An Algebraic Geometry Approach to Protein Structure Determination from NMR Data

Lincong Wang* **Ramgopal R. Mettu*** **Bruce Randall Donald** *,†,‡,§,¶

Abstract

Our paper describes the first provably-efficient algorithm for determining protein structures de novo*, solely from experimental data. We show how the global nature of a certain kind of NMR data provides quantifiable complexity-theoretic benefits, allowing us to classify our algorithm as running in* polynomial time*. While our algorithm uses NMR data as input, it is the first polynomial-time algorithm to compute high-resolution structures* de novo *using* any *experimentally-recorded data, from either NMR spectroscopy or X-Ray crystallography.*

Improved algorithms for protein structure determination are needed, because currently, the process is expensive and time-consuming. For example, an area of intense research in NMR methodology is automated assignment of nuclear Overhauser effect (NOE) restraints, in which structure determination sits in a tight inner-loop (cycle) of assignment/refinement. These algorithms are very time-consuming, and typically require a large cluster. Thus, algorithms for protein structure determination that are known to run in polynomial time and provide guarantees on solution accuracy are likely to have great impact in the long-term. Methods stemming from a technique called "distance geometry embedding" do come with provable guarantees, but the $\mathcal{NP}$-hardness of these problem formulations implies that in the worst case these techniques cannot run in polynomial time. We are able to avoid the $\mathcal{NP}$-hardness by (a) some mild assumptions about the protein being studied, (b) the use of residual dipolar couplings (RDCs) instead of a dense network of NOEs, and (c) novel algorithms and proofs that exploit the biophysical geometry of (a) and (b), drawing on a variety of computer science, computational geometry, and computational algebra techniques.

In our algorithm, RDC data, which gives global restraints on the orientation of internuclear bond vectors, is used in conjunction with very sparse NOE data to obtain a polynomial-time algorithm for protein structure determination. An implementation of our algorithm has been applied to 6 different real biological NMR data sets recorded for 3 proteins. Our algorithm is combinatorially precise, polynomial-time, and uses much less NMR data to produce results that are as good or better than previous approaches in terms of accuracy of the computed structure as well as running time. In practice approaches such as restrained molecular dynamics and simulated annealing, which lack both combinatorial precision and guarantees on running time and solution quality, are commonly used. Our results show that by using a different "slice" of the data, an algorithm that is polynomial time and that has guarantees about solution quality can be obtained. We believe that our techniques can be extended and generalized for other structure-determination problems such as computing side-chain conformations and the structure of nucleic acids from experimental data.

1 Introduction[1]

Protein structure is the key to understanding protein function, and is also the starting point for structure-based drug design. One of the key tools used to study protein structure and function in solution is NMR spectroscopy. Traditionally, nuclear Overhauser effect (NOE) spectroscopy has been used to obtain approximate interproton distance restraints, which, in turn, have been used for structure determination. Due to the sparsity of the data and experimental error, however, the problem of structure determination using experimental NOE data is NP-hard [36, 31, 6], and rigorous approaches to structure determination based on solving this problem, such as the distance geometry method [15, 14], require exponential time. In practice, the most commonly used structure determination protocols use experimental NMR data along with techniques such as molecular dynamics (MD) and simulated

*Dartmouth Computer Science Department, Hanover, NH 03755, USA.

†Dartmouth Chemistry Department, Hanover, NH 03755, USA.

‡Dartmouth Department of Biological Sciences, Hanover, NH 03755, USA.

§Corresponding author: 6211 Sudikoff Laboratory, Dartmouth Computer Science Department, Hanover, NH 03755, USA. Phone: 603-646-3173. Fax: 603-646-1672. Email: brd@cs.dartmouth.edu

¶This work is supported by the following grants to B.R.D.: National Institutes of Health (R01 GM 65982), and National Science Foundation (EIA-0305444 and EIA-9802068).

annealing (SA). These approaches, however, lack combinatorial precision, guarantees on running time, as well as guarantees on solution quality. Additionally, NOE data is tedious and time-consuming to interpret due to the difficulty of *assigning* the distance restraints. In practice, traditional NOE-based structure determination approaches are not suited for high-throughput structure determination, since it may take months to assign a sufficient number of NOEs, especially those involving sidechain protons, to compute an accurate NMR structure [11].

In recent years, residual dipolar coupling (RDC) data has been used to provide global orientational restraints on the protein structure [40, 41, 39, 19, 33]. RDC data gives *global* orientational restraints on, for example, backbone NH bond vectors with respect to a global coordinate frame. Additionally, RDCs can be recorded and assigned much faster (e.g., in a few hours) than the NOEs required by traditional NMR structure determination methods. Existing structure determination approaches do use RDCs, along with other experimental restraints such as chemical shifts or sparse NOEs [1, 17, 21, 28, 35, 39], yet remain heuristic in nature, without guarantees on solution quality or running time. In this paper, we make the biophysically reasonable assumption that the protein under consideration is globular and contains regular secondary structure. Globular proteins comprise the majority of proteins in nature, and are far more abundant than fibrous proteins (e.g., collagen or coiled-coil oligomers). This assumption implies that each secondary structure element has length bounded by a constant (which, for implementation purposes, is straightforward to check in linear time). Under this assumption, previous formulations of the structure determination problem remain NP-hard. We show that our formulation of the structure determination problem, given RDC data, sparse NOEs and experimentally-determined secondary structure types, can be solved in polynomial time. Unlike previous approaches, which have either no theoretical guarantees or run in exponential time, we show that it is possible to exploit the global nature of RDC data to develop an algorithm that runs in polynomial time and computes the structure that agrees best with the given experimental RDC and NOE data. While our algorithm uses NMR data as input, it is the first polynomial-time algorithm to compute high-resolution structures *de novo* using *any* experimentally recorded data, from either NMR spectroscopy or X-Ray crystallography.

Our formulation of the structure determination problem assumes that we are given the following experimental NMR data: (a) 2 RDCs of backbone vectors per residue (e.g., assigned NH RDCs in two media or NH and CH RDCs in a single medium), (b) identified α-helices and β-sheets with known hydrogen bonds (H-bonds) between paired strands, and (c) a few NOE distance restraints. The implementation discussed in Sec. 5 uses this experimental data, and allows for missing data as well. The secondary structure types of backbone residues can be determined by NMR from experimentally recorded HNHA [10, pages 524–528] data, or J-doubling [16] data for larger proteins. NMR chemical shifts [48, 50, 49, 30] or automated assignment [2] can also be used. Hydrogen bonds can be determined by NMR from experimentally recorded data [13, 46], or, e.g., by using backbone resonance assignment programs such as JIGSAW [2]. Additionally, it is relatively straightforward to rapidly obtain the few (3 or 4), unambiguous NOEs required for our algorithm using, for example, the labeling strategy of Kay and coworkers [20]. The user of our algorithm has a choice, to record either (a) NH RDCs in two aligning media, or (b) 2 RDCs per residue (e.g., NH and CH) in one medium. This flexibility allows our algorithm to be applied to a wider range of proteins. In the remainder of the paper, we present our algorithm assuming that we are given assigned NH RDCs in two media. Our results also hold for the case of NH and CH RDCs in one medium with slight modifications to the equations in Sec. 3 (see [43]).

A key building block of our algorithm makes use of *exact*, low-degree polynomial equations [44] that relate the experimental RDCs to the backbone (ϕ, ψ) dihedral angles, which determine the protein backbone geometry. These equations, however, do not yield a unique solution for the (ϕ, ψ) angles since they are low-degree (at most 4) polynomials; furthermore, error in the experimentally recorded RDCs also makes it possible that these equations are not solvable. Thus, we formulate and exactly solve a semi-algebraic optimization problem to compute the conformation of the secondary structure elements that optimally fit the experimental data. Since RDCs give *global* restraints on internuclear vectors, the conformation of the secondary structure elements can be computed with respect to a global coordinate frame. Thus, given the optimal conformation of secondary structure elements, we must next find only their relative translations to compute the backbone structure. To do this, we require sparse, assigned NOEs between successive pairs of secondary structure elements; we formulate and solve an optimization problem which asks us to find the translation that maximizes agreement with the experimental NOE data. Our approach to solving these optimization problems makes use of the *theory of real closed fields* [22, 3], which gives algorithms for deciding first-order sentences on sets of polynomial inequalities. The running time of these algorithms is parameterized by the degree, number of variables and number of alternations in the input sentences; we show that our optimization problems can be formulated such that we can find the optimal solution in polynomial time. Finally, since our algorithm

[1] Abbreviations used: NMR, nuclear magnetic resonance; RDC, residual dipolar coupling; 3D, three-dimensional; H^N, amide proton; NH, backbone amide bond vector; NC_α, backbone bond vector between N and C_α atoms; C_α, backbone α-carbon atom; H_α, backbone C_α proton; HNHA, an NMR experiment to measure the 3-bond scalar coupling H^N–^{15}N–H_α; RMSD, root-mean squared distance; SA, Simulated Annealing; MD, Molecular Dynamics; MC, Monte-Carlo; NOE, nuclear Overhauser effect; $SO(3)$, special orthogonal (rotation) group in 3D; POF, principal order frame; SVD, singular value decomposition.

is based on low-degree polynomials that relate the experimental RDCs directly to NH vector orientations, our algorithm is the first approach to structure determination that makes it possible to *analytically* quantify the effect of experimental error on the resulting backbone structure. We also show that an implementation based on our algorithm, given only RDCs, sparse NOEs, hydrogen bonds, and secondary structure types, is able to quickly compute structures that are as good or better, in terms of RMSD accuracy, than structures produced by previous techniques. Under our assumption that the protein is globular, this implementation runs in polynomial time.

