux workstation.

Comparison with other structure determination protocols

To investigate whether traditional SA/MD-based structure determination protocols can compute accurate loop conformations using sparse data, we ran XPLOR-NIH[53] on the same input used by POOL for ubiquitin, FF2, DinI, and GB3. Table II summarizes the results. In Figure 6, a comparison is made between the results obtained by applying POOL versus those obtained by applying XPLOR-NIH. The loops computed by POOL have much smaller (three- to six-fold less for longer loops) backbone RMSD versus the reference structures than those computed using XPLOR-NIH. For example, for ubiquitin loop 50–65, the loop computed by POOL has backbone RMSD 0.66 Å, a six-fold decrease versus the loop computed by XPLOR-NIH (3.94 Å). This shows that when given sparse data, our algorithm is able to compute more accurate loop conformations than the SA/MD-based protocols.

Further, to compare our method with other sparse data protocols, we used the CS-ROSETTA[75,106,107] protocol with ROSETTA-3.2,[108] with the same set of NMR restraints including the RDCs used by POOL, in addition

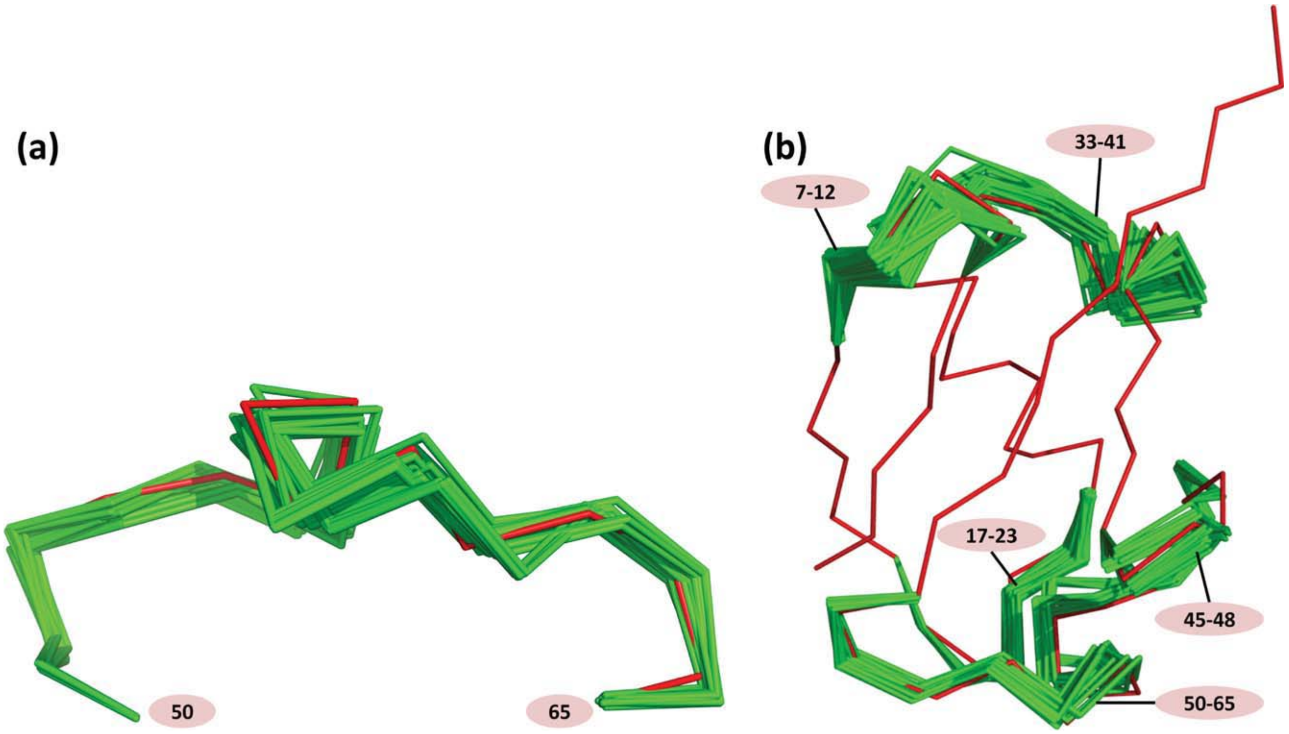


Figure 5

(**a**) Overlay of an ensemble of 48 loop conformations (green) for the loop 50–65 of ubiquitin computed by POOL using C^{α}–H^{α} and N–H^{N} RDCs and filtered against C^{α}–C′ RDCs, versus the corresponding loop (red) in the NMR reference structure (1d3z Model 1) without any structural alignment. (**b**) Overlay of the ensembles of all five ubiquitin loops (green) computed by POOL using C^{α}–H^{α} and N–H^{N} RDCs and filtered against C^{α}–C′ RDCs, versus the NMR reference structure (1d3z Model 1) shown in red without any structural alignment. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

to the backbone chemical shifts required by CS-ROSETTA. Since POOL and CS-ROSETTA are both sparse-data algorithms, we believe that the comparison made here sheds some light on the limits on the sparsity of the experimental NMR data which can be used to determine loop conformations, and for structure determination in general, in addition to assessing the relative performance of these two algorithms based on two completely different algorithmic techniques. For FF2, ubiquitin, and GB3, experimental chemical shifts were used. For DinI, since the chemical shifts are not available in the BMRB,[105] the backbone chemical shifts were simulated using SHIFTX2.[109] Experimental RDCs were used for all four proteins. In each case, 3000 ROSETTA models were generated, and the lowest-energy models were selected and aligned with the respective NMR reference models by the SSEs. The backbone RMSDs between CS-ROSETTA-computed loops and the reference NMR loops for the loop regions were then computed. Table II summarizes the results for the loops computed by CS-ROSETTA for

Table III

Summary of the POOL-Ensembles of Loops for Ubiquitin Computed Using C^{α}–H^{α} and N–H^{N} RDCs and Filtered Against C^{α}–C′ RDCs, in One Alignment Medium

Ubiquitin Loop	Ensemble Size	RMSD of reference loop to the ensemble (min,max)	Average RMSD of reference loop to the ensemble	Average RMSD to mean coordinates
Ubiquitin 7–12	66	0.64, 1.64	1.30	1.17
Ubiquitin 17–23	214	0.60, 1.40	0.93	0.53
Ubiquitin 33–41	67	0.89, 1.95	1.45	0.95
Ubiquitin 45–48	28	0.27, 0.85	0.59	0.72
Ubiquitin 50–65	48	0.66, 1.50	0.94	0.88

The backbone RMSDs and their averages for the POOL-ensembles were computed without any structural alignment. The reference loops are obtained from the NMR reference structure (PDB id: 1d3z) Model 1. The reference loops and the POOL-ensembles are in the same POF.

