

A HAUSDORFF-BASED NOE ASSIGNMENT ALGORITHM USING PROTEIN BACKBONE DETERMINED FROM RESIDUAL DIPOLAR COUPLINGS AND ROTAMER PATTERNS

Jiayang (Michael) Zeng and Chittaranjan Tripathy

*Department of Computer Science, Duke University,
Durham, NC 27708, USA*

Pei Zhou

*Department of Biochemistry, Duke University Medical Center,
Durham, NC 27708, USA*

Bruce R. Donald^{*†}

*Department of Computer Science, Duke University,
Department of Biochemistry, Duke University Medical Center,
Durham, NC 27708, USA*

**Email: www.cs.duke.edu/brd*

High-throughput structure determination based on solution Nuclear Magnetic Resonance (NMR) spectroscopy plays an important role in structural genomics. One of the main bottlenecks in NMR structure determination is the interpretation of NMR data to obtain a sufficient number of accurate distance restraints by assigning nuclear Overhauser effect (NOE) spectral peaks to pairs of protons. The difficulty in automated NOE assignment mainly lies in the ambiguities arising both from the resonance degeneracy of chemical shifts and from the uncertainty due to experimental errors in NOE peak positions. In this paper we present a novel NOE assignment algorithm, called *HAusdorff-based NOE Assignment (HANA)*, that starts with a high-resolution protein backbone computed using only two residual dipolar couplings (RDCs) per residue^{37, 39}, employs a Hausdorff-based pattern matching technique to deduce similarity between experimental and back-computed NOE spectra for each rotamer from a statistically diverse library, and drives the selection of optimal position-specific rotamers for filtering ambiguous NOE assignments. Our algorithm runs in time $O(tn^3 + tn \log t)$, where t is the maximum number of rotamers per residue and n is the size of the protein. Application of our algorithm on biological NMR data for three proteins, namely, human ubiquitin, the zinc finger domain of the human DNA Y-polymerase Eta (pol η) and the human Set2-Rpb1 interacting domain (hSRI) demonstrates that our algorithm overcomes spectral noise to achieve more than 90% assignment accuracy. Additionally, the final structures calculated using our automated NOE assignments have backbone RMSD $< 1.7 \text{ \AA}$ and all-heavy-atom RMSD $< 2.5 \text{ \AA}$ from reference structures that were determined either by X-ray crystallography or traditional NMR approaches. These results show that our NOE assignment algorithm can be successfully applied to protein NMR spectra to obtain high-quality structures.

1. INTRODUCTION

High-throughput structure determination based on X-ray crystallography and Nuclear Magnetic Resonance (NMR^a) spectroscopy are key steps towards the era of structural genomics. Unfortunately, structure determination by either approach is generally time-consuming. In X-ray crystallography, growing a good quality crystal is in general a difficult task, while in NMR structure determination, the bottle-

neck lies in the processing and analysis of NMR data, and in interpreting a sufficient number of accurate distance restraints from experimental Nuclear Overhauser Enhancement Spectroscopy (NOESY) spectra, which exploit the dipolar interaction of nuclear spins, called *nuclear Overhauser effect* (NOE), for through-space correlation of protons. The intensity (or volume) of an NOE peak in a NOESY spectrum is converted into a distance restraint by calibrating the

^{*}Corresponding author.

[†]This work is supported by the following grant to B.R.D.: National Institutes of Health (R01 GM-65982).

^aAbbreviations used: NMR, Nuclear Magnetic Resonance; ppm, parts per million; RMSD, root mean square deviation; NOESY, Nuclear Overhauser Enhancement Spectroscopy; HSQC, Heteronuclear Single Quantum Coherence spectroscopy; NOE, Nuclear Overhauser Effect; RDC, Residual Dipolar Coupling; PDB, Protein Data Bank; pol η , zinc finger domain of the human DNA Y-polymerase Eta; hSRI, human Set2-Rpb1 interacting domain; POF, Principal Order Frame; CCD, Cyclic Coordinate Descent; SA, Simulated Annealing; MD, Molecular Dynamics; □, Q.E.D.; SM, Supplementary Material.

intensity (or volume) vs. distance curve or classifying all NOESY peaks into different bins.^{12, 16, 38} Traditional NMR structure determination approaches use NOE distance restraints as the main source of information to compute the structure of a protein, a problem known to be strongly NP-hard,³⁰ essentially due to the local nature of the restraints. Rigorous approaches to solve this problem using NOE data, such as the distance geometry method,¹⁰ require exponential time in the worst-case (see discussion in Ref. 39). While substantial progress has been made to design practical algorithms for structure determination,^{3, 12–14, 24, 28, 31} most algorithms still rely on heuristic techniques such as molecular dynamics (MD) and simulated annealing (SA), which use NOE data plus other NMR data to compute a protein structure. The NOE distances used by these distance-based structure determination protocols must be obtained by *assigning* NOE data, i.e., for every NOE, we must determine the associated pair of interacting protons in the primary sequence. This is called the *NOE assignment problem*.

While much progress has been made in automated NOE assignment,^{12, 14, 16, 21, 24, 27, 28} most NOE assignment algorithms have a SA/MD-based or a distance geometry-based structure determination protocol sitting in a tight inner loop, which is invoked many times to filter ambiguous assignments. Since distance geometry methods have exponential worst-case time complexity, and SA/MD-based structure determination protocols lack combinatorial precision and have no guarantees on solution quality or running time, these NOE assignment algorithms suffer from the same drawbacks, in addition to the inherent difficulties in the interpretation of NOESY spectra. Therefore, it is natural to ask if there exists a *provably polynomial-time* algorithm for the NOE assignment problem, which can guarantee solution quality—this will pave new ways for better understanding and interpretation of experimental data, and for developing robust protocols with *both* theoretical guarantees and good practical performance.

In Ref. 39, a new linear time algorithm was developed, based on Refs. 37 and 36, to determine protein backbone structure accurately using a minimum amount of residual dipolar coupling (RDC) data. RDCs provide *global* orientational restraints

on internuclear vectors, for example, backbone NH and CH bond vectors with respect to a global frame of reference. The algorithm in Refs. 37, 36, and 39 computes the backbone conformation by solving, in closed form, systems of low-degree polynomial equations formulated using the RDC restraints. The algorithm is combinatorially-precise and employs a systematic search strategy to compute the backbone structure in polynomial time. The accurately-computed backbone conformations enable us to propose a new strategy for NOE assignment. In Ref. 38, for example, an NOE assignment algorithm was proposed to filter ambiguous NOE assignments based on an ensemble of distance intervals computed using intra-residue vectors mined from a rotamer database, and inter-residue vectors from the backbone structure determined from Refs. 37, 36, and 39. The algorithm in Ref. 38 uses a *triangle-like inequality* between the intra-residue and inter-residue vectors to prune incorrect assignment for side-chain NOEs. However, the algorithm in Ref. 38 has the following deficiencies: (a) it does not exploit the diversity of the rotamers in the library, (b) uncertainty in NOE peak position, and other inherent difficulties in interpreting NOESY spectra suggest a probabilistic model with provable properties which Ref. 38 does not capture, and (c) it does not exploit rotamer pattern structure in NOESY spectra.