Our result is consistent with previous observations [40, 41, 39, 19, 33, 1, 17, 21, 28, 35, 47] that, empirically, RDCs increase the speed and accuracy of biomacromolecular structure determination, and formally quantitates the complexity-theoretic benefits of employing globally-referenced angular data on internuclear bond vectors. In summary, our main contributions in this paper are:

1. To show that low-degree polynomial equations can be solved *exactly* and in *constant time* to give solutions for backbone (ϕ, ψ) angles from experimentally recorded RDCs.
2. The first *combinatorially precise*, *polynomial-time* algorithm for structure determination using RDCs, secondary structure type, and very sparse NOEs.
3. The first polynomial-time algorithm for *de novo* backbone protein structure determination solely from experimental data (of any kind).
4. An implementation of our algorithm that is as good or better in terms of accuracy and speed, but requires much less data, than existing NMR structure determination techniques.
5. Testing and results of our algorithm on real biological NMR data.

1.1 Related Work

Previously-studied theoretical formulations of the structure determination problem use local distance restraints, e.g. NOEs, as the only constraint on the structure. We note this problem is not as straightforward as reconstructing a set of n points with a complete and exact distance matrix; this problem can be solved exactly using SVD in $O(n^3)$ time. Berger *et al.* [4] assume $\Omega(n^2)$ distances are given but study the problem of reconstructing a set of n points where some of the distances are missing or erroneous (and the errors are not known). They give a randomized $O(n \log n)$-time algorithm to enumerate all point sets consistant with these distances, where the given distance matrix has at most $(1/2 - \epsilon)n$ errors per row. They also showed that under a certain random error model they can correct errors of the same density in a sparse matrix, where only $\beta > 0$ fraction of the entries in each row are given.

In practice, far fewer than $\binom{n}{2}$ NOEs are observed experimentally: for example, even in an ideal case, it is in general possible to obtain only about $15n = O(n)$ NOE-derived distance restraints. Furthermore, it is unrealistic to assume that some NOE restraints encode perfect distances, while others are arbitrarily corrupted; it is more realistic to assume that all of the NOE data is subject to bounded experimental error. Saxe [36] viewed the structural model as a graph where the vertices represent atoms and edge weights represent distance constraints. The *molecule problem* asks whether such a graph, given a sparse set of edges with perfect distances, can be embedded in $\mathbb{R}^3$ while preserving the edge weights; Saxe showed that this problem is NP-hard. Hendrickson [26, 27] studies conditions under which embedding such a graph is even possible, and gives (super-polynomial time) algorithms for the problem. Crippen and Havel [15] studied the *distance geometry* problem; in this problem, we must use distance intervals, rather than scalar distance restraints, to construct a point set that satisfies the restraints imposed by the intervals. This problem has application in NOE-based structure determination since it can be used to find a consistent interpretation of noisy experimental NOEs. However, the NP-hardness of this problem follows from the results of Saxe [36, 31, 6], and existing algorithms for solving the distance geometry problem require exponential time in the worst-case [15, 6].

Traditional NMR structure determination algorithms such as [5, 23] were initially designed to use NOE-derived distance restraints, but these methods are neither combinatorially precise nor polynomial time. Table 2 in Sec. 5 gives a detailed summary of existing methods for structure determination, including the experimental data requirements and accuracies of the resulting structure. Finally, we note that although [44, 43] provide some building blocks for this paper, those algorithms are neither combinatorially precise nor polynomial time. Furthermore, they do not compute loop or turn structures, which we show can be done with our algorithm (see Sec. 5).

2 Preliminaries

The equation for the RDC r associated with an internuclear bond vector $\mathbf{v}$ can be written [40, 41] as a quadratic form:

$$r = D_{max} \mathbf{v}^T \mathbf{S} \mathbf{v}, \tag{1}$$

where D_{max} is the dipolar interaction constant, $\mathbf{v}$ is the bond vector of interest with respect to an arbitrary global coordinate frame, and $\mathbf{S}$ is the 3×3 *Saupe* order matrix, or *alignment tensor*, which specifies the orientation of the protein in the laboratory frame (i.e, magnetic field in the NMR spectrometer) with respect to the aligning medium. Our goal is to determine the orientation of vector $\mathbf{v}$ given an experimentally recorded RDC. It is common practice to record multiple sets of RDCs to further constrain $\mathbf{v}$, and we assume that 2 independent sets of RDCs have been recorded. The user of our algorithm has a choice, to record either (a) NH RDCs in two aligning media, or (b) 2 RDCs

per residue (e.g., NH and CH) in one medium. This flexibility allows our algorithm to be applied to a wider range of proteins. In the remainder of the paper, we present our results assuming that we are given assigned NH RDCs in two media. Our results also hold for the case of NH and CH RDCs in one medium with slight modifications to the equations in Sec. 3 [43]. Given an alignment tensor for each aligning medium, our problem specification asks us, informally, to find a conformation vector such that its backbone (ϕ, ψ) angles fit the experimental RDC data as closely as possible. Additionally, we ask that the (ϕ, ψ) values are as close as possible to the average (ϕ_a, ψ_a) angles over the PDB for the corresponding secondary structure type. Then, after determining the conformation of the secondary structure elements, we must translate the secondary structure elements using a set of sparse NOEs to obtain the final backbone structure. Finding this translation requires only a constant number of NOEs for each secondary structure element, since RDCs give an orientation of the entire protein with respect to a global coordinate frame and thus the global orientations of the secondary structure elements are known once their conformations have been computed.

We now formalize the structure determination problem discussed above. First, let $\mathcal{A}$ denote a secondary structure element with length c. Let $D_1 = (r_{1,1}, r_{1,2}, \ldots, r_{1,c})$ and $D_2 = (r_{2,1}, r_{2,2}, \ldots, r_{2,c})$ denote the recorded RDC values in the first and second medium, respectively. Let (ϕ_i, ψ_i) denote the backbone dihedral angles for the $i+1^{\text{st}}$ residue, $1 \leq i \leq c-1$, and let $w(\phi)$ (resp., $w(\psi)$) denote the unit vector $(\cos\phi, \sin\phi)$ (resp., $(\cos\psi, \sin\psi)$). Let $\mathcal{C}_i = (w(\phi_1), w(\psi_1), \ldots, w(\phi_i), w(\psi_i))$. Each conformation of $\mathcal{A}$ can be specified by the orientation of the first peptide plane and the conformation vector $\mathcal{C} = \mathcal{C}_{c-1}$. Finally, for any RDC r, let $G(r)$ denote the interval $[r-1, r+1]$, which represents an experimental error range of ± 1 Hz.

It has been shown that, due to experimental error, experimentally-recorded RDCs cannot in general be fit to a secondary structure element unless they are perturbed (within some error window) [44]. To account for error in the experimentally recorded RDCs, we parameterize the experimental RDCs in our objective function by defining the following sets. Let $\mathcal{G}(D_j)$ denote the set $G(r_{j,1}) \times G(r_{j,2}) \times \ldots \times G(r_{j,c})$ for two aligning media $j = 1, 2$. Then, for each secondary structure element, we seek to minimize the following objective functions on the orientation of the first peptide plane and backbone (ϕ, ψ) angles. Let $b_{j,1}(\mathbf{R}) = D_{max}\mathbf{v}_i(\mathbf{R})^T\mathbf{S}_j\mathbf{v}_i(\mathbf{R})$ and $b_{j,i}(\mathbf{R}, \mathcal{C}_{i-1}) = D_{max}\mathbf{v}_i(\mathbf{R}, \mathcal{C}_{i-1})^T\mathbf{S}_j\mathbf{v}_i(\mathbf{R}, \mathcal{C}_{i-1})$ for $2 \leq i \leq c$ be the back-computed RDCs under the alignment tensor $\mathbf{S}_j$. Here, $\mathbf{R}$ is the rotation matrix that defines the orientation of the first peptide plane of $\mathcal{A}$ and $\mathbf{v}_i(\mathbf{R}, \mathcal{C}_{i-1})$ is the orientation of the i^{th} backbone NH vector, which can be specified uniquely by $\mathbf{R}$ and $\mathcal{C}_{i-1}$. We note that the first NH vector, and thus the first back-computed RDC, is defined slightly differently since it depends only on the orientation of the first peptide plane (see Sec. 3 for further discussion). For notational convenience, we will write $b_{j,1} = b_{j,1}(\mathbf{R})$ and $b_{j,i} = b_{j,i}(\mathbf{R}, \mathcal{C}_{i-1})$ for $2 \leq i \leq c$ and $j = 1, 2$.

