IEEE Computational Systems Bioinformatics (CSB) Conference. Stanford, CA 2005; pp. 189–202.

An Efficient and Accurate Algorithm for Assigning Nuclear Overhauser Effect Restraints Using a Rotamer Library Ensemble and Residual Dipolar Couplings

Lincong Wang [*] and Bruce Randall Donald [*,†,‡,§,¶]

Abstract

Nuclear Overhauser effect (NOE) distance restraints are the main experimental data from protein nuclear magnetic resonance (NMR) spectroscopy for computing a complete three dimensional solution structure including sidechain conformations. In general, NOE restraints must be assigned before they can be used in a structure determination program. NOE assignment is very time-consuming to do manually, challenging to fully automate, and has become a key bottleneck for high-throughput NMR structure determination. The difficulty in automated NOE assignment is ambiguity*: there can be tens of possible different assignments for an NOE peak based solely on its chemical shifts. Previous automated NOE assignment approaches rely on an ensemble of structures, computed from a subset of all the NOEs, to iteratively filter ambiguous assignments. These algorithms are heuristic in nature, provide no guarantees on solution quality or running time, and are slow in practice. In this paper we present an accurate, efficient NOE assignment algorithm. The algorithm first invokes the algorithm in [30, 29] to compute an accurate backbone structure using only two backbone residual dipolar couplings (RDCs) per residue. The algorithm then filters ambiguous NOE assignments by merging an ensemble of intra-residue vectors from a protein rotamer database, together with internuclear vectors from the computed backbone structure. The protein rotamer database was built from ultra-high resolution structures (<1.0 Å) in the Protein Data Bank (PDB). The algorithm has been successfully applied to assign more than 1,700 NOE distance restraints with better than 90% accuracy on the protein human ubiquitin using real experimentally-recorded NMR data. The algorithm assigns these NOE restraints in less than* one second *on a single-processor workstation.*

1 Introduction[1]

Among the biggest challenges of the post-genomic era are the determination of functions for gene sequences and the development of drugs using genomic information. For a protein-coding gene sequence, the three-dimensional structure of the expressed protein is useful not only for learning gene function, but also as the starting point for structure-based drug design. Although several approaches exist for *ab initio* prediction of protein backbone folds from a primary sequence [19], rarely do the predicted folds have better than 5.0 Å backbone RMSD from the experimentally-determined structure [33]. Further, *ab initio* prediction of an accurate complete structure with all the sidechain conformations, starting with a primary sequence, is still very challenging at present. Accurate, high-resolution complete structure can only be obtained through experimental techniques, mainly, x-ray diffraction and nuclear magnetic resonance (NMR). However, structure determination by either technique is, in general, very time-consuming. For x-ray, the difficulty is to grow a good quality crystal while for NMR, the bottleneck is NOE assignment and to a lesser extent, resonance assignment. To compute a well-defined structure, months of time may be required to manually assign enough NOE distance restraints. Thus, the automa-

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

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

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

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

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

[1] Abbreviations used: NMR, nuclear magnetic resonance; RDC, residual dipolar coupling; NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser effect spectroscopy; 2D, two-dimensional; 3D, three-dimensional; 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); SA, simulated annealing; MC, Monte-Carlo; CH, the vector between C_α and H_α, H-D exchange experiment; hydrogen-deuterium exchange experiment; FID, free-induction decay.

tion of NOE assignment is critical to high-throughput NMR structure determination. However, previous automatic NOE assignment algorithms rely on heuristic methods, require very high quality input data, and consume many hours to weeks of computation time. In this paper, we present an efficient NOE assignment algorithm that is capable of correctly assigning more than 90% of all the NOE cross-peaks in less than *one second*. One novel feature of our algorithm is the use of an ensemble of intra-residue vectors between the backbone atoms and sidechain protons to reduce the ambiguity in the assignment of NOE restraints involving sidechain protons, especially aliphatic protons. The ensemble of intra-residue vectors was mined from a protein rotamer library database built from ultra-high resolution structures (< 1.0 Å) in the Protein Data Bank (PDB). The algorithm has been demonstrated successfully using an NOE cross-peak list picked from two different, real NOESY spectra recorded on the protein human ubiquitin. Together with our efficient algorithm for computing backbone structure using only 2 residual dipolar couplings (RDCs) per residue [30, 29], we have developed an efficient protocol for computing a complete NMR structure with all the sidechain conformations starting from *resonance assignment*. Since each RDC data set can be acquired in less than one hour, and RDC data can be assigned immediately given chemical shift resonance assignment, our protocol should be important for structural genomics. Our main contributions in this paper are:

1. Rotamer ensemble-based techniques for assigning NOEs involving sidechain protons.
2. The first *provably-efficient* algorithm for assigning both backbone and sidechain NOEs.
3. An implementation of our algorithm that is much faster than existing NOE assignment algorithms.
4. Successful application of the algorithm to real biological protein NMR spectra.

