JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 13, Number 7, 2006
© Mary Ann Liebert, Inc.
Pp. 1267–1288

A Polynomial-Time Algorithm for *De Novo* Protein Backbone Structure Determination from Nuclear Magnetic Resonance Data

LINCONG WANG,[1] RAMGOPAL R. METTU,[2] and BRUCE RANDALL DONALD[3,4,5]

ABSTRACT

We describe an efficient algorithm for protein backbone structure determination from solution Nuclear Magnetic Resonance (NMR) data. A key feature of our algorithm is that it finds the conformation and orientation of secondary structure elements as well as the global fold in *polynomial time*. This is the first polynomial-time algorithm for *de novo* high-resolution biomacromolecular structure determination using experimentally recorded data from either NMR spectroscopy or X-ray crystallography. Previous algorithmic formulations of this problem focused on using local distance restraints from NMR (e.g., nuclear Overhauser effect [NOE] restraints) to determine protein structure. This approach has been shown to be NP-hard, essentially due to the local nature of the constraints. In practice, approaches such as molecular dynamics and simulated annealing, which lack both combinatorial precision and guarantees on running time and solution quality, are used routinely for structure determination. We show that residual dipolar coupling (RDC) data, which gives global restraints on the orientation of internuclear bond vectors, can be used in conjunction with very sparse NOE data to obtain a polynomial-time algorithm for structure determination. Furthermore, an implementation of our algorithm has been applied to six different real biological NMR data sets recorded for three 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.

Key words: Nuclear Magnetic Resonance, structural biology, protein structure, residual dipolar couplings, algorithms, computational molecular biology, NP-completeness.

1. INTRODUCTION

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

[1] CABM Structural Bioinformatics Laboratory, Rutgers University, Piscataway, NJ.

[2] Department of Electrical and Computer Engineering, University of Massachusetts at Amherst.

[3] Department of Computer Science, Duke University, Durham, NC.

[4] Department of Biochemistry, Duke University Medical Center, Durham, NC.

[5] *Corresponding author*: Bruce Randall Donald, P.O. Box 90129, Department of Computer Science, Duke University, Durham, NC 27708-0129.

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 (Saxe, 1979), and rigorous approaches to structure determination based on solving this problem, such as the distance geometry method (Crippen, 1991; Crippen and Havel, 1988), require exponential time. These lower-bound arguments are based on showing that for certain counterexamples, distance geometry algorithms require time exponential in the size of the protein (assuming $P \neq NP$). Distance geometry can perform better in practice due to the use of additional restraints that are available *a priori* (e.g., bond angles and lengths). However, it is interesting to ask if there is a *provably* polynomial-time algorithm to determine protein structure from experimental data, since this would improve our understanding of the data and be useful in devising practical approaches to structure determination that come with worst-case guarantees on *both* running time and solution quality. The most commonly used structure determination protocols use experimental NMR data along with techniques such as molecular dynamics (MD) and simulated annealing (SA). These approaches, however, lack combinatorial precision, guarantees on running time, as well as guarantees on solution quality. The interatomic distance restraints used by distance-based structure determination algorithms must be obtained by *assigning* NOE data. The *NOE assignment problem* asks us to determine, for every NOE restraint, the associated pair of protons in the primary sequence. In its unassigned form, an NOE restraint gives the information that two nuclei are approximately d Å apart ($1.8 \leq d \leq 6$) but not the identity of the two nuclei (in the primary sequence). Automated methods can be used to quickly obtain *resonance* assignments, but automated *NOE assignment* typically requires hours to weeks, since NOE assignment often sits in a tight inner loop of structure refinement (e.g., ARIA [Nilges et al., 1997], CANDID [Herrmann et al., 2002], AUTO-STRUCTURE [Huang et al., 2003], and PSAD [Juszewski et al., 2004]). Furthermore, it is not uncommon to need manual intervention (e.g., to assign side-chain NOEs [Clore, 2000]) to obtain enough distance restraints (in conjunction with *a priori* restraints such as bond angles and bond lengths) to compute an accurate NMR structure. Since our algorithm uses RDC data, it requires only a minimal set of distance restraints, and thus relaxes the requirement for a complete assignment of NOE restraints.

In recent years, residual dipolar coupling (RDC) (Tjandra and Bax, 1997; Tolman et al., 1995) data has been used to provide global orientational restraints on the protein structure (Fowler et al., 2000; Prestegard et al., 2004; Tian et al., 2001). 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 (Andrec et al., 2001; Delaglio et al., 2000; Giesen et al., 2003; Hus et al., 2001; Rohl and Baker, 2002; Tian et al., 2001), 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 with regular secondary structure comprise the a large fraction of proteins in nature, and are far more abundant than fibrous proteins (e.g., collagen or coiled-coil oligomers). If we consider proteins with regular secondary structure, this assumption implies that the secondary structure elements have 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.

There is a tradition in computer science to measure the performance of an algorithm by the worst-case asymptotic complexity of its running time as a function of input size. Globular proteins with regular secondary structure have a natural size limitation throughout the biosphere, and NMR techniques are similarly limited in the size of protein they can deal with. However, it is interesting to ask if there is a provably polynomial-time algorithm to determine protein structure from experimental data. Such an algorithm is of interest to the NMR community, since it would quantify what NMR experiments are *neccessary* (existing approaches record a *sufficient* amount of data) for structure determination; additionally, it would have practical implications, since structure determination is often used a subroutine in other applications (Nilges et al., 1997; Herrmann et al., 2002; Huang et al., 2003; Juszewski et al., 2004). 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) two RDCs of backbone internuclear 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 Section 6 uses this experimental data, and allows for missing data as well. In contrast to NOE assignment, RDCs can be recorded and assigned on the order of hours. Additionally, it is relatively straightforward to rapidly obtain the few (three or four), unambiguous NOEs required for our algorithm from a standard NOESY spectrum, or by using, for example, the labeling strategy of Gardner and Kay (1997). The secondary structure types of residues along the backbone can be determined by NMR from experimentally-recorded scalar coupling HNHA (Cavanagh et al., 1995, pages 524–528) data, or J-doubling (del Rio-Portílla et al., 1994) data for larger proteins (these experiments report on the ϕ backbone angles). NMR chemical shifts (Marin et al., 2004; Wishart and Sykes, 1994; Wishart et al., 1991, 1992) or automated assignment (Bailey-Kellogg et al., 2000a) can also be used. Hydrogen bonds can be determined by NMR from experimentally recorded data (Cordier et al., 1999; Wang et al., 1999), or, e.g., by using backbone resonance assignment programs such as JIGSAW (Bailey-Kellogg et al., 2000a). The user of our algorithm has a choice, to record either (a) one type of backbone RDC (such as NH RDCs) in two aligning media, or (b) two types of backbone RDCs (such as NH and CH RDCs) in a single medium. This flexibility allows our algorithm to be applied to a wider range of proteins. NH RDCs in two media allows the experimental RDCs to be collected on an ^{15}N-labelled sample, which is an order of magnitude cheaper to prepare than a doubly-labelled ^{15}N/^{13}C sample. However, it is not always straightforward to find two aligning media for a protein; in this case our algorithm can also use NH and CH RDCs in a single medium since recording an extra set of RDCs in the same medium requires only slightly more spectrometer time. 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 Section 3 (Wang and Donald, 2004a). Additionally, while our implementation requires hydrogen bond information to impose additional constraint on β-sheets, we omit the discussion of incorporating hydrogen bond information for the sake of brevity. Our problem definition needs to be modified only slightly to incorporate this data, and all our theoretical results still hold (see Section 2 for references and further discussion).

A key building block of our algorithm makes use of *exact*, low-degree polynomial equations (Wang and Donald, 2004b) 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 four) 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 fits 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 uses the *theory of real closed fields* (Basu et al., 2003; Grigor'ev, 1988), 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 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 using many more restraints. Under our assumption that the protein is globular and has regular secondary structure, our algorithm runs in polynomial time. We note that using our techniques, we can also obtain a polynomial-time algorithm even when the secondary structure elements are allowed to be arbitrarily long; the tradeoff being that more experimental NOEs are required (see Section 5). We have previously analyzed the running time of an implementation of our algorithm (Wang and Donald, 2004b) when the length of a secondary structure element is a parameter k, $1 \leq k \leq n$. In this case, the worst-case running time is exponential in k. In practice however, our algorithm is still quite fast in terms of expected running time; an average case analysis (Wang and Donald, 2004a) shows that the base of exponential term in the running time is quite small, about 1.1.