Let (ϕ_a, ψ_a) denote the average values for the backbone (ϕ, ψ) dihedral angles for the secondary structure type of $\mathcal{A}$ over the PDB. Then, let

$$\sigma(D_1', D_2', \mathbf{R}, \mathcal{C}) = \sum_{i=1}^{c-1} \|w(\phi_i) - w(\phi_a)\|^2 + \|w(\psi_i) - w(\psi_a)\|^2 + \sum_{i=1}^{c} \left((b_{1,i} - r_{1,i})^2 + (b_{2,i} - r_{2,i})^2 \right). \quad (2)$$

Our goal is to find $D_1' \in \mathcal{G}(D_1)$, $D_2' \in \mathcal{G}(D_2)$, a rotation $\mathbf{R} \in SO(3)$, and conformation $\mathcal{C}$ so that $\sigma(D_1', D_2', \mathbf{R}, \mathcal{C})$ is minimized. Note that $w(\phi_i)$ and $w(\psi_i)$ are elements of $\mathcal{C}_i$ (for $1 \leq i < c$), and that $b_{j,i}$ is a function of $\mathcal{C}_{i-1}$ and $\mathbf{R}$ (for $j = 1, 2$ and $1 < i < c$; $b_{j,1}$ is a function of $\mathbf{R}$ only). All elements of $\mathcal{C}$ are roots of polynomials whose coefficients are completely determined by D_1', D_2' and $\mathbf{R}$. The minima of Eq. (2) represent the conformations for the given secondary structure element that agree best with both the experimental RDCs and the secondary structure type. We note that as written Eq. (2) is underconstrained. Given 2 RDCs for residue i, the NH bond vector must lie in a finite set, defined by a quartic monomial [44]. This, in turn, constrains (ϕ_i, ψ_i) to lie in a finite algebraic set, defined by backbone kinematics [44]. Hence, the optimization[2] in Eq. (2) is performed over a finite algebraic subset of a $2(c-1)$-torus (see Sec. 3 for further discussion).

Given conformations of the secondary structure elements, we must next compute the backbone fold by computing the relative translations of the elements. We emphasize that our algorithm (and our formulation of the problem) does not simply 'pack' ideal helix/strand geometries. The solution structure is computed with respect to all of the RDCs (rather than any individual RDC) using the score function σ. Therefore, individual dihedral angles of a solved helix/strand computed by our algorithm may differ from the average values by as much as 29° (See Figure 6 of [44, page 234]). To compute relative translations, we require at least 3 Euclidean distances between three (non-collinear) nuclei between each pair of successive secondary structure elements. NOE restraints provide this information, but are subject, like RDCs, to experimental error. Informally, given experimentally recorded NOE restraints between a pair of successive secondary structure elements, we

[2]For simplicity of analysis, we have omitted the distinction between α-helices and β-sheets in the definition of Eq. (2). The objective function for β-sheets has an extra additive term that accounts for hydrogen bonds between β-strands and provides additional constraint on the conformation of the β-sheet. This modification for β-sheets can be incorporated easily by the algorithm and analysis given in Sec. 4; this additional term in the objective function is discussed in detail in [44]. To handle hydrogen bond geometry in β-sheets, we use Equation (9) in [44, page 228] as the additional term and make use of the techniques of Lemma 2 to cope with the additional term in the objective function (see Sec. 4.2).

wish to find a translation between the secondary structure elements that agree best with the NOE restraints. More formally, for each oriented pair of successive secondary structure elements $\mathcal{A}$ and $\mathcal{B}$, let $A = \{a_1, a_2, \ldots, a_\ell\}$ (resp., $B = \{b_1, b_2, \ldots, b_\ell\}$) be the 3D coordinates of the ℓ nuclei in $\mathcal{A}$ (resp., $\mathcal{B}$) for which we are given distances (derived from NOE restraints) $N = (n_1, n_2, \ldots, n_\ell)$. Then, we wish to find a translation $x \in \mathbb{R}^3$ that minimizes

$$\sigma_{NOE}(x) = \sum_{i=1}^{\ell} \left(\|a_i - b_i + x\| - n_i \right)^2 . \quad (3)$$

The minima of Eq. (3) represent relative translations between a successive pair of secondary structures that agree as closely as possible with the experimental NOE restraints.

3 Equations for computing successive (ϕ, ψ) angles from RDCs

In this section, we present an exact, constant time (per residue) method to compute backbone dihedral angles from RDCs in two aligning media. We show that it is possible to derive, from the physics of RDCs, low-degree monomials (with degree at most 4) whose solutions give the backbone (ϕ, ψ) angles. Due to space constraints, we only present the details of these equations that are relevant to Sec. 4; further exposition is provided in Appendix B. For simplicity we assume that the dipolar interaction constant D_{max} is equal to 1. By considering a global coordinate frame which diagonalizes the alignment tensor, Eq. (1) becomes:

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

where S_{xx}, S_{yy} and S_{zz} are the three diagonal elements of a diagonalized Saupe matrix $\mathbf{S}$ (the alignment tensor), and x, y and z are, respectively, the $x, y, z-$components of the unit vector $\mathbf{v}$ in a principal order frame (POF) which diagonalizes $\mathbf{S}$. In order to make our problem algebraic, we write x, y and z in terms of variables t and u, where $x = a \sin t$, $y = b \cos t$, and $u = \cos 2t$. Now, $\mathbf{S}$ is a 3×3 symmetric, traceless matrix with five independent elements [40, 41]. Given NH RDCs in two aligning media, the associated NH vector $\mathbf{v}$ must lie on the intersection of two conic curves [37, 47]. We show

Proposition 1 *Given the diagonal Saupe elements S_{xx} and S_{yy} for medium 1, S'_{xx} and S'_{yy} for medium 2, and a relative rotation matrix $\mathbf{R}_{12}$ between the POFs of medium 1 and 2, the square of the x-component of the unit vector $\mathbf{v}$ satisfies a monomial quartic equation $f_4u^4 + f_3u^3 + f_2u^2 + f_1u + f_0 = 0$.*

The proof of Prop. 1 is provided for in Appendix B; the full expressions for the coefficients a, b, f_0, f_1, f_2, f_3, f_4 are given in [44]. Since $u = 1 - 2(\frac{x}{a})^2$, the equation in Prop. 1 above is also quartic in x^2. Given solutions for u, the y-component of $\mathbf{v}$ can be computed directly from Eq. (4) and the change of the variables given above. Due to two-fold symmetry in the RDC equation the number of real solutions for $\mathbf{v}$ is at most 8. Now, let NC_α denote the bond vector between the N and C_α atoms along the backbone. We show that:

Proposition 2 *Given the* NH *unit vectors $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$ of residues i and $i+1$ and the NC_α vector of residue i, the sines and cosines of the intervening backbone dihedral angles (ϕ, ψ) satisfy the trigonometric equations $\sin(\phi + a_1) = b_1$ and $\sin(\psi + a_2) = b_2$, where a_1 and b_1 are constants depending on $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$, and a_2 and b_2 depend on $\mathbf{v}_i$, $\mathbf{v}_{i+1}$, $\sin\phi$ and $\cos\phi$. Furthermore, exact solutions for $\sin(\phi)$ and $\cos(\phi)$ can be computed from a quadratic equation by the substitution $w = \tan\frac{\phi}{2}$, $\sin\phi = 2w/(1+w^2), \cos\phi = (1-w^2)/(1+w^2)$; equations for $\sin\psi$ and $\cos\psi$ can be obtained and solved exactly by a similar substitution.*

The proof of Prop. 2, as well as definitions of a_1, b_1, a_2, b_2, are provided in Appendix B.

