Analysis of a Systematic Search-Based Algorithm for Determining Protein Backbone Structure from a Minimum Number of Residual Dipolar Couplings

Lincong Wang[*]

Bruce Randall Donald [*,†,‡,§,¶]

Abstract

We have developed an ab initio *algorithm for determining a protein backbone structure using global orientational restraints on internuclear vectors derived from residual dipolar couplings (RDCs) measured in one or two different aligning media by solution nuclear magnetic resonance (NMR) spectroscopy [14, 15]. Specifically, the conformation and global orientations of individual secondary structure elements are computed, independently, by an exact solution, systematic search-based minimization algorithm using only 2 RDCs per residue. The systematic search is built upon a quartic equation for computing,* exactly *and* in constant time*, the directions of an internuclear vector from RDCs, and linear or quadratic equations for computing the sines and cosines of backbone dihedral* (ϕ, ψ) *angles from two vectors in consecutive peptide planes. In contrast to heuristic search such as simulated annealing (SA) or Monte-Carlo (MC) used by other NMR structure determination algorithms, our minimization algorithm can be analyzed rigorously in terms of expected algorithmic complexity and the coordinate precision of the protein structure as a function of error in the input data. The algorithm has been successfully applied to compute the backbone structures of three proteins using real NMR data.*

1 Introduction[1]

The increasing gap between the speed of DNA sequencing and protein structure determination requires the development of efficient algorithms for computing 3-dimensional structures as accurately as possible using a minimum number of restraints obtainable rapidly by experimental techniques. One way to achieve this is to develop algorithms whose key components are *analytic* expressions computable *in constant time*. Our contribution to NMR structure determination [14] is the development of an exact solution, systematic search based-deterministic minimization algorithm for computing a protein backbone structure using only *two* residual dipolar couplings (RDCs) per residue in either one or two media and sparse distance restraints. Our newly derived low-degree monomials can compute, *exactly and in constant time*, the sines and cosines of individual backbone (ϕ, ψ) angles from RDCs, which makes the previous grid search-based methods obsolete. Further, our minimization algorithm searches over all the possible conformations consistent with the input data (experimental RDCs) and employs a provable pruning strategy that guarantees pruned conformations need not be considered further. If the input data are perfect (without any experimental error), systematic search is guaranteed to find a global minimum, which differs fundamentally from heuristic search such as simulated annealing (SA) and Monte-Carlo (MC) used by other NMR structure determination algorithms [1, 6, 11]. The latter can only sample the conformation space stochastically. Furthermore, in contrast to heuristic search, our minimization algorithm can be analyzed rigorously. In our algorithm, the solution (conformation) space of the systematic search is pruned by the following three filters: a real solution filter,

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

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

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

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

[¶] This work is supported by the following grants to B.R.D.: National Institutes of Health (R01 GM 65982) and National Science Foundation (IIS-9906790, EIA-0102710, EIA-0102712, EIA-9818299, EIA-0305444 and EIA-9802068).

[1] Abbreviations used: NMR, nuclear magnetic resonance; RDC, residual dipolar coupling; NOE, nuclear Overhauser effect; SVD, singular value decomposition; DFS, depth-first search; RMSD, root mean square deviation; POF, principal order frame; PDB, protein data bank; MD, molecular dynamics; H-bond, hydrogen bond; NH, the internuclear vector between amide nitrogen (N) and amide proton (H); PDF, probability distribution function; SA, simulated annealing; MC, Monte-Carlo; MD, molecular dynamics; CH, the vector between C_α and H_α.

Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)
0-7695-2194-0/04 $20.00 © 2004 IEEE

a Ramachandran filter, and a geometric filter. We first quantify the contributions of the three filters to the efficiency of our systematic search. We then show that the performance of the minimization algorithm can be profitably modeled using branching processes [4] when real data with experimental errors are used as input. The analysis concludes that although the algorithm is exponential in the worst case, it is quite fast in the average case due to the pruning provided by the above three filters. Taken together, the analysis provides a mathematical basis for the efficiency of our minimization algorithm.

1.1 Organization of the paper

We begin, in section 2 with a description of existing algorithms for backbone structure determination using RDCs, including a discussion of their limitations. Section 3 outlines our algorithm. Section 4 presents the results of applying our algorithm to three proteins using real NMR data and discusses the significance of our algorithm for structural genomics. Section 5 present an analysis of our algorithm including (a) the complexity of the algorithm and its performance in practice, (b) the quantification of three filters used for pruning, (c) a stochastic model for analyzing the performances of our minimization algorithm, and (d) the comparison between the theory and computational experimental results.

2 Previous work

Traditional NMR structure determination algorithms [1, 6] were designed to use distance restraints derived from nuclear Overhauser effect (NOE) experiments. Months of time may be required to extract enough NOE distance restraints to compute a well-defined structure. In contrast, RDC orientational restraints can be obtained much faster experimentally, and thus are more suitable for developing high-throughput structure determination algorithms. However, previous algorithms using RDCs rely on either heuristic search [5, 8] or a structural database (the PDB) [2, 11] and require more than three RDCs per residue in order to compute a well-defined backbone structure. Brown and coworkers [5] applied molecular dynamics (MD) and SA to compute the fold of human ubiquitin using three RDCs per residue in two media (six RDCs per residue in total). Blackledge and coworkers [8] employed least-square fitting followed by MD/SA to compute the backbone structure of ubiquitin using five RDCs per residue in two media. Baker and coworkers [11] incorporated RDCs into their *ab initio* structure prediction algorithm ROSETTA where RDCs were employed to select structural fragments from the PDB. The final structures constructed from the chosen fragments were further refined by ROSETTA using MC search. One limitation is that the accuracy of the computed structure was rather poor when fewer than 3 RDCs per residue were employed. The database-based algorithms were first developed by Bax and coworkers [2], who use five RDCs per residue in two media plus chemical shifts to select seven-residue fragments from the PDB. One concern about such algorithms is that the computed fold may be biased toward those deposited in the PDB. The above algorithms use either heuristic search such as SA [5, 8] or MC [11] to find a best solution consistent with the input data. A heuristic search samples the search (conformation) space stochastically so there is no guarantee that the computed best solution is a true global minimum. Furthermore, it is very difficult to analyze and consequently to optimize heuristic algorithms. For example, when a heuristic search is employed it is not easy or even possible to quantify the contributions of the pruning to the performance of the algorithm. Finally, an algorithm using a heuristic search may require more restraints in order to achieve an accuracy in the computed structure similar to that obtained by systematic search.

3 The Algorithm

The inputs to our algorithm are: (a) 2 RDCs of backbone vectors per residue (e.g., 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. In the following, we briefly describe the algorithm for NH and CH RDCs measured in a single medium. Interested readers can see our previous paper [14] for an algorithm for computing a protein backbone structure using NH RDCs in two media. The algorithm is divided into three stages: (I) *alignment tensor computation* by singular value decomposition (SVD), (II) *computation of the orientation and conformation of a secondary structure element* by systematic search, and (III) *backbone structure determination* by rigid-body minimization. The alignment tensors from stage I are used in stage II to compute the directions of both CH and NH vectors in a common principal order frame (POF) of NH and CH RDCs by solving quartic equations using RDCs exclusively. Given CH and NH directions, the sines and cosines of individual (ϕ, ψ) angles are computed exactly by solving a linear equation. In two media, the sines and cosines of individual (ϕ, ψ) angles are computed by solving quadratic equations [14, 15]. The (ϕ, ψ) angles of an entire structural fragment are computed in stage II by a minimization algorithm as detailed in Section 3.1. The relative positions of the fragments computed from stage II are determined in stage III by rigid-body minimization using H-bonds and a few NOE restraints.