Our result is consistent with previous observations (Andrec et al., 2001; Delaglio et al., 2000; Fowler et al., 2000; Giesen et al., 2003; Hus et al., 2001; Prestegard et al., 2004; Rohl and Baker, 2002; Tjandra and Bax, 1997; Tian et al., 2001; Tolman et al., 1995; Wedemeyer et al., 2002) that, empirically, RDCs increase the speed and accuracy of biomacromolecular structure determination, and formally quantitates the 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 use low-degree polynomial equations that 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 provably 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, previous 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 from with a complete and exact distance matrix; this problem can be solved exactly using SVD in $O(n^3)$ time. Recent work (Dong and Wu, 2002) gives an $O(n)$-time algorithm for this problem. Berger et al. (1999) 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 consistent 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. Thus, distance-based structure determination approaches also use *a priori* restraints, such as bond angles and bond lengths, to ensure the problem is not underconstrained. A number of theoretical studies have been undertaken to examine the relationship between distance restraints (with error) and the time complexity of structure determination. Saxe (1979) 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 (1992, 1995) studies conditions under which embedding such a graph is even possible, and gives (super-polynomial time) algorithms for the problem. Crippen and Havel (1988) 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 (1979), and existing algorithms for solving the distance geometry problem require exponential time in the worst-case (Crippen and Havel, 1988).

Traditional NMR structure determination algorithms such as Brünger (1993) and Güntert et al. (1997) were initially designed to use NOE-derived distance restraints. Even recently developed RDC-based structure determination approaches rely on heuristic approaches such as simulated annealing or molecular dynamics (Giesen et al., 2003; Hus et al., 2001) or a structural database (the PDB) (Delaglio et al., 2000; Rohl and Baker, 2002) in order to compute a well-defined backbone structure. In Delaglio et al. (2000), the computed structure is obtained using a gradient-descent approach, while in Rohl and Baker (2002), a Monte-Carlo–based algorithm is used; both of these approaches can only guarantee that the given objective function value achieves a local minimum. Table 2 in Section 6 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 Wang and Donald (2004b, 2004a) 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 Sections 5 and 6).

2. PRELIMINARIES AND FORMAL PROBLEM DEFINITION

Informally, our formulation of the structure determination problem assumes that we are given the following experimental NMR data: (a) two 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. Our goal is to find a backbone conformation that is the best-fit to the given experimental RDCs and NOEs. Figure 1 shows the relationship between the given experimental data and protein backbone. We first discuss the input experimental data and our approach, and then present a formal problem definition.

The equation for the RDC r associated with an internuclear bond vector $\mathbf{v}$ can be written (Saupe, 1968) 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). 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 two 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) two 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 Section 3 (Wang and Donald, 2004a).

Given an alignment tensor, 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. Only the relative translation for each pair of secondary structure elements that best fit the given NOE restraints must be 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 and/or dynamics, experimentally-recorded RDCs cannot in general be fit to a secondary structure element unless they are perturbed (within some error window) (Wang and Donald, 2004b). 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 $b_{j,1}(\mathbf{R}) = D_{max}\mathbf{v}_1(\mathbf{R})^T\mathbf{S}_j\mathbf{v}_1(\mathbf{R})$ 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 $2 \le i \le c$ 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 Section 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 $2 \le i \le c$ 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

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

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 $1 \le i < c$), 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 Equation (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 Equation (2) is underconstrained. Given 2 RDCs for residue i, the NH bond vector $\mathbf{v}_i$ must lie in a finite set, defined by a quartic monomial (Wang and Donald, 2004b). This, in turn, constrains (ϕ_i, ψ_i) to lie in a finite algebraic set, defined by backbone kinematics (Wang and Donald, 2004b). Hence, the optimization[1] in Equation (2) is performed over a finite algebraic subset of a $2(c-1)$-torus (see Section 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 Fig. 6 of Wang and Donald [2004b, page 234]). To compute relative translations, we require at least three 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 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

[1] For simplicity of analysis, we have omitted the distinction between α-helices and β-sheets in the definition of Equation (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 Section 4; this additional term in the objective function is discussed in detail elsewhere. To handle hydrogen bond geometry in β-sheets, we use Equation (9) in Wang and Donald (2004b, page 228) as the additional term and make use of the techniques of Lemma 4.2 to cope with the additional term in the objective function (see Section 4.2).

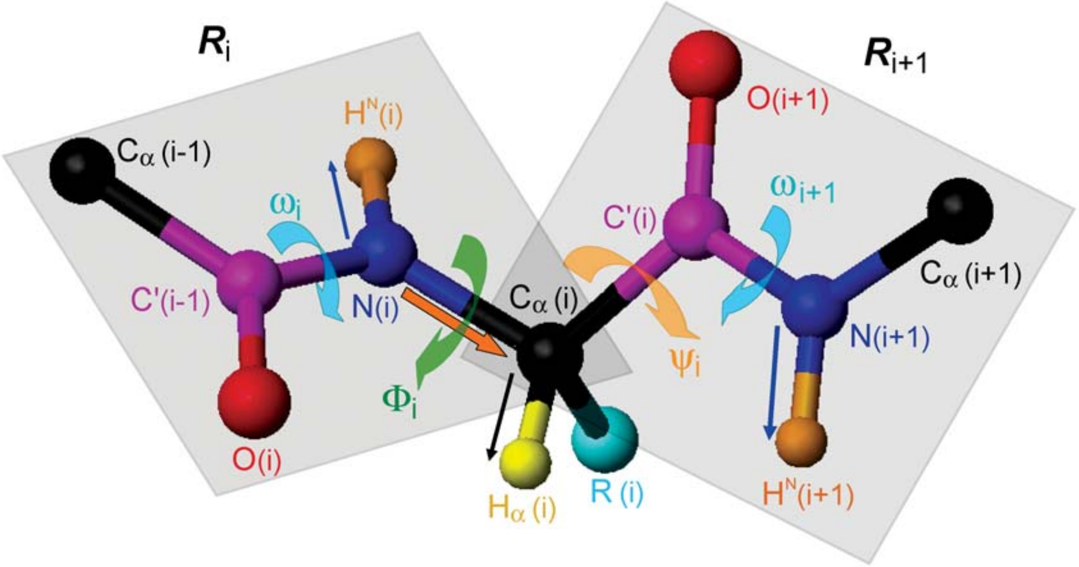

FIG. 1. The protein backbone structure illustrating RDCs used for computing bond vector orientation and backbone ϕ and ψ angles. Our algorithm uses either one type of backbone RDC (such as NH RDCs) measured in two different aligning media or two types of RDCs (such as CH and NH RDCs) measured in a single medium. The bond vectors whose RDCs are used in our algorithm are indicated by thin arrows. In our algorithm, we use typical values for bond length and angle as well as the peptide plane dihedral angle (ω). The orientation of NH vector in principal order frame (POF) can be computed exactly from NH RDCs in two media by solving a quartic equation (see Section 3). For NH RDCs in two media, the sine and cosine of the backbone ϕ_i and ψ_i angles can be computed from the orientation of the two consecutive NH vectors by solving a quadratic equation. Furthermore, given $\mathbf{R}_i$ as well as the ϕ_i and ψ_i angles, the orientation of peptide plane $i+1$ in the POF, specified by the rotation matrix, $\mathbf{R}_{i+1}$ can be determined exactly.

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

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

The minima of Equation (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 BACKBONE DIHEDRAL 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 four) whose solutions give the backbone (ϕ, ψ) angles. We give statements of these results; proof sketches are given in Appendix A, and full details of the proofs and equations can be found in Wang and Donald (2004b). 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, Equation (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 three diagonal elements of a diagonalized Saupe matrix **S** (the alignment tensor), and x, y and z are, respectively, the $x, y, z-$components of the unit vector **v** in a principal order frame (POF) which diagonalizes **S**. Recall that **S** is a 3×3 symmetric, traceless matrix with five independent elements. Given NH RDCs in two aligning media, the associated NH vector **v** must lie on the intersection of two conic curves (Skrynnikov and Kay, 2000; Wedemeyer et al., 2002). We state the two propositions needed for Sections 3.1 and 4 below.