Props. 1 and 2 show that the sines and cosines of (ϕ, ψ) angles can be computed *exactly*, and in constant-time, from RDCs. This in turn implies that candidate conformations for the protein backbone structure can be built using the sines and cosines of (ϕ, ψ) angles. There are only two independent solutions for the (ϕ, ψ) angles of residue i given the NH vectors for residues i and $i+1$ if the orientation of the i^{th} peptide plane is also known. We can define the i^{th} peptide plane by two vectors: an NH vector solved from the quartic equation in Prop. 1, and an NC_α vector. The rotation matrix $\mathbf{R}_i$ defines the relative rotation between a POF and a coordinate system in the i^{th} peptide plane. The rotation matrix $\mathbf{R}_1$ defining the first peptide plane can be determined by solving an optimization problem (see Sec. 4). This matrix is denoted $\mathbf{R}$ in Eq. (2) above; below, we let $\mathbf{R}_1 = \mathbf{R}$. Let $F_R(\mathbf{R}_i, \phi_i, \psi_i)$ be the algebraic function for computing the matrix $\mathbf{R}_{i+1}$ from ϕ_i, ψ_i and $\mathbf{R}_i$; F_R can be derived from backbone kinematics [44]. In summary, Props. 1 and 2 show that given the rotation $\mathbf{R}_i$, ϕ_i and ψ_i for residue i can be computed, *exactly and in constant time*, from two low-degree polynomial equations

$$F_{\phi_i}(r_{1,i}, r_{2,i}, r_{1,i+1}, r_{2,i+1}, \mathbf{R}_i) = 0 \quad (5)$$
$$F_{\psi_i}(r_{1,i}, r_{2,i}, r_{1,i+1}, r_{2,i+1}, \mathbf{R}_i, w(\phi_i)) = 0, \quad (6)$$

where $r_{1,i}$, $r_{1,i+1}$ and $r_{2,i}$ and $r_{2,i+1}$ are NH RDCs measured for residue i and $i+1$ in medium 1 and 2, respectively. The roots of F_{ϕ_i} (resp., F_{ψ_i}) are the vectors $w(\phi_i)$ (resp., $w(\psi_i)$). The algebraic function F_R has degree 2 with 4 variables. Eqs. (5) and (6) both have degree 4 and have 3 and 4 variables, respectively. We note that analogous low-degree polynomial equations can also be derived for NH and CH RDCs measured in a single aligning medium [43].

Given experimentally-measured RDCs $Z_i = \{r_{1,i}, r_{1,i+1}, r_{2,i}, r_{2,i+1}\}$, and the rotation matrix $\mathbf{R}_i$,

for $1 \leq i < c$, the solutions to F_{ϕ_i}, and F_{ψ_i} above define a discrete, finite, algebraic subset $Y_i(Z_i, \mathbf{R}_i)$ of the 2-torus $S^1 \times S^1$, containing at most 16 points, in which the backbone dihedral angles (ϕ_i, ψ_i) must lie. By Eqs. (5) and (6) for $w(\phi_i)$ and $w(\psi_i)$, $Y_i(Z_i, \mathbf{R}_i)$ can be computed exactly, in closed-form, and in constant-time. Hence, the conformation $\mathcal{C}$ of each secondary structure element must lie in a discrete, finite, algebraic subset of the $2(c-1)$-torus $(S^1)^{2(c-1)}$, and is defined by $\mathcal{Y}(D_1, D_2, \mathbf{R}_1) = \Pi_{i=1}^{c-1} Y_i(Z_i, \mathbf{R}_i)$. Each set $Y_i(Z_i, \mathbf{R}_i)$ is described by the polynomial equations for ϕ_i (of degree 4 with 3 variables), ψ_i (of degree 4 with 4 variables), and $\mathbf{R}_i$ (of degree 2 with 4 variables). Since the equations for (ϕ_i, ψ_i) utilize the rotation $\mathbf{R}_i$, $Y_i(Z_i, \mathbf{R}_1)$ requires $2(c-1)$ equations with degree $O(c)$ in $2(c-1)+4 = 2c+2$ variables. We will exploit the fact that the backbone conformation lies in a discrete, finite, algebraic set in the next section, where we present an algorithm to find the conformation that optimizes Eq. (2), subject to the constraint $\mathcal{Y}(D_1, D_2, \mathbf{R}_1)$.

4 A Polynomial-Time Algorithm for Protein Structure Determination

In Sec. 3, we presented low-degree polynomial equations that relate RDCs to backbone dihedral angles. However, the equations for a given pair of (ϕ, ψ) angles depend on the corresponding experimental RDC values as well as the orientation of the previous peptide plane. These equations are not guaranteed to have a unique solution and thus there may be multiple (ϕ, ψ) pairs that are consistent with the experimental RDC value; this is a consequence of the degree of the equations for F_{ϕ_i} and F_{ψ_i} in Sec. 3. Furthermore, in order to account for experimental error, we must interpret our RDCs as being in a range rather than being a fixed value, and there is no guarantee that the entire range yields solvable polynomials for the (ϕ, ψ) angles. Thus, these equations do not immediately yield a unique conformation, and a search algorithm is needed to compute the optimal conformation inside the cross-product ($\mathcal{Y}$) of the discrete solution choices for the backbone (ϕ, ψ) angles. In this section we present an algorithm that uses these equations to find the optimal conformation, with respect to the objective functions given in Sec. 2, in *polynomial time*. Throughout the presentation of the algorithm and analysis, we will assume that our protein has n residues and m secondary structure elements. Recall that we assumed that our protein was globular; this implies that $m = O(n)$ and that $c = O(1)$.

4.1 Algorithm

In this section, we give our algorithm for structure determination. We give a high-level description of the algorithm, and give a detailed description of some of the key steps in Sec. 4.2 below. In Sec. 5, we show that all these minimization steps can in fact be implemented in practice and performed efficiently to rapidly compute accurate structures given real, experimental NMR data as input. Our algorithm consists of three phases. We describe the first two phases, for simplicity, for a single secondary structure element. In the first phase, we compute the alignment tensor for the protein. We assume without loss of generality that D_1 and D_2 correspond to an α-helix with $c \geq 5$ residues. To compute alignment tensors $\mathbf{S}_1$ and $\mathbf{S}_2$ for each medium we use SVD [29] to fit the RDCs to the NH vectors of a c-residue α-helix with ideal geometry. The running time of this phase is $O(c^3)$.

In the second phase, we determine the conformation and global orientation of each secondary structure element, and in the third phase, we determine the relative translations of the secondary structure elements to obtain the backbone fold. We find $D_1' \in \mathcal{G}(D_1)$ and $D_2' \in \mathcal{G}(D_2)$, $\mathbf{R}$, and $\mathcal{C} \in \mathcal{Y}(D_1, D_2, \mathbf{R})$ that minimize Eq. (2), subject to $\mathcal{Y}$ (see Sec. 3 for definition) simultaneously by deciding, and finding a witness for, a sentence in the first-order theory of real closed fields [22, 3]. We show this minimization procedure is polynomial-time in Sec. 4.2 below.

In the third phase, we are given sparse NOEs between successive pairs of secondary structure elements, and must compute their relative translation. Fix two successive secondary structure elements $\mathcal{A}$ and $\mathcal{B}$, and let $N = (n_1, n_2, \ldots, n_\ell)$ be the Euclidean distances between ℓ pairs of nuclei from $\mathcal{A}$ and $\mathcal{B}$ derived from the sparse experimental NOE restraints. We compute a translation $x \in \mathbb{R}^3$ between $\mathcal{A}$ and $\mathcal{B}$, that minimizes Eq. (3) by deciding, and finding a witness for, a sentence in the first-order theory of real closed fields. Computing this translation is sufficient since RDCs are global restraints and thus all bond vectors are determined in a common coordinate frame; the second phase explicitly determines the global orientation of secondary structure fragments. The time required for this phase is $O(m) = O(n)$ times the cost to compute an optimal translation for each pair of secondary structure elements; we show that the running time of the latter is polynomial in n.

4.2 Analysis of Running Time

In this section, we show that the key optimization steps in the algorithm of Sec. 4.1 can be performed in polynomial time. At a high level, our proof relies on the observation that the objective functions being minimized in the algorithm can be cast into sentences in the first-order theory of real closed fields. This allows us to apply the algorithm of [3, Chapter 14] to obtain the desired minima.

There has been much study of how efficiently a first-order predicate on polynomial inequalities can be decided [38, 12, 22, 8, 25, 34]. We use a result of Basu *et al.* [3], which has an improved asymptotic running time. We now restate their result:

Theorem 1 (Basu *et al.* [3, page 507]) *Let $\mathcal{P}$ be a first-order predicate over s polynomials of degree at most d in k variables with coefficients bounded by 2^C and a alternately quantified blocks of $k_1, k_2, \ldots, k_a$ variables. The truth of $\mathcal{P}$, along with a witness if $\mathcal{P}$ is true, can be determined in $O(C \cdot s^{(k_1+1)\ldots(k_a+1)} \cdot d^{O(k_1)\ldots O(k_a)})$ time.*

We will show that, for our purposes, we only require a constant number of quantifiers over polynomials of constant degree whose coefficients are bounded by a constant and have a constant number of variables. In Sec. 4.1 we gave an algorithm which requires several objective functions to be minimized; we formulate these objective functions as sentences in the first-order theory of real closed fields and apply Theorem 1 to obtain the optimal parameters to these objective functions. We note that the first-order sentences constructed in all of the lemmas in fact are guaranteed to be satisfiable, since all of our objective functions are guaranteed to have at least one set of parameter values for which they are minimized.