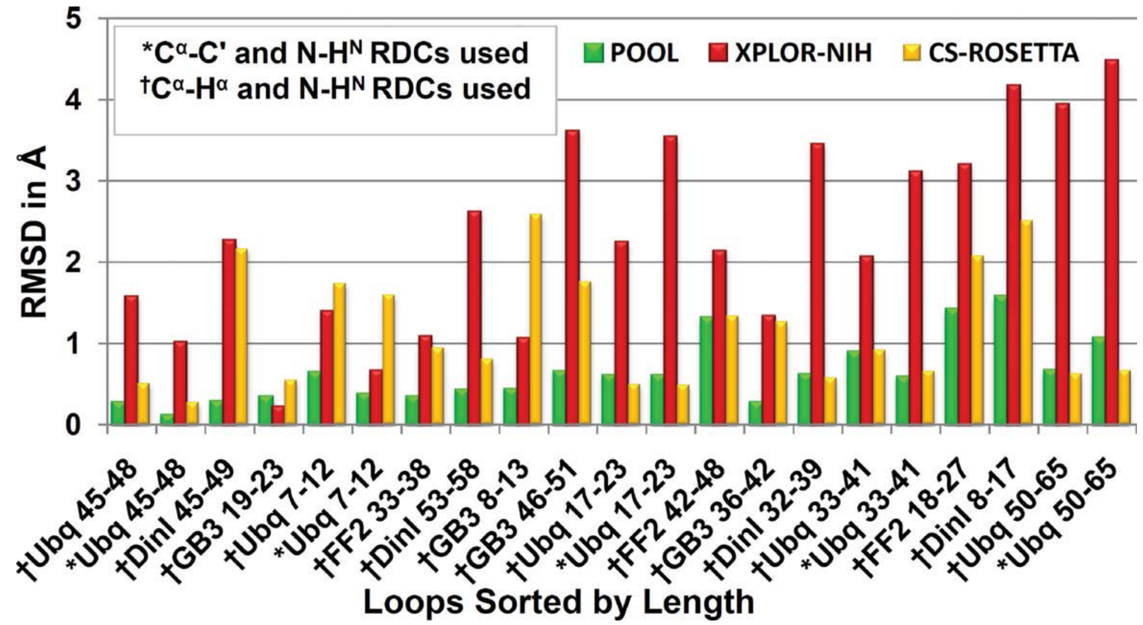


Figure 6

The loops computed by POOL achieve up to six-fold improvement in backbone RMSD compared to loops computed by XPLOR-NIH. The loops computed by POOL are more accurate than those computed by CS-ROSETTA in 16 out of 21 cases. CS-ROSETTA used the same RDCs as POOL, plus the backbone chemical shifts. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

comparison with POOL. While the POOL-computed loops have RMSDs within the range of 0.11–1.57 Å versus the reference NMR structures, CS-ROSETTA-computed loops have RMSDs within a bigger range of 0.28–2.59 Å. The average RMSD over all the loops computed by POOL is 0.64 Å, whereas the average RMSD over all the loops computed by CS-ROSETTA is 1.17 Å. This shows that while CS-ROSETTA performed comparably to POOL for the ubiquitin loops in four out of 10 cases, and for one loop in DinI, for other ubiquitin loops, and other proteins, POOL performed better (in total, for 16 out of 21 cases; see Fig. 6). These results show that our algorithm POOL, using a novel polynomial equation-based approach to compute loop conformations from sparse RDC constraints, performs better than CS-ROSETTA, for the above test cases.

Comparison with loop prediction algorithms

We compared the performance of POOL with three other loop prediction algorithms including the CCD method by Canutescu and Dunbrack,[28] the CSJD algorithm by Coutsias *et al.*,[26] and the self-organizing superimposition (SOS) algorithm by Liu *et al.*[78] Furthermore, we compared the performance of our algorithm with the kinematic closure (KIC) protocol by Mandell *et al.*,[13] which uses a resultant-based analytic loop closure method[27] to serve as the loop conformation generator within the ROSETTA[108,110] framework. Unlike these algorithms, which do not use any data, POOL is a sparse data-driven algorithm. Although CCD, CSJD, SOS, and KIC algorithms have applications in protein structure prediction,[11–13] none of them is specifically designed to incorporate geometric restraints from experimental NMR data. Our algorithm POOL exploits this opportunity, and provides an approach to compute loops using sparse NMR data, specifically, RDCs.

In our study, we used the same test set as in Refs. 26,28,78. This set consists of 10 loops each with 4, 8, and 12 residues chosen from a set of nonredundant X-ray crystallographic structures from the PDB. Since there is no experimental RDC data available for these proteins, we simulated the RDCs using PALES.[111,112] Gaussian noise of 1 Hz was added to the RDCs to simulate experimental error. Details of the RDC simulation are described in Supporting Information Appendix B. The alignment tensor, the RDC data, and the two anchors of the loop were used by POOL to compute the loop conformations.

Table IV summarizes the results for POOL, CCD, CSJD, and SOS algorithms. Figure 8 shows a graphical comparison of the results obtained by these algorithms. In Figure 7, examples of minimum RMSD loop conformations determined by POOL are shown. For the 4-residue loops the average minimum RMSD of the computed loops by POOL is larger than that for SOS, but smaller than that for CSJD and CCD. This can be explained by the fact that SOS allows slight deviations from standard protein geometry. For the 8- and 12-residue loops POOL computes more accurate loops than other algorithms. For example, for

Table IV

The Minimum RMSD (Å) from X-ray Structures for these Four Algorithms

4-Residue loops					8-Residue loops					12-Residue loops				
Loop	POOL	SOS	CSJD	CCD	Loop	POOL	SOS	CSJD	CCD	Loop	POOL	SOS	CSJD	CCD
1dvjA_20	0.74	0.23	0.38	0.61	1cruA_85	0.72	1.48	0.99	1.75	1cruA_358	1.54	2.39	2.00	2.54
1dysA_47	0.25	0.16	0.37	0.68	1ctqA_144	0.91	1.37	0.96	1.34	1ctqA_26	0.65	2.54	1.86	2.49
1eguA_404	0.42	0.16	0.36	0.68	1d8wA_334	0.28	1.18	0.37	1.51	1d4oA_88	1.83	2.44	1.60	2.33
1ej0A_74	0.18	0.16	0.21	0.34	1ds1A_20	0.70	0.93	1.30	1.58	1d8wA_46	0.93	2.17	2.94	4.83
1i0hA_123	0.27	0.22	0.26	0.62	1gk8A_122	0.87	0.96	1.29	1.68	1ds1A_282	1.50	2.33	3.10	3.04
1id0A_405	0.63	0.33	0.72	0.67	1i0hA_122	0.45	1.37	0.36	1.35	1dysA_291	0.76	2.08	3.04	2.48
1qnrA_195	0.47	0.32	0.39	0.49	1ixh_106	0.68	1.21	2.36	1.61	1eguA_508	1.25	2.36	2.82	2.14
1qopA_44	0.36	0.13	0.61	0.63	1lam_420	0.42	0.90	0.83	1.60	1f74A_11	0.76	2.23	1.53	2.72
1tca_95	0.12	0.15	0.28	0.39	1qopB_14	0.87	1.24	0.69	1.85	1qlwA_31	1.27	1.73	2.32	3.38
1thfD_121	0.25	0.11	0.36	0.50	3chbD_51	0.96	1.23	0.96	1.66	1qopA_178	0.87	2.21	2.18	4.57
Average	**0.37**	**0.20**	**0.40**	**0.56**	**Average**	**0.69**	**1.19**	**1.01**	**1.59**	**Average**	**1.14**	**2.25**	**2.34**	**3.05**

The loops computed by POOL using only one ϕ-defining and one ψ-defining RDC per residue simulated as described in Supporting Information Appendix B. SOS, CSJD, and CCD results were obtained from Table 1, Table 1 and 2 of Refs. 78, 26, and 28, respectively. These three methods do not use any experimental NMR data.

the 12-residue loops, the average minimum RMSD of the loops are 1.14, 2.25, 2.34, and 3.05 Å for POOL, SOS, CSJD, and CCD, respectively, which shows a two-fold improvement in accuracy by POOL. For five of these loops, POOL computed loops with sub-angstrom accuracy. In a recent work by Lee *et al.*,[32] the authors developed an approach based on a combination of fragment assembly and analytical loop closure to model loops. The average backbone RMSDs of the loops computed by their algorithm, which can be found in Table II of Ref. 32, for the same set of loops in Table IV and Figure 8 of length 4, 8, and 12 are 0.22, 0.72, and 1.81 Å, respectively. When compared with the results from POOL (0.37, 0.69, and 1.14 Å for loops of length 4, 8, and 12, respectively), it illustrates that POOL can compute loops with better accuracy for longer loops.