3.1 Exact solution and systematic search-based minimization algorithm

The RDC, d, between two nuclei is related to the direction of the corresponding internuclear unit vector $\mathbf{v} = (x, y, z)$ by [12],

$$d = S_{xx}x^2 + S_{yy}y^2 + S_{zz}z^2 \tag{1}$$

where S_{xx}, S_{yy} and S_{zz} are three diagonal elements of a diagonalized Saupe matrix $\mathbf{S}$ (the alignment tensor) specifying the ensemble-averaged anisotropic orientation of a molecule in the laboratory frame; x, y and z are, respectively, the $x, y, z-$components of $\mathbf{v}$ in a POF which diagonalizes $\mathbf{S}$. $\mathbf{S}$ is a 3×3 symmetric, traceless matrix with five independent elements. Note that $x^2 + y^2 + z^2 = 1$ and $S_{xx} + S_{yy} + S_{zz} = 0$. Thus, when projected onto the XY-plane of the POF of RDCs, Eq. (1) represents two ellipses with major axis of $\sqrt{\frac{d-S_{zz}}{S_{xx}-S_{zz}}}$ and minor axis of $\sqrt{\frac{d-S_{zz}}{S_{yy}-S_{zz}}}$ assuming $|S_{zz}| \geq |S_{xx}| \geq |S_{yy}|$. Such a projected ellipse will be called either NH or CH *RDC ellipse*, for brevity.

The minimization algorithm is divided into three phases. The first phase is the sampling of both NH and CH RDCs based on Gaussian distributions about the experimental data values (the measured experimental RDC value and the experimental error define, respectively, the mean and variance of the Gaussian distribution). The perturbation of the experimental values is necessary for computing a structure with backbone (ϕ, ψ) angles in the favorable Ramachandran regions. The second phase is the computation of an optimal peptide plane for the first residue *and* an optimal conformation vector from the two sets of sampled RDCs. A conformation vector for an m-residue fragment is defined as $(\phi_1, \psi_1, \phi_2, \psi_2, \cdots, \phi_{m-1}, \psi_{m-1})$, where (ϕ_i, ψ_i) are the dihedral angles of residue i. The third phase is the construction of a backbone model from the first peptide plane, and the optimal conformation vector. Interested readers can see our previous paper [14] for an algorithm for computing the optimal first peptide plane. The computation of an optimal conformation vector proceeds as follows. First, the set of all the plausible conformational vectors for the fragment are computed by depth-first search (DFS) over the cross product of all the sets of CH and NH directions computed from RDCs. A *plausible* conformation vector is defined as a vector with all its $m - 1$ (ϕ, ψ) angles in the favorable Ramachandran region for the corresponding secondary structure type. Such a favorable region defines our *Ramachandran filter*. Next, an optimal conformational vector is computed from the set of all the plausible conformational vectors by minimizing the following target function, $\sum_{i=1}^{m-1}((\phi_i - \phi_{\mu_a})^2 + (\psi_i - \psi_{\mu_a})^2) + \sum_{i=1}^{m}((d'_{1,i} - d_{1,i})^2 + (d'_{2,i} - d_{2,i})^2)$, where (ϕ_i, ψ_i) are the computed angles from solving the quartic and linear equations for residue i and $(\phi_{\mu_a}, \psi_{\mu_a})$ are the average (ϕ, ψ) angles over the PDB for the corresponding secondary structure type, $d_{1,i}$ and $d_{2,i}$ are, respectively, the experimental NH and CH RDC values and $d'_{1,i}$ and $d'_{2,i}$ are the sampled NH and CH RDC values. This minimization is a search over a *finite* set, namely the (ϕ, ψ) angles obtained by *exactly* solving low-degree monomials. The data structure utilized for DFS is a dynamically constructed search tree (DFS-tree for brevity). The maximum height of the tree is m for an $m-$residue fragment and the first residue corresponds to depth 1. With this convention, the residue number i is the same as the depth i of the DFS-tree. Each node at depth i corresponds to a solution for the backbone (ϕ_i, ψ_i) angles of residue i computed, respectively, from the CH RDC of residue i and the NH RDC of residue $i + 1$. At each step of the search, the algorithm prunes the computed multiple (ϕ, ψ) angles using a Ramachandran filter. During DFS the orientation of the peptide plane of residue $i + 1$ is computed from the orientation of the peptide plane i and the intervening (ϕ_i, ψ_i) angles.

4 Biological results and significance

Our algorithm has been successfully applied to compute the backbone structures of three proteins. We first applied the algorithm to the 76-residue protein human ubiquitin using NH RDCs in two media [14, 15] or NH and CH RDCs in a single medium (manuscript in preparation), plus twelve hydrogen bonds and four NOE distances. The backbone RMSD between the RDC-derived backbone substructure consisting of an α-helix (N25–E34) and a β-sheet with five strands, and the corresponding portion of the X-ray structure (PDB ID, 1UBQ) [13], is only 0.97 Å in a single medium or 1.23 Å in two media. We have also applied our algorithm to compute the backbone substructures of two other proteins, 81-residue DNA-damage-inducible protein I (PDB ID, 1GHH) and 56-residue immunoglobulin binding protein G (PDB ID, 3GB1), using NH RDCs in two media and sparse distance restraints. The backbone RMSDs between the substructures computed by our algorithm and the corresponding portions of NMR structures are, respectively, 1.55 Å for DNA-damage-inducible protein I and 0.96 Å for immunoglobulin binding protein G. Note that the NMR structures to which we compared were computed by MD/SA [1] 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). All the experimental data were downloaded from the PDB. For ubiquitin the restraints are extracted from the PDB file 1D3Z.

Compared with heuristic algorithms [5, 8, 2, 11] for computing backbone structures using RDCs, our algorithm achieves similar or better accuracy but requires less data, is much less biased toward the PDB and does not depend on MD. The success of our algorithm using only *two* RDCs per residue and sparse distance restraints shows that solution NMR spectroscopy can play a major role in determining protein backbone fold rapidly and inexpensively, which should be important in structural genomics.

5 Analysis

In this section, we first analyze the complexity of our algorithm. Then we quantify three filters for systematic search: the real solution filter, the Ramachandran filter and the geometric filter. Finally, we propose a stochastic model for analyzing our minimization algorithm. We also present computational experimental results and compare them with the theoretical analysis.

5.1 Algorithmic complexity and performance

The complexity analysis of the algorithm in section 3 is as follows. The alignment tensor, the coefficients of the quartic and the linear or quadratic equations and their solutions (stage I) can all be computed in $O(m)$ time for an m-residue fragment. The search for an optimal first peptide plane (stage II) [14], takes $O(mk_1^3)$ time on a $k_1 \times k_1 \times k_1$ grid for three Euler angles. In practice, it takes less than one minute on a $180 \times 90 \times 180$ grid on a Pentium 4 (2.4GHz) Linux workstation. The search for relative positions among RDC-derived structure elements using NOE distances (stage III) takes $O(lk_2^2)$ time on a $k_2 \times k_2$ grid for the polar angles ϕ and θ with l NOEs. In practice, it takes several seconds on a 90×180 grid. The computation of the conformation and global orientation of an m-residue fragment by the systematic search-based minimization (stage II), takes $O(k16^m)$ time in the worst case, where k is the resolution of a grid search over a Gaussian distribution about the experimental data [14], and $16 = 4 \times 4$ where 4 is the maximum number of solutions for either the ϕ angle from a CH RDC or a ψ angle from an NH RDC. The number 16 is also the maximum branching factor B_{max} of the DFS-tree at each node. In summary, the total run time of the algorithm is $O(n(m + lk_2^2 + mk_1^3 + k16^m))$ for n $m-$residue fragments. However, despite the worst-case exponential running time the systematic search-based minimization takes, in practice, only several minutes for computing either a helix or strand. Furthermore, our algorithm takes only 30-40 minutes to compute an entire backbone substructure consisting of α-helices and β-sheets. The running time depends on the size of the protein and the quality of the experimental data: the biggest effects are due to the size of the experimental error and the number of missing RDC data. In the following we prove that the average case complexity is indeed much faster than the worst-bound.