Lemma 1 *The sets of RDCs $D_1^* \in \mathcal{G}(D_1)$, $D_2^* \in \mathcal{G}(D_2)$, the rotation $\mathbf{R}^* \in SO(3)$, and the conformation $\mathcal{C}^* \in \mathcal{Y}(D_1^*, D_2^*, \mathbf{R}^*)$ that minimize Eq. (2) can be found in $c^{O(c^3)}$ time.*

Proof: Minimizing Eq. (2) subject to $\mathcal{Y}$ (as defined in Sec. 3) is equivalent to finding witnesses $D_1^* \in \mathcal{G}(D_1)$, $D_2^* \in \mathcal{G}(D_2)$, $\mathbf{R}^* \in SO(3)$, and $\mathcal{C}^* \in \mathcal{Y}(D_1^*, D_2^*, \mathbf{R}^*)$ for the first-order sentence:

$$\begin{aligned}
&\exists D_1^* \in \mathcal{G}(D_1), \exists D_2^* \in \mathcal{G}(D_2), \exists \mathbf{R}^* \in SO(3), \\
&\exists \mathcal{C}^* \in \mathcal{Y}(D_1^*, D_2^*, \mathbf{R}^*) : \forall D_1' \in \mathcal{G}(D_1), \forall D_2' \in \mathcal{G}(D_2), \\
&\forall \mathbf{R} \in SO(3), \forall \mathcal{C} \in \mathcal{Y}(D_1, D_2, \mathbf{R}) :: \\
&\sigma(D_1^*, D_2^*, \mathbf{R}^*, \mathcal{C}^*) \leq \sigma(D_1', D_2', \mathbf{R}, \mathcal{C}); \qquad (7)
\end{aligned}$$

recall that σ is defined by Eq. (2). We now analyze the running time of solving Eq. (7) by applying Theorem 1. First, we observe that Eq. (7) has degree $O(c)$, the same as that of Eq. (2); we will also argue below that the quantified sets are all of degree $O(c)$ as well. Recall that we argued in Sec. 3 that $\mathcal{Y}$ has degree $O(c)$. As stated, Eq. (7) has the same number of variables on the left and right hand side; we will now account for these variables. First, the set D_1^* (resp., D_2^*, D_1', and D_2') can be represented succinctly since we are only concerned with scalar error; that is, we can simply represent $r_{1,i}^*$ (resp., $r_{2,i}^*$, $r_{1,i}^*$, $r_{2,i}^*$) with a variable $\varepsilon_{1,i}$ with $-1 \leq \varepsilon_{1,i} \leq 1$ (resp., $\varepsilon_{2,i}$ with $-1 \leq \varepsilon_{2,i} \leq 1$, etc.) for $1 \leq i \leq c$. The variables $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$ add c equations of degree 1 and $2c$ variables to the first-order sentence, giving a total of $2c$ equations and $4c$ variables for both sides of the inequality. The variables $\mathbf{R}^*$ and $\mathbf{R}$ can be represented by using a quaternion representation of rotations; a quaternion can be represented using 4 variables and a quadratic equation. As mentioned in Sec. 3, the backbone (ϕ, ψ) angles in $\mathcal{Y}$ for both $\mathcal{C}^*$ and $\mathcal{C}$ in Eq. (7) are the roots of the polynomial equations for the unit vectors $w(\phi)$ and $w(\psi)$, which have degree $O(c)$ (due to the rotation $\mathbf{R}_i$ that must be applied to compute ϕ_i and ψ_i) and $2c$ variables. Since the i^{th} NH orientation can be written as a quartic equation (as described in Sec. 3), the summation in Eqs. (2) and (7) involving $b_{j,i}$, for $1 \leq i \leq c$, $j = 1, 2$, has degree $O(c)$ as well (due to the rotation $\mathbf{R}_i$ that must be applied and the square in each term of the summation) and $6c$ variables.

Thus, we have $3c$ equations, 1 inequality, and blocks of $4c + 5$, $2c$, and $6c + 5$ quantified variables. Note that the coefficients in our polynomial inequalities are a function of the experimental RDCs and the parameters of the alignment tensor, and that these coefficients are all bounded by constants. The maximum degree of the inequalities is $O(c)$, thus by Theorem 1 we can find the witnesses D_1^*, D_2^*, $\mathbf{R}^*$, and $\mathcal{C}^*$ to Eq. (7) in $c^{O(c^3)} \cdot O((3c+1)^{(4c+6)(2c+2)(6c+6)}) = c^{O(c^3)}$ time. ∎

Lemma 1 allows us to not only compute the conformation of a secondary structure element, but also constructs these conformations such that every secondary structure element has the correct global orientation. We can construct a predicate for the objective function Eq. (3) in a manner similar to Lemma 1 and apply Theorem 1 to find the required relative translation of any pair of successive secondary structure elements. The complete proof of Lemma 2 is provided in Appendix B.1.

Lemma 2 *For any successive pair of secondary structure elements, we can find a translation $x \in \mathbb{R}^3$ that minimizes Eq. (3) in $O(1)$ time.*

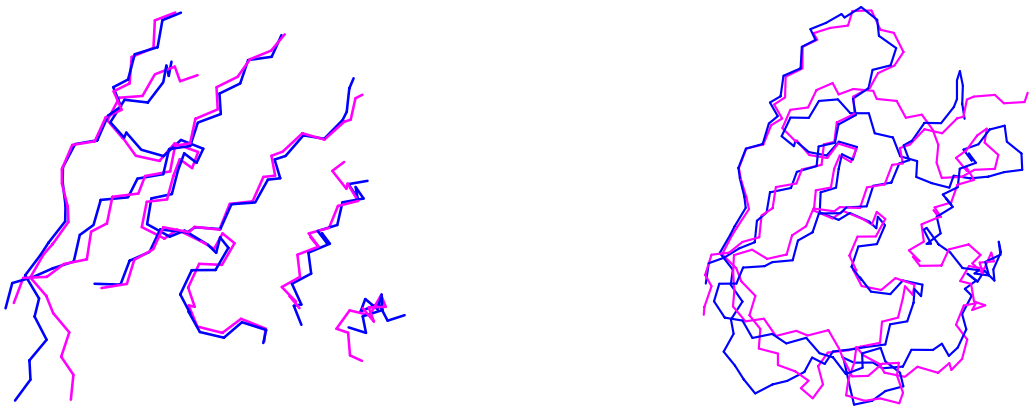
The first phase of the algorithm can be performed in $O(c^3)$, since a secondary structure element has size at most c. By Lemma 1, the second phase can be performed in $c^{O(c^3)}$ time for each secondary structure element, giving a total of $m \cdot c^{O(c^3)}$ time. The third phase runs in $O(m)$ time, since we can orient each successive pair of secondary structure elements in $O(1)$ time by Lemma 2. We then obtain the following:

Theorem 2 *The algorithm of Sec. 4.1 runs in $mc^{O(c^3)}$ time.*

Since in globular proteins $c = O(1)$ and $m = O(n)$, the running time of our algorithm is polynomial in n.

5 Experimental Results

As shown in Sec. 4, for globular proteins, our algorithm for structure determination provably runs in polynomial time. In this section, we discuss an implementation of our algorithm that has been applied to 6 real biological NMR data sets recorded for three structurally distinct proteins. Our implementation is quite fast compared to existing software; it took about 20 minutes per protein on average (over all datasets) on a single-processor Pentium-4 class machine. The NMR data used in our experiments on ubiquitin was taken from the PDB entry for 1D3Z (RDCs)



(a) (b)

Figure 1. Computed structures of ubiquitin. (a) The ubiquitin backbone structure excluding loop regions (blue) was computed by our algorithm using 37 NH and 39 CH RDCs, 12 hydrogen bonds, and 4 NOEs. Our structure has an RMSD of 0.97 Å when compared to the high-resolution X-ray structure (PDB ID: 1UBQ, in magenta) [42]. (b) We have also extended our algorithm to handle loop regions. The structure for ubiquitin (blue) was computed using 59 NH and 58 CH RDCs (117 out of 137 possible RDCs, 20 are missing), 12 H-bonds and 2 unambiguous NOEs. Our structure has a backbone RMSD of 1.45 Å with the high-resolution X-ray structure (magenta). The depicted structures consist of residues from Met1 to Arg72, since the C-terminal four residues of ubiquitin do not have a well-defined structure in solution.

and from NMR spectra from the Driscoll Lab [24] (NOEs). We first applied the algorithm to the protein human ubiquitin using 78 NH RDCs in two media [44, 43] or 76 NH and CH RDCs in a single medium, plus twelve hydrogen bonds and four NOE distances. For NH RDCs in 2 media, we obtained a structure with an RMSD of 1.23 Å when compared to the high-resolution X-ray structure (PDB ID: 1UBQ, in magenta) [42]. For CH and NH RDCs in 1 medium, we obtained a structure with an RMSD of 0.97 Å (Fig. 1a). We have also applied our algorithm to compute the backbone substructures using 4 other experimental data sets for two proteins, DNA-damage-inducible protein I and immunoglobulin binding protein G, using NH RDCs in two media (or NH and CH RDCs in one medium) and sparse distance restraints. Experimental RDC data for Dini (PDB ID: 1GHH) and Protein G (PDB ID: 3GB1) was taken from the Protein Data Bank (PDB). The backbone RMSDs between the substructures computed by our algorithm and the corresponding portions of previously-solved NMR structures are, respectively, 1.55 Å for DNA-damage-inducible protein I and 0.98 Å for immunoglobulin binding protein G. Table 1 gives a summary of the types and amount of experimental data used, as well as the accuracies of the computed backbone structures.