In a recent work by Mandell *et al.*,[13] the authors developed a robotics-inspired conformational sampling and loop reconstruction method, called KIC, and showed that the KIC protocol frequently samples conformational space within 1.0 Å from the X-ray crystallographic reference loops. To assess the added value of the NMR restraints, we used the same set of twenty 12-residue long loops published in Ref. 13, and simulated RDCs as described in Supporting Information Appendix B. POOL was then used to compute the loop conformations using these synthetic RDCs. POOL computed better loop conformations than the previous methods. The mean backbone RMSD of the loop conformations computed by POOL is 0.89 Å (Table V and Fig. 9), which is more than two-fold improvement in accuracy compared with the standard ROSETTA[110] and KIC de novo protocols. In Ref. 13, the authors also computed loop conformations starting from a set of loop conformations, called therein "perturbed," that began with the starting loop conformations that are sampled away from the native X-ray loop conformations published in Ref. 113. Starting with perturbed X-ray loops may be viewed as providing KIC with a structural

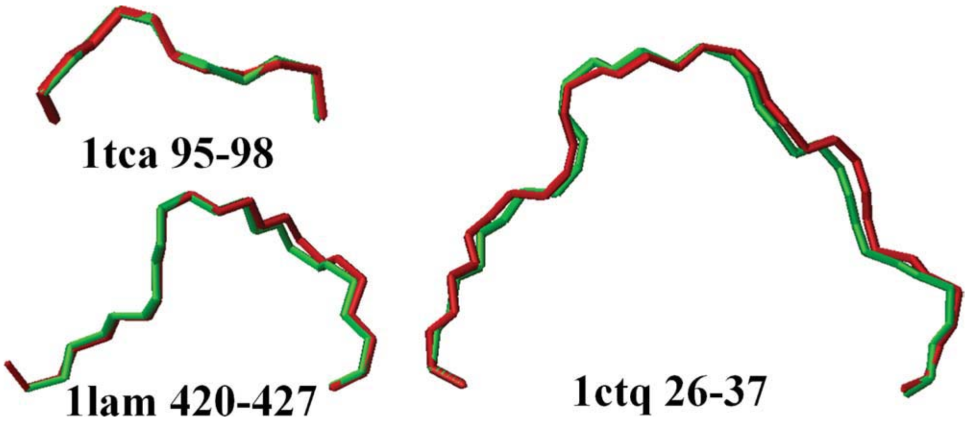


Figure 7

Overlay of the lowest RMSD loops (green) computed by POOL for 4-, 8-, and 12-residue loops versus the X-ray structures of the reference loops (red) without any structural alignment. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

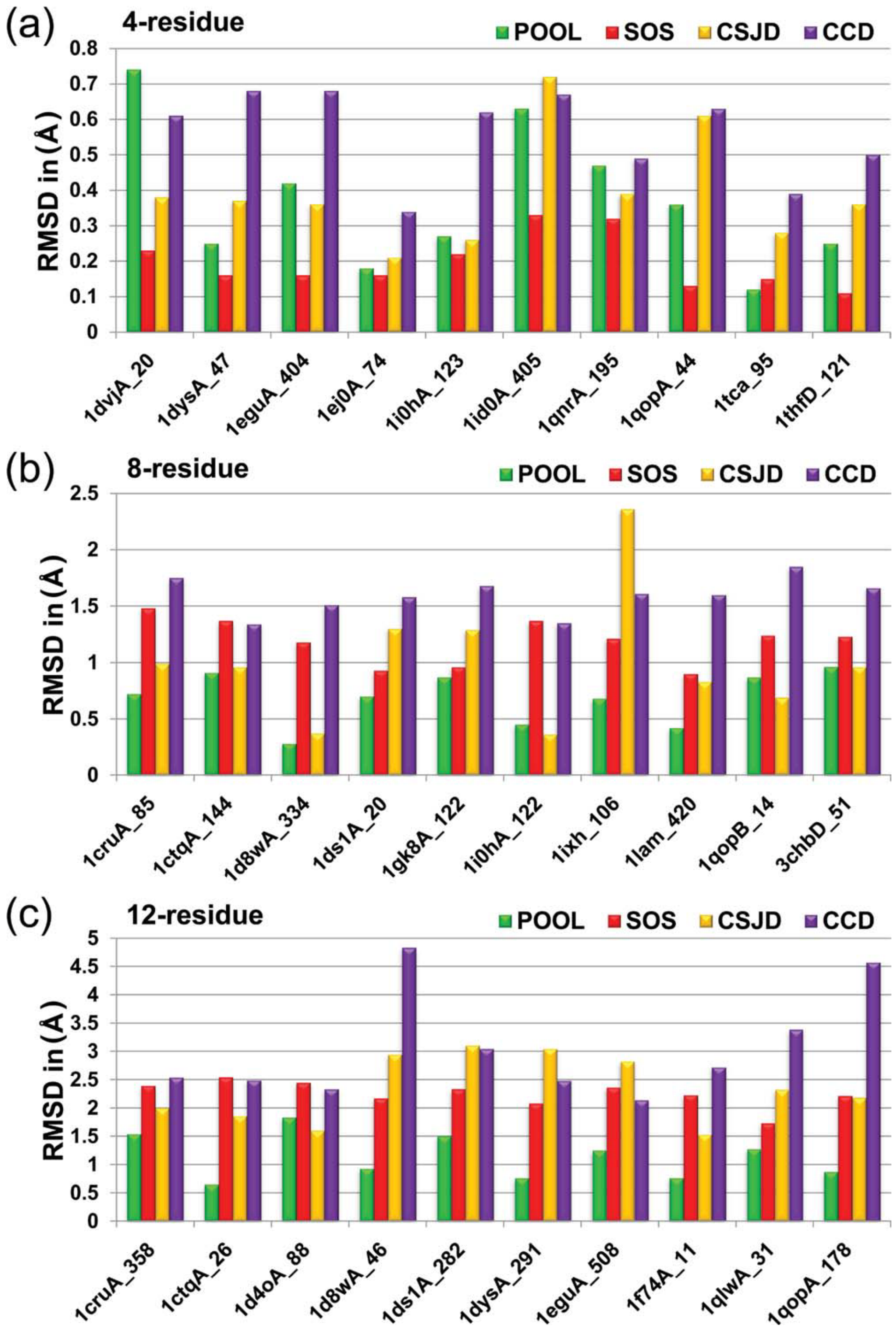


Figure 8

Comparison of the results from POOL, SOS, CSJD, and CCD algorithms when applied on (**a**) 4-residue loops, (**b**) 8-residue loops, (**c**) 12-residue loops. The data comes from Table IV. For 8- and 12-residue loops POOL computes loops with higher accuracy than other methods. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

prior. However, the KIC de novo protocol does not use any structural prior. KIC performed better (than KIC de novo) when started with the perturbed X-ray loop conformations, with a mean backbone RMSD of 1.6 Å. Our algorithm POOL, in contrast, does not use any such structural prior, but computes the loop conformations de novo from the data-derived orientational restraints, and performed 1.8-fold better than KIC. Fifteen out of the

Table V

Performance in Terms of Backbone RMSD (Å) of POOL, KIC Protocol[13] and Standard ROSETTA Protocol[110] on a Set of X-ray Reference Loops of Length 12 Published in Refs. 13, 113