5.2 Ellipse equations for backbone CH and NH vectors

In this section, we state and prove four Propositions, which provide a basis for quantifying the Ramachandran and geometric filters used for pruning the solution (conformation) space of the systematic search.

Proposition 5.1 *The backbone CH unit vector of residue i, when projected onto the XY-plane of any global coordinate frame in space, lies on an ellipse. The resulting ellipse equation can be represented in a parametric form with the backbone ϕ_i angle as the parameter. The coefficients of the ellipse equation are determined by the orientation of peptide plane i in the global frame.*

Proposition 5.2 *The backbone NH unit vector of residue $i + 1$, when projected onto the XY-plane of any global coordinate frame in space, lies on an ellipse. The resulting ellipse equation can be represented in a parametric form with the backbone ψ_i angle as the parameter. The coefficients of the ellipse equation are determined by the ϕ_i angle and the orientation of peptide plane i in the global frame.*

We sketch a proof for Proposition 5.1. Proposition 5.2 can be proved similarly. In the following we let the global coordinate frame be the POF of RDCs.

Proof. From protein backbone geometry we have

$$\begin{aligned}\mathbf{vM} &= (x, \quad y, \quad z)\mathbf{M}\\ &= (C_x \cos\phi_i + C_z \sin\phi_i,\ C_y,\\ &\quad - C_x \sin\phi_i + C_z \cos\phi_i),\end{aligned} \tag{2}$$

where $\mathbf{v} = (x, y, z)$ is a backbone CH unit vector in the POF of RDCs, and (C_x, C_y, C_z) is a unit vector known from the fixed protein backbone geometry taken from table 4 in [14]. The 3×3 matrix $\mathbf{M}$ is defined as follows

$$\mathbf{M}(\alpha_i, \beta_i, \gamma_i) = \mathbf{R}_B \mathbf{R}_G(\alpha_i, \beta_i, \gamma_i), \tag{3}$$

where $\mathbf{R}_G(\alpha_i, \beta_i, \gamma_i)$ is the rotation matrix between the global POF of RDCs and a local coordinate frame defined in the peptide plane i, and $(\alpha_i, \beta_i, \gamma_i)$ are three Euler angles. The 3×3 matrix $\mathbf{R}_B$ is known from the fixed backbone geometry [14].

Defining a new unit vector $\mathbf{w} = (x', y', z') = \mathbf{vM}$ and letting $C_d = \sqrt{1 - C_y^2}$, Eq. (2) can be written as

$$\begin{aligned}&(x', \quad y', \quad z') =\\ &(C_d \sin(\phi_i + \phi_0),\ C_y,\ C_d \cos(\phi_i + \phi_0))\end{aligned} \tag{4}$$

where $\sin\phi_{0} = \frac{C_x}{\sqrt{1-C_y^2}}$ and $\cos\phi_{0} = \frac{C_z}{\sqrt{1-C_y^2}}$. Eq. (4) represents a curve (in fact, a circle) in the XZ-plane of a coordinate frame with +Z axis in peptide plane i and +Y axis along the $\overrightarrow{\mathrm{NC}_\alpha}$ vector. In the POF of RDCs the curve becomes

$$\begin{aligned}
x &= M_{11}x' + M_{21}y' + M_{31}z' \\
&= C_d(M_{11}\sin(\phi_i+\phi_{0}) + M_{31}\cos(\phi_i+\phi_{0})) + M_{21}C_y \\
y &= M_{12}x' + M_{22}y' + M_{32}z' \\
&= C_d(M_{12}\sin(\phi_i+\phi_{0}) + M_{32}\cos(\phi_i+\phi_{0})) + M_{22}C_y \\
z &= M_{13}x' + M_{23}y' + M_{33}z \\
&= C_d(M_{13}\sin(\phi_i+\phi_{0}) + M_{33}\cos(\phi_i+\phi_{0})) + M_{23}C_y.
\end{aligned}$$

When a rectangular Ramachandran filter for the ϕ angle is applied, the arc length L of this curve is

$$L(\phi_l,\phi_h) = \int_{\phi_l}^{\phi_h} \sqrt{(\frac{dx}{d\phi_i})^2 + (\frac{dy}{d\phi_i})^2 + (\frac{dz}{d\phi_i})^2}\ d\phi_i\ . \tag{5}$$

where ϕ_h and ϕ_l are the range of ϕ angle defined by the filter. By the orthogonality property of the matrix $\mathbf{M}$, the expression for L can be simplified as

$$L(\phi_l,\phi_h) = C_d\ (\phi_h - \phi_l)\ . \tag{6}$$

Eq. (6) is consistent with the geometric intuition that in space the curve is still a circle. However, the projection of this circle onto the XY-plane of the POF is, in general, an ellipse, which will be called the CH *vector ellipse* for brevity. Letting $x_{0} = M_{21}C_y$, $y_{0} = M_{22}C_y$, $\sin\theta_{X} = \frac{M_{11}}{\sqrt{M_{11}^2+M_{31}^2}}$, $\cos\theta_{X} = \frac{M_{31}}{\sqrt{M_{11}^2+M_{31}^2}}$ and $\sin\theta_{Y} = \frac{M_{12}}{\sqrt{M_{12}^2+M_{32}^2}}$, $\cos\theta_{Y} = \frac{M_{32}}{\sqrt{M_{12}^2+M_{32}^2}}$, we have

$$\begin{aligned}
x - x_{0} &= a\sin(\phi_i + \phi_{0} + \theta_{X}) \\
y - y_{0} &= b\cos(\phi_i + \phi_{0} + \theta_{Y})\ .
\end{aligned} \tag{7}$$

Eq. (7) is an ellipse equation in parametric form with ϕ_i as the parameter, and major axis $a = C_d\sqrt{(M_{11}^2 + M_{31}^2)}$ and minor axis $b = C_d\sqrt{(M_{12}^2 + M_{32}^2)}$. Letting $\phi_a = \phi_i + \phi_{0} + \theta_{X}, \phi_b = \phi_i + \phi_{0} + \theta_{Y}$, the arc length L becomes

$$\begin{aligned}
&L(\phi_l,\phi_h,\alpha_i,\beta_i,\gamma_i) \\
&= \int_{\phi_l}^{\phi_h} \sqrt{(\frac{dx}{d\phi_i})^2 + (\frac{dy}{d\phi_i})^2}\ d\phi_i \\
&= \int_{\phi_l}^{\phi_h} \sqrt{b^2\sin^2\phi_b + a^2\cos^2\phi_a}\ d\phi_i
\end{aligned} \tag{8}$$

According to Eqs. (7, 8), the length L, the center of the ellipse, (x_{0}, y_{0}), and the axes a, b all are functions of the rotation matrix $\mathbf{R}_{G}$ between a local frame defined in peptide plane i and the global POF since $\mathbf{M}$ is a function of $\mathbf{R}_{G}(\alpha_i,\beta_i,\gamma_i)$ (Eq. (3)). Eq. (8) is an elliptic integral, which can be represented as a Legendre elliptic integral of the second kind by algebraic manipulation. Letting $t = \phi_a$ and $\theta = \theta_{Y} - \theta_{X}$ the integrand in Eq. (8) can be written as

$$b^2\sin^2\phi_b + a^2\cos^2\phi_a = a_1\sin 2t + b_1\cos 2t + c_1, \tag{9}$$

where $a_1 = \frac{1}{2}b^2\sin 2\theta, b_1 = \frac{1}{2}(a^2 + b^2\cos 2\theta)$ and $c_1 = \frac{1}{2}(a^2 + b^2\cos 2\theta + 2b^2\cos^2\theta)$. Letting $r_1 = \sqrt{a_1^2 + b_1^2}$, $\cos t_{0} = \frac{b_1}{r_1}$ and $\sin t_{0} = \frac{a_1}{r_1}$ we have

$$b^2\sin^2\phi_b + a^2\cos^2\phi_a = (r_1 + c_1)(1 - K\sin^2(t - \frac{t_{0}}{2})), \tag{10}$$

where $K = \frac{2r_1}{r_1+c_1}$. It can be shown that $0 \leq K \leq 1$, thus, Eq. (8) is a Legendre elliptic integral of the second kind. The integral (Eq. (8)) can be computed accurately using quickly convergent expansion [3, 10]. ▮