Table 2 shows the data requirements and accuracy of our algorithm versus other approaches, for ubiquitin. We note that the NMR structures we compared with were computed by MD/SA [5] using about 15 restraints per residue (including both NOE and RDC restraints). In contrast, our backbone structures have been computed using about 2.4 restraints per residue (2 RDCs and 0.4 distance restraints per residue). The fact that our algorithm needs very little RDC data (only 2 restraints per residue), and is still able to compute accurate structures is important for high-throughput applications as well as for structure-based drug design.

Finally, we have successfully extended our algorithm to compute a complete backbone structure, including turns and loops (connecting the secondary structure elements) using only NH and CH RDCs in a single medium (i.e., only 2 RDCs per residue) and 2 unambiguous NOEs. This algorithm, which computes the structure of the turn and loop regions also runs in polynomial-time for a globular protein with one additional assumption. We assume that our globular protein has $O(n)$ loop and turn regions, each with length $c_\ell = O(1)$; a majority of globular proteins indeed have short (constant-length) turn and loop regions. The final ubiquitin backbone structure computed by this algorithm has a 1.45 Å backbone RMSD from the X-ray structure; see Fig. 1b. This accuracy is similar to that of the ubiquitin backbone structure computed by a commonly-used heuristic approach [21] (see Table 2). The latter is the previous best result obtained for ubiquitin structure when using 6 or fewer RDCs per residue. Our accuracy is also better than the ubiquitin structure computed by [35]; they use 3 RDCs per residues plus 5 chemical shifts per residue as input to their algorithm. Furthermore, our algorithm is capable of handling up to 15% missing RDC data (Fig. 1b). Further

Protein[a]	α or β residues[b]	RDCs[c]	Type of RDCs[d]	Hydrogen bonds[e]	NOEs[f]	RMSD[g]
ubiquitin	39/75	78	NH in two media	12	4	1.23 Å
ubiquitin	41/75	76	NH, CH in one medium	12	4	0.97 Å
Dini	41/81	75	NH in two media	6	9	1.55 Å
Dini	41/81	80	NH, $C_\alpha C'$ in one medium	6	9	1.35 Å
Protein G	29/56	53	NH in two media	9	4	0.98 Å
Protein G	33/56	61	NH, $C_\alpha C'$ in one medium	9	4	1.30 Å

Table 1. Results of our algorithm. (a) experimental RDC data for ubiquitin (PDB ID: 1D3Z), Dini (PDB ID: 1GHH) and Protein G (PDB ID: 3GB1) were taken from the Protein Data Bank (PDB). (b) number of residues in α-helices or β-sheets versus total number of residues. (c) total number of RDCs used. (d) RDCs from different experimental datasets (for different bond vectors) were used. (e) number of hydrogen bonds used. (f) number of NOEs used. (g) RMSD computed between the oriented and translated secondary structure elements computed by our algorithm to existing structures: ubiquitin to a high-resolution X-ray structure (PDB ID:1UBQ); Dini to an NMR structure (PDB ID: 1GHH); and Protein G to an NMR structure (PDB ID: 3GB1).

Reference[a]	Program	Technique[b]	Restraints Per Residue[c]	Accuracy[d]
Brown *et al.* [21]	X-plor	MD/SA	6 RDCs	1.45 Å
Blackledge *et al.* [28]	SCULPTOR	MD/SA	11 RDCs,	1.00 Å
Bax *et al.* [17]	MFR	Database	10 RDCs, 5 Chemical shifts	1.21 Å
Baker *et al.* [35]	RosettaNMR	DataBase/MC	3 RDCs, 5 Chemical shifts	1.65 Å
Baker *et al.* [35]	RosettaNMR	DataBase/MC	1 RDC	2.75 Å
Our algorithm	–	Exact Equations	2 RDCs	1.45 Å

Table 2. Comparison of results for ubiquitin with existing approaches. (a) References to previously-computed ubiquitin backbone structures, (b) Algorithmic technique; (c) Data requirements; (d) Backbone RMSD of computed structure compared to the X-ray structure (PDB ID: 1UBQ) [42]; our algorithm used NH and CH RDCs in a single medium and 2 unambiguous NOEs.

details of the extended algorithm to handle loops and turns can be found in [45, Appendix C].

6 Conclusion

In this paper, we have shown that the global properties of residual dipolar couplings can be used to develop a polynomial-time algorithm for *de novo* high-resolution protein structure determination. This is the first polynomial-time algorithm for *de novo* structure determination from any type of experimental data. Furthermore, we have shown that in practice, on real biological NMR data, that our algorithm is as good or better in terms of accuracy and speed, and requires less data, than existing NMR structure determination techniques.

Our polynomial-time backbone structure determination algorithm can be extended to compute *complete* protein structures (including side-chains), since exact equations analogous to Eqs. (5) and (6) can be derived *mutatis mutandis* to compute the side-chain dihedral angles $\chi_1, \chi_2, \ldots$ from experimentally-recorded side-chain RDCs. In this case, the average angles ϕ_a and ψ_a in Eq. (2) would be replaced with side-chain rotamer angles $\chi_{a,1}, \chi_{a,2}, \ldots$. Finally, our algorithm might also be extended to speed up the structure determination of nucleic acids, since similar exact equations (from DNA and RNA RDCs) can easily be derived to compute the backbone torsion and χ angles in nucleic acids.

References

[1] M. Andrec, P. Du, and R. M. Levy. *J. Biomol. NMR*, 21(4):335–347, 2001.

[2] C. Bailey-Kellogg, A. Widge, J. J. Kelley, M. J. Berardi, J. H. Bushweller, and B. R. Donald. *J. Comput. Biol.*, 7(3–4):537–558, 2000.

[3] S. Basu, R. Pollack, and M.-F. Roy. *Algorithms in Real Algebraic Geometry*. Springer-Verlag, 2003.

[4] B. Berger, J. Kleinberg, and F. T. Leighton. *Journal of the ACM*, 46(2):212–235, 1999.

[5] A. T. Brünger. *XPLOR: A system for X-ray crystallography and NMR*. Yale University Press: New Haven, 1993.

[6] S. Cabello, E. Demaine, and G. Rote. In *Proc. of the 11th Symposium on Graph Drawing*, pages 283–294, 2003.

[7] J. Canny. In *Proc. of the 28th IEEE Conference on Foundations of Computer Science*, pages 39–48, 1987.

[8] J. Canny. *Computer Journal*, 36:409–418, 1993.

[9] J. Canny, B. R. Donald, and G. Ressler. In *Proc. ACM Symposium on Computational Geometry*, pages 251–260, Berlin, June 1992.

[10] J. Cavanagh, Fairbrother W. J., A. G. Palmer III, and N. J. Skelton. *Protein NMR Spectroscopy: Principles and Practice*. Academic Press, 1995.

[11] G. M. Clore. *Proc. Nat. Acad. Sci. USA*, 97:9021–9025, 2000.

[12] G. E. Collins. In *Springer Lecture Notes in Computer Science*, volume 33, pages 515–532, 1975.

[13] F. Cordier, M. Rogowski, S. Grzesiek, and A. Bax. *J. Magn. Reson.*, 40(2):510–512, 1999.

[14] G. Crippen. *J. Math. Chem.*, 6:307–324, 1991.

[15] G. M. Crippen and T. F. Havel. *Distance Geometry and Molecular Conformations*. Wiley, 1988.

[16] F. del Rio-Portílla, V. Blechta, and R. Freeman. *J. Magn. Reson.*, 111a:132–135, 1994.

[17] F. Delaglio, G. Kontaxis, and A. Bax. *J. Am. Chem. Soc.*, 122(9):2142–2143, 2000.

[18] B. R. Donald, D. Kapur, and J. L. Mundy. *Symbolic and Numerical Computation for Artificial Intelligence*. Academic Press, Boston, MA, 1992.

[19] A. C. Fowler, F. Tian, H. M. Al-Hashimi, and J. H. Prestegard. *J. Mol. Biol.*, 304(3):447–460, 2000.

[20] K. H. Gardner and L. E. Kay. *J. Am. Chem. Soc.*, 119(32):7599–7600, 1997.

[21] A. W. Giesen, S. W. Homans, and J. M. Brown. *J. Biomol. NMR*, 25:63–71, 2003.

[22] D.Y. Grigor'ev. *Journal of Symbolic Computation*, 5(1–2):65–108, February/April 1988.