To address the shortcomings in Ref. 38 and other previous work, our algorithm, *Hausdorff-based NOE Assignment* (HANA), uses a novel pattern-directed framework for NOE assignment, that combines a combinatorially-precise, algebraic geometry-based approach for computing *high-resolution* protein backbones from residual dipolar coupling data, with a framework that uses a statistically diverse library of rotamers and the Hausdorff distance to measure similarity between experimental and back-computed NOE spectra, and drives the selection of optimal position-specific rotamers to prune ambiguous NOE assignments. Our Hausdorff-based framework views the NOE assignment problem as a *pattern-recognition* problem, where the objective is to establish a match by choosing the correct rotamers between the experimental NOESY spectrum and the back-computed NOE pattern. By explicitly modeling the uncertainty in NOE peak positions

and the probability of mismatches between NOE patterns, we provide a rigorous means of analyzing and evaluating the algorithmic benefits and the quality of assignments.

We first compute a high-resolution protein backbone from RDC data using the algorithms in Refs. 37, 36, and 39. Using this backbone structure, an assigned resonance list, and a library of rotamers²⁵, the NOE pattern for each rotamer can be back-computed (Figure 1B). By measuring the match of the back-computed NOE patterns with experimental NOESY spectrum, we choose an ensemble of top rotamers according to the match scores for each residue. Then, we construct an initial low-resolution protein structure by combining the high-resolution backbone and the chosen approximate rotamers together. The low-resolution structure is then used to *filter* ambiguous NOE assignments. Finally, our NOE assignments are fed to a structure calculation program, e.g., XPLOR/CNS³ which outputs the final ensemble of structures. The experimental results, based on our NMR data for three proteins, viz., human ubiquitin, the zinc finger domain of the human DNA Y-polymerase Eta (pol η) and the human Set2-Rpb1 interacting domain (hSRI) show that HANA achieves an assignment accuracy of more than 90%. In summary, our main contributions in this paper are:

- (1) Development of a novel framework that combines a combinatorially-precise, algebraic geometry-based linear time algorithm for high-resolution backbone structure determination with the Hausdorff distance measure, and exploits the statistical diversity of a rotamer library to infer accurate NOE assignments for both backbone and side-chain NOEs from 2D and 3D NOESY spectra.
- (2) Introduction of Hausdorff distance-based pattern matching technique to measure the similarity between experimental NOE spectra and back-computed NOE spectra, and modeling uncertainties arising both from false random matches and from experimental deviations in NOE peak positions.

- (3) A fully-automated $O(tn^3 + tn \log t)$ time NOE assignment algorithm, where t is the maximum number of rotamers in a residue and n is the number of residues in the protein.
- (4) Derivation of provable properties, viz. soundness in rotamer selection.
- (5) Application of our algorithm on three real biological NMR data sets to demonstrate high assignment accuracy ($> 90\%$), and fast running times (< 2 minutes).

2. PRELIMINARIES AND PROBLEM DEFINITION

In NMR spectra, each proton or atom is identified by its *chemical shift* (or *resonance*), which is obtained by mapping atom names in the known primary sequence of the protein to the corresponding frequencies from triple-resonance or other NMR spectra; this process is referred to as *resonance assignment*. Substantial progress has been made in designing efficient algorithms^{1, 20, 22, 26} for automatic resonance assignment. Given the chemical shift of each proton, the *NOE assignment problem* in two dimensions^b is to assign each NOESY peak to each pair of protons that are correlated through a dipole-dipole NOE interaction.

Formally, let $\{a_1, \dots, a_q\}$ denote the set of proton names (e.g., H $^\alpha$ of Arg56), where $q = \Theta(n)$ is the total number of protons and n is the number of residues in a protein. Let $\omega(a_i)$ denote the chemical shift for proton a_i determined from resonance assignment, $1 \leq i \leq q$. An NOE peak (a.k.a. *cross-peak*) with respective frequencies x and y for a pair of protons, is denoted by the point (x, y) on the plane of NOESY spectrum. Given a set of known chemical shifts $L = \{\omega(a_i), \dots, \omega(a_q)\}$ for all protons $\{a_1, \dots, a_q\}$ and a list of NOESY peaks (i.e., a set of points on the plane of NOESY spectrum), the *NOE assignment problem* is to map each NOE cross-peak (x, y) to an interacting proton pair (a_i, a_j) such that $\|\omega(a_i) - x\| \leq \delta_x$ and $\|\omega(a_j) - y\| \leq \delta_y$, where δ_x and δ_y encode the uncertainty in the peak position due to experimental errors.

^bThe problem for 3D and 4D cases can be defined in an analogous manner. Here the 2D case is explained for clarity. Our NOE assignment algorithm has been tested on both 2D and 3D spectra, and extends easily to handle 4D NOESY spectra.

In a hypothetical ideal case without any experimental error and noise, this would be an easy problem. However, for most proteins, two pairs of interacting protons can produce overlapping NOE peaks in a NOESY spectrum. The chemical shift differences of different protons are often too small to resolve experimentally, a phenomenon often referred to as *chemical shift degeneracy*. Also, due to experimental noise, artifact NOE peaks might occur from either manual or automated peak picking. These factors lead to more than one possible NOE assignment for a 2D NOESY spectrum which are called *ambiguous NOE assignments*.^{12, 21} Hence, one or more additional dimensions are generally introduced to relieve the congestion of NOE peaks. In a 3D NMR experiment, for example, each NOE peak is labeled with chemical shifts of a triple of atoms, viz., dipole-dipole interacting protons plus the heavy atom nucleus such as ¹⁵N or ¹³C bonded to the second proton. Even for 3D spectra, the interpretation and assignment of NOESY cross-peaks still remains hard, and poses a difficult computational challenge to obtain a unique NOE assignment. Manual assignment of NOESY peaks take months of time on average, requires significant expertise, and is prone to human errors. In structure determination, even a few incorrect NOE assignments can result in incorrect structures.⁵ Hence, it is critical to develop highly efficient and fully automated NOE assignment algorithms to aid high-throughput NMR structure determination.