To prove Proposition 5.2, we note that in the POF of RDCs, given ϕ_i, the NH vector of residue $i+1$ lies on a circle with ψ_i as a parameter. Similarly to Eq. (6), when a Ramachandran filter for the ψ angle is applied, the arc length of this circle can be written as

$$L(\psi_l,\psi_h) = D_y\ (\psi_h - \psi_l) \tag{11}$$

where D_y is a constant known from the fixed backbone geometry [14], and ψ_h and ψ_l are the range of ψ angle defined by the filter. When projected onto the XY-plane of the POF, the circle becomes an ellipse, which will be called the NH *vector ellipse*. The length $L(\psi_h,\psi_l)$, the center and both axes of the NH vector ellipse, all are functions of both the ϕ_i angle and the rotation matrix $\mathbf{R}_{G}(\alpha_i,\beta_i,\gamma_i)$.

By Propositions 5.1 and 5.2 and the RDC equation (Eq. (1)) we can easily prove the following two Propositions for computing, respectively, the sines and cosines of ϕ and ψ angles from the CH and NH RDCs measured in a single medium. Interested readers can see our previous paper [14] for similar Propositions for computing (ϕ,ψ) angles from NH RDCs measured in two media.

Proposition 5.3 *Given the orientation of peptide plane i in the POF of RDCs, the x-component of the CH unit vector* $\mathbf{v}$ *of residue i, in the POF, can be computed by solving a quartic monomial in x derived from the CH RDC ellipse equation (Eq. (1)) and the corresponding CH vector ellipse equation (Eq. (7)). Given the x-component, the y-component can be computed from either Eq. (1) or Eq. (7), and the z-component from $x^2 + y^2 + z^2 = 1$. Given* $\mathbf{v}$*, the sine and cosine of the ϕ_i angle can be computed by solving a linear equation.*

Proposition 5.4 *Given the orientation of peptide plane i in the POF of RDCs, the x-component of the NH unit vector*

$\mathbf{v}$ *of residue* $i+1$*, in the POF, can be computed by solving a quartic monomial in* x *derived from the NH RDC ellipse equation (Eq. (1)) and the corresponding NH vector ellipse equation. Given the* x*-component, the* y*-component can be computed from Eq. (1), and the* z*-component from* $x^2+y^2+z^2=1$*. Given* $\mathbf{v}$*, the sine and cosine of the* ψ_i *angle can be computed by solving a linear equation.*

Geometrically, the solutions to a quartic equation are just the intersections of the corresponding RDC and vector ellipses.

5.3 A probabilistic model for pruning

We begin our analysis of the average-case complexity of the minimization algorithm with the quantification of filters contributing to the reduction of branching factors of the DFS-tree for each systematic search. As stated previously (Section 5.1), even though the branching factor, B_i, at a node of depth i can be 16 in the worst case, a much smaller B_i is observed in practice. In the following we show how B_i is reduced. We represent each filter as a probability that a solution passes that filter. Hence, a maximum probability of 1 corresponds to no filter (all solutions pass), and 0 would be the most restrictive filter (no solutions pass). For convenience, we do not normalize the corresponding probability distribution. In this paper, we present our computational experimental results for NH and CH RDCs in a single medium and compare them with the theoretical analysis. A similar analysis can be applied to NH RDCs in two media since the practical performance of the search in both cases is very similar. Since the running time is directly proportional to the total number, S_i, of paths (conformations) at depth i of the DFS-tree (Fig. 1), the theoretical and computational experimental results are compared using B_i and S_m, where S_m is the total number of paths at the leaves of an m-depth DFS-tree. Note that $S_m = \prod_{i=1}^{m} B_i$ and $B_i = \frac{S_i}{S_{i-1}}$ where $S_0 = 1$. Furthermore, we argue that it is more proper to analyze the performance of the search by studying the value of B_i, rather than to analyze its asymptotic behavior when the number of residues in an α-helix or a β-sheet gets very large, since the number of residues in a typical α-helix or β-sheet is, most likely to be small (< 20 for a large majority of proteins and almost never > 50). Note that according to our convention (Section 3.1), residue number i is the same as depth i of the DFS-tree. In the following, depending on which filters are used, a different branching factor B_i is obtained. In sections $5.3.1 - 5.3.3$ the branching factor B_i is analyzed for systematic searches using some of the 3 filters either in isolation (i.e., with all other filters turned off), and in pairs. Later we also analyze the branching factor associated with the Gaussian distribution. In each case, we will state which filters are being used when we calculate B_i.

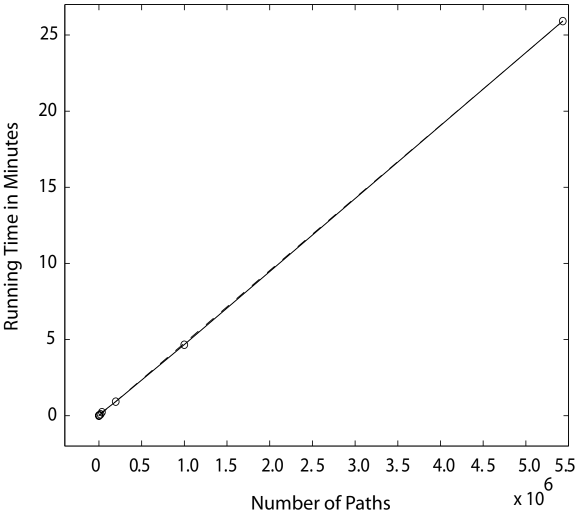

Figure 1. The running time vs. the number of paths (S_i). The x-axis is the number of paths of a systematic search over a ten-residue ideal α-helix using the back-computed CH and NH RDCs of the protein ubiquitin. The y-axis is the running time in minutes. At the lower end of the line, there are 7 overlapping data points.

5.3.1 Real solution filter

It has been shown in algebraic geometry [9] that an r-degree monomial with random real coefficients has, on average, $\sqrt{r}$ real solutions. In our case, $r=4$ since both ϕ and ψ are computed by solving quartic equations. Thus, we expect that there are only $\sqrt{r} \times \sqrt{r} = 2 \times 2$ solutions for (ϕ, ψ) angles for any residue. In practice, with a real solution filter alone we observe an average $B_i = 4.19$ for all the residues when the input to the systematic search are RDCs back-computed from an ideal α-helix (Fig. 2A). Back-computed RDCs mean the RDCs simulated from an ideal α-helix using Saupe matrices computed in stage I of our algorithm (Section 3). We call this reduction in B_i the *real solution filter*. We represent this filter as $P_{M} = \frac{\sqrt{r}}{r} \times \frac{\sqrt{r}}{r} = \frac{2}{4} \times \frac{2}{4} = \frac{1}{4}$. Without the filter, in contrast, $P_{M} = 1$.