[23] P. Güntert, C. Mumenthaler, and K. Wüthrich. *J. Mol. Biol.*, 273:283–298, 1997.

[24] R. Harris. The ubiquitin NMR Resource Page, BBSRC Bloomsbury Center for Structural Biology. `http://www.biochem.ucl.ac.uk/bsm/nmr/ubq/`, 2002.

[25] J. Heintz, M.-F. Roy, and P. Solernò. *Bull. Soc. Math France*, 118:101–126, 1990.

[26] B. Hendrickson. *SIAM Journal on Computing*, 21:65–84, 1992.

[27] B. Hendrickson. *SIAM Journal on Optimization*, 5:835–857, 1995.

[28] J. C. Hus, D. Marion, and M. Blackledge. *J. Am. Chem. Soc.*, 123:1541–1542, 2001.

[29] J. A. Losonczi, M. Andrec, M. W. Fischer, and J. H. Prestegard. *J. Magn. Reson.*, 138(2):334–342, 1999.

[30] A. Marin, T. E. Malliavin, P. Nicolas, and M. A. Delsuc. *J. Biomol. NMR*, 30(1):47–60, 2004.

[31] J. J. Moré and Z. Wu. In P. M. Pardalos, D. Shalloway, and G. Xue, editors, *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*. American Mathematical Society, 1995.

[32] J. Ponce and D. J. Kriegman. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*. Academic Press, Boston, MA, 1992.

[33] J. H. Prestegard, C. M. Bougault, and A. I. Kishore. *Chem. Rev.*, 104(8):3519–3540, 2004.

[34] J. Renegar. *Journal of Symbolic Computation*, 13(3):255–352, 1992.

[35] C. A. Rohl and D. Baker. *J. Am. Chem. Soc.*, 124(11):2723–2729, 2002.

[36] J. B. Saxe. In *Proc. of the 17th Allerton Conference on Communications, Control, and Computing*, pages 480–489, 1979.

[37] N. R. Skrynnikov and L. E. Kay. *J. Biomol. NMR*, 18(3):239–252, 2000.

[38] A. Tarski. *A decision method for elementary algebra and geometry*. University of California Press, 1951.

[39] F. Tian, H. Valafar, and J. H. Prestegard. *J. Am. Chem. Soc.*, 123(47):11791–11796, 2001.

[40] N. Tjandra and A. Bax. *Science*, 278:1111–1114, 1997.

[41] J. R. Tolman, J. M. Flanagan, M. A. Kennedy, and J. H. Prestegard. *Proc. Natl. Acad. Sci. USA*, 92:9279–9283, 1995.

[42] S. Vijay-Kumar, C. E. Bugg, and W. J. Cook. *J. Mol. Biol.*, 194:531–544, 1987.

[43] L. Wang and B. R. Donald. In *IEEE Computational Systems Bioinformatics Conference*, pages 319–330, 2004.

[44] L. Wang and B. R. Donald. *J. Biomol. NMR*, 29:223–242, 2004.

[45] L. Wang and B. R. Donald. An Efficient and Accurate Algorithm for Assigning Nuclear Overhauser Effect Restraints Using a Rotamer Library Ensemble and Residual Dipolar Couplings. In *IEEE Computational Systems Bioinformatics Conference*, 2005. To appear.

[46] Y. X. Wang, J. Jacob, F. Cordier, P. Wingfield, S. J. Stahl, S. Lee-Huang, D. Torchia, S. Grzesiek, and A. Bax. *J. Biomol. NMR*, 14(2):181–184, 1999.

[47] W. J. Wedemeyer, C. A. Rohl, and H. A. Scheraga. *J. Biomol. NMR*, 22:137–151, 2002.

[48] D.S. Wishart and B.D. Sykes. *J. Biomol. NMR*, 4:171–180, 1994.

[49] D.S. Wishart, B.D. Sykes, and F. M. Richards. *J. Mol. Biol.*, 222(2):311–33, 1991.

[50] D.S. Wishart, B.D. Sykes, and F. M. Richards. *Biochemistry*, 31(6):1647–1651, 1992.

Appendix

In Appendix A, we describe our implementation in further detail. In Appendix B, we give proofs of Lemma 2 and Props. 1 and 2 from Sec. 3 of the text.

A Implementation

Practical algorithms for quantifier elimination and the existential theory of real closed fields have been efficiently implemented [7, 32] to find the minima of objective functions that are similar to Eqs. (2) and (3). In our implementation, the second phase of the algorithm was implemented as a systematic depth-first search along with a pruning criterion that only considers (ϕ, ψ) angles that are in

the algebraic subset defined by $\mathcal{Y}$ and in the Ramachandran region of the current secondary structure type. While there is a long history of validating exact algorithms using implementations that contain numerical subroutines,[3] these codes must be tested on real data to verify robustness and accuracy. Our algorithm is combinatorially precise and uses exact algebraic numbers; to test it in practice we implemented some subroutines exactly (i.e., the closed-form exact solutions for internuclear NH and CH bond vectors and backbone (ϕ, ψ) angles, and used a discrete, combinatorial tree-search over the algebraic cross-product $\mathcal{Y}$ of possible solutions) and some numerically (i.e., we used a grid search over $SO(3)$ for the orientation of the first peptide plane and over $\mathbb{R}^3$ to find translations between successive secondary structure elements) for both implementation speed and to avoid some technical issues in approximating rational rotations [18, pages 1–23] [9]. In practice, the implementation took about 20 minutes on average over all datasets on a single-processor Pentium-4 class machine.

B Proofs

In this section, we give the details of proofs omitted from the text. In Appendix B.1, we give a proof of Lemma 2, and in Appendix B.2 we give a more detailed presentation of Props. 1 and 2.

B.1 Computing Relative Translations

In this section, we give the proof of Lemma 2. While it is similar to, and simpler than, the proof of Lemma 1, we include it here for completeness.

Lemma 2 *For any successive pair of secondary structure elements, we can find a translation $x \in \mathbb{R}^3$ that minimizes Eq. (3) in $O(1)$ time.*

Proof: Consider a successive pair of secondary structure elements $\mathcal{A}$ and $\mathcal{B}$ and without loss of generality fix ℓ, $2 \leq \ell \leq c$, and the distances N derived from the experimental NOE restraints N. Let $A = \{a_1, a_2, \ldots, a_\ell\}$ (resp., $B = \{b_1, b_2, \ldots, b_\ell\}$) be the 3D coordinates of the ℓ nuclei in $\mathcal{A}$ (resp., $\mathcal{B}$) that correspond to the distances in N. Minimizing Eq. (3) is equivalent to finding a witness x^* such that:

$$\exists x^* \in \mathbb{R}^3 : \forall x \in \mathbb{R}^3 :: \sum_{i=1}^{\ell} \left(\|a_i - b_j + x^*\| - n_i \right)^2 \leq \sum_{i=1}^{\ell} \left(\|a_i - b_j + x\| - n_i \right)^2 . \tag{8}$$

This predicate has degree at most 4, and 2 blocks of 6 quantified variables. In this predicate, the largest coefficient is at most the square of the maximum distance in N. We note that there is an inherent upper bound on NOE restraints of about 6 Å, thus the coefficients are all bounded by a constant. The running time of finding a witness x^* for Eq. (8) is then $O(2^{7 \cdot 7}) = O(1)$. (Remark: Without this bound on the NOE distance restraints, the coefficients in the inequalities are bounded by the diameter of the protein, which would increase the running time by a factor logarithmic in the protein diameter.) ∎

B.2 Equations for computing backbone dihedral angles from RDCs

In this section, we give a more detailed presentation of the method to compute backbone dihedral angles from RDCs in two aligning media exactly and in constant time per residue. We show that it is possible to derive, from the physics of RDCs, low-degree monomials (with degree at most 4) whose solutions give the backbone (ϕ, ψ) angles. Due to space considerations, we sketch the proofs here; the interested reader can refer to [44] for further details of the proofs and equations. As before, we assume that the dipolar interaction constant D_{max} is equal to 1. By considering a global coordinate frame which diagonalizes the alignment tensor, Eq. (1) becomes:

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

where S_{xx}, S_{yy}, and S_{zz} are the three diagonal elements of a diagonalized Saupe matrix $\mathbf{S}$ (the alignment tensor), and x, y, and z are, respectively, the $x, y, z-$components of the unit vector $\mathbf{v}$ in a principal order frame (POF) which diagonalizes $\mathbf{S}$. Now, $\mathbf{S}$ is a 3×3 symmetric, traceless matrix with five independent elements [40, 41]. Given NH RDCs in two aligning media, the associated NH vector $\mathbf{v}$ must lie on the intersection of two conic curves [37, 47]. We show

Proposition 1 *Given the diagonal Saupe elements S_{xx} and S_{yy} for medium 1, S'_{xx} and S'_{yy} for medium 2, and a relative rotation matrix $\mathbf{R}_{12}$ between the POFs of medium 1 and 2, the square of the x-component of the unit vector $\mathbf{v}$ satisfies a monomial quartic equation.*

The following is a sketch of the proof. The methods for the computation of the seven parameters (S_{xx}, S_{yy}, S'_{xx}, S'_{yy}, and $\mathbf{R}_{12}$) and the full expressions for the polynomial

[3] M. A. Erdmann and T. Lozano-Perez, *Algorithmica*, 2(4):477–521, 1987; J. Canny and B. R. Donald, *Discrete and Computational Geometry*, 3(3):219–236, 1988; K.-F. Böhringer, B. R. Donald, and N. MacDonald, In *Proc.* International Workshop on the Algorithmic Foundations of Robotics (WAFR), 1996; B. R. Donald, *Algorithmica*, 5(3):353–382, 1990; B. R. Donald, *IEEE Trans. on Robotics and Automation*, 8(2), 1992; B. R. Donald, *Algorithmica*, 10(2/3/4):91–101, 1993; R. G. Brown and B. R. Donald, *Algorithmica*, 26(3/4):515–559, 2000; M. A. Erdmann, In *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology*, pages 195–204, 2004.

coefficients and temporary variables (a_2, b_2, c_1, etc.) can be found in [44].