3. PREVIOUS WORK

Protein structure determination using NOE distance restraints is strongly NP-hard,³⁰ essentially due to sparsity of the experimental data and local nature of the constraints. While rigorous approaches to solve this problem using distance intervals from NOE data, such as the distance geometry method,¹⁰ require exponential time in the worst-case; heuristic approaches such as SA/MD, while providing practical ways of solving this problem, lack combinatorial precision, and have no guarantees on running time or solution quality. Previous approaches for NOE assignment^{12, 14, 16, 21, 24, 27, 28} follow an iterative strategy, in which an initial set of relatively unambiguous NOEs is used to generate an ensemble of structures, which are then used to filter ambiguous

and inconsistent NOE assignments. This iterative assignment process is repeated until no further improvements in NOE assignments or structures can be obtained. What makes such approaches loose guarantees on the running time and assignment accuracy is their tight coupling with a heuristic structure determination protocol, which sits in a tight inner-loop of the assignment algorithm.

NOAH,^{27, 12} for example, uses the structure determination package DYANA,¹⁴ and follows the previously mentioned iterative strategy starting with an initial set of NOE assignments with supposedly one or two possible assignments. ARIA^{28, 24} and CANDID¹⁴ improved on NOAH by incorporating better modeling of ambiguous distance constraints. In AUTO-STRUCTURE¹⁶ more experimental data such as dihedral angle restraints from TALOS⁸ and slow H-D exchange data are used to improve assignment accuracy. In PASD²¹ several strategies were proposed to reduce the chance of invoking the structure calculation into a biased path due to the incorrect initial global fold. Since all these iterative NOE assignment programs invoke SA/MD-based structure determination protocols such as XPLOR/CNS³, they may converge to a local, but not a global minimum to obtain a best-fit of the data; therefore, the NOE assignments might not be correct.

An alternative approach for automated NOE assignment proposed by Wang and Donald in Ref. 38, based on Refs. 37, 36, and 39, uses a rotamer ensemble and residual dipolar couplings, and is the first polynomial-time algorithm for automated NOE assignment. However, Ref. 38 does not exploit the pattern structure of NOESY spectrum to model the uncertainty in peak positions probabilistically using a library of rotamers; therefore, assignment accuracy is reduced while processing NOESY spectra with many noisy peaks.

Our algorithm HANA retains the paradigm of Ref. 38, and develops a novel framework using the algebraic geometry-based linear time algorithm developed in Ref. 39 to compute high-resolution protein backbones from residual dipolar couplings, and then uses this backbone and a library of rotamers to do NOE assignments. Viewing the NOE assignment problem as a pattern-recognition problem, our algorithm uses an extended Hausdorff distance-

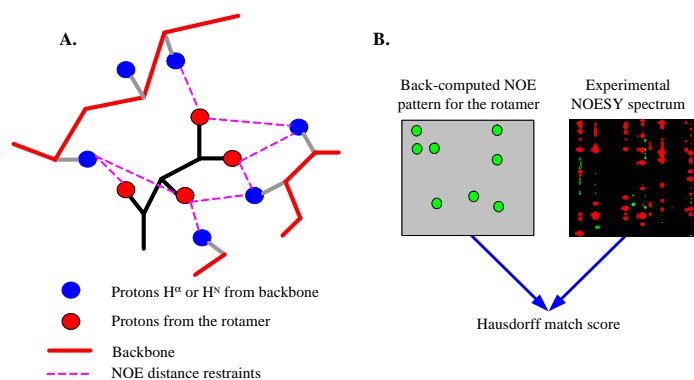


Fig. 1. Schematic illustration of the NOE assignment approach.

based probabilistic framework to model the uncertainties in NOE peak positions and the probability of mismatches between NOE patterns. In contrast to previous heuristic algorithms^{12, 14, 16, 21, 24, 27, 28} for NOE assignment, HANA has the advantages of being combinatorially precise with a running time of $O(tn^3 + tn \log t)$, where t is the maximum number of rotamers per residue and n is the size of the protein, and runs extremely fast in practice to compute high quality NOE assignments (> 90% assignment accuracy).

4. NOE ASSIGNMENT BASED ON ROTAMER PATTERNS

4.1. Overview of our approach

Our goal is to assign pairs of proton names^c to cross-peaks in NOESY data. Figure 1 illustrates the basic idea of our algorithm. The NOE assignment process can be divided into three phases, viz. initial NOE assignment (phase 1), rotamer selection (phase 2), and filtration of ambiguous NOE assignments (phase 3). The initial NOE assignment (phase 1) is done by considering all pairs of ambiguous NOEs assigned to a NOESY cross peak if the resonances of corresponding atoms fall within a tolerance window around the NOE peak. In the rotamer selection phase, we first compute the backbone structure from RDCs (see Section 4.2), and then place all the rotamers at each residue into backbone and compute all expected NOEs within the upper-bound limit of NOE

distance (Figure 1A). Based on the set of all expected NOEs and the resonance assignment list, we back-compute the expected NOE peak pattern for each rotamer (Figure 1B). By matching the back-computed NOE pattern with the experimental NOESY spectrum using an extended model of the Hausdorff distance,^{17, 19} we measure how well a rotamer fits the real side-chain conformation when interpreted in terms of the NOESY data. We then select the top k rotamers with highest fitness scores at each residue, and obtain a “low-resolution” structure,^d by combining the high-resolution backbone structure and the approximate ensemble of side-chain conformations at each residue. The low-resolution structure is then used (in phase 3) to filter ambiguous NOE assignments. The details of filtering ambiguous NOE assignments using the low-resolution structure are provided in Supplementary Material (SM) **Section 4** available online in Ref. 40.

4.2. Protein backbone structure determination from residual dipolar couplings

Residual dipolar coupling^{33, 34} data provide global orientational restraints on the internuclear bond vectors, such as, backbone NH and CH bond vectors with respect to a global coordinate frame. In solution NMR, RDCs can be recorded with high precision, and assigned much faster than NOEs. In Refs. 39 and 37, the authors proposed the first polynomial-

^cWe will use terms *proton name* and *proton* interchangeably in this paper.

^dThe “low resolution” structure generally has approximately 2.0–3.0 Å (all heavy atom) RMSD from the reference structures solved by X-ray or traditional NMR approaches.

time *de novo* algorithm, which we henceforth refer to as RDC-EXACT, to compute high-resolution protein backbone structures from RDC data. RDC-EXACT takes as input (a) two RDCs per residue (e.g., assigned NH RDCs in two media or NH and CH RDCs in a single medium), (b) delimited α -helices and β -sheets with known hydrogen bond information between paired strands, and a few unambiguous NOEs (used to pack the helices and strands). Note that, these sparse set of NOEs used by RDC-EXACT can usually be assigned using chemical shift information alone^{37, 39} without requiring any sophisticated NOE assignment algorithm. Our algorithm HANA uses the high-resolution backbones computed by RDC-EXACT. Loops with missing RDCs are computed using an enhanced version of robotics-based cyclic coordinate descent (CCD) algorithm.^{4, 32} The details of RDC-EXACT and modeling of loops (in case of missing RDCs) are provided in SM⁴⁰ **Section 1**.