Proposition 3.1. *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* **v** *satisfies a monomial quartic equation.*

Proposition 3.2. *Given the* NH *unit vectors $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$ of residues i and $i+1$ and the* NC$_\alpha$ *vector of residue i, the sines and cosines of the intervening backbone dihedral angles (ϕ, ψ) satisfy quadratic equations.*

3.1. Successive computation of (ϕ, ψ) angles of a structure element from RDCs

Propositions 3.1 and 3.2 shows 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. Recall that $w(\phi)$ (resp., $w(\psi)$) denotes the unit vector $(\sin\phi, \cos\phi)$ (resp., $(\sin\psi, \cos\psi)$). 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 Proposition 3.1, and an NC$_\alpha$ vector. The rotation matrix $\mathbf{R}_i$ defines the relative rotation between a POF and a coordinate system in the i^{th} peptide plane (Fig. 1). The rotation matrix $\mathbf{R}_1$ defining the first peptide plane can be determined by solving an optimization problem (see Section 4). This matrix is denoted **R** in Equation (2) above; below, we let $\mathbf{R}_1 = \mathbf{R}$. Let $F_R(\mathbf{R}_i, \phi_i, \psi_i)$ be an algebraic function for computing the rotation matrix $\mathbf{R}_{i+1}$ from ϕ_i, ψ_i and $\mathbf{R}_i$; that is, $\mathbf{R}_{i+1} = F_R(\mathbf{R}_i, \phi_i, \psi_i)$. F_R can be easily derived from backbone kinematics (Wang and Donald, 2004b). In summary, Propositions 3.1 and 3.2 show that given the rotation $\mathbf{R}_i$, and (ϕ_i, ψ_i) for residue i can be computed, *exactly and in constant time*, from two low-degree polynomial equations

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

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

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 two with four variables. Equations (5) and (6) both have degree four and have three and four 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 (Wang and Donald, 2004a).

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 $1 \le i < c$, 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 Equations (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 four with three variables), ψ_i (of degree four with four variables), and $\mathbf{R}_i$ (of degree two with four variables). Since the equations for (ϕ_i, ψ_i) utilize the rotation $\mathbf{R}_i$, $Y_i(Z_i, \mathbf{R}_i)$ 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 Equation (2), subject to the constraint $\mathcal{Y}(D_1, D_2, \mathbf{R}_1)$.

4. A POLYNOMIAL-TIME ALGORITHM FOR PROTEIN STRUCTURE DETERMINATION

In Section 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; furthermore, the 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 and F_ψ in Section 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 Section 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 and had regular secondary structure; 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 Section 4.2 below. In Section 6, 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 (Losonczi et al., 1999) to fit the RDCs to the NH vectors of an 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 Equation (2), subject to $\mathcal{Y}$ (see Section 3.1 for definition) simultaneously by deciding, and finding a witness for, a sentence in the first-order theory of real closed fields (Basu et al., 2003; Grigor'ev, 1988). We show this minimization procedure is polynomial-time in Section 4.2 below.

We now describe the third phase, in which we are given sparse NOEs between successive pairs of secondary structure elements, and must compute their relative translation. For two successive secondary structure elements $\mathcal{A}$ and $\mathcal{B}$, 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 Equation (3) by deciding, and finding a witness for, a sentence in the first-order theory of real closed fields. Section 4.2 below shows how to find the translation x that minimizes Equation (3). 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. Thus, we require only that $\ell \geq 3$ in order to compute the correct translation between oriented secondary structure elements. 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 Section 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 Chapter 14 in Basu et al. (2003) to obtain the desired minima.

There has been much study of how efficiently a first-order predicate on polynomial inequalities can be decided. Tarski (1951) first showed that the problem was indeed decidable, although the complexity of his algorithm is not elementary recursive. Collins (1975) gave the first reasonable worst-case time bound for this problem. Grigor'ev and Vorobjov (1988) gave the first algorithm that was sub-doubly-exponential in the number of variables, and a number of following results improved the complexity in various ways

(Canny, 1993; Heintz et al., 1990; Renegar, 1992). We use a result of Basu et al. (2003), which has an improved asymptotic running time. We now restate their result:

Theorem 1 (From Basu et al. (2003, 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 Section 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 4.1. *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 Equation (2) can be found in $c^{O(c^3)}$ time.*

Proof. Minimizing Equation (2) subject to $\mathcal{Y}$ (as defined in Section 3.1) 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:

$$\exists D_1^* \in \mathcal{G}(D_1), \exists D_2^* \in \mathcal{G}(D_2), \exists \mathbf{R}^* \in SO(3), \exists \mathcal{C}^* \in \mathcal{Y}(D_1^*, D_2^*, \mathbf{R}^*) :$$

$$\forall D_1' \in \mathcal{G}(D_1), \forall D_2' \in \mathcal{G}(D_2), \forall \mathbf{R} \in SO(3), \forall \mathcal{C} \in \mathcal{Y}(D_1, D_2, \mathbf{R}) ::$$

$$\sigma(D_1^*, D_2^*, \mathbf{R}^*, \mathcal{C}^*) \leq \sigma(D_1', D_2', \mathbf{R}, \mathcal{C}); \tag{7}$$

recall that σ is defined by Equation (2) in Section 2. We now analyze the running time of solving Equation (4.2) by applying Theorem 1. First, we observe that Equation (4.2) has degree $O(c)$, the same as that of Equation (2); we will also argue below that the quantified sets are all of degree $O(c)$ as well. Recall that we argued in Section 3 that $\mathcal{Y}$ has degree $O(c)$. As stated, Equation (4.2) 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 $r_{1,i}^* \in D_1^*$ (resp., $r_{2,i}^* \in D_2^*$, $r_{1,i}' \in D_1'$, $r_{2,i}' \in D_2'$) 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 Section 3, the backbone (ϕ, ψ) angles in $\mathcal{Y}$ for both $\mathcal{C}^*$ and $\mathcal{C}$ in Equation (4.2) 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 Section 3), the summation in Equations (2) and (4.2) 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 Equation (4.2) in $c^{O(c^3)} \cdot O((3c+1)^{(4c+6)(2c+2)(6c+6)}) = c^{O(c^3)}$ time. ■

Lemma 4.2. *For any successive pair of secondary structure elements, we can find a translation $x \in \mathbb{R}^3$ that minimizes Equation (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 $N = \{n_1, n_2, \ldots, n_\ell\}$ derived from the experimental NOE restraints. 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 Equation (3) is equivalent to finding a witness x^* such that:

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

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 Equation (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.) ■

The above lemmas show that each of the phases of the algorithm in Section 4.1 can be performed in polynomial 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 4.1, 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 Lemma 4.2. We then obtain the following:

Theorem 2. *The algorithm of Section 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. LIMITATIONS AND EXTENSIONS

In this section, we discuss limitations of, and extensions to, the algorithm presented in Section 4. Section 3 showed that it is possible to compute successive backbone dihedral angles directly from RDCs. These equations are used in the algorithm of Section 4 to compute the optimal backbone conformations of secondary structure elements; we show in Section 5.1 below that similar equations can be derived for loop and turn regions of a given protein. The derivations are similar to those in Section 3, however, we are able to show that for short loop regions it is possible to pin down the backbone (ϕ, ψ) angles exactly *without* additional constraints (such as, e.g., secondary structure type). Section 4 showed that given secondary structure types, RDCs and sparse NOEs, it is possible to compute the optimal backbone structure for a globular protein with regular secondary structure in polynomial time. In Section 5.2, we show that the algorithm of Section 4 can be modified to also handle non-globular proteins (i.e., proteins with arbitrarily long secondary structure elements) in polynomial time, with only a constant factor increase in the number of NOE restraints required (i.e., we still only require $\Theta(n)$ NOEs).