Loop	POOL	KIC de novo protocol	KIC starting from perturbed X-ray loops[a]	Standard ROSETTA de novo	Standard ROSETTA starting from perturbed X-ray loops[b]
1a8d_155	0.93	6.9	0.6	5.4	5.3
1arb_182	0.69	1.0	1.4	1.6	5.1
1bhe_121	0.92	0.8	0.7	7.1	4.9
1bn8_298	1.28	0.8	0.6	2.5	1.7
1c5e_83	0.62	0.5	0.4	0.8	5.1
1cb0_33	0.93	0.6	0.7	1.0	1.1
1cnv_188	0.34	1.4	2.1	2.3	2.8
1cs6_145	1.79	3.0	3.0	2.5	4.0
1dqz_209	0.70	0.7	2.6	1.9	1.8
1exm_291	1.51	0.9	0.9	0.6	2.8
1f46_64	1.34	2.5	2.3	2.1	0.7
1i7p_63	0.57	2.7	0.4	0.7	0.8
1m3s_68	0.64	6.3	5.6	3.6	2.2
1ms9_529	0.95	0.4	1.0	2.5	2.8
1my7_254	0.95	2.3	2.3	2.0	0.6
1oth_69	0.65	0.6	0.6	0.6	1.9
1oyc_203	0.79	4.0	3.9	3.2	1.7
1qlw_31	1.27	1.0	0.9	3.3	5.0
1t1d_127	0.56	0.8	0.8	0.5	0.6
2pia_30	0.45	1.0	0.9	1.1	1.0
Average	**0.89**	**1.9**	**1.6**	**2.3**	**2.6**

The loops were computed by POOL using only one ϕ-defining and one ψ-defining RDC per residue simulated as described in Supporting Information Appendix B. The minimum RMSD loop from the computed ensemble is reported here. Results for KIC and ROSETTA protocols were obtained from the Supplementary Table 2 of Ref. 13.

[a,b]Simulations reported in Ref. 13, called therein "perturbed," that began with the starting loop conformations that are sampled away from their X-ray conformations published in Ref. 113. KIC and ROSETTA protocols do not use any experimental NMR data.

twenty loops computed by POOL had backbone RMSD less than 1.0 Å, which shows its ability to consistently compute near-native loop conformations using sparse RDC data.

Further, the reference loops in Tables IV and V have deviations from standard protein geometry; therefore, the RDCs simulated on them inherit these deviations, in addition to a Gaussian noise of 1 Hz added to account for experimental errors. These results suggest that POOL is robust to both experimental uncertainties in RDCs, and minor deviations from standard protein geometry assumptions. Therefore, POOL can be useful to compute longer loops with high accuracy using a minimal amount of RDC data.

Computing near-native loop conformations in the presence of dynamics

RDCs provide a sensitive probe to protein conformational dynamics[114–117] from nanosecond to millisecond timescales. Loop regions of proteins are usually more flexible, and their motional fluctuations often contribute to their recognition dynamics. To study the ability of our algorithm to compute loop conformations in the presence of a moderate level of dynamics, we chose to work on ubiquitin, which is an important protein involved in various biological processes such as protein degradation and signal transduction. While the structure and function of ubiquitin have been well studied, its dynamics and implications for function have been an active area of research.[117–119] In general, RDCs in at least five independent alignment media are required to extract dynamics information from RDCs.[120] In a recent study using RDCs measured in 36 different alignment media, Lange *et al.*[117] demonstrated that unbound ubiquitin samples conformations similar to those found in ubiquitin complexes; thus, providing evidence of conformational selection, rather than induced-fit motion, for the binding process of ubiquitin.

Since POOL is a sparse-data algorithm, it is not possible to probe protein dynamics directly, since the sparse amount of data leads to an underdetermined system from the dynamics viewpoint. However, we show that in the presence of moderate dynamics, such as those found in ubiquitin loop regions, POOL can compute near-native ensembles of loop conformations with high accuracy. Only RDCs in one alignment medium were used in our study. We focused on two loops of ubiquitin, between residues Thr7–Thr12 and Phe45–Lys48. To reduce the effect of rigid loop anchors obtained from ubiquitin NMR Model 1, we extended the loop region by one residue in either end; therefore, POOL computed an eight- and a six-residue loop, (Lys6–Ile13 and Ile44–Gln49, respectively). Henceforth, we refer to these two loops as the 6–13 and 44–49 loops, respectively. POOL used C^{α}–H^{α} and N–H^{N} RDCs to compute the analytic solutions for the backbone dihedrals to build the conformation tree of loops that satisfy both the loop closure criterion and the RDCs. Sampling RDCs from the experimental error window alone is not sufficient to account for the variations in RDC magnitudes due to internal dynamics causing fluctuations of the internuclear vector orientations.[117] Therefore, we sampled the RDCs from a normal distribution with wider Gaussian intervals, with standard deviations σ of 1.5 and 2.0 Hz for N–H^{N} and C^{α}–H^{α} RDCs, respectively. Roughly 32% of the sampled RDC values were expected to deviate from the recorded values by more than σ, and can be in the range of 3σ, or more.[69] An additional set of C^{α}–C' RDCs was used as a filter to prune the conformation tree.[71] For the 6–13 and 44–49 loops, ensembles of 232 and 229 loop conformations were respectively computed by POOL after filtering against C^{α}–C' RDCs. To analyze the conformational variations and properties of these loop ensembles, we obtained the 46 ubiquitin X-ray structures from the PDB that were used in the study of Ref. 117. First, these X-ray structures were protonated using the REDUCE module of MOLPROBITY,[121–123] and then aligned with ubiquitin NMR Model 1 by their SSEs only. Then, the loops from these X-ray structures were extracted. As the X-ray-ensemble of

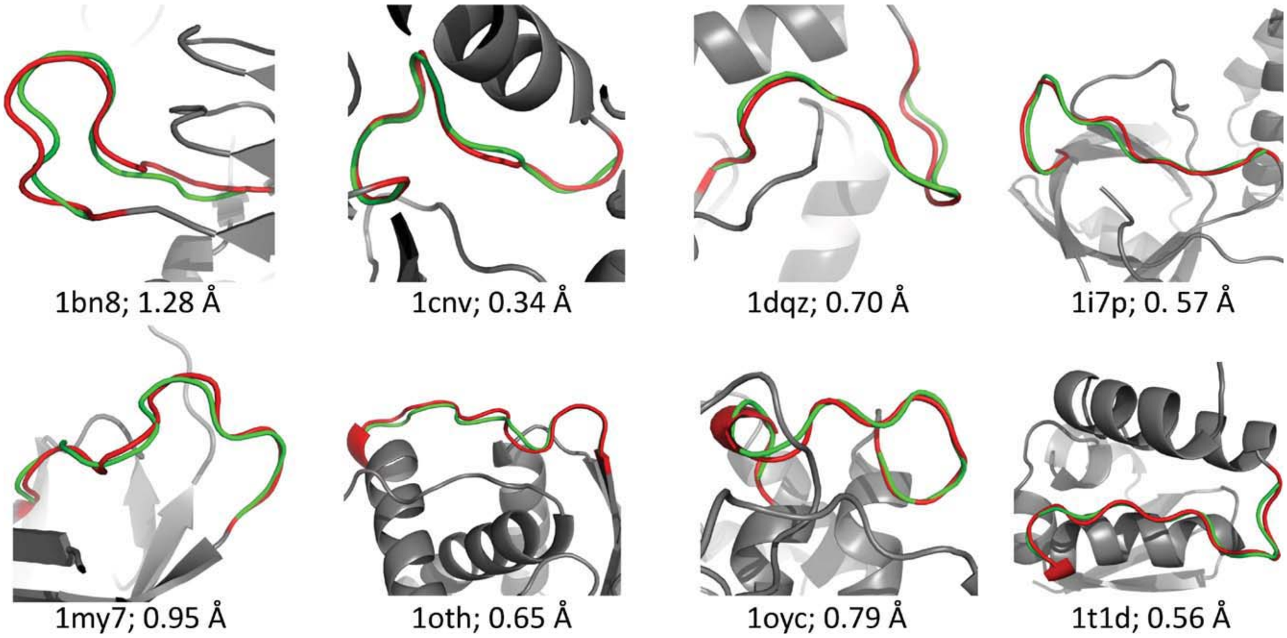


Figure 9

A representative set of eight 12-residue loops (see Table V) computed by POOL are shown in green. Shown in red are the corresponding loops in the reference X-ray crystallographic structures. The corresponding PDB ids and the backbone RMSDs between the POOL-computed loops and the reference X-ray loops are shown. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

loops and the POOL-ensemble of loops were in the same reference frame and aligned, this allowed us to directly compare the POOL-ensemble of loops with the X-ray-ensemble.