Proof Sketch: Fix a backbone NH vector $\mathbf{v}$ along the backbone and let r and r' be the experimental RDCs for $\mathbf{v}$ in the first and second medium, respectively. From Eq. (4) we have

$$r = S_{xx}x^2 + S_{yy}y^2 + S_{zz}z^2$$
$$r' = S'_{xx}x'^2 + S'_{yy}y'^2 + S'_{zz}z'^2$$
$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

where r is the RDC value, x, y, z are the x, y, z-components of $\mathbf{v}$ in a POF of medium 1, r' and x', y', z' are the corresponding variables for medium 2. Eliminating x', y', and z', we have

$$r_2 = a_2x^2 + b_2y^2 + c_1xy + c_2xz + c_3yz, \quad (9)$$
$$r_1 = a_1x^2 + b_1y^2, \quad (10)$$

where $a_2 = (S'_{xx} - S'_{zz})(R_{11}^2 - R_{13}^2) + (S'_{yy} - S'_{zz})(R_{21}^2 - R_{23}^2)$ and $c_2 = 2(S'_{xx} - S'_{zz})R_{11}R_{13} + 2(S'_{yy} - S'_{zz})R_{21}R_{23}$, and $b_2, c_1, c_2, c_3, a_1, b_1$ are similar constants; full details are given in [44].

Eliminating z from Eq. (9) we obtain

$$d_8x^4 + d_7x^3y + d_6x^2y^2 - d_5x^2 + d_4xy^3 - d_3xy - d_2y^2 + d_1y^4 + d_0 = 0, \quad (11)$$

where $d_8 = a_2^2 + c_2^2$, and $d_7, d_6, ..., d_0$ are analogously defined; these are defined fully in [44]. Eq. (11) is a degree 8 monomial in x after direct elimination of y using Eq. (10). However, it can be reduced to a quartic equation by substitution since only the terms with the degrees of 0, 2, 4, and 8 appear in it. Introducing new variables t and u such that

$$x = a\sin t, \qquad y = b\cos t, \qquad u = \cos 2t, \quad (12)$$

and through algebraic manipulation we finally obtain

$$f_4u^4 + f_3u^3 + f_2u^2 + f_1u + f_0 = 0. \quad (13)$$

The full expressions for coefficients a, b and f_0, f_1, f_2, f_3, f_4 are given in [44]. Since $u = 1 - 2(\frac{x}{a})^2$ Eq. (13) is also a quartic equation in x^2. ■

The y-component of $\mathbf{v}$ can be computed directly from Eq. (12). Due to two-fold symmetry in the RDC equation the number of real solutions for $\mathbf{v}$ is at most 8. We will refer to the bond vector between the N and C_α atoms as the NC_α vector. Given two unit vectors in consecutive peptide planes we can use backbone kinematics to derive quadratic equations to compute the sines and cosines of the (ϕ, ψ) angles:

Proposition 2 *Given the* NH *unit vectors* $\mathbf{v}_i$ *and* $\mathbf{v}_{i+1}$ *of residues* i *and* $i+1$ *and the* NC_α *vector of residue* i*, the sines and cosines of the intervening backbone dihedral angles* (ϕ, ψ) *satisfy the trigonometric equations* $\sin(\phi + a_1) = b_1$ *and* $\sin(\psi + a_2) = b_2$*, where* a_1 *and* b_1 *are constants depending on* $\mathbf{v}_i$ *and* $\mathbf{v}_{i+1}$*, and* a_2 *and* b_2 *depend on* $\mathbf{v}_i$*,* $\mathbf{v}_{i+1}$*,* $\sin\phi$ *and* $\cos\phi$*. Furthermore, exact solutions for* $\sin(\phi)$ *and* $\cos(\phi)$ *can be computed from a quadratic equation by the substitution* $w = \tan\frac{\phi}{2}$*,* $\sin\phi = 2w/(1+w^2), \cos\phi = (1-w^2)/(1+w^2)$*; equations for* $\sin\psi$ *and* $\cos\psi$ *can be obtained and solved exactly by a similar substitution.*

The following is a sketch of the proof. Full expressions for the polynomial coefficients and temporary variables $(x_1, y_1, z_1, x_2, y_2, z_2, a_1, b_1, a_2, b_2)$ introduced in the proof are given in [44].

Proof Sketch: Following a procedure similar to kinematics the two NH vectors $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$ can be related by 8 rotation matrices between two coordinate systems in peptide planes i and $i+1$:

$$\mathbf{v}_i = \mathbf{R}_x(\theta_7)\mathbf{R}_y(\theta_6)\mathbf{R}_x(\theta_5)\mathbf{R}_z(\psi+\pi)\cdot \mathbf{R}_x(\theta_3)\mathbf{R}_y(\phi)\mathbf{R}_y(\theta_8)\mathbf{R}_x(\theta_1)\mathbf{v}_{i+1}. \quad (14)$$

The definitions of the coordinate systems, the expressions for the rotation matrices $\mathbf{R}_x$, $\mathbf{R}_y$, and $\mathbf{R}_z$ and the definitions of the six backbone angles $(\theta_1, \theta_3, \theta_5, \theta_6, \theta_7$ and $\theta_8)$ are given in [44]. The backbone (ϕ, ψ) angles are defined according to the standard convention. Given the values of these six angles $\mathbf{R}_l = \mathbf{R}_x(\theta_7)\mathbf{R}_y(\theta_6)\mathbf{R}_x(\theta_5)$ and $\mathbf{R}_r = \mathbf{R}_y(\theta_8)\mathbf{R}_x(\theta_1)$ are two 3×3 *constant* matrices. Define two new vectors $\mathbf{w}_1 = (x_1, y_1, z_1) = {\mathbf{R}_l}^{-1}\mathbf{v}_i$ and $\mathbf{w}_2 = (x_2, y_2, z_2) = \mathbf{R}_r\mathbf{v}_{i+1}$ to obtain

$$\begin{aligned} x_1 =& -(\cos\phi\cos\psi + \sin\theta_3\sin\phi\sin\psi)\, x_2 - \\ & \cos\theta_3\sin\psi\, y_2 + (\cos\psi\sin\phi - \cos\phi\sin\theta_3\sin\psi)\, z_2 \\ y_1 =& (\cos\phi\sin\psi - \sin\theta_3\sin\phi\cos\psi)\, x_2 - \\ & \cos\theta_3\cos\psi\, y_2 - (\sin\phi\sin\psi + \cos\phi\sin\theta_3\cos\psi)\, z_2 \\ z_1 =& \cos\theta_3\sin\phi\, x_2 - \sin\theta_3\, y_2 + \\ & \cos\theta_3\cos\phi\, z_2. \end{aligned} \quad (15)$$

By Eq. (15) we can then obtain a simple trigonometric equation:

$$\sin(\phi + a_1) = b_1 \quad (16)$$

where $b_1 = \frac{z_1 + y_2\sin\theta_3}{\sqrt{(x_2\cos\theta_3)^2 + (z_2\cos\theta_3)^2}}$, and a_1 is a similar constant; see [44] for details. $\sin\phi$ and $\cos\phi$ can be computed from a quadratic equation by the substitution $w = \tan\frac{\phi}{2}, \sin\phi = \frac{2w}{1+w^2}, \cos\phi = \frac{1-w^2}{1+w^2}$. Substituting the computed $\sin\phi$ and $\cos\phi$ into Eq. (16) we can obtain another simple trigonometric equation:

$$\sin(\psi + a_2) = b_2. \quad (17)$$

$\sin\psi$ and $\cos\psi$ can be computed similarly from a quadratic equation where both a_2 and $b_2 \leq 1$ are computed from $y_1, x_2, y_2, z_2, \theta_3$ and $\sin\phi$ and $\cos\phi$. ■