4.3. NOE pattern matching based on the Hausdorff distance measure

Given two finite sets of points $B = \{b_1, \dots, b_m\}$ and $Y = \{y_1, \dots, y_n\}$ in Euclidean space, the Hausdorff distance between B and Y is defined as $H(B, Y) = \max\{h(B, Y), h(Y, B)\}$, where $h(B, Y) = \max_{b \in B} \min_{y \in Y} \|b - y\|$, and $\|b - y\|$ measures the normed distance (e.g., L_2 -norm) between points b and y . Intuitively, the Hausdorff distance $H(B, Y)$ finds the point in one set that is farthest from any point in the other set, and thus measures the degree of mismatch between the two point sets B and Y . The Hausdorff distance has been widely used in the image processing and computer vision problems, such as visual correspondence,¹⁷ pattern recognition,¹⁹ and shape matching,¹⁸ etc. Unlike many other pattern-recognition algorithms, Hausdorff-based algorithms are combinatorially precise, and provide a robust method for measuring the similarity between two point sets or image patterns^{18, 19} in the presence of noise and positional uncertainties.

In the NOE assignment problem, let B denote a back-computed NOE pattern, i.e., the set of back-computed NOE peaks, and let Y denote the set of experimental NOESY peaks. Generally, the size of a back-computed NOE pattern is much smaller than the total number of experimen-

tal NOESY peaks. Therefore, we only consider the *directed* Hausdorff distance from B to Y , namely, $h(B, Y) = \max_{b \in B} \min_{y \in Y} \|b - y\|$. We apply an extended model of Hausdorff distance^{18, 19, 17} to measure the match between the back-computed NOE pattern and experimental NOESY spectrum. Below, we assume 3D NOESY spectra without loss of generality.

Given the back-computed NOE pattern B with m peaks, and the set of NOESY peaks Y with w peaks, the τ -th Hausdorff distance from B to Y is defined as

$$h_\tau(B, Y) = \tau\text{th} \min_{b \in B} \min_{y \in Y} \|b - y\|,$$

where τ th is the τ -th largest of m values. We call $f = \tau/m$ the *similarity score* between the back-computed NOE pattern B and the experimental peak set Y , after fixing the Hausdorff distance $h_\tau(B, Y) = \delta$, which is the error tolerance in the NOESY spectra. The similarity score for a rotamer given δ can be computed using a scheme similar to Ref. 17:

$$s = \frac{|B \cap Y_\delta|}{|B|}, \quad (1)$$

where Y_δ denotes the union of all balls obtained by replacing each point in Y with a ball of radius δ , $B \cap Y_\delta$ denotes the intersection of sets B and Y_δ , and $|\cdot|$ denotes the size of a set.

We incorporate two types of uncertainty in the calculation of the similarity score in Equation (1) for the match between the back-computed NOE pattern and experimental NOESY spectrum: (a) possibility of a false random match¹⁷ in the NOESY spectra; (b) uncertainty of NOE peak positions due to experimental noise.

(a) Possibility of a false random match.¹⁷ A *false random match* between the back-computed NOE pattern and the experimental NOESY spectrum is defined as a match when $h_\tau(B, Y) \leq \delta$ occurs at random. We calculate the probability of a false random match and use it as a weighting factor for the similarity score in Equation (1). Let p be the probability for a back-computed NOE peak to randomly match to an experimental peak in Y_δ . Let θ be the probability of a false random match, which can be estimated using the following asymptotic approximation from Ref. 17:

$$\theta \approx \frac{1}{2} \left(\Phi\left(\frac{(1-p)m}{\rho}\right) - \Phi\left(\frac{(s-p)m}{\rho}\right) \right),$$

where $\rho = \sqrt{2mp(1-p)}$, and $\Phi(\cdot)$ is the Gauss error function.

(b) Uncertainty from the NOE peak positions. Let $b_i = (\omega(a_1), \omega(a_2), \omega(a_3))$ denote the back-computed NOE peak for an NOE (a_1, a_2, a_3) in a 3D NOESY spectrum. The likelihood for a back-computed peak $b_i = (\omega(a_1), \omega(a_2), \omega(a_3))$ in the NOE pattern B to match an experimental NOESY peak within the distance δ in Y_δ can be defined as

$$\mathcal{N}_i(b_i) = \prod_{j=1}^3 \mathcal{N}(|\omega(a_j) - p_j|, \sigma_j),$$

where (p_1, p_2, p_3) is the experimental NOESY peak matched to $(\omega(a_1), \omega(a_2), \omega(a_3))$ according to the Hausdorff distance measure, and $\mathcal{N}(|x - \mu|, \sigma)$ is the probability of observing the difference $|x - \mu|$ in a normal distribution with mean μ and standard deviation σ . Here we assume that the noise distribution of peak positions at each dimension is independent of each other. We note that the normal distribution and other similar distribution families have been widely and efficiently used to approximate the noise in the NMR data, e.g., see Refs. 29 and 22.

Then the expected number of peaks in $B \cap Y_\delta$ can be bounded by $|B \cap Y_\delta| = \sum_{i=1}^m \mathcal{N}_i(b_i)$. Thus, we have the following equation for the similarity score:

$$s = \frac{1}{m} \sum_{i=1}^m \mathcal{N}_i(b_i). \quad (2)$$

After considering both possibility from a false random match and uncertainty from the NOE peak positions, we obtain the following fitness score for a rotamer

$$s' = (1 - \theta)s = \frac{1 - \theta}{m} \sum_{i=1}^m \mathcal{N}_i(b_i). \quad (3)$$

For each rotamer, the computation of its similarity score s' can be computed in $O(mw)$ time, where m is the number of back-computed NOE peaks, and w is the total number of cross peaks in the experimental NOESY spectrum. The detailed pseudocodes for computing the similarity score and for HANA are provided in **SM Sections 3-4** available in Ref. 40.