5.1. Computation of loops and turns

The algorithm for turns and loops is built upon the following two propositions:

Proposition 5.1. *Given the orientation of peptide planes i and $i+2$ and the backbone dihedral angle ϕ_i, the sines and cosine of the backbone dihedral angles ψ_i, ϕ_{i+1} and ψ_{i+1} can be computed exactly and in constant time.*

Proof. In the following, small and capital bold letters denote, respectively, vectors (column vectors) and matrices. All the vectors are 3D vectors and all the matrices are 3D rotation matrices. Let $\mathbf{v}_1$, $\mathbf{v}_3$ and

$\mathbf{w}_1$, $\mathbf{w}_3$ denote, respectively, the NH and NC$_\alpha$ vectors of peptide planes i, and $i+2$. From protein backbone kinematics we have

$$\begin{aligned} \mathbf{LG}_1\mathbf{w}_3 &= \mathbf{R}_z(\psi_i)\mathbf{R}\mathbf{R}_y(\phi_{i+1})\mathbf{R}_x(\theta_3)\mathbf{R}_z(\psi_{i+1})\mathbf{c}_w \\ \mathbf{LG}_1\mathbf{v}_3 &= \mathbf{R}_z(\psi_i)\mathbf{R}\mathbf{R}_y(\phi_{i+1})\mathbf{R}_x(\theta_3)\mathbf{R}_z(\psi_{i+1})\mathbf{c}_v \end{aligned} \tag{9}$$

where $\mathbf{R}$ is a constant matrix, and $\mathbf{c}_w$ and $\mathbf{c}_v$ are two constant vectors. Given the backbone angle ϕ_i, the matrix $\mathbf{L}$ is known. The matrix $\mathbf{G}_1$ is the rotation matrix from the POF of RDCs to a coordinate frame defined in the peptide plane i. From Equation (9), through algebraic manipulation we can derive the following three simple trigonometric equations satisfied by the ψ_i, ϕ_{i+1} and ψ_{i+1} angles:

$$a_1 \sin\phi_{i+1} + b_1 \cos\phi_{i+1} = c_1, \qquad a_2 \sin\psi_{i+1} + b_2 \cos\psi_{i+1} = c_2, \qquad a_3 \sin\phi_i + b_3 \cos\phi_i = c_3$$

where a_1, b_1, c_1 are constants from the matrix $\mathbf{R}$, and the six variables, $a_2, b_2, c_2, a_3, b_3, c_3$, are simple trigonometric function of the ϕ_{i+1} angle. ■

Proposition 5.2. *Given the orientation and position of peptide planes i and $i+3$ in a POF of RDCs, the six backbone dihedral angles ϕ_i, ψ_i, ϕ_{i+1}, ψ_{i+1}, ϕ_{i+2} and ψ_{i+2} can be computed exactly and in constant time.*

Recall that for Equation (2), we used the observed averages ϕ_a and ψ_a for the backbone (ϕ, ψ) angles, where ϕ_a and ψ_a were the average backbone dihedral angles for the secondary structure type under consideration. For loop regions, these observed averages are not meaningful; we can, however, fit the structure to the data and avoid steric clash. For any conformation $\mathcal{C}$ (as defined in Section 2), let $\mathbf{B}(\mathcal{C})$ be the atom positions in the backbone defined by $\mathcal{C}$, according to standard backbone geometry. Then, let $d_{x,y}$ be the distance between the backbone atoms $x, y \in \mathbf{B}(\mathcal{C})$; note that there are $k = O(c)$ atoms in this conformation by definition, since each residue type has a constant number of atoms. Let $\delta_{x,y}$ be the sum of the van der Waals radii of the two atoms $x, y \in \mathbf{B}(\mathcal{C})$, and let $CO(\mathcal{C})$ be a predicate that is *true* if $\mathcal{C}$ has steric clash and *false* otherwise (i.e., *true* if all distances $d_{x,y}$ for all $x, y \in \mathbf{B}(\mathcal{C})$ are greater than $\delta_{x,y}$). Finally, let $F = \{\mathcal{C} \mid \neg CO(\mathcal{C})\}$, the set of all conformations $\mathcal{C}$ that do not have steric clash. Then, we wish to compute a conformation $\mathcal{C} \in F$ such that the following objective function is minimized:

$$\sigma_L(D_1', D_2', \mathbf{R}, \mathcal{C}) = \sum_{i=1}^{c} \left((b_{1,i} - r_{1,i})^2 + (b_{2,i} - r_{2,i})^2 \right). \tag{10}$$

Recall that this objective function also appears as the first term in Equation (2). In fact, we can use the techniques presented in Section 4 to find a conformation without steric clash that efficiently optimizes Equation (10) over F for a short (constant-length) loop region:

Lemma 5.1. *The conformation of a loop region of length c that optimizes Equation (10), and does not have steric clash, can be found in $c^{O(c^3)}$ time.*

Proof. Note that the objective function is a simplified version of the one optimized in Lemma 4.1; however we must also ensure that the witness conformation that is identified does not have steric clash. The key observation is that $CO(\cdot)$ can be specified with semi-algebraic constraints, and thus we can formulate a predicate whose truth, along with a witness, can be found in polynomial time, as in Lemmas 4.1 and 4.2. We wish to find a witness for following predicate:

$$\begin{gathered} \exists D_1^* \in \mathcal{G}(D_1), \exists D_2^* \in \mathcal{G}(D_2), \exists \mathbf{R}^* \in SO(3), \exists \mathcal{C}^* \in \mathcal{Y}(D_1^*, D_2^*, \mathbf{R}^*) : \\ \forall D_1' \in \mathcal{G}(D_1), \forall D_2' \in \mathcal{G}(D_2), \forall \mathbf{R} \in SO(3), \forall \mathcal{C} \in \mathcal{Y}(D_1, D_2, \mathbf{R}) :: \\ \neg CO(\mathcal{C}^*) \wedge \neg CO(\mathcal{C}) \wedge \sigma_L(D_1^*, D_2^*, \mathbf{R}^*, \mathcal{C}^*) \leq \sigma_L(D_1', D_2', \mathbf{R}, \mathcal{C}). \end{gathered} \tag{11}$$

First, we note that in any conformation $\mathcal{C}$ for which $CO(\mathcal{C})$ is *false*, $\mathbf{B}(\mathcal{C})$ defines a set of spheres that do not overlap. $\mathbf{B}(\mathcal{C})$ can be defined uniquely for a given conformation $\mathcal{C}$, and has size $O(c)$. Then, the required predicate $CO(\cdot)$ that defines F can be written using $O(c^2)$ inequalities of degree 2; the inequalities ensure the distances $d_{x,y}$ all exceed the van der Waals distances $\delta_{x,y}$. The inequalities for minimizing σ_L are identical to those given in Lemma 4.1 for minimizing the first sum term in Equation (2). Minimizing σ_L requires $c^{O(c^3)}$ time as before. The additional $O(c^2)$ equations that ensure that steric clash does not occur increase the base in this running time, but the exponent increases only by a factor of two, and thus the overall running time is $c^{O(c^3)}$. ■

As before, if $c = O(1)$, then this optimization step runs in $O(1)$ time for each loop region; this optimization step would be performed in the same phase of the algorithm in which the conformation of secondary structure elements is computed. Since there are at most n loop regions, the time needed to compute the overall backbone structure of the protein (including loop regions of constant-length) is $O(n)$. The computation of relative translations proceeds as before; we note that there is no increase in asymptotic running time in this step, since even including loop regions, there are $O(n)$ successive pairs of conformations that must be positioned relative to one another. Thus the algorithm of Section 4 can be modified to compute the optimal conformation of loop regions without increasing the overall asymptotic running time. As mentioned in Section 6, we have successfully incorporated Proposition 5.2 into our algorithm to compute the turns and loops for the protein human ubiquitin using NH and CH RDCs in a single medium. Two short turns, Leu8–Gly10 and Gly47–Lys48, could be computed without using any experimental RDCs, since they are less than 3 residues long (Proposition 5.2). The two loops, Glu18–Thr22 and Gly35–Glu41, connecting the helix (Leu23–Glu34) to the single sheet (consisting of five strands), can also be computed using only NH and CH RDCs in a single medium. The conformations of these two loops determine the relative position between the helix and sheet. The most difficult problem is the computation of the long loop, Glu51–Lys63, connecting two β-strands in the sheet. Two long-range backbone NOE distances, H_N(Thr22)↔H_N(Thr55) and H_N(Ile23)↔H_α(Leu56), automatically-assigned based on chemical shift alone are required for improving the accuracy of the conformation. The complete backbone structure computed by our algorithm (Fig. 4) has a 1.45 Å backbone RMSD (computed using C_α, N, and C$'$ backbone atoms) from the corresponding X-ray backbone structure (PDB ID 1UBQ) (Vijay-Kumar et al., 1987).