1.1 Organization of the paper

We begin, in section 2 with a statement of the NOE assignment problem. Section 3 describes existing heuristic approaches for automatic NOE assignment. Section 4 describes our efficient algorithm. Section 5 presents the results of applying our algorithm to assign real, experimental NOE cross-peaks picked from a 3-dimensional (3D) ^{15}N-edited NOESY and a 3D ^{13}C-edited NOESY spectrum of the protein human ubiquitin. We also present NMR solution structures of ubiquitin computed from our automatically-assigned NOE restraints, and discuss the significance of our algorithm for structural genomics. Section 6 analyzes the complexity of the algorithm and describes its performance in practice.

2 NOE assignment problem

NOE distance restraints are the main input for NMR structure determination, especially for the determination of a complete structure with all the sidechain conformations. An NOE restraint is computed from the intensity (or volume) of an NOE peak in an NMR spectrum. The dipole-dipole interaction between two nuclei, $\mathbf{a}$ and $\mathbf{b}$ (e.g. two protons), gives rise to the NOE peak. Appendix A contains a more detailed description of NOE experiments. In order to use these restraints as input, the identity of the two interacting nuclei must be assigned: this is the *NOE assignment problem*. In practice, an NMR structural biologist first assigns the chemical shift resonances of NMR observable nuclei (*resonance assignment*), then assigns NOE peaks. Resonance assignment first groups together a set of peaks (whose frequency coordinates are *labeled by* chemical shifts) belonging to the same amino acid from one or more NMR spectra, then maps that set to a residue in the protein sequence. Several automatic resonance assignment algorithms [1, 20, 34, 2, 22, 21, 32] are available at present. In contrast, in *NOE assignment*, one must map the two interacting nuclei, $\mathbf{a}$ and $\mathbf{b}$, to two or more specific nuclei in the sequence. More than one pair of interacting nuclei may produce a single NOE peak (this is called *peak overlap*) if their chemical shift difference is too small to be resolved by NMR experiments. Here, for ease of exposition, each peak is labeled by three chemical shifts, c_n, c_a and c_b, corresponding to 3D NMR experiments where c_n, c_a and c_b are, respectively, the chemical shifts of an ^{15}N or ^{13}C nucleus, the proton attached to the ^{15}N or ^{13}C nucleus, and the second proton interacting through-space with the attached proton. The difficulty of NOE assignment is mainly due to *chemical shift degeneracy*, that is, except for rare cases, there will be a set, $\mathbf{S}_a$ for proton $\mathbf{a}$, whose chemical shifts fall into the interval $c_a \pm \delta_c$, where δ_c is the error in chemical shift. The same is true for proton $\mathbf{b}$. Heteronuclear multi-dimensional NMR experiments [5] such as 3D ^{15}N-edited or ^{13}C-edited NOESY are designed to keep the size of the set $\mathbf{S}_a$ small, usually one for most peaks and a few for other remaining ones. However, the size of $\mathbf{S}_b$ can be quite large, up to a few dozens. In general, the size of $\mathbf{S}_a$ and $\mathbf{S}_b$ for aliphatic protons is much larger than the size of these sets for backbone protons. Furthermore, the size of both sets increases rapidly with the size of the protein. Without resonance assignment, the identities of the elements in both sets, $\mathbf{S}_a$ and $\mathbf{S}_b$, are unknown, and NOE assignment must select (assign) the correct elements in both sets. With resonance assignment, the identities of $\mathbf{S}_a$'s elements are given, and only the correct element or elements in $\mathbf{S}_b$ need to be assigned. In summary, due to peak overlap and chemical shift degeneracy, for most NOE peaks, the chemical shifts alone, do not provide enough information for a unique assignment. For example, for a medium-sized protein with about 150 residues, less than 10% of all the experimentally-observed NOE peaks can typically be assigned based on chemical shift alone. The percentage decreases rapidly with the size of the protein but increases with the dimensionality of NMR experiments. Increasing the dimensionality of an NMR experiment will reduce the size of both $\mathbf{S}_a$ and $\mathbf{S}_b$. However, the addition of even one more dimension to an NMR experiment will at least double the spectrometer time.