Figure 10(a) and (e) show the overlay of the X-ray-ensemble versus the POOL-ensemble for 6–13 and 44–49 loops, respectively. We used the following simple measure of similarity between the X-ray-ensemble and the POOL-ensemble. For the loop 6–13, for each loop in the X-ray-ensemble, we computed the nearest loop (in terms of backbone RMSD) in the POOL-ensemble. This is shown in Figure 10(b), which plots the number of X-ray loops (out of 46) on the *y*-axis within the distance interval measured from the POOL-ensemble on the *x*-axis. For example, 13 X-ray loops are within 0.6–0.8 Å, and 19 X-ray loops are within 0.8–1.0 Å from the POOL-ensemble. The distance of the POOL-ensemble from the X-ray-ensemble was similarly computed [Fig. 10(c)]. Since a large fraction of loops in each ensemble have low-RMSD nearest loops in the other ensemble, and vice versa, Figure 10(b) and (c) together indicate that the two ensembles have considerable similarity. An identical analysis for the loop 44–49, shows the similarity between the X-ray-ensemble and POOL-ensemble [Figure 10(f) and (g)]. Figure 10(d) and (h) plot the spread of both the X-ray-ensemble and the POOL-ensemble with respect to each of the X-ray loops, which show that the POOL-ensemble covers the X-ray-ensemble well, and therefore, captures the conformational diversity. In Figure 10(h), it can be seen that the POOL-ensemble contains most of the X-ray-ensemble, but is larger. This can be explained, in part from the fact that the sparse amount of RDCs and backbone kinematics here allowed the exploration of a larger conformation space, and in part by the solution-state dynamics sampling a larger conformation space than the frozen X-ray structures. In addition, we believe that the relatively higher backbone RMSD for some of the loops in both the ensembles is due to the kinematic constraints imposed on the POOL-ensemble by the rigid anchors, which can limit the exploration of certain regions of the conformational space.

The ubiquitin experimental RDCs should reflect an ensemble average over an ensemble of loop conformations that is subject to conformational selection during binding and other protein functions. However, these loop conformations may have very different populations, and the relative population weighting affects the RDC ensemble average. Although these relative populations for the ubiquitin loops are not known from experimental measurements, in the following computational experiment, we used the aforementioned 46 X-ray loops for the ubiquitin loop 6–13, and computed a simulated ensemble-averaged RDC dataset assuming they had equal populations. The goal was to see how much of the correct ubiquitin ensemble could be recovered from this simulated RDC data. This is a challenging loop reconstruction problem since the back-

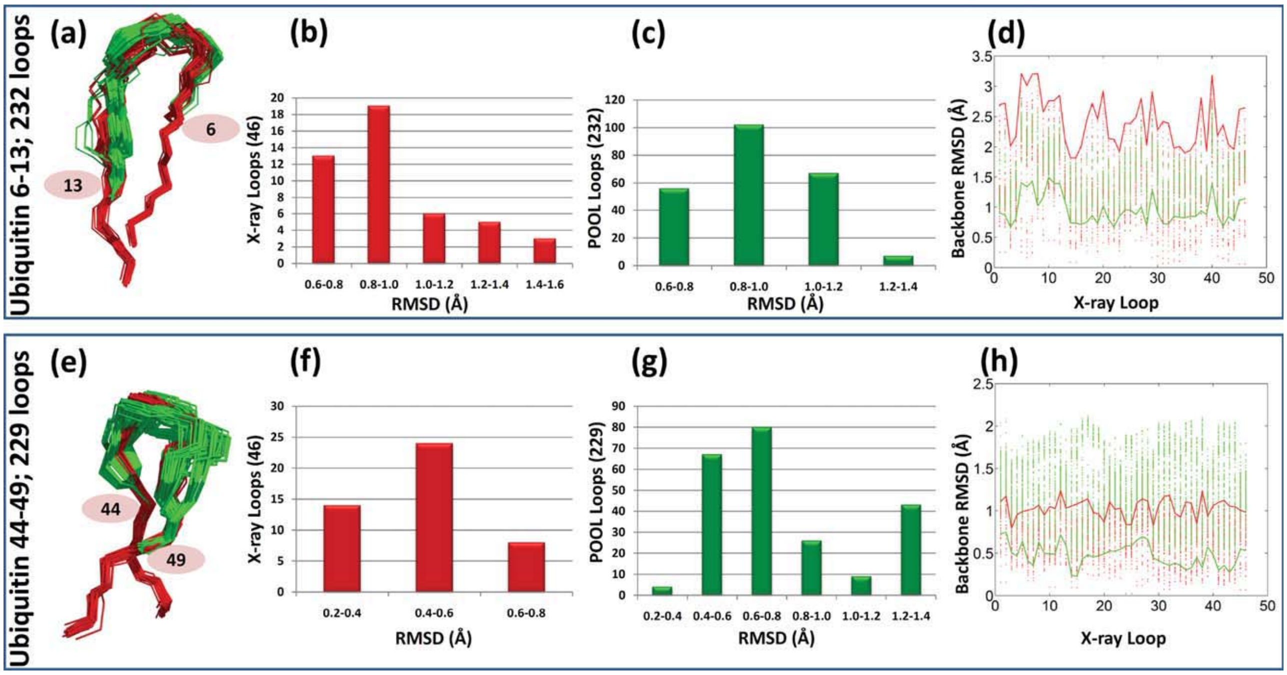


Figure 10

(**a**) Overlay of the loop ensemble containing 232 loops computed by POOL (green) for the loop 6-13 of ubiquitin versus the loops extracted from the 46 X-ray structures of ubiquitin (red). (**e**) Overlay of the loop ensemble containing 229 loops computed by POOL (green) for the loop 44-49 of ubiquitin versus the loops extracted from the 46 X-ray structures of ubiquitin (red). (**b** and **f**) The number of X-ray loops in *y*-axis within the backbone RMSD interval measured from POOL-ensemble in *x*-axis. (**c** and **g**) The number of loops from POOL-ensemble in *y*-axis within the backbone RMSD interval measured from X-ray-ensemble in *x*-axis. (**d** and **h**) For each of the 46 X-ray loops, the backbone RMSDs of the X-ray loops are shown using red dots, and the backbone RMSDs of the loops from the POOL-ensemble are shown using green dots. The minimal backbone RMSDs for POOL-ensemble of loops are shown as green lines. The maximal backbone RMSDs for X-ray loops are shown as red lines.

bone RMSDs of the 46 X-ray loops vary up to 3.2 Å [see the red line in Figure 10(d)] in span, which is large for a 8-residue segment. This simulated ensemble-averaged RDC dataset was used by POOL to compute an ensemble of 295 loop conformations. By doing a similar analysis as earlier, we observed that the computed POOL-ensemble covers the X-ray-ensemble, and each member in the POOL-ensemble was within 1.2 Å from the X-ray-ensemble, suggesting that POOL can be used successfully to compute an ensemble of loops from ensemble-averaged RDCs representing a dynamic ensemble.