5.2. *Non-globular proteins*

In Section 4, we presented an algorithm that computes the optimal backbone structure (see Section 2 for our optimization criteria) for globular proteins with regular secondary structure. We note that it can be checked easily, in $O(n)$ time given the secondary structure types of residues, whether a protein is globular or not, and whether the secondary structure elements are of constant length. The complexity of our algorithm is parameterized by n, m and c. For our algorithm, we must have $m \cdot c \leq n$, or, more precisely, $\sum_{i=1}^{m} c_i = n$, where c_i is the length of the i^{th} secondary structure element. In Section 4, we let $c = \max\{c_1, c_2, \ldots, c_m\}$ and handle the case where $c = O(1)$. We note that this assumption is reasonable, since secondary structure element length appears to be bounded by a constant as protein size grows. Figure 2 verifies this statement for β-strands by showing a plot of protein length versus maximum β-strand length for proteins in the PDB Finder (Hooft et al., 1996) database; a similar trend holds for α-helices.

If we wish to apply our algorithm to a protein that is not globular (as mentioned above, we can check this in $O(n)$ time), we can use a modified version of the algorithm from Section 4 with slight modifications that requires additional NOE restraints. We formalize this approach with Theorem 3 below. Informally, the idea behind the modified algorithm is to partition arbitrarily long secondary structure elements into *fragments* of constant length, apply our minimization technique to find optimal conformations for these fragments, and assemble the fragments as in the last phase of our original algorithm. The global nature of RDCs guarantees us that the relative orientations[2] of the fragments are correct after computing their conformations. Recall that the assembly procedure required three NOEs for every successive pair of secondary structure elements; in the modified algorithm we will require three NOEs for every successive

[2] See Wang and Donald (2004b, p. 234) for how the sparse NOEs can be used, within the same time bound, to resolve the relative orientational degeneracy in one medium due to the symmetry of the dipolar operator.

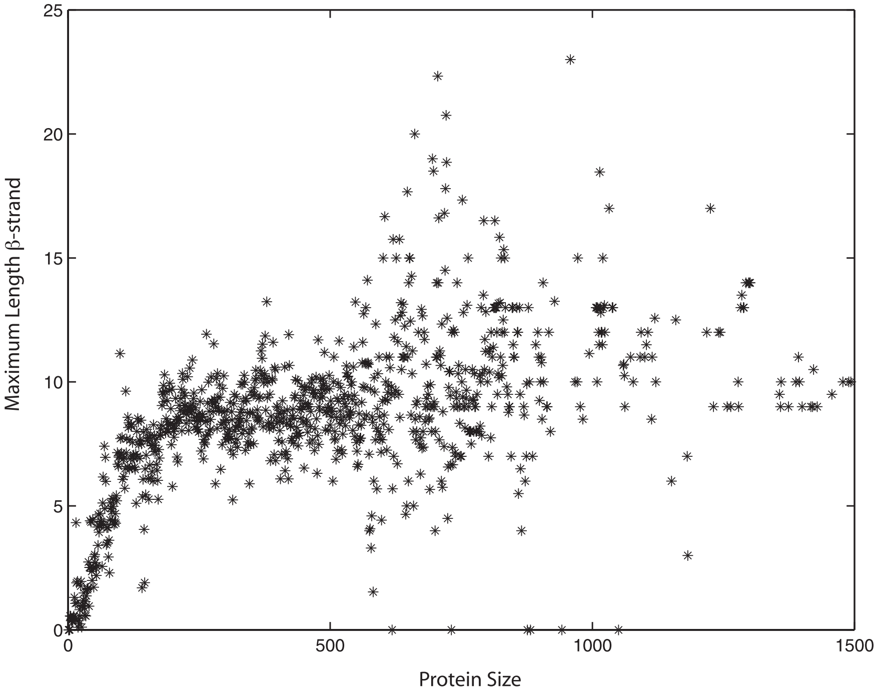

FIG. 2. A plot of β-strand lengths for proteins in the PDBFinder (Hooft et al., 1996) database; the x-axis is the length of a protein (number of residues), and the y-axis is the maximum length of any β-strand in the given protein.

pair of fragments. In non-globular proteins, secondary structure elements can have length $\Theta(n)$, and thus this approach could yield $\Theta(n)$ fragments and requires $\Theta(n)$ NOE restraints. Note that, asymptotically, the required number of NOE restraints is not increased.

Theorem 3. *For a non-globular protein with n residues, we can compute an optimal backbone structure satisfying Equations (2) and (3) subject to the constraint $\mathcal{Y}(D_1, D_2, \mathbf{R}_1)$ in $O(n)$ time.*

Proof. Let γ be a constant, and let the protein under consideration have n residues. Our modified algorithm works as follows. First, we partition the secondary structure elements of the protein into fragments of length at most γ. We can now apply Lemma 4.1 to find the conformation of each of these fragments in $O(\gamma^{O(\gamma^3)}) = O(1)$ time, since $\gamma = O(1)$. Note that every pair of these fragments has the correct relative *orientation*, and to correctly determine the backbone structure, it suffices to compute the relative *position* of successive pairs of fragments. To do this, we require NOE restraints between every successive pair of the chosen fragments; this requirement is the only constraint on the fragment length γ. Furthermore, all fragments do not need to be the same length, as long as their length is bounded by γ. We can then apply Lemma 4.2 to each successive pair of fragments to obtain the backbone structure. There are at most n fragments, and each application of Lemma 4.2 requires $O(1)$ time, yielding a total running time of $O(n)$ to construct the backbone structure from the conformations of the computed fragments. Thus, we obtain an overall running time of $O(n)$. ■

We note the above algorithm is a strict generalization of the algorithm presented in Section 4; in other words, the two algorithms are equivalent if we set $\gamma = c$. Given the experimental data described in Section 1, page 1269, including the additional NOEs described in the proof of Theorem 3 above, the algorithm presented in this section can compute the backbone structure of globular and non-globular proteins in polynomial time. In Section 6, we present experimental results for a version of our algorithm that also computes the intervening loop regions in the backbone structure; Section 5.1 above detailed the equations needed to compute NH orientations in loop regions. We note that the generalized algorithm above can also be applied in conjunction with the results in Section 5.1 above to compute the complete backbone structure, including loops and turns, of both globular or non-globular proteins.

6. EXPERIMENTAL RESULTS

As shown in Section 4, for globular proteins with regular secondary structure, our algorithm for structure determination provably runs in polynomial time. While 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 (Canny et al., 1992; Donald and Xavier, 1995a, 1995b; Donald et al., 1992, pages 1–23). In practice, the implementation took about 20 minutes on average on a single-processor Pentium-4 class machine. Table 1 gives the results of using our algorithm to compute the backbone secondary structure elements for six real experimental data sets for three proteins.