5. ANALYSIS

5.1. Analysis of rotamer selection based on NOE patterns

Given a back-computed NOE peak $b_i = (\omega_{i1}, \omega_{i2}, \omega_{i3})$ in the NOE pattern of a rotamer, suppose that it finds a matched experimental peak in Y^δ with probability $g(\omega_{i1}, \omega_{i2}, \omega_{i3}, Y^\delta)$. Finding such a matched experimental NOESY peak for b_i can be regarded as a Poisson trial with success probability $g(\omega_{i1}, \omega_{i2}, \omega_{i3}, Y^\delta)$. We present the following result about the expected number of matched peaks for the back-computed NOE pattern of a rotamer.

Lemma 5.1. *Let X_i be an indicator random variable which is equal to 1 if the back-computed NOE peak b_i of a rotamer \mathbf{r} finds a matched experimental peak; 0 otherwise. Let $X = \sum_{i=1}^m X_i$, where m is the total number of back-computed NOE cross-peaks for the rotamer \mathbf{r} . Then the expected number of back-computed NOE peaks that find matched experimental peaks is given by*

$$E(X) = \sum_{i=1}^m E(X_i) = \sum_{i=1}^m g(\omega_{i1}, \omega_{i2}, \omega_{i3}, Y^\delta).$$

Let \mathbf{r}_t denote the rotamer closest to the real side-chain conformation for a residue, and let \mathbf{r}_f denote another rotamer in the library for the same residue. We call \mathbf{r}_t the *true* rotamer, and \mathbf{r}_f the *false* rotamer. Let X_i and Y_i be indicator random variables as defined in Lemma 5.1 for each back-computed NOE peak in the true rotamer \mathbf{r}_t and the false configuration \mathbf{r}_f respectively. Let m_t and m_f denote the numbers of back-computed NOE peaks for the true rotamer \mathbf{r}_t and the false rotamer \mathbf{r}_f . Let $X = \sum_{i=1}^{m_t} X_i$ and $Y = \sum_{i=1}^{m_f} Y_i$ denote the number of back-computed NOE peaks that find matched experimental peaks for rotamers \mathbf{r}_t and \mathbf{r}_f respectively. Let $\mu_t = E(X)$ and $\mu_f = E(Y)$ denote the expectations of X and Y . For simplicity of our theoretical analysis, we use Equation (1) to measure the fitness between the back-computed NOE pattern of a rotamer and the experimental spectrum in our theoretical model.

To measure the accuracy of the rotamer chosen based on our scoring function, we calculate the probability that the algorithm chooses the wrong rotamer \mathbf{r}_f rather than the true rotamer \mathbf{r}_t , and show how it is bounded by certain threshold. The following the-

orem formally states this result. The proof of this theorem can be found in **SM⁴⁰ Section 5**.

Theorem 5.1. *Suppose that $m_f\mu_t - m_t\mu_f \geq \max(m_f, \sqrt{m_fm_t}) \cdot 4\sqrt{\mu_t \ln m_t}$. Then with probability at least $1 - m_t^{-1}$, our algorithm chooses the true rotamer \mathbf{r}_t rather than the false rotamer \mathbf{r}_f .*

Theorem 5.1 indicates that if the difference between the expected numbers of matched NOE peaks for two rotamers is larger than certain threshold, we are able to distinguish these two rotamers based on the Hausdorff distance measure with certain probability bound. By Theorem 5.1, we have the following result on the bound of the probability of picking the correct rotamer from the library based on the Hausdorff distance measure, if we select top k rotamers with highest similarity scores.

Theorem 5.2. *Let t denote the maximum number of rotamers for a residue. Suppose that $m_f\mu_t - m_t\mu_f \geq 4 \max(m_f, \sqrt{m_fm_t}) \cdot \sqrt{\mu_t \ln m_t}$ and $m_t > t - k$ hold for the true rotamer \mathbf{r}_t and every false rotamer \mathbf{r}_f . Then with probability at least $1 - \frac{t-k}{m_t}$, our algorithm chooses the correct rotamer.*

Proof. Since the total number of rotamers in a residue is t , by Theorem 5.1 the probability that the similarity score of the true rotamer is larger than that of at least $t - k$ rotamers is at least $(1 - \frac{1}{m_t})^{t-k}$. According to the fact $(1 + x)^a \geq 1 + ax$ for $x > -1$ and $a \geq 1$, we have $(1 - \frac{1}{m_t})^{t-k} \geq 1 - \frac{t-k}{m_t}$. Thus, the probability for the algorithm to choose the right rotamer is at least $1 - \frac{t-k}{m_t}$. \square

Theorem 5.2 shows that if the discrepancy of the expected number of matched NOE peaks between the true rotamer and every other rotamer, and the number of back-computed NOE peaks are sufficiently large, the ensemble of top k rotamers with highest similarity scores will contain the true rotamer.

5.2. Time complexity analysis

The following theorem states that HANA runs in polynomial time.

Theorem 5.3. *HANA runs in $O(tn^3 + tn \log t)$ time, where t is the maximum number of rotamers at a residue and n is the total number of residues in the protein sequence.*

The detailed derivation of the time complexity can be found in **SM⁴⁰ Section 6**. We note that in practice, our NOE assignment algorithm HANA runs in 1-2 minutes on a 3 GHz single-processor Linux workstation.

6. RESULTS

HANA takes as input (a) protein sequence, (b) 3D NOESY-HSQC or 2D NOESY peak list, (c) assigned resonance list, (d) backbone computed by using the RDC-EXACT algorithm^{37, 39} (Section 4.2), and (e) Xtalview rotamer library.²⁵ HANA was tested on experimental NMR data for human ubiquitin,^{35, 9} zinc finger domain of the human DNA Y-polymerase Eta (pol η)² and human Set2-Rpb1 interacting domain (hSRI).²³ The high-resolution structures of these three proteins have been solved either by X-ray crystallography³⁵ or by traditional NMR approaches using both distance restraints from NOE data and orientational restraints from scalar and dipolar couplings.^{9, 2, 23} We used these solved structures, which are also in the Protein Data Bank (PDB), as the reference structures to compare and check the quality of NMR structures determined from our NOE assignment tables. The NMR data for hSRI and pol η were recorded using Varian 600 and 800 MHz spectrometers at Duke University. Ubiquitin NMR data was obtained from Ref. 15 and from the PDB (ID: 1D3Z).

6.1. Robustness of Hausdorff distance and NOE assignment accuracy

To check the robustness of the Hausdorff distance measure for NOE pattern matching, we first computed a low-resolution structure of ubiquitin by combining the backbone determined from RDC-EXACT,^{37, 36, 39} and rotamers selected based on the Hausdorff distance measure using patterns for backbone-sidechain NOEs. This low-resolution NMR structure is not the final structure, but is used to filter ambiguous NOE assignments (including backbone-backbone, backbone-sidechain and sidechain-sidechain NOE assignments). Our result shows that the low-resolution structure of ubiquitin obtained from our algorithm has a backbone RMSD 1.58 Å and an all-heavy-atom RMSD 2.85 Å from the corresponding X-ray structure (PDB ID: 1UBQ). Using this low-resolution structure, HANA was able to

Table 1. NOE assignment results for ubiquitin, pol η and hSRI.