While our algorithm POOL has been shown to work in the presence of a moderate level of dynamics, as such it does not characterize protein dynamics explicitly, specifically, due to the fact that the system it solves is underdetermined from the standpoint of probing dynamics. Therefore, model bias cannot be ruled out. However, results obtained for ubiquitin loops 6–13 and 44–49 suggest a future direction, to probe protein dynamics using our polynomial equation-based approach with RDCs measured in a relatively fewer number of alignment media. We envision that while such an approach can lead to a new methodological development, it would complement the current methods by providing an alternative way of exploiting the geometric information from the RDC data.

CONCLUSIONS

While the global fold of a protein can often be determined from experimental NMR data,[59,61,62,64,71] determining loop conformations from sparse NMR data is a difficult problem. We described a novel, efficient, and practical deterministic algorithm, POOL, that determines accurate loop conformations from sparse RDC data. Empirical comparisons with traditional structure determination protocols,[53] and also with previous sparse-data protocols,[75] demonstrate that POOL performs better than these approaches when using sparse data.

Previous approaches, such as Ref. 32, SOS, CSJD, and CCD randomly sample the conformation space of the loop to compute an ensemble of loop conformations in a generate-and-test fashion, that can subsequently be filtered against experimental data. In contrast, POOL takes a complementary approach, and uses constraint posting in a rigorous algorithmic framework to restrict the solution space by exploiting the algebra of RDCs and protein kinematics, to compute a complete set of loop conformations that satisfy the RDC data. While a minimal amount of RDCs are required by POOL, additional distance, orientational, and torsion-angle constraints, whenever available, can be directly incorporated into our framework.

Since an accurate and complete protein backbone is a prerequisite for NOE-assignment algorithms[50,71] and side-chain resonance assignment methods[70] in many NMR structure determination protocols, POOL will be useful in high-resolution protein structure determination. Whenever RDCs can be collected for proteins with known X-ray structures containing missing loops, POOL can be used to determine the loop conformations.

POOL has been shown to compute ensembles of near-native loop conformations, from RDCs in one alignment medium, in the presence of modest levels of dynamics in protein loops. Since RDCs provide sensitive probes to protein conformational dynamics[114–117] from nanosecond to millisecond timescales, it will be interesting to extend our algorithm to capture and characterize the motional fluctuations, and deconvolve the dynamics from measured RDCs. In such cases, the ensemble of loops computed by POOL will effectively define a normal distribution of conformations centered at the experimentally measured RDCs, and as such encode a dynamic ensemble about a protein's native fold. Our algorithm can even be a stepping stone to computing ensembles reflecting more complex dynamics.

Availability

The source code of our algorithm is available by contacting the authors, and is freely distributed open-source under the GNU Lesser General Public License (Gnu, 2002).

ACKNOWLEDGMENTS

The authors benefited from an implementation of the CCD algorithm generously provided by Prof. Roland Dunbrack. They thank Prof. Jane and Prof. Dave Richardson, Dr. A. Yershova, Dr. A. Yan, Dr. V. Chen, Mr. J. MacMaster, and all members of the Donald, Zhou and Richardson Labs for helpful discussions and comments. They would like to thank the anonymous reviewers for helpful comments and suggestions.

REFERENCES

1. Pesce S, Benezara R. The loop region of the helix-loop-helix protein Id1 is critical for its dominant negative activity. Mol Cell Biol 1993;13:7874–7880.
2. Buchbinder JL, Fletterick RJ. Role of the active site gate of glycogen phosphorylase in allosteric inhibition and substrate binding. J Biol Chem 1996;271:22305–22309.
3. Greenwald J, Le V, Butler SL, Bushman FD, Choe S. The mobility of an HIV-1 integrase active site loop is correlated with catalytic activity. Biochemistry 1999;38:8892–8898.
4. Shi L, Javitch JA. The second extracellular loop of the dopamine D_2 receptor lines the binding-site crevice. Proc Natl Acad Sci USA 2004;101:440–445.
5. Gorczynski MJ, Grembecka J, Zhou Y, Kong Y, Roudaia L, Douvas MG, Newman M, Bielnicka I, Baber G, Corpora T, Shi J, Sridharan M, Lilien R, Donald BR, Speck NA, Brown ML, Bushweller JH. Allosteric inhibition of the protein-protein interaction between the leukemia-associated proteins Runx1 and CBFβ. Chemistry & Biology 2007;14:1186–1197.
6. Velloso LM, Bhaskaran SS, Schuch R, Fischetti VA, Stebbins CE. A structural basis for the allosteric regulation of non-hydrolysing UDP-GlcNAc 2-epimerases. EMBO Rep 2008;9:199–205.
7. Santiago C, Celma ML, Stehle T, Casasnovas JM. Structure of the measles virus hemagglutinin bound to the CD46 receptor. Nat Struct Mol Biol 2010;17:124–129.
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourn PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
9. Gō N, Scheraga HA. Ring closure and local conformational deformations of chain molecules. Macromolecules 1970;3:178–187.
10. Xiang Z, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci USA 2002;99:7432–7437.
11. Wang C, Bradley P, Baker D. Protein-protein docking with backbone flexibility. J Mol Biol 2007;373:503–519.
12. Hu X, Wang H, Ke H, Kuhlman B. High-resolution design of a protein loop. Proc Natl Acad Sci USA 2007;104:17668–17673.
13. Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods 2009;6:551–552.
14. Rufino SD, Donate LE, Canard LHJ, Blundell TL. Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling. J Mol Biol 1997;267:352–367.
15. van Vlijmen HWT, Karplus M. PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 1997;267:975–1001.
16. Tosatto SCE, Eckart Bindewald JH, Männer R. A divide and conquer approach to fast loop modeling. Protein Eng 2002;15:279–286.
17. Du P, Andrec M, Levy RM. Have we seen all structures corresponding to short protein fragments in the protein data bank? An update. Protein Eng 2003;16:407–414.
18. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B. Loop modeling: sampling, filtering, and scoring. Proteins: Struct Funct Bioinform 2008;70:834–843.
19. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–1773.
20. Shenkin PS, Yarmush DL, Fine RM, Wang H, Levinthal C. Predicting antobody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers 1987; 26:2053–2085.
21. Koehl P, Delarue M. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. Nat Struct Biol 1995;2:163–170.
22. Wang L-CT, Chen CC. A combined optimization method for solving the inverse kinematics problem of mechanical manipulators. IEEE Trans Rob Autom 1991;7:489–499.
23. Cortés J, Siméon T, Remaud-Siméon M, Tran V. Geometric algorithms for the conformational analysis of long protein loops. J Comput Chem 2004;25:956–967.
24. Milgram RJ, Liu G, Latombe J-C. On the structure of the inverse kinematics map of a fragment of protein backbone. J Comput Chem 2008;29:50–68.
25. Wedemeyer WJ, Scheraga HA. Exact analytical loop closure in proteins using polynomial equations. J Comput Chem 1999;20:819–844.