3 Previous work

As stated in section 2, information in addition to chemical shifts is, generally, required for assigning most NOE peaks, especially the peaks involving aliphatic protons, even with known resonance assignments. The resonance assignment programs, Jigsaw [1] and NVR [20], are also capable of assigning the backbone NOEs. Except for Jigsaw and NVR, all the previous formulations of the problem cast the NOE assignment problem as a structure-determination problem on-the-fly: they all start with an initial set of assigned NOE restraints (a fraction of the total number of NOE restraints) to compute an ensemble of structures, and then use the newly-computed ensemble to make more assignments. This assignment-structure-assignment process (cycle) is repeated many times until certain criteria are satisfied, for example, a target function has achieved a minimum. These previous approaches rely on heuristic methods such as simulated annealing (SA) and Monte-Carlo (MC) search, and typically need many hours to weeks of computer time. These approaches differ in whether resonance assignment is given explicitly and how the initial set of NOE restraints is generated, as described below.

Two approaches [10, 24] have been developed for NOE assignment and structure determination without performing an explicit resonance assignment step. CLOUDS uses a relaxation matrix analysis of NOE peaks to compute distance restraints from NOE peak volumes. An NOE restraint computed by a relaxation matrix method is supposed to be more accurate than that computed using the two-isolated spin assumption (Appendix A). CLOUDS is bootstrapped with three groups of protons (methyl, amide and other protons), distinguished from one another by line-shape analysis and H-D exchange experiments. CLOUDS requires that both $\mathbf{S}_a$ and $\mathbf{S}_b$ are singletons. Further, a large number of anti-distance constraints (ADCs) between backbone H_N-H_N and H_N-H_α proton pairs must be added. An ADC specifies that the two nuclei *must* be more than 5.0 Å apart in the structure, and is identified if the spectrum does not yield any detectable NOESY peak. The above requirements are unrealistic for NMR spectra recorded on a vast large majority of proteins. CLOUDS relies on molecular dynamics (MD) and SA to compute structures. Another approach [24], built upon a structure prediction program, ROSETTA, has been developed by Meiler and Baker. In this approach, the structure most consistent with the unassigned NOE restraints is selected from an ensemble of structures predicted by ROSETTA through MC search. For peaks from real NMR spectra, it is often quite difficult to distinguish real weak peaks from the noise peaks. CLOUDS does not handle the noise peaks.

Robust NOE assignment approaches [25, 15, 16, 18] using real NOE data require rather complete resonance assignment. Their performance depends critically on the accuracy of the backbone fold from the first cycle of structure computation, which in turn relies on the number and quality of NOE restraints in the initial set. ARIA [25] bootstraps the assignment-structure-assignment cycle with a fair amount of manually-assigned NOE peaks ($>$ 5 NOE restraints per residue), where the structures are computed by the program XPLOR/CNS [4] using MD/SA techniques. From these NOEs a fairly accurate backbone fold ($<$3.0 Å backbone RMSD to the final structure) could be computed. Thus, ARIA has been used, principally, in combination with manual NOE assignment to speed up structure determination once a correct initial fold has been computed.

In CANDID [15], an initial set of NOE restraints are first identified by chemical shifts, then the restraints are ranked by network-anchoring and constraint-combination techniques. Network-anchoring makes sure that individual NOE peaks in the set of assigned NOEs are weighted by the extent to which they can be embedded into the network formed by all other NOE assignments. Constraint-combination can be viewed as a noise-reduction technique to limit the deleterious impact of noise peaks on the accuracy of the computed structures, particularly the structures computed from the initial set of NOE restraints. The structures are computed by the program DYANA [11] using a MD/SA technique. Compared with ARIA, CANDID needs much less human intervention. However, it is still not robust for assigning real NOE spectra since it requires $> 90\%$ complete resonance assignments (corresponding to 87% complete side chain resonances), almost complete aromatic sidechain assignments, a low percentage of noise peaks, and small chemical shift variations, that is, small errors in chemical shifts with a $\delta_c \leq 0.02$ ppm for 2D and ≤ 0.03 ppm for 3D NOESY spectra. Note a smaller δ_c, in general, greatly reduces the size of the set $\mathbf{S}_b$ and thus increases the size of initial NOE peak set. However, it can be physically impractical to reduce δ_c, and in general, more time is required to record a spectrum with a smaller δ_c.