Proteins	# of residues	# of NOESY peaks [§]	# of compatible assignments [†]	# of incompatible assignments [†]	Assignment accuracy
ubiquitin [*]	76	1580	901	93	90.6%
pol η ^{**}	39	1386	590	65	90.1%
hSRI ^{***}	112	5916	1429	119	92.3%

^{*} The ubiquitin backbone calculated from the RDC data using RDC-EXACT has RMSD 1.58 Å from the X-ray reference structure (PDB ID: 1UBQ) (residues 2-71).

^{**} The pol η backbone calculated from the RDC data using RDC-EXACT has RMSD 1.28 Å for the secondary structure regions and RMSD 2.71 Å for both secondary structure and loop regions (residues 8-36) from the NMR reference structure (PDB ID: 2I5O).

^{***} The hSRI backbone calculated from the RDC data using RDC-EXACT has RMSD 1.62 Å from the NMR reference structure (PDB ID: 2A7O) for the secondary structure regions (residues 15-34, 51-72, 82-97).

[§] The NOESY peak list contains diagonal and symmetric cross peaks.

[†] Redundant symmetric NOE restraints have been removed from the final NOE assignment table.

resolve the NOE assignment ambiguity caused from the chemical shift degeneracy, and prune a sufficient number of ambiguous NOE assignments, as we will discuss next.

To measure the assignment accuracy of HANA, we define a *compatible* NOE assignment as one in which the distance between the assigned pair of NOE protons in the reference structure is within NOE distance bound of 6.0 Å. Otherwise, we call it an *incompatible* NOE assignment. The number of compatible NOE assignments can be larger than the number of total NOESY peaks, since it is possible that multiple compatible NOEs can be assigned to a single NOESY cross peak. Next, the *assignment accuracy* is defined as the fraction of compatible assignments in the final assignment table output by HANA.

As summarized in Table 1, our NOE assignment algorithm achieved above 90% assignment accuracy for all three proteins. We note that the fraction of assigned peaks of hSRI is less than the other two proteins. This is because we only used backbones in the secondary structure regions (residues 15-34, 51-72, 82-97) for pruning ambiguous NOE assignments for hSRI. Presently we are developing new algorithms to solve long loops. We believe that with more accurate loop backbone structures, we will be able to improve the accuracy of our NOE assignment algorithm, while assigning more NOE peaks. We note that the ubiquitin ¹³C NOESY data from Ref. 15 are quite degenerate, thus we carefully picked a subset of NOESY peaks for assigning NOEs. Presently we are re-collecting a completely new set of ubiquitin NMR data including four-dimensional NOESY spectra for further testing of our algorithm.

Since the long-range NOEs, in which the spin-interacting protons are at least four residues away, play an important role in the structure determination, we also checked the fraction of incompatible long-range NOE assignments from our algorithm. We found that less than 3% of total assignments were from incompatible long-range NOEs in our computed assignments. As we will discuss next, such a small fraction of incompatible long-range NOE assignments can be easily resolved after one iteration of structure calculation.

6.2. Evaluation of structures from our NOE assignment tables

To test the quality of our NOE assignment results for structure determination, we fed the NOE assignment tables into the standard structure calculation program XPLOR.³ The input files for the structure calculation include protein sequence, NOE assignment table, and dihedral restraints. Compared with Refs. 2 and 23, in which RDCs are incorporated along with NOE restraints into the final structure calculation, here we only used RDCs to compute the initial backbone fold. From an algorithmic point of view, our structure determination using only NOEs can be considered as a good “control” test of the quality of our NOE assignment. The structure calculation was performed in two rounds. After the first round of structure calculation, the NOE violations larger than 0.5 Å among top 10 structures with lowest energies were removed from the NOE assignment table. Then the refined NOE table was fed into the XPLOR program for the second-round structure calculation.