26. Coutsias EA, Seok C, Jacobson MP, Dill KA. A kinematic view of loop closure. J Comput Chem 2004;25:510–528.
27. Coutsias EA, Seok C, Wester MJ, Dill KA. Resultants and loop closure. Int J Quantum Chem 2006;106:176–189.
28. Canutescu AA, Dunbrack RL, Jr. Cyclic coordinate descent: a robotics algorithm for protein loop closure. Protein Sci 2003; 12:963–972.
29. Al-Nasr K, He J. An effective convergence independent loop closure method using forward-backward cyclic coordinate descent. Int J Data Min Bioin 2009;3:346–361.
30. Yao P, Dhanik A, Marz N, Propper R, Kou C, Liu G, van den Bedem H, Latombe J-C, Halperin-Landsberg I, Altman RB. Efficient algorithms to explore conformation spaces of flexible protein loops. IEEE/ACM Trans Comput Bioinform 2008;5:534–545.
31. Kolodny R, Guibas L, Levitt M, Koehl P. Inverse kinematics in biology: the protein loop closure problem. Int J Rob Res 2005; 24:151–163.
32. Lee J, Lee D, Park H, Coutsias EA, Seok C. Protein loop modeling by using fragment assembly and analytical loop closure. Proteins Struct Funct Bioinform 2010;78:3428–3436.
33. Sudarsanam S, DuBose RF, March CJ, Srinivasan S. Modeling protein loops using a ϕ_{i+1}, ψ_i dimer database. Protein Sci 1995; 4:1412–1420.
34. Janardhan A, Vajda S. Selecting near-native conformations in homology modeling: the role of molecular mechanics and solvation terms. Protein Sci 1998;7:1772–1780.
35. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol 1987;196:901–917.
36. Martin ACR, Thornton JM. Structural families in loops of homologous proteins: automatic classification, modeling, and application to antibodies. J Mol Biol 1996;263:800–815.
37. Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C. Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. Proteins Struct Funct Bioinform 1986;1:342–362.
38. Bruccoleri RE, Karplus M. Conformational sampling using high temperature molecular dynamics. Macromolecules 1990;29:1847–1862.
39. Cheng X, Wang H, Grant B, Sine SM, McCammon JA. Targeted molecular dynamics study of C-loop closure and channel gating in nicotinic receptors. PLoS Comput Biol 2006;2:e134.
40. Collura V, Higo J, Garnier J. Modeling of protein loops by simulated annealing. Protein Sci 1993;2:1502–1510.
41. Rosenbach D, Rosenfeld R. Simultaneous modeling of multiple loops in proteins. Protein Sci 1995;4:496–505.
42. Jiang H, Blouin C. Ab initio construction of all-atom loop conformations. J Mol Model 2006;12:221–228.
43. Donald BR. Algorithms in structural molecular biology. Cambridge, MA: The MIT Press; 2011.
44. Manocha D, Canny JF. Efficient inverse kinematics for general 6r manipulators. IEEE Trans Rob Autom 1994;10:648–657.
45. Chirikjian GS. General methods for computing hyper-redundant manipulator inverse kinematics. In: Proceedings of the 1993 IEEE/RSJ international conference on intelligent robots and systems, Yokohama, Japan, vol. 2, 1993. pp 1067–1073.
46. van den Bedem H, Lotan I, Latombe J-C, Deacona AM. Real-space protein-model completion: an inverse-kinematics approach. Acta Crystallogr D Biol Crystallogr 2005;61:2–13.
47. Shehu A, Clementi C, Kavraki LE. Modeling protein conformational ensembles: from missing loops to equilibrium fluctuations. Proteins Struct Funct Bioinform 2006;65:164–179.
48. Saxe JB. Embeddability of weighted graphs in *k*-space is strongly NP-hard. In: Proceedings of the 17th Allerton conference on communications, control, and computing, Monticello, IL, USA. 1979, pp 480–489.
49. Clore GM, Gronenborn AM, Tjandra N. Direct structure refinement against residual dipolar couplings in the presence of rhombicity of unknown magnitude. J Magn Reson 1998;131:159–162.
50. Güntert P. Automated NMR protein structure determination. Prog Nucl Magn Reson Spectrosc 2003;43:105–125.
51. Mumenthaler C, Güntert P, Braun W, Wüthrich K. Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. J Biomol NMR 1997;10:351–362.
52. Kuszewski J, Schwieters CD, Garrett DS, Byrd RA, Tjandra N, Clore GM. Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments. J Am Chem Soc 2004;126:6258–6273.
53. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The Xplor-NIH NMR molecular structure determination package. J Magn Reson 2003;160:65–73.
54. Brünger AT. X-PLOR, version 3.1. A system for X-ray crystallography and NMR. New Haven, CT: Yale University Press; 1992.
55. Delaglio F, Kontaxis G, Bax A. Protein structure determination using molecular fragment replacement and NMR dipolar couplings. J Am Chem Soc 2000;122:2142–2143.
56. Rohl CA, Baker D. De novo determination of protein backbone structure from residual dipolar couplings using rosetta. J Am Chem Soc 2002;124:2723–2729.
57. Bouvignies G, Meier S, Grzesiek S, Blackledge M. Ultrahigh-resolution backbone structure of perdeuterated protein GB1 using residual dipolar couplings from two alignment media. Angew Chem Int Ed Engl 2006;118:8346–8349.
58. Tian F, Valafar H, Prestegard JH. A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. J Am Chem Soc 2001;123: 11791–11796.
59. Giesen AW, Homans SW, Brown JM. Determination of protein global folds using backbone residual dipolar coupling and long-range NOE restraints. J Biomol NMR 2003;25:63–71.
60. Andrec M, Du P, Levy RM. Protein backbone structure determination using only residual dipolar couplings from one ordering medium. J Biomol NMR 2004;21:335–347.
61. Wang L, Donald BR. Exact solutions for internuclear vectors and backbone dihedral angles from NH residual dipolar couplings in two media, and their application in a systematic search algorithm for determining protein backbone structure. J Biomol NMR 2004; 29:223–242.
62. Wang L, Mettu RR, Donald BR. A polynomial-time algorithm for de novo protein backbone structure determination from NMR data. J Comput Biol 2006;13:1276–1288.
63. Donald BR, Martin J. Automated NMR assignment and protein structure determination using sparse dipolar coupling constraints. Prog Nucl Magn Reson Spectrosc 2009;55:101–127.
64. Yershova A, Tripathy C, Zhou P, Donald BR. Algorithms and analytic solutions using sparse residual dipolar couplings for high-resolution automated protein backbone structure determination by NMR. Presented at the The Ninth International Workshop on the Algorithmic Foundations of Robotics, Singapore. 2010. pp 355–372.
65. Frey KM, Georgiev I, Donald BR, Anderson AC. Predicting resistance mutations using protein design algorithms. Proc Natl Acad Sci USA 2010;107:13707–13712.
66. Chen C-Y, Georgiev I, Anderson AC, Donald BR. Computational structure-based redesign of enzyme activity. Proc Natl Acad Sci USA 2009;106:3764–3769.
67. Georgiev I, Lilien RH, Donald BR. The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. J Comput Chem 2008;29:1527–1542.
68. Lilien RH, Stevens BW, Anderson AC, Donald BR. A novel ensemble-based scoring and search algorithm for protein redesign, and its application to modify the substrate specificity of the gramicidin synthetase A phenylalanine adenylation enzyme. J Comput Biol 2005;12:740–761.