Practical algorithms for quantifier elimination and the existential theory of real closed fields have been efficiently implemented (Canny, 1987; Ponce and Kriegman, 1992) to find the minima of objective functions that are similar to Equations (2) and (3). In our implementation, the second phase of the algorithm was implemented with 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 (Böhringer et al., 1996; Brown and Donald, 2000; Canny and Donald, 1988; Donald, 1990, 1992, 1993; Erdmann, 2004; Erdmann and Lozano-Perez, 1987), these codes must be tested on real data to verify robustness and accuracy. Our algorithm has been successfully implemented and applied to real protein NMR data to compute the backbone substructures (oriented and translated secondary structure elements) of three structurally distinct proteins. We first applied the algorithm to the protein human ubiquitin using NH RDCs in two media (Wang and Donald, 2004b, 2004a) or NH and CH RDCs in a single medium (Table 1), plus 12 hydrogen bonds and four NOE distances (Fig. 3). For our experiments, we used the PDB assignment of secondary structure as input, although we note that the secondary structure assignment is evident from the NMR chemical shift index (CSI) (Wishart and Sykes, 1994) of backbone atoms. We have also applied our algorithm to compute the backbone substructures of two other 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. The RMSD, computed using backbone C_α, N, and C′ atoms, between the backbone structures (excluding loop regions) computed by our algorithm and the corresponding portions of previously-solved NMR structures is, respectively, 1.55 Å for DNA-damage-inducible protein I and 0.96 Å for immunoglobulin binding protein G. Note that the NMR structures we compared with are computed by MD/SA (Brünger, 1993)

TABLE 1. RESULTS OF OUR ALGORITHM

Protein[a]	*α/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

[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 the total number of residues.

[c] The total number of experimental RDCs (note that RDCs are missing for some residues).

[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 (for C_α, N, and C′ backbone atoms) between the oriented and translated secondary structure elements (excluding loop regions) 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).

TABLE 2. COMPARISON WITH EXISTING APPROACHES

Reference[a]	*Program*	*Technique*[b]	*Restraints per residue*[c]	*Accuracy*[d]
Giesen et al. (2003)	X-plor	MD/SA	6 RDCs	1.45 Å
Hus et al. (2001)	SCULPTOR	MD/SA	11 RDCs,	1.00 Å
Delaglio et al. (2000)	MFR	Database	10 RDCs, 5 chemical shifts	1.21 Å
Rohl and Baker (2002)	RosettaNMR	Database/MC	3 RDCs, 5 chemical shifts	1.65 Å
Rohl and Baker (2002)	RosettaNMR	Database/MC	1 RDC	2.75 Å
Our algorithm	—	Exact equations	2 RDCs	1.45 Å

[a] References to previously-computed ubiquitin backbone structures (including loop regions).

[b] Algorithmic technique.

[c] Data requirements.

[d] Backbone RMSD (for C_α, N, and C′ backbone atoms) of the structure computed by our algorithm (including loops and turns) compared to the X-ray structure (PDB ID: 1UBQ) (Vijay-Kumar et al., 1987).

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 two restraints per residue) is important for high-throughput applications such as structure-based drug design. This is because, in practice, it is difficult and time-consuming to measure more than two RDCs per residue for many proteins due to their dynamic behavior in solution.

Finally, as mentioned in Section 5 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 two RDCs per residue) and two unambiguous NOEs. This algorithm, which also computes the structure of the turn and loop regions also runs in polynomial-time for a globular protein with regular secondary structure if we assume that our globular protein has $O(n)$ loop and turn regions each with length $c_\ell = O(1)$; an overwhelming majority of globular proteins indeed have short (constant-length) turn and loop regions (see Section 5 for further discussion). When tested on ubiquitin, the final backbone structure computed by this algorithm has a 1.45 Å backbone RMSD (for all backbone atoms) from the X-ray structure (Fig. 4). This accuracy is similar to that of the ubiquitin backbone structure computed by a commonly-used heuristic approach (Giesen et al., 2003) (Table 2). The latter is the previous best result obtained for ubiquitin structure when using six or fewer RDCs per residue. Our accuracy is also better than the ubiquitin structure computed by Rohl and Baker (2002); they use three RDCs per residues plus five chemical shifts per residue as input to their algorithm. Furthermore, our algorithm is capable of handling up to 15% missing RDC data (Fig. 4).

7. CONCLUSION

In this paper, we have shown that the global nature of RDC data 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* high-resolution 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.

A key feature of our approach is that we establish an exact relationship between the experimental data and the computed protein structure (i.e., Propositions 3.1 and 3.2 relate NH orientations *exactly* to RDCs). For example, it is easy to compute the contribution of each NH orientation chosen by our algorithm to the optimal value of Equation (2). Furthermore, the effect of error in an experimental RDC can also be expressed as an exact, algebraic function of NH orientation using Proposition 3.1, which then allows us to quantitate the effect of a single RDC on the final structure. For secondary structure elements, our algorithm finds the NH orientations that optimize Equation (2), but it would be straightforward to treat any subset of the RDCs as parameters in the quartic equation derived in Proposition 3.1. We can then analytically solve for the NH orientations that satisfy Proposition 3.1 and hence “cover” the range of experimental values of the RDCs.

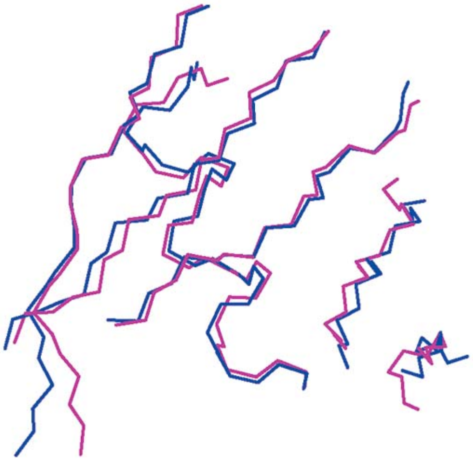

FIG. 3. Structure of ubiquitin backbone without loops. The ubiquitin backbone structure (blue) was computed by our algorithm using 37 NH and 39 CH RDCs, 12 hydrogen bonds, and four NOEs. Our structure has an RMSD of 0.97 Å when compared to the high-resolution X-ray structure (PDB ID: 1UBQ, in magenta) (Vijay-Kumar et al., 1987). 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.

Our algorithm and implementation can easily be extended to output a set of k conformations, for any k, rather than a single best-fit structure. After computing the best-fit structure in polynomial time, the existential predicates used in Lemmas 4.1 and 4.2 can be modified to find an additional distinct conformation; this procedure can be repeated k times to find the k top-scoring structures for the given experimental data. We note that the overall running time is increased only by a factor of k, the desired number of conformations. (Remark: It is interesting to point out that if $k = O(1)$, it is also relatively straightforward to find k best-fit conformations in a manner similar to Lemma 4.1, by using a different set of variables for each of the k distinct conformations. The total running time for both methods of generating k conformations is $O(n)$ for $k = O(1)$.) Let δ_0 be the combined cost (Equations (2) and (3)) of the optimal conformation returned by our algorithm. The predicates in Equations (4.2) and (8) can be modified to represent the set S_ε of structures whose combined score is at most $\delta_0 + \varepsilon$, for all $\varepsilon > 0$. Therefore, S_ε is also a semi-algebraic set that can be decided in polynomial-time. Thus, we can also easily

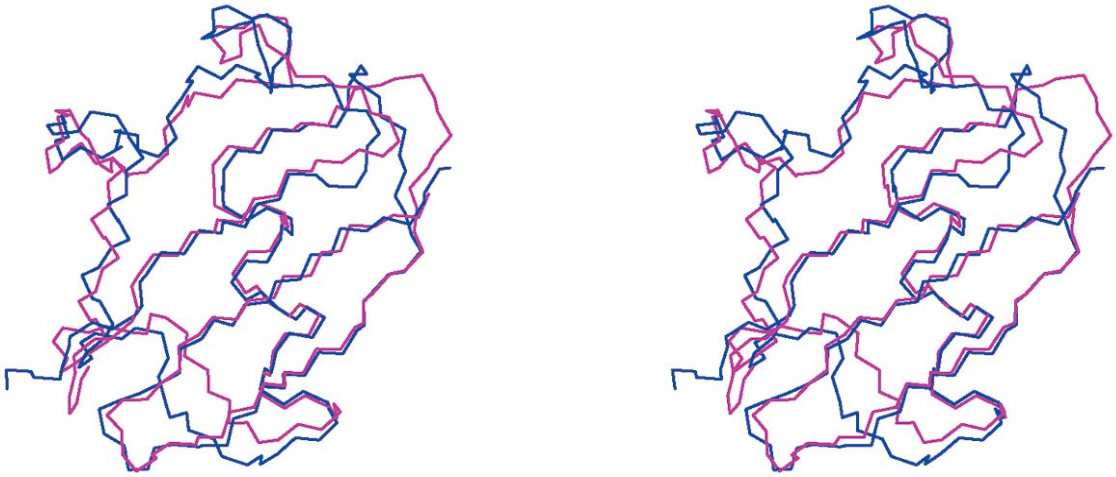

FIG. 4. Structure of ubiquitin backbone with loops. The ubiquitin backbone structure (blue) was computed by extending our algorithm to handle loop regions along the protein backbone. The structure was computed using 59 NH and 58 CH RDCs (117 out of 137 possible RDCs, 20 are missing), 12 H-bonds and two unambiguous NOEs. Our structure has a backbone RMSD of 1.45 Å with the high-resolution X-ray structure (PDB ID: 1UBQ, in magenta) (Vijay-Kumar et al., 1987).

specify the range of total cost these k conformations allowed to span (i.e., that none of the computed conformations exceeds the optimal cost by more than an additive factor of ε, for any $\varepsilon > 0$). Analogously, the tree-search based implementation can easily be modified to return the k top scoring conformations.