5.3.2 Ramachandran filter

We assume that the number of (ϕ, ψ) solutions, that is, B_i, computed from the quartic and linear equations is, on average, a function of the arc lengths of the two intersecting ellipses. Hence, as shown in the proofs of Propositions 5.1-5.2 (Section 5.2), in the POF of RDCs, a Ramachandran filter defined by $[\phi_l, \phi_h], [\psi_l, \psi_h]$ can be represented as a probability $P_{R}(i) = P(\phi_i)P(\psi_i)$, where $P(\phi_i)$ is a probability representing the pruning of ϕ_i solutions by the filter, and $P(\psi_i)$ is a probability representing the pruning of ψ_i solutions by the filter. Formally, given a single RDC value, the distributions for $P(\phi_i)$ and $P(\psi_i)$ can each be defined as a conditional probability distribution using the Lebesgue-Stieltjes integral for a discrete random variable on a curve [7]. Here, the curve is an ellipse (Sec. 5.2). With continuously varying RDC values, the distributions for $P(\phi_i)$ and $P(\psi_i)$ can each be defined as a Lebesgue integral on a 2D band (area) on the surface of a sphere. We associate the same probabilities with the Ramachandran filter when the curves or bands are projected onto the XY-plane of the POF of RDCs. As an approximation, we will assume that

$$P(\phi_i) = \frac{\phi_h - \phi_l}{2\pi}, \quad \text{and} \quad P(\psi_i) = \frac{\psi_h - \psi_l}{2\pi}, \qquad (12)$$

that is, the probability, $P(\phi_i)$, is simply the ratio between the length of a circular arc defined by $[\phi_l, \phi_h]$ and the perimeter of the unit circle (2π). Similarly, the probability, $P(\psi_i)$, is simply the ratio between the length of a circular arc defined by $[\psi_l, \psi_h]$ and the perimeter of the unit circle. Without such an approximation, the probability can not be written as simply as Eq. (12) since, in principle, the arc length of an ellipse must be represented as an elliptic integral, which is a function of the rotation matrix $\mathbf{R}_G(\alpha, \beta, \gamma)$ (Eq. (8)). Our computational experimental results show that it is reasonable to make such an assumption. Indeed, the results in Fig. 2B show that the average B_i is reduced to about 1.11 by using a Ramachandran filter of $[\phi_l, \phi_h] = [-\pi, 0]$ and $[\psi_l, \psi_h] = [-\pi, 0]$, which is approximately one quarter of the average $B_i = 4.19$ (using only the real solution filter).

The Ramachandran filter is very effective in reducing the number of paths explored by the systematic search (Fig. 3). A conservative filter such as $[\phi_l, \phi_h] = [-\pi, 0]$ and $[\psi_l, \psi_h] = [-\pi, 0]$ can reduce the running time of the systematic search considerably. For example, at depth 10, without a Ramachandran filter, $S_{10} = 5,433,078$, and it takes 25 minute to compute all these conformations, while with the above filter $S_{10} = 32$ and it takes less than 0.002 minutes to compute them. We also assume that the Ramachandran filters for depth i and $i+1$, $P_R(i+1)$ and $P_R(i)$, are independent of one another.

5.3.3 Geometric filter

It is well known that in a typical α-helix or β-sheet, both the orientations of peptide planes and the directions of backbone vectors have certain periodicity along the backbone. For example, the peptide planes in a typical α-helix make a turn every 3.6 residues. Thus, the rotation matrices $\mathbf{R}_G(\alpha, \beta, \gamma)$ (Eq. (3)), along the backbone of an α-helix, oscillate with a 3.6 period, which will induce a similar oscillation in the coefficients of both the CH and NH vector ellipses (Eq. (7)) since these coefficients are functions of $\mathbf{R}_G(\alpha, \beta, \gamma)$. Indeed, in practice, we observed that average arc lengths of CH vector ellipses have a 3.6 period along the backbone of an ideal α-helix (Fig. 4A). Similarly, according to Eq. (1), the oscillation in the direction of backbone vectors is expected to induce similar oscillation in RDC values, and further to induce oscillation in the coefficients of both CH and NH RDC ellipse equations (Eq. (1)). Indeed, we found that the major axes of both NH and CH RDC ellipses have a 3.6 period along the backbone of an α-helix (Fig. 4). Note that the major axis of an RDC ellipse is a function of its RDC value (Section 3.1). According to Propositions 5.3-5.4, the (ϕ, ψ) angles are computed from the quartic equations obtained from an RDC ellipse equation and the corresponding vector ellipse equation. Therefore, the coefficients of the quartic equations will oscillate with a 3.6 period for a typical α-helix. As is well known in algebra, the number of real solutions to a monomial is a function of the coefficients of the monomial. Thus, in the present case, the number of (ϕ, ψ) solutions to the quartic equations will oscillate with a 3.6 period. The average number of (ϕ, ψ) solutions for residue i is the same as the branching factor, B_i, at depth i of the DFS-tree. We conclude that B_i will oscillate along the backbone of an α-helix with a 3.6 period. We call such an oscillation in the B_i value along the backbone the *geometric filter* for brevity. In practice, the oscillation is especially obvious when back-computed RDCs are used (the magenta line in Fig. 5). For real experimental RDCs the oscillation is less significant but still discernible (the red line in Fig .5). The reduction in the amplitude of the oscillation, when real experimental RDCs are used, is possibly caused by the large RMSDs between the RDCs back-computed from an ideal helix (E24–E34) and the real experimental RDCs. The RMSD is, respectively, 4.16 Hz for CH RDCs and 1.25 Hz for NH RDCs. Our Gaussian samplings are centered on the mean (the experimental value). Thus, the difference between the backbone (ϕ, ψ) angles computed using the sampled CH and NH RDCs close to their means and the average ideal (ϕ, ψ) angles is relatively large. A helix built with the computed (ϕ, ψ) angles will have a period different from 3.6. The B_i's are computed by averaging over the size of the Gaussian sampling with each successful sampling corresponding to a helix with a different period. Consequently, the amplitude of the observed oscillation can be reduced. By a similar argument to the above, for a β-strand, the oscillation is expected to be sinusoidal with a period of 2. Note that the oscillations in coefficients and B_i do not depend on where you start in the helix: they are a function of the orientations of the peptide planes, not a result of the algorithm's starting conditions. The geometric filter is really a discrete function of depth i with a 3.6 period for a typical α-helix, not a proper probability. Here, we represent it formally as a probability for