69. Martin JW, Yan AK, Bailey-Kellogg C, Zhou P, Donald BR. A graphical method for analyzing distance restraints using residual dipolar couplings for structure determination of symmetric protein homo-oligomers. Protein Sci 2011;20:970–985.
70. Zeng J, Zhou P, Donald BR. Protein side-chain resonance assignment and NOE assignment using RDC-defined backbones without TOCSY data. J Biomol NMR 2011;50:371–395.
71. Zeng J, Boyles J, Tripathy C, Wang L, Yan A, Zhou P, Donald BR. High-resolution protein structure determination starting with a global fold calculated from exact solutions to the RDC equations. J Biomol NMR 2009;45:265–281.
72. Wang L, Donald BR. An efficient and accurate algorithm for assigning nuclear overhauser effect restraints using a rotamer library ensemble and residual dipolar couplings. Presented at the IEEE Computational Systems Bioinformatics Conference, Stanford, CA. 2005. pp. 189–202.
73. Zeng J, Tripathy C, Zhou P, Donald BR. A Hausdorff-based NOE assignment algorithm using protein backbone determined from residual dipolar couplings and rotamer Patterns. In Proceedings of the 7th annual international conference on computational systems bioinformatics, Stanford CA, USA 2008. pp. 169–181.
74. Higman VA, Boyd J, Smith LJ, Redfield C. Residual dipolar couplings: are multiple independent alignments always possible? J Biomol NMR 2011;49:53–60.
75. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A. Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 2008;105:4685–4690.
76. Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 1999;13:289–302.
77. Shen Y, Delaglio F, Cornilescu G, Bax A. TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 2009;44:213–223.
78. Liu P, Zhu F, Rassokhin DN, Agrafiotis DK. A self-organizing algorithm for modeling protein loops. PLoS Comput Biol 2009; 5:e1000478, 08.
79. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
80. Langmead CJ, Donald BR. 3D structural homology detection via unassigned residual dipolar couplings. In: Procedings of 2003 IEEE computational systems bioinformatics conference, Stanford, CA, USA 2003. pp 209–217.
81. Langmead CJ, Donald BR. High-throughput 3D structural homology detection via NMR resonance assignment. In Procedings of 2004 IEEE computational systems bioinformatics conference, Stanford, CA, USA 2004. pp 278–289.
82. Losonczi JA, Andrec M, Fischer MWF, Prestegard JH. Order matrix analysis of residual dipolar couplings using singular value decomposition. J Magn Reson 1999;138:334–342.
83. Saupe A. Recent results in the field of liquid crystals. Angew Chem 1968;7:97–112.
84. Tjandra N, Bax A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. Science 1997;278:1111–1114.
85. Tolman JR, Flanagan JM, Kennedy MA, Prestegard JH. Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. Proc Natl Acad Sci USA 1995;92:9279–9283.
86. Prestegard JH, Bougault CM, Kishore AI. Residual dipolar couplings in structure determination of biomolecules. Chem Rev 2004;104:3519–3540.
87. Ramirez BE, Bax A. Modulation of the alignment tensor of macromolecules dissolved in a dilute liquid crystalline medium. J Am Chem Soc 1998;120:9106–9107.
88. Casey J. On cyclides and sphero-quartics. Proc R Soc Lond 1871;XIX:495–497.
89. Salmon G. A treatise on the analytic geometry of three dimensions, 5th ed. London: Longmans, Green and Company; 1912.
90. Salmon G. A treatise on conic sections, 6th ed. New York: Chelsea Publishing Company; 1960.
91. Yoshikawa T. Foundations of robotics: analysis and control. Cambridge, MA: The MIT Press; 1990.
92. Chirikjian GS, Burdick JW. A hyper-redundant manipulator. IEEE Rob Autom Mag 1994;1(4):22–29.
93. Lovell SC, Davis IW, Arendall WB, III, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure Validation by Cα Geometry: ϕ,ψ, and Cβ deviation. Proteins: Struct Funct Genet 2003;50:437–450.
94. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 1999;285:1711–1733.
95. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 1995;6:277–293.
96. Johnson BA, Blevins RA. NMRView: a computer program for the visualization and analysis of NMR data. J Biomol NMR 1994; 4:603–614.
97. Bartels C, Xia T, Billeter M, Güntert P, Wüthrich K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR 1995;6:1–10.
98. Coggins BE, Zhou P. PACES: protein sequential assignment by computer-assisted exhaustive search. J Biomol NMR 2003;26:93–111.
99. Ottiger M, Delaglio F, Bax A. Measurement of *J* and dipolar couplings from simplified two-dimensional NMR spectra. J Mag Reson 1998;131:373–378.
100. Ball G, Meenan N, Bromek K, Smith BO, Bella J, Uhrín D. Measurement of one-bond ^{13}C$^{\alpha}$–^{1}H$^{\alpha}$ residual dipolar coupling constants in proteins by selective manipulation of C$^{\alpha}$H$^{\alpha}$ spins. J Magn Reson 2006;180:127–136.
101. Permi P, Rosevear PR, Annila A. A set of HNCO-based experiments for measurement of residual dipolar couplings in ^{15}N, ^{13}C, (^{2}H)-labeled proteins. J Biomol NMR 2000;17:43–54.
102. Cornilescu G, Marquardt JL, Ottiger M, Bax A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J Am Chem Soc 1998;120:6836–6837.
103. Ramirez BE, Voloshin ON, Camerini-Otero RD, Bax A. Solution structure of DinI provides insight into its mode of RecA inactivation. Protein Sci 2000;9:2161–2169.
104. Ulmer TS, Ramirez BE, Delaglio F, Bax A. Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. J Am Chem Soc 2003;125:9179–9191.
105. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao H, Markley JL. BioMagResBank. Nucleic Acids Res 2008;36(Database issue):D402–D408.
106. Shen Y, Vernon R, Baker D, Bax A. De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 2007;43:63–78.
107. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D. NMR structure determination for larger proteins using backbone-only data. Science 2010;327:1014–1018.
108. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban YEA, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popovic Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 2011;487:545–574.

109. Han B, Liu Y, Ginzinger S, Wishart D. SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 2011; 50:43–57.

110. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D. Progress in modeling of protein structures and interactions. Science 2005;310:638–642.

111. Zweckstetter M, Bax A. Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. J Am Chem Soc 2000;122:3791–3792.

112. Zweckstetter M. NMR: prediction of molecular alignment from structure using the PALES software. Nat Protoc 2008;3:679–690.

113. Sellers BD, Zhu K, Zhao S, Friesner RA, Jacobson MP. Toward better refinement of comparative models: predicting loops in inexact environments. Proteins Struct Funct Bioinform 2008;72: 959–971.

114. Tolman JR, Flanagan JM, Kennedy MA, Prestegard JH. NMR evidence for slow collective motions in cyanometmyoglobin. Nat Struct Biol 1997;4:292–297.

115. Markwick PRL, Bouvignies G, Salmon L, McCammon JA, Nilges M, Blackledge M. Toward a unified representation of protein structural dynamics in solution. J Am Chem Soc 2009;131:16968–16975.

116. Salmon L, Bouvignies G, Markwick P, Lakomek N, Showalter S, Li D-W, Walter K, Griesinger C, Brüschweiler R, Blackledge M. Protein conformational flexibility from structure-free analysis of NMR dipolar couplings: quantitative and absolute determination of backbone motion in ubiquitin. Angew Chem Int Ed Engl 2009; 48:4154–4157.

117. Lange OF, Lakomek N-A, Fares C, Schroder GF, Walter KFA, Becker S, Meiler J, Grubmuller H, Griesinger C, de Groot BL. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science 2008;320:1471–1475.

118. Wlodarski T, Zagrovic B. Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. Proc Natl Acad Sci USA 2009;106:19346–19351.

119. Long D, Brüschweiler R. In silico elucidation of the recognition dynamics of ubiquitin. PLoS Comput Biol 2011;7:e1002035.

120. Tolman JR, Ruan K. NMR residual dipolar couplings as probes of biomolecular dynamics. Chem Rev 2006;106:1720–1736.

121. Chen VB, Arendall WB, III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 2010;66:12–21.

122. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, III, Snoeyink J, Richardson JS, Richardson DC. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res 2007:35(Web Server issue);W375–W383.

123. Chen VB, Arendall WB, III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Laura WM, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 2010;66:12–21.