Additionally, because our structure determination approach is based on an exact relationship between experimental data and the resulting structure, it also compatible with approaches that seek to characterize the likelihood of a computed structure, i.e., an objective *figure of merit*, with respect to experimental parameters (Rieping et al., 2005). In other words, we can assign likelihoods to the NH vectors and backbone atom positions computed by our algorithm that are based on their agreement with the input experimental RDC and NOE data.

Furthermore, our polynomial-time backbone structure determination algorithm can be extended to compute *complete* protein structures (including side-chains), since exact equations analogous to Equations (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 Equation (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.

APPENDIX

A. 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. We sketch the proofs here; the interested reader can refer to (Wang and Donald, 2004b) 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, Equation (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 (Tjandra and Bax, 1997; Tolman et al., 1995). Given NH RDCs in two aligning media, the associated NH vector $\mathbf{v}$ must lie on the intersection of two conic curves (Skrynnikov and Kay, 2000; Wedemeyer et al., 2002). 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 coefficients and temporary variables (a_2, b_2, c_1, etc.) can be found in (Wang and Donald, 2004b).

Proof. 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 Equation (4) we have

$$r = S_{xx}x^2 + S_{yy}y^2 + S_{zz}z^2, \qquad r' = S'_{xx}x'^2 + S'_{yy}y'^2 + S'_{zz}z'^2,$$

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \mathbf{R_{12}} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

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

$$r_2 = a_2x^2 + b_2y^2 + c_1xy + c_2xz + c_3yz \tag{12}$$

$$r_1 = a_1x^2 + b_1y^2 \tag{13}$$

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 Wang and Donald (2004b).

Eliminating z from Equation (12) we obtain

$$d_8x^4 + d_7x^3y + d_6x^2y^2 - d_5x^2 + d_4xy^3 - d_3xy - d_2y^2 + d_1y^4 + d_0 = 0 \tag{14}$$

where $d_8 = a_2^2 + c_2^2$, and $d_7, d_6, \ldots, d_0$ are analogously defined; these are defined fully in Wang and Donald (2004b). Equation (14) is a degree 8 monomial in x after direct elimination of y using Equation (13). 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

$$x = a \sin t, \qquad y = b \cos t, \qquad u = \cos 2t \tag{15}$$

and through algebraic manipulation we finally obtain

$$f_4u^4 + f_3u^3 + f_2u^2 + f_1u + f_0 = 0. \tag{16}$$

The full expressions for coefficients a, b and f_0, f_1, f_2, f_3, f_4 are given in Wang and Donald (2004b). Since $u = 1 - 2(\frac{x}{a})^2$ Equation (16) is also a quartic equation in x^2. ■

The y-component of $\mathbf{v}$ can be computed directly from Equation (15). 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 Wang and Donald (2004b).

Proof. 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 $i+1$:

$$\mathbf{v}_i = \mathbf{R}_x(\theta_7)\mathbf{R}_y(\theta_6)\mathbf{R}_x(\theta_5)\mathbf{R}_z(\psi + \pi)\mathbf{R}_x(\theta_3)\mathbf{R}_y(\phi)\mathbf{R}_y(\theta_8)\mathbf{R}_x(\theta_1)\mathbf{v}_{i+1}. \tag{17}$$

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 θ_8) are given in Wang and Donald (2004b).

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

$$\begin{aligned} x_1 &= -(\cos\phi\cos\psi + \sin\theta_3\sin\phi\sin\psi)\ x_2 - \cos\theta_3\sin\psi\ \ y_2 + (\cos\psi\sin\phi - \cos\phi\sin\theta_3\sin\psi)\ z_2 \\ y_1 &= (\cos\phi\sin\psi - \sin\theta_3\sin\phi\cos\psi)\ x_2 - \cos\theta_3\cos\psi\ \ y_2 - (\sin\phi\sin\psi + \cos\phi\sin\theta_3\cos\psi)\ z_2 \\ z_1 &= \cos\theta_3\sin\phi\ x_2 - \sin\theta_3\ y_2 + \cos\theta_3\cos\phi\ z_2 \end{aligned} \tag{18}$$

By Equation (18) we can then obtain a simple trigonometric equation:

$$\sin(\phi + a_1) = b_1 \tag{19}$$

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 Wang and Donald (2004b) 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 Equation (19) we can obtain another simple trigonometric equation:

$$\sin(\psi + a_2) = b_2 \tag{20}$$

$\sin\psi$ 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 $\cos\phi$. ■

ACKNOWLEDGMENTS

This work was done while L.W. and R.M. were postdoctoral fellows in the laboratory of B.R.D. We would like to thank S. Apaydin for making Figure 2, and for technical help with Section 5.2. We thank all members of the Donald laboratory, as well as J. Hoch, C. Bailey-Kellogg, S. Potluri, P. Zhou, G. Wagner, and J.J. Chou for helpful discussions and feedback. 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).

REFERENCES

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

Bailey-Kellogg, C., Widge, A., Kelley, J.J., et al. 2000a. The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J. Comput. Biol.* 7, 537–558.

Bailey-Kellogg, C., Widge, A., Kelley, III, J.J., et al. 2000b. The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *4th Ann. Int. Conf. Comput. Mol. Biol.* 33–44.

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

Berger, B., Kleinberg, J., and Leighton, F.T. 1999. Reconstructing a three-dimensional model with arbitrary errors. *J. ACM* 46, 212–235.

Böhringer, K.-F., Donald, B.R., and MacDonald, N. 1996. Upper and lower bounds for programmable vector fields with applications to MEMS and vibratory plate parts feeders. *Proc. Int. Workshop Algorithmic Found. Robot. (WAFR).*

Brown, R.G., and Donald, B.R. 2000. Mobile robot self-localization without explicit landmarks. *Algorithmica* 26, 515–559.

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

Canny, J. 1987. A new algebraic method for robot motion planning and real geometry. *Proc. 28th IEEE Conf. Found. Comput. Sci.* 39–48.

Canny, J. 1993. Improved algorithms for sign determination and existential quantifier elimination. *Comput. J.* 36, 409–418.

Canny, J., and Donald, B.R. 1988. Simplified Voronoi diagrams. *Discrete Comput. Geom.* 3, 219–236.

Canny, J., Donald, B.R., and Ressler, G. 1992. A rational rotation method for robust geometric algorithms. *Proc. 8th ACM Symp. Comput. Geom.* 251–260.

Cavanagh, J., Fairbrother, W.J., Palmer, III, A.G., et al. 1995. *Protein NMR Spectroscopy: Principles and Practice*. Academic Press, San Diego.

Clore, G.M. 2000. Accurate and rapid docking of protein-protein complexes on the basis of intermolecular nuclear Overhauser enhancement data and dipolar couplings by rigid body minimization. *Proc. Natl. Acad. Sci. USA* 97, 9021–9025.

Collins, G.E. 1975. Quantifier elimination for real closed fields by cylindrical algebraic decomposition. *Lect. Notes Comput. Sci.* 33, 515–532.

Cordier, F., Rogowski, M., Grzesiek, S., et al. 1999. Observation of through-hydrogen-bond 2hJHC′ in a perdeuterated protein. *J. Magnet. Reson.* 40, 510–512.