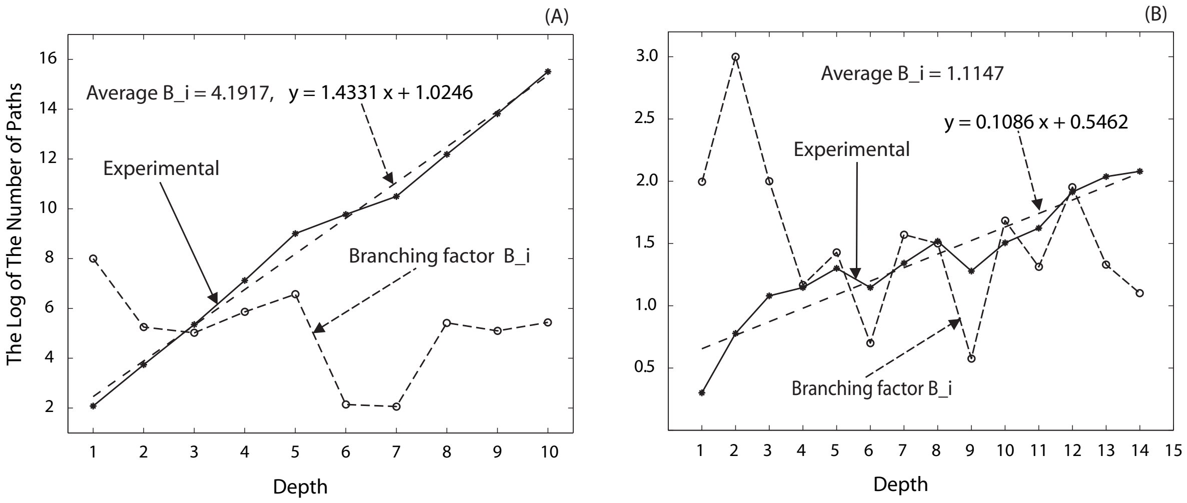

Figure 2. The Logarithm of the number of paths (S_i) with respect to the depth of the DFS-tree. The x-axis is the residue number i along the backbone (also the depth of the DFS-tree). The y-axis is the logarithm of the number of paths (S_i). The y-axis also gives the branching factor, B_i, using the same scale. (A) The solid line is S_i vs. depth i for the systematic search with the real solution filter but no Ramachandran filter, the two dashed lines are, respectively, a fit linear line and the branching factor B_i. (B) The same as (A) but with both the real solution and Ramachandran filters ($[-\pi, 0], [-\pi, 0]$) applied. The label “Experimental” means the data from computational experiments using back-computed RDCs. The data are presented for the protein ubiquitin using back-computed NH and CH RDCs.

convenience. Quantitatively, it can be approximated by

$$P_{\scriptscriptstyle G}(i) = c_1 \sin\frac{2\pi i}{3.6} + c_2 \sin\frac{2\pi i}{18} \tag{13}$$

where i is the residue number or depth of the DFS-tree, and c_1 and c_2 are constants. The number 3.6 is the period for an ideal α-helix while $18 = 5 \times 3.6$ is another period for the helix. In most cases, c_2 is close to 0.

In our implementation, the real solution and Ramachandran filters are represented *explicitly* (although clearly, the real solution filter is an intrinsic property of a polynomial with real coefficients). The geometric filter can be viewed as an analysis tool to describe the behavior of the algebraic system, namely, how the number of solutions varies with the geometry of the protein backbone. It is also an intrinsic property of the peptide plane geometry of regular secondary structures and is *implicitly* represented in the code as the modulation of the combined action of the real solution and Ramachandran filters.

In summary, assuming independence of the three filters (which is consistent with the computational experimental results), it is possible to assign a probability $P(i)$ for depth i of the DFS-tree (corresponding to residue i) with $P(i) = P_{\scriptscriptstyle M}(i)P_{\scriptscriptstyle R}(i)P_{\scriptscriptstyle G}(i)$, where $P_{\scriptscriptstyle M}(i), P_{\scriptscriptstyle R}(i)$ and $P_{\scriptscriptstyle G}(i)$ correspond, respectively, to the real solution, Ramachandran and geometric filters defined above. Then, with the above three filters $B_i = B_{max}P(i)$, where $B_{max} = 16$ is the theoretical worst-case branching factor without pruning by any filter (Section 5.1). As an approximation, it is reasonable to model the pruning of the three filters for the systematic search over an entire m-residue fragment by assigning a probability as follows:

$$P(\phi_1, \psi_1, \ldots, \phi_m, \psi_m) = \prod_{i=1}^{m} P_{\scriptscriptstyle M}(i)P_{\scriptscriptstyle R}(i)P_{\scriptscriptstyle G}(i). \tag{14}$$

5.4 Using a stochastic model to analyze a deterministic algorithm

The above results and analysis for a single systematic search shows that, on average, B_i can be reduced from the maximum value, $B_{max} = 16$, to about 1.11 using back-computed RDCs and all the above three filters. As stated previously (Section 3.1), it is necessary to perturb the real RDC data due to experimental errors and we model this perturbation with a Gaussian (normal) distribution, $\mathcal{N}(\mu, \sigma)$, where μ is the experimental (or back-computed) RDC and σ is the experimental error. In the implementation of our minimization algorithm, instead of using a grid search over a Gaussian distribution, we employ Gaussian sampling. The interested readers can see [14] for a detailed discussion of the Gaussian sampling. In the following we will present and analyze computational experimental results for the minimization, but first we will show that the

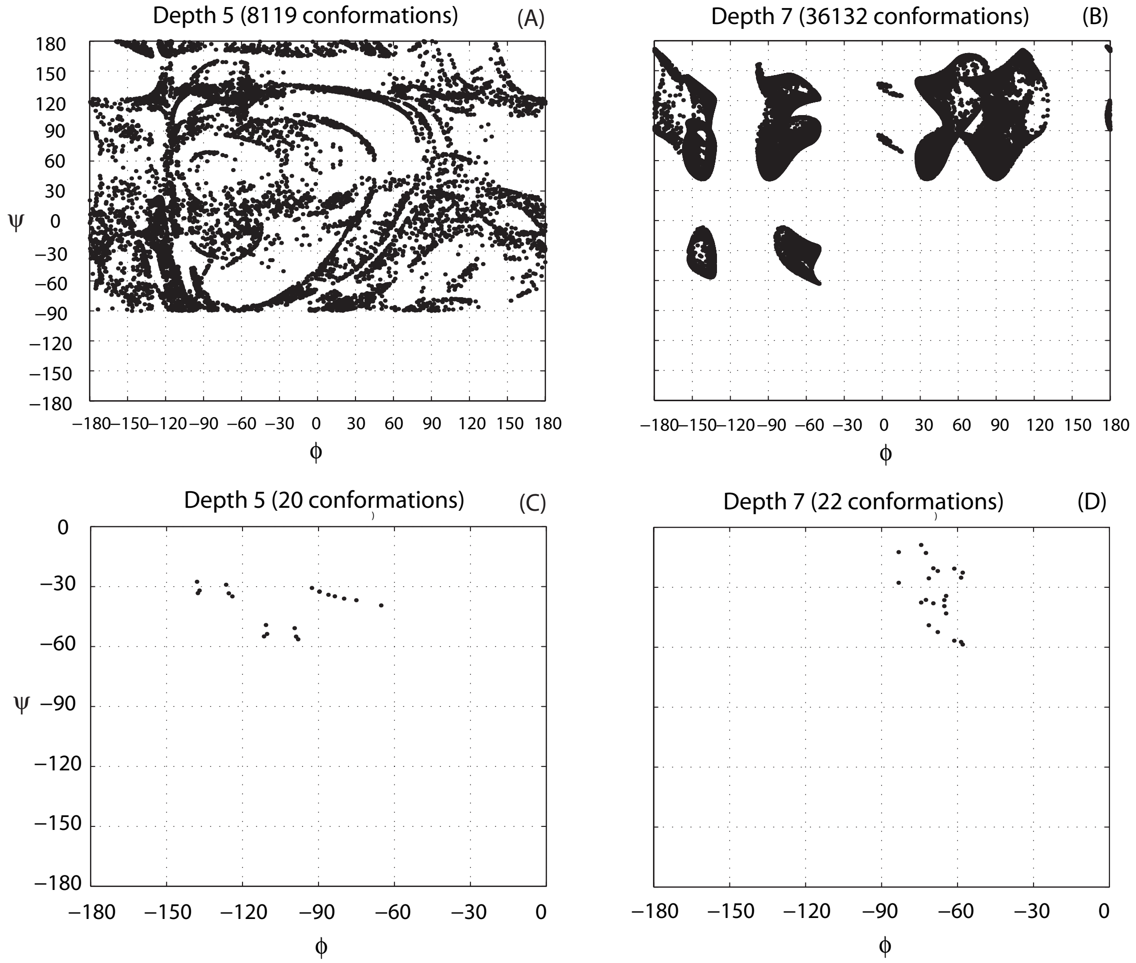

Figure 3. The distributions of (ϕ, ψ) solutions with respect to the depth of a DFS-tree without a Ramachandran filter (A, B) and with a Ramachandran filter (C, D). The x-axis is backbone ϕ angles, the y-axis is backbone ψ angles. The applied Ramachandran filter is $([-\pi, 0], [-\pi, 0])$. Also shown are the numbers of paths (conformations) at the corresponding depth. The data are presented for the protein ubiquitin using back-computed NH and CH RDCs.