In AUTO-STRUCTURE [16], in addition to the initial set of NOE restraints assigned from chemical shift information alone, the input to the first cycle of structure computation requires identified secondary structures and restraints for backbone dihedral angles generated from chemical shifts by the program TALOS [6]. Secondary structures (including alignments between β-strands) are identified by a combined pattern analysis of secondary-structure-specific NOE contacts, chemical shifts, scalar coupling constants, and slow H-D exchange experiments. Aided by these additional data, AUTO-STRUCTURE has less strict requirements on the quality of the NOE peak list than CANDID. For example, it requires the resonance assignment to be $> 85\%$ complete, compared to $> 90\%$ as demanded by CANDID. AUTO-STRUCTURE has been interfaced with MD/SA based-structure determination programs DYANA and XPLOR/CNS.

Recently, Clore and coworkers presented a fault-tolerant structure determination program, PASD [18], which is capable of computing a correct structure even with up to 80% incorrect *long-range* NOE restraints in the initial NOE peak list generated from an automatic peak-picking program. PASD requires extensive NOE spectra including both 3D and 4D NOESY spectra, almost complete resonance assignments, a database-based pseudo-energy term, and rather tight restraints for backbone dihedral angles generated from the program TALOS. How tolerant PASD is to the incompleteness of resonance assignment is unknown. Higher-dimensional spectra reduce the sizes of both $\mathbf{S}_a$ and $\mathbf{S}_b$ but at least double the spectrometer time. All the above iterative approaches are computationally intensive. In general, many iterative cycles (at least 20, but up to 70 cycles) are required since each cycle is only able to assign a small percentage of the NOE peaks. Furthermore, an ensemble of structures, rather than a single structure, must be computed by these MD/SA methods to sample enough of the search space. These approaches are routinely run on large clusters in order to assign enough NOE restraints to compute a well-defined structure. Among them, PASD is the most computationally expensive approach: it is infeasible to run PASD on a single-processor workstation. As stated above, what makes these NOE assignment approaches expensive is their reliance on using MD/SA for structure determination in a tight inner cycle. The requirement of many cycles of structure computation makes these approaches rather inefficient in practice.

4 An efficient algorithm for NOE assignment

Our algorithm begins with known resonance assignments and an accurate backbone structure computed using only 2 RDCs per residue. The efficient and accurate algorithm for computing protein backbone structure using only 2 RDCs per residue has been described previously [30, 29]. We employ this structure determination algorithm as a first stage, namely, given the RDC data, the algorithm in [30, 29] is called as a subroutine to compute an accurate backbone structure. This backbone structure, in turn, is used to bootstrap the automated assignment of NOEs: the assignment proceeds by *filtering* the experimentally-measured NOEs based on consistency with the backbone structure. In order to fully understand our automatic assignment algorithm, these details are given in Appendix C. We note that in principle, another structure determination could have been used in the place of [30]; however, it is the only prior algorithm that can compute a complete protein backbone structure *de novo* using only two RDCs per residue.

4.1 Assignment of backbone NOEs and computation of NOE restraints from peak intensity

The algorithm begins by performing the assignment of backbone H_{N}-H_{N} and H_{N}-H_{α} NOEs. As defined in section 2, there are two sets, $\mathbf{S}_a$ and $\mathbf{S}_b$, associated with each NOE peak. In heteronuclear, multidimensional NMR, the set $\mathbf{S}_a$ is usually small. In fact, all but a few peaks in the ubiquitin heteronuclear single quantum correlation (HSQC) spectrum are well-resolved. The set $\mathbf{S}_a$ for a peak picked from the 3D ^{15}N-edited NOESY spectrum is a singleton if the corresponding HSQC peak is well-resolved. However, even for ubiquitin NOEs involving only backbone protons (*backbone NOEs*), the set $\mathbf{S}_b$ may have up to 8 elements. For a majority of $\mathbf{S}_b$'s involving backbone protons, the correct element can be assigned by simply checking whether such an NOE peak could be detected by comparing with the calculated inter-proton distance, d_b, computed from the backbone structure. The assignment of *backbone NOEs* proceeds as follows:

1. Sort the list of all protons (^{1}H) chemical shifts.
2. For every backbone NOE peak:
   (a) Search through the sorted chemical shift list to select all the protons that have chemical shifts in the interval $c_b \pm \delta_c$, where δ_c is the error in ^{1}H chemical shift, and insert them into the set $\mathbf{S}_b$.
   (b) For each element in $\mathbf{S}_b$, compute the expected inter-proton distance, d_b, using the backbone structure. If the distance $d_b > 6$ Å, then the element is deleted.
   (c) If only one element remains in $\mathbf{S}_b$, it is selected as the correct assignment.

The resulting assigned unambiguous backbone NOE peaks become the initial assigned NOE peak list S_I. Next, the algorithm best fits the intensities (I) of the assigned peaks, S_{I} to the corresponding distances, d_b, computed from the backbone structure using the following function:

$$I = a_1 + \frac{a_2}{d_b^p} \qquad (1)$$

where a_1, a_2 and p are the best-fit parameters (Fig. 1A). Eq. (1) is subsequently used to compute the NOE distance d_n from the intensity of unassigned peaks the algorithm will process next. Please note that the empirical parameter p can differ from the ideal value of 6 (see Eq. (5) of Appendix A) due to protein dynamics in solution. Appendix A contains a more detailed description of NOE experiments.

4.2 Intra-residue vectors between the C$_\beta$ nucleus and sidechain protons

The protein backbone structure computed using RDCs [30] has coordinates for only the backbone atoms: N, H$_N$, C$_\alpha$, C$_\beta$, C′, O and H$_\alpha$; no coordinates for sidechain protons are present, except for glycine where, instead of C$_\beta$ we have proton H$_{\alpha 2}$ and instead of H$_\alpha$ we have proton H$_{\alpha 1}$. To compute the possible positions of sidechain protons other than H$_\beta$, given the position for the atom C$_\beta$, we mine a small protein database consisting of 23 ultra-high resolution (<1.0 Å) X-ray structures containing coordinates for protons (see Table 1). The idea is similar to mining the PDB for a rotamer library. As is well known, the sidechain conformation can be specified by dihedral angles χ_1, χ_2, $\ldots$, depending on the amino acid type. According to physical chemistry, the *staggered* conformation for a σ-bond between two carbons has the lowest conformation energy, while the *eclipsed* has the highest and the *gauche* has an intermediate energy. Thus, each sidechain dihedral angle has preferred values. Such preferences have been confirmed experimentally and constitute the physical basis for the protein sidechain rotamer library. We apply the same insight to mine *vectors* between backbone nuclei and sidechain protons from the PDB. We call these vectors mined from the PDB *rotamer ensemble-based intra-residue vectors*. The sidechain protons here mean the protons other than H$_N$, H$_\alpha$, and H$_\beta$. Given a backbone structure and by exploiting the sidechain kinematics we can prove the following:

Proposition 1 *Given a backbone structure and the sidechain dihedral angles* χ_1, χ_2, χ_3 *and* χ_4, *let* H_{γ_i}, H_{δ_i} *and* H_{ϵ_i} *be particular* H_γ, H_δ, H_ϵ *protons (i = 1, 2, 3, depending on the amino acid). Then the three vectors from the atom* C_β *to the three respective sidechain protons,* H_{γ_i}, H_{δ_i} *and* H_{ϵ_i}, *can be computed analytically by*

$$\begin{aligned}\mathbf{v}_{\beta\gamma_i} &= \mathbf{p}_1(\chi_1, \chi_2)\\ \mathbf{v}_{\beta\delta_i} &= \mathbf{p}_2(\chi_1, \chi_2, \chi_3)\\ \mathbf{v}_{\beta\epsilon_i} &= \mathbf{p}_3(\chi_1, \chi_2, \chi_3, \chi_4),\end{aligned} \tag{2}$$

where $\mathbf{p}_1$, $\mathbf{p}_2$ *and* $\mathbf{p}_3$ *are vector-valued trigonometric polynomials.*

Below, the lengths of