Crippen, G. 1991. Chemical distance geometry: current realization and future projection. *J. Math. Chem.* 6, 307–324.

Crippen, G.M., and Havel, T.F. 1988. *Distance Geometry and Molecular Conformations*. Wiley, New York.

del Rio-Portílla, F., Blechta, V., and Freeman, R. 1994. Measurement of poorly-resolved splittings by J-doubling in the frequency domain. *J. Magnet. Reson.* 111a, 132–135.

Delaglio, F., Kontaxis, G., and Bax, A. 2000. Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *J. Am. Chem. Soc.* 122, 2142–2143.

Donald, B.R. 1990. The complexity of planar compliant motion planning with uncertainty. *Algorithmica* 5, 353–382.

Donald, B.R. 1992. Robot motion planning. *IEEE Trans. Robot. Autom.* 8.

Donald, B.R. 1993. Computational robotics: manipulation, planning, and control. *Algorithmica* 10, 91–101.

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

Donald, B.R., and Xavier, P. 1995a. Provably good approximation algorithms for optimal kinodynamic planning for Cartesian robots and open chain manipulators. *Algorithmica* 14, 443–479.

Donald, B.R., and Xavier, P. 1995b Provably good approximation algorithms for optimal kinodynamic planning: robots with decoupled dynamics bounds. *Algorithmica* 14, 480–530.

Dong, Q., and Wu, Z. 2002. A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances. *J. Global Optim.* 22, 365–375.

Erdmann, M.A. 2004. Protein similarity from knot theory and geometric convolution. *Proc. 8th Annu. Int. Conf. Comput. Mol. Biol.* 195–204.

Erdmann, M.A., and Lozano-Perez, T. 1987. On multiple moving objects. *Algorithmica* 2, 477–521.

Fowler, A.C., Tian, F., Al-Hashimi, H.M., et al. 2000. Rapid determination of protein folds using residual dipolar couplings. *J. Mol. Biol.* 304, 447–460.

Gardner, K.H., and Kay, L.E. 1997. Production and incorporation of ^{15}N, ^{13}C, ^{2}H (^{1}H-δ1 methyl) isoleucine into proteins for multidimensional NMR studies. *J. Am. Chem. Soc.* 119, 7599–7600.

Giesen, A.W., Homans, S.W., and Brown, J.M. 2003. Determination of protein global folds using backbone residual dipolar coupling and long-range NOE restraints. *J. Biomol. NMR* 25, 63–71.

Grigor'ev, D.Y. 1988. Complexity of deciding Tarski algebra. *J. Symbolic Comput.* 5, 65–108.

Grigor'ev, D.Y., and Vorobjov, Jr., N.N. 1988. Solving systems of polynomial inequalities in sub-exponential time. *J. Symbolic Comput.* 5, 65–108.

Güntert, P., Mumenthaler, C., and Wüthrich, K. 1997. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* 273, 283–298.

Heintz, J., Roy, M.-F., and Solernò, P. 1990. Sur la complexité du principe de Tarski-Seidenberg. *Bull. Soc. Math France* 118, 101–126.

Hendrickson, B. 1992. Conditions for unique graph realizations. *SIAM J. Comput.* 21, 65–84.

Hendrickson, B. 1995. The molecule problem: exploiting structures in global optimization. *SIAM J. Optim.* 5, 835–857.

Herrmann, T., Güntert, P., and Wüthrich, K. 2002. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* 319, 209–227.

Hooft, R.W.W., Sander, C., and Vriend, G. 1996. The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. *Comput. Appl. Biosci.* 12, 525–529.

Huang, Y.J., Swapna, G.V., Rajan, P.K., et al. 2003. Solution NMR structure of ribosome–binding factor a (RbfA), a cold-shock adaptation protein from *Escherichia coli*. *J. Mol. Biol.* 327, 521–536.

Hus, J.C., Marion, D., and Blackledge, M. 2001. Determination of protein backbone using only residual dipolar couplings. *J. Am. Chem. Soc.* 123, 1541–1542.

Juszewski, K., Schwieters, C.D.S., Garrett, D.S., et al. 2004. Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear Overhauser enhancement spectra and chemical shift assignments. *J. Am. Chem. Soc.* 126, 6258–6273.

Losonczi, J.A., Andrec, M., Fischer, M.W., et al. 1999. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J. Magnet. Reson.* 138, 334–342.

Marin, A., Malliavin, T.E., Nicolas, P., et al. 2004. From NMR chemical shifts to amino acid types: investigation of the predictive power carried by nuclei. *J. Biomol. NMR* 30, 47–60.

Nilges, M., Macias, M., Odonoghue, S., et al. 1997. Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from β-spectrin. *J. Mol. Biol.* 269, 408–422.

Ponce, J., and Kriegman, D.J. 1992. Elimination theory and computer vision: recognition and positioning of 3D objects from range, intensity, or contours. In Donald, B.R., Kapur, D., and Mundy, J.L., eds. *Symbolic and Numerical Computation for Artificial Intelligence*. Academic Press, Boston.

Prestegard, J.H., Bougault, C.M., and Kishore, A.I. 2004. Residual dipolar couplings in structure determination of biomolecules. *Chem. Rev.* 104, 3519–3540.

Renegar, J. 1992. On the computational complexity and geometry of the first-order theory of reals. *J. Symbolic Comput.* 13, 255–352.

Rieping, W., Habeck, M., and Nilges, M. 2005. Inferential structure determination. *Science* 209, 303–306.

Rohl, C.A., and Baker, D. 2002. *De novo* determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc.* 124, 2723–2729.

Saupe, A. 1968. Recent results in the field of liquid crystals. *Angew. Chem.* 7, 97–112.

Saxe, J.B. 1979. Embeddability of weighted graphs in k-space is strongly NP-hard. *Proc. 17th Allerton Conf. Comm. Control Comput.* 480–489.

Skrynnikov, N.R., and Kay, L.E. 2000. Assessment of molecular structure using frame-independent orientational restraints derived from residual dipolar couplings. *J. Biomol. NMR* 18, 239–252.

Tarski, A. 1951. *A Decision Method for Elementary Algebra and Geometry*. University of California Press, Berkeley.

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

Tjandra, N., and Bax, A. 1997. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* 278, 1111–1114.

Tolman, J.R., Flanagan, J.M., Kennedy, M.A., et al. 1995. Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proc. Natl. Acad. Sci. USA* 92, 9279–9283.

Vijay-Kumar. S., Bugg, C.E., and Cook, W.J. 1987. Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* 194, 531–544.

Wang, L., and Donald, B.R. 2004a. Analysis of a systematic search-based algorithm for determining protein backbone structure from a minimal number of residual dipolar couplings. *IEEE Comput. Sys. Bioinform. Conf.* 319–330.

Wang, L., and Donald, B.R. 2004b. 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* 29, 223–242.

Wang, Y.X., Jacob, J., Cordier, F., et al. 1999. Measurement of 3hJNC$'$ connectivities across hydrogen bonds in a 30-kDa protein. *J. Biomol. NMR* 14, 181–184.

Wedemeyer, W.J., Rohl, C.A., and Scheraga, H.A. 2002. Exact solutions for chemical bond orientations from residual dipolar couplings. *J. Biomol. NMR* 22, 137–151.

Wishart, D.S., and Sykes, B.D. 1994. The 13C chemical shift index: a simple method for the identification of protein secondary structure using 13C chemical shift data. *J. Biomol. NMR* 4, 171–180.

Wishart, D.S., Sykes, B.D., and Richards, F.M. 1991. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J. Mol. Biol.* 222, 311–333.

Wishart, D.S., Sykes, B.D., and Richards, F.M. 1992. The 13C chemical shift index: A fast and simple method for the identification of protein secondary structure using 13C chemical shift data. *Biochemistry* 31, 1647–1651.

Address correspondence to:
Bruce Randall Donald
Department of Computer Science
and Department of Biochemistry
Duke University Medical Center
D101 Levine Science Research Center
Research Drive, P.O. Box 90129
Duke University
Durham, NC 27708-0129

E-mail: brd@cs.duke.edu