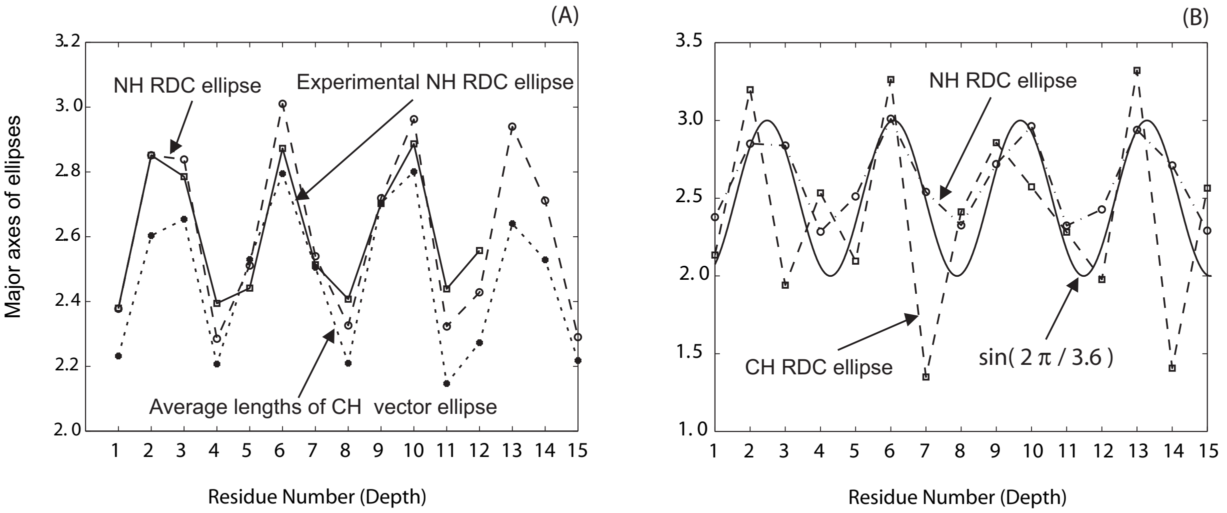

Figure 4. The major axes of NH and CH RDC ellipses (A) and the average arc lengths of CH vector ellipses along the backbone of an α-helix of the protein ubiquitin (B). The x-axis is the residue number along the α-helix. (A) The major axes of back-computed NH RDC ellipses (Eq. (1)) (dashed line) and real experimental NH RDC ellipses (solid line), as well as the average arc lengths of CH vector ellipses (dashed line) computed by Eq. (8). (B) The major axes of back-computed NH and CH RDC ellipses (dashed lines) and a sinusoidal curve (solid line) with a 3.6 period.

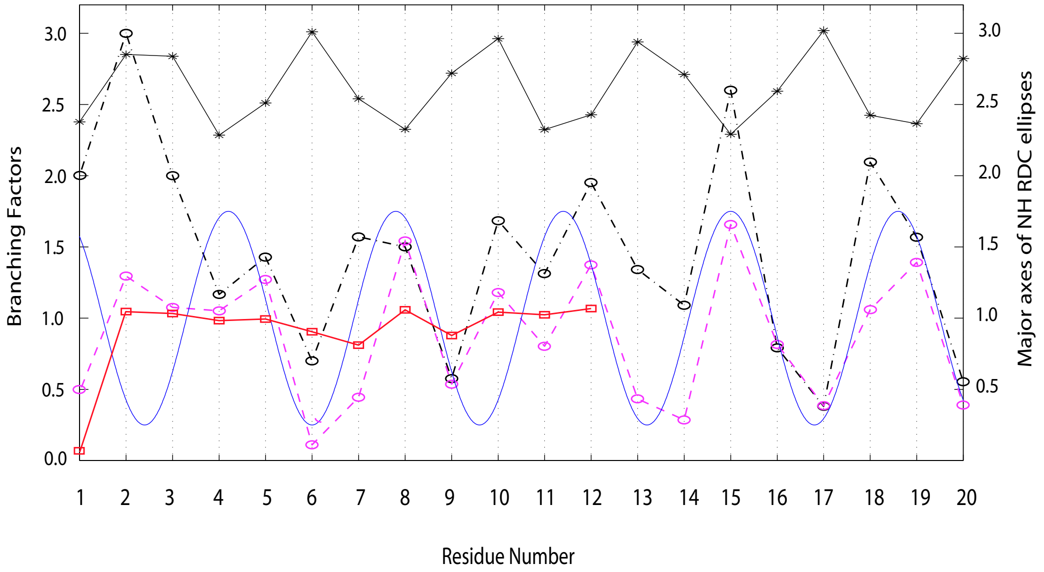

Figure 5. Branching factors (B_i) along the backbone of an α-helix of the protein ubiquitin. The x-axis is the residue number along the α-helix. The y-axis is the branching factor (left), and the major axes of NH RDC ellipses (right). The black solid line is the major axes of back-computed NH RDC ellipses. The dashed black line is the B_i of the systematic search using back-computed NH and CH RDCs with a Ramachandran filter of $[-\pi, 0]$, $[-\pi, 0]$ but without Gaussian sampling. The dashed magenta line is the average B_i of the systematic search-based minimization using back-computed NH and CH RDCs with the same Ramachandran filter *and* Gaussian sampling. The red line is the B_i using real experimental NH and CH RDCs with the same Ramachandran filter *and* Gaussian sampling. The B_1 of depth 1 for both back-computed (the dashed magenta line) and real experimental (the red line) NH and CH RDCs with a Ramachandran filter *and* Gaussian sampling is computed from the number of (ϕ, ψ) solutions divided by the size of the sampling. Excluding B_1, the average branching factor is 0.96 for back-computed (the dashed magenta line) RDCs, and 0.95 for real experimental RDCs. The solid blue line is a sinusoidal curve with a 3.6 period.

systematic search using a set of sampled CH and NH RDCs can be modeled as a branching process [4].

Empirically, if our algorithm is run with ideal back-computed RDCs, the average $B_i = 1.11$ (Section 5.3). If it is run on RDCs sampled from a Gaussian distribution away from the ideal back-computed value (the mean of the distribution), the average $B_i = 0.96$ (excluding B_1) (Fig. 5). Intuitively, this is because more conformations are pruned since they deviate from the ideal helix geometry that generated the back-computed RDCs. This reduction in average B_i can be explained, partially, by the following consideration: in order to have a real solution from the quartic equation the sampled RDC, d', must be in the following range: $|d'| \leq \sqrt{2(S_{yy}^2 + S_{zz}^2 + S_{yy}S_{zz})}$. Note that the range can be derived directly from Eq. (1) by applying the Cauchy-Schwarz inequality. We define a probability $P_{\scriptscriptstyle N}(i)$ to describe such a reduction in average B_i at depth i, that is, the B_i with a Gaussian sampling and the above three filters becomes

$$B_i = B_{max} P_{\scriptscriptstyle M}(i) P_{\scriptscriptstyle R}(i) P_{\scriptscriptstyle G}(i) P_{\scriptscriptstyle N}(i), \tag{15}$$