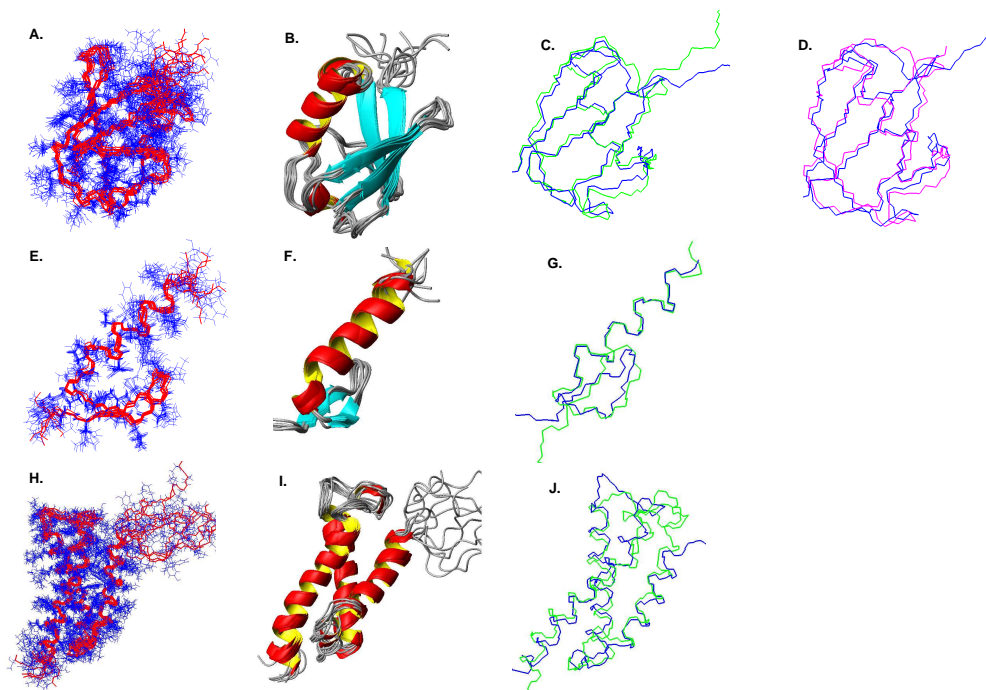


Fig. 2. The NMR structures of ubiquitin, pol η and hSRI computed from our automatically-assigned NOEs. Panels A, B, C and D in first row show the structures of ubiquitin, Panels E, F and G in the middle row show the structures of pol η , and Panels H, I and J in the bottom row show the structures of hSRI. Panels A, E and H show the ensemble of 10 best NMR structures with minimum energies. The backbones are shown in red while the side-chains are shown in blue. Panels B, F and I show the ribbon view of the ensemble of structures. Panel D shows the backbone overlay of the mean structures (in blue color) of ubiquitin with its X-ray reference structures³⁵ (in magenta color). The RMSD between the mean structure and the x-ray structure of ubiquitin is 1.23 Å for backbone atoms and 2.01 Å for all heavy atoms. Panels C, G and J show the backbone overlay of the mean structures (in blue color) with corresponding NMR reference structures (in green color) that have been deposited into the Protein Data Bank (PDB ID of ubiquitin⁹: 1D3Z; PDB ID of pol η ²: 2I5O; PDB ID of hSRI²³: 2A7O). The backbone RMSDs between the mean structures and the reference structures are 1.20 Å for ubiquitin, 1.38 Å for pol η , and 1.71 Å for hSRI. The all-heavy-atom RMSDs between the mean structures and the reference structures are 1.92 Å for ubiquitin, 2.39 Å for pol η , and 2.43 Å for hSRI.

Figure 2 illustrates final NMR structures of ubiquitin, pol η and hSRI calculated from XPLOR using our NOE restraint tables. For all three proteins, only a small number 18–60 (which is 1 – 4% of the total number of NOE assignments) of NOE violations larger than 0.5 Å occurred after the first round of structure calculation. All final structures converged to an ensemble of low-energy structures with small RMSDs from the reference structure solved either by the X-ray crystallography or by traditional NMR approaches. For all three test cases, the mean structure of final top 10 structures with lowest energies had a backbone RMSD less than 1.7 Å and an all-heavy-atom RMSD less than 2.5 Å from the refer-

ence structure. This implies that our NOE assignment algorithm has provided a sufficient number of accurate distance restraints for protein structure determination. In particular, we examined the structure quality in secondary structure and loop regions. We found that the secondary structure regions have better RMSD from the reference structure than the loop regions. After the final structure calculated by XPLOR using our NOE assignment table output by HANA, the RMSD of secondary structure regions in pol η is 0.81 Å for backbone atoms and 1.74 Å for all heavy atoms, and the RMSD of secondary structure regions in ubiquitin is 0.93 Å for backbone atoms and 1.59 Å for all heavy atoms. These results show that

the initial fold of secondary structure regions solved using the RDC-EXACT algorithm is accurate enough to combine with chosen rotamers from NOE patterns to resolve the NOE assignment ambiguities. In addition, we also found that the short loop regions of final structures can achieve about the same RMSD from the reference structure as the secondary structure regions. This indicates that the CCD algorithm with filtering of loops based on RDC fit can provide accurate short loops for our NOE assignment algorithm.

Our structure calculation protocol only requires one iteration, while other traditional NMR approaches in general take 7–10 iterations between NOE assignment and structure calculation. In addition, our NOE assignment algorithm only takes 1–2 minutes, versus hours to weeks for other methods. This efficiency is consistent with the proofs of correctness and time complexity of our algorithm. Therefore, the structure calculation framework based on our NOE assignment algorithm is more efficient than all other previous approaches in both theory and practice.

7. CONCLUSION

We have described a novel automated NOE assignment algorithm, HANA, that is combinatorially precise, and runs in polynomial time. To our knowledge, HANA is the first NOE assignment algorithm that simultaneously exploits the accurate algebraic geometry-based high-resolution backbone computation from RDC data,^{37, 39} the statistical diversity of rotamers from a rotamer library,²⁵ and the robust Hausdorff measure^{17, 19} for comparing the back-computed NOE patterns with the experimental NOE spectra and choosing accurate rotamers, to finally compute the NOE assignments with high accuracy. Owing to its simplicity, HANA runs extremely fast in practice. Furthermore, when applied to real biological NMR spectra for three proteins, our algorithm yields high assignment accuracy (> 90%) in each case suggesting its ability to play a role in high-throughput structure determination.

Although our current implementation of HANA uses 2D and 3D NOESY spectra, HANA is general and can be easily extended to use higher-dimensional (e.g., 4D) NOESY data.^{6, 7} In addition, it would be interesting to extend the current version of HANA for

NOE assignment with missing resonances. In general, acquisition of complete resonance assignment can require selective labeling of proteins, and is time-consuming. On the other hand, selection of correct rotamers can help the resonance assignment for side-chains. In principle, HANA can be extended to accommodate the NOE assignment with a partially assigned resonance list, as long as the back-computed NOE patterns with missing peaks are sufficient to identify accurate rotamers. Finally, it would be interesting to explore the use of side-chain rotamer packing algorithms¹¹ to choose rotamers that fit the data.

Acknowledgements

We thank Dr. L. Wang, Mr. A. Yan, Dr. S. Apaydin, Mr. J. Boyles, Prof. J. Richardson, Prof. D. Richardson, and all members of the Donald and Zhou Labs for helpful discussions and comments. We are grateful to Ms. M. Bomar for helping us with pol η NMR data.

References

1. C. Bailey-Kellogg, S. Chainraj, and G. Pandurangan. A random graph approach to nmr sequential assignment. *Journal of Computational Biology*, 12(6):569–583, 2005.
2. M. G. Bomar, M. Pai, S. Tzeng, S. Li, and P. Zhou. Structure of the ubiquitin-binding zinc finger domain of human DNA Y-polymerase η . *EMBO reports*, 8:247–251, 2007.
3. A. T. Brünger. X-PLOR, Version 3.1: a system for X-ray crystallography and NMR. *Journal of the American Chemical Society*, 1992.
4. A. A. Canutescu and R. L. Dunbrack Jr. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science*, 12:963–972, 2003.
5. G. M. Clore, J. G. Omichinski, K. Sakaguchi, N. Zambano, H. Sakamoto, E. Appella, and A. M. Gronenborn. Interhelical angles in the solution structure of the oligomerization domain of the tumour suppressor p53. *Science*, 267:1515–1516, 1995.
6. B. E. Coggins, R. A. Venters, and P. Zhou. Filtered Back-projection for the Reconstruction of a High-Resolution (4,2)D CH₃-NH NOESY Spectrum on a 29 kDa Protein. *J Am Chem Soc*, 127:11562–11563, 2005.
7. B. E. Coggins and P. Zhou. PR-CALC: A Program for the Reconstruction of NMR Spectra from Projections. *J Biomol NMR*, 34:179–95, 2006.
8. G. Cornilescu, F. Delaglio, and A. Bax. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *Journal of Biomolecular NMR*, 13:289–302, 1999.
9. G. Cornilescu, J. L. Marquardt, M. Ottiger, and A. Bax. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline

- Phase. *Journal of the American Chemical Society*, 120:6836–6837, 1998.
10. G. M. Crippen and T. F. Havel. Distance Geometry and Molecular Conformations. Wiley, New York., pages 635–642, 1988.
 11. I. Georgiev, R. H. Lilien, and B. R. Donald. The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. *Journal of Computational Chemistry*, [Epub ahead of print] PMID: 1829329, Feb 21, 2008.
 12. P. Güntert. Automated NMR Protein Structure Determination. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 43:105–125, 2003.
 13. P. Güntert. Automated NMR protein structure calculation with CYANA. *Meth. Mol. Biol.*, 278:353–378, 2004.
 14. T. Herrmann, P. Güntert, and K. Wüthrich. Protein NMR Structure Determination with Automated NOE Assignment Using the New Software CANDID and the Torsion Angle Dynamics Algorithm DYANA. *Journal of Molecular Biology*, 319(1):209–227, 2002.
 15. R. Harris. The ubiquitin NMR resource page, BBSRC Bloomsbury Center for Structural Biology, <http://www.biochem.ucl.ac.uk/bsm/nmr/ubq/>. Jun 2007.
 16. Y. J. Huang, R. Tejero, R. Powers, and G. T. Montelione. A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins: Structure Function and Bioinformatics*, 62(3):587–603, 2006.
 17. D. P. Huttenlocher and E. W. Jaquith. Computing visual correspondence: Incorporating the probability of a false match. In *Proceedings of the Fifth International Conference on Computer Vision (ICCV 95)*, pages 515–522, 1995.
 18. D. P. Huttenlocher and K. Kedem. *Distance Metrics for Comparing Shapes in the Plane*. In B. R. Donald and D. Kapur and J. Mundy, editors, Symbolic and Numerical Computation for Artificial Intelligence, pages 201–219, Academic press, 1992.
 19. D. P. Huttenlocher, G. A. Klanderman, and W. Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(9):850–863, 1993.
 20. H. Kamisetty, C. Bailey-Kellogg, and G. Pandurangan. An efficient randomized algorithm for contact-based nmr backbone resonance assignment. *Bioinformatics*, 22(2):172–180, 2006.
 21. J. Kuszewski, C. D. Schwieters, D. S. Garrett, R. A. Byrd, N. Tjandra, and G. M. Clore. Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments. *J. Am. Chem. Soc.*, 126(20):6258–6273, 2004.
 22. C. J. Langmead, A. K. Yan, R. H. Lilien, L. Wang, and B. R. Donald. A polynomial-time nuclear vector replacement algorithm for automated nmr resonance assignments. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, pages 176–187, 2003.
 23. M. Li, H. P. Phatnani, Z. Guan, H. Sage, A. L. Greenleaf, and P. Zhou. Solution structure of the Set2-Rpb1 interacting domain of human Set2 and its interaction with the hyperphosphorylated C-terminal domain of Rpb1. *Proceedings of the National Academy of Sciences*, 102:17636–17641, 2005.
 24. J. P. Linge, M. Habeck, W. Rieping, and M. Nilges. ARIA: Automated NOE assignment and NMR structure calculation. *Bioinformatics*, 19(2):315–316, 2003.
 25. S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. The Penultimate Rotamer Library. *Proteins: Structure Function and Genetics*, 40:389–408, 2000.
 26. G. T. Montelione and H. N. B. Moseley. Automated analysis of NMR assignments and structures for proteins. *Curr. Opin. Struct. Biol.*, 9:635–642, 1999.
 27. C. Mumenthaler, P. Güntert, W. Braun, and K. Wüthrich. Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *J. Biomol. NMR*, 10(4):351–362, 1997.
 28. M. Nilges, M. J. Macias, S. I. O’Donoghue, and H. Oschkinat. Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from β -spectrin. *Journal of Molecular Biology*, 269(3):408–422, 1997.
 29. W. Rieping, M. Habeck, and M. Nilges. Inferential Structure Determination. *Science*, 309:303–306, 2005.
 30. J. B. Saxe. Embeddability of weighted graphs in k -space is strongly NP-hard. *Proc. 17th Allerton Conf. Commun. Control Comput.*, pages 480–489, 1979.
 31. C. D. Schwieters, J. J. Kuszewski, N. Tjandra, and G. M. Clore. The Xplor-NIH NMR molecular structure determination package. *J Magn Reson*, 160:65–73, 2003.
 32. A. Shehu, C. Clementi, and L. E. Kavvaki. Modeling protein conformational ensembles: from missing loops to equilibrium fluctuations. *Proteins: Structure, Function, and Bioinformatics*, 65(1):164–79, 2006.
 33. N. Tjandra and A. Bax. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science*, 278:1111–1114, 1997.
 34. J. R. Tolman, J. M. Flanagan, M. A. Kennedy, and J. H. Prestegard. Nuclear magnetic dipole interactions in field-oriented proteins: Information for structure determination in solution. *Proc. Natl. Acad. Sci. USA*, 92:9279–9283, 1995.
 35. S. Vijay-Kumar, C. E. Bugg, and W. J. Cook. Structure of ubiquitin refined at 1.8 Å resolution. *Journal of Molecular Biology*, 194:531–44, 1987.
 36. L. Wang and B. R. Donald. Analysis of a Systematic Search-Based Algorithm for Determining Protein Backbone Structure from a Minimal Number of Residual Dipolar Couplings. In *Proceedings of The IEEE Computational Systems Bioinformatics Conference (CSB), Stanford CA (August, 2004)*, 2004.
 37. 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. *Jour. Biomolecular NMR*, 29(3):223–242, 2004.
 38. L. Wang and B. R. Donald. An Efficient and Accurate Algorithm for Assigning Nuclear Overhauser Effect Restraints Using a Rotamer Library Ensemble and Residual Dipolar Couplings. *The IEEE Computational Systems Bioinformatics Conference (CSB), Stanford CA. (August, 2005)*, pages 189–202, 2005.

39. L. Wang, R. Mettu, and B. R. Donald. A Polynomial-Time Algorithm for De Novo Protein Backbone Structure Determination from NMR Data. *Journal of Computational Biology*, 13(7):1276–1288, 2006.
40. J. Zeng, C. Tripathy, P. Zhou, and B. R. Donald. A Hausdorff-Based NOE Assignment Algorithm Using Protein Backbone Determined from Residual Dipolar Couplings and Rotamer Patterns – Supplementary Material. Department of Computer Science, Duke University, [online]. Available: <http://www.cs.duke.edu/donaldlab/Supplementary/csb08/>. May, 2008.