assuming that the sampling is independent of the three filters. Although the mathematical expression and estimation for $P_{\scriptscriptstyle N}(i)$ in terms of $\mathcal{N}(\mu_i, \sigma_i)$ are complicated, in practice, we observe that with Gaussian samplings using the back-computed RDCs as their means, $P_{\scriptscriptstyle N}(i)$ decreases with increasing depth: $P_{\scriptscriptstyle N}(1) = 0.50, P_{\scriptscriptstyle N}(2) = 0.22, P_{\scriptscriptstyle N}(3) = 0.12, ..., P_{\scriptscriptstyle N}(10) = 0.003$. Here, the experimental errors (σ_i) are set to be 3.0 Hz for CH RDCs and 1.5 Hz for NH RDCs. When the real experimental RDCs are used as the means, the probability, $P_{\scriptscriptstyle N}(i)$, is further reduced about 100-fold compared with the corresponding $P_{\scriptscriptstyle N}(i)$ using the back-computed RDCs as the means since for the latter, most of the sampled points are close to the mean, μ_i, while for the former, most of the sampled points are away from μ_i. As stated previously, when the branching factor for depth 1, B_1, is excluded, the average $B_i = 0.96$ (Fig. 5) when the back-computed RDCs are used as the means. Similarly, excluding B_1, the average $B_i = 0.95$ (Fig. 5) when the real experimental RDCs are used as the means. In both cases, $P_{\scriptscriptstyle N}(i)$ is smaller than 1 (excluding B_1) and is much smaller than 1 (including B_1). Thus, with B_i computed by Eq. (15), each systematic search over the DFS-tree with Gaussian samplings and our three filters can be modeled as a branching process [4]. As is well known in the theory of branching process [4] such a search (process) will extinct with probability one if B_i is less than 1, and often the process stops rather soon. In practice, for back-computed RDCs we found that about 85% of all the processes started at depth 1 stopped before depth 8. For real experimental RDCs, only 1% of the processes can start at depth 1, and among those started at depth 1, only 20% can go up to depth 8.

Finally, we are in a position to analyze our minimization algorithm. We assume that each search (process), j, starting successfully at depth 1 is itself a branching process described by a random variable $\mathbf{S}_m(j)$ (the total number of paths at depth m of an m-depth DFS-tree) [4] with an average branching factor well below 1, where $1 \leq j \leq k$ with k being the size of the sampling, that is, the resolution of a grid search over a Gaussian distribution at depth 1. Then, the number of paths at depth $m > 1$, $\mathbf{N}_p(m)$, can be represented as the sum of all the $\mathbf{S}_m(j)$'s, that is, $\mathbf{N}_p(m) = \sum_{j=1}^{k} \mathbf{S}_m(j)$. Thus, according to the central limit theorem [4], the total number of paths at depth m should have a normal distribution. In practice, we found that the number of paths with respect to the depth of the DFS-tree of the systematic search for both the back-computed RDCs and the real experimental RDCs, can be fit reasonably well with normal distributions (Fig. 6). In other words, the computational experimental data and our analysis show that asymptotically (as the sampling gets large), our minimization algorithm is a linear time algorithm and should be efficient in practice, consistent with our experimental observations.

5.5 Conclusion

We have developed an exact solution and systematic search-based minimization algorithm to compute the protein backbone structure using only *two* RDCs per residue and very sparse distance restraints. The algorithm has been implemented and demonstrated on three proteins using real experimental NMR data. Furthermore, we show that the average case complexity of our minimization algorithm with pruning provided by the real solution, Ramachandran and the geometric filters can be profitably modeled as branching processes. The analysis concludes that despite the worst-case exponential running time, our minimization algorithm should be quite efficient in practice, consistent with our experimental observations.

Acknowledgments

We would like to thank all the members of the Donald lab for helpful discussions and critical reading of the manuscript.

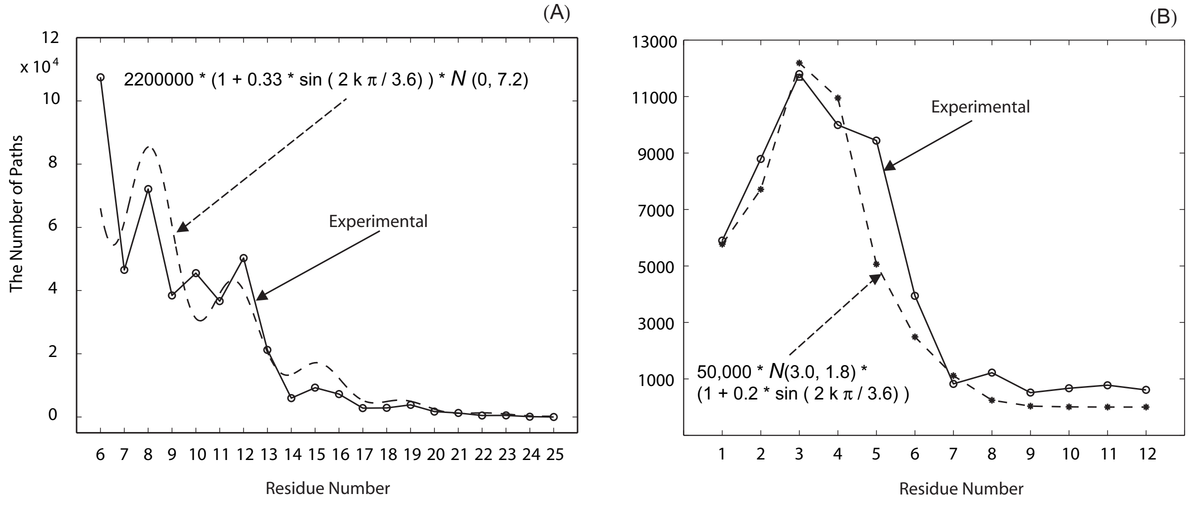

Figure 6. The number of paths ($\mathbf{N}_p(i)$) along protein backbone. The x-axis is the depth, i, of the DFS-tree, the y-axis is the total number of paths, $\mathbf{N}_p(i)$, as defined in the main text. (A) With back-computed NH and CH RDCs and a Ramachandran filter of $([-\pi, 0], [-\pi, 0])$. The solid line is the computational experimental data while the dashed line is a normal PDF, $\mathcal{N}(\mu, \sigma)$, multiplied by a sinusoidal function with a 3.6 period. (B) With real experimental NH and CH RDCs and a Ramachandran filter of $([-\pi, 0], [-\pi, 0])$. The solid line is the computational experimental data, and the dashed line is a normal PDF multiplied by a sinusoidal function with a 3.6 period. The data are presented for the protein ubiquitin with a sampling size = 512×1024. Note no normalization has been applied.

References

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

[2] F. Delaglio, G. Kontaxis, and A. Bax. Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *J. Am. Chem. Soc.*, 122(9):2142–2143, 2000.

[3] A. R. DiDonato and A. V. Hershey. New formulas for computing incomplete elliptic integrals of the first and second kind. *J. ACM*, 6(4):515–526, 1959.

[4] W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, Inc, New York, 1970.

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

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

[7] P. R. Halmos. *Measure Theory*. D. Van Nostrand Co., Inc., Princeton, N. J., 1950.

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

[9] M. Kac. On the average number of real roots of a random algebraic equation (ii). *Proc. London Math. Soc.*, 50:390–408, 1948.

[10] G. E. Lee-Whiting. Formulas for computing incomplete elliptic integrals of the first and second kinds. *J. ACM*, 10(2):126–130, 1963.

[11] C. A. Rohl and D. Baker. De Novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc.*, 124(11):2723–2729, 2002.

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

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

[14] L. Wang and B. R. Donald. 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. *Journal of Biomolecular NMR*, 2004. In Press. A PDF preprint can be obtained at: http://www.cs.dartmouth.edu/brd/.

[15] L. Wang, R. Mettu, R. Lilien, and B. R. Donald. An exact algorithm for determining protein backbone structure from NH residual dipolar couplings. In *IEEE Computer Society Bioinformatics Conference*, pages 611–612, Stanford University, CA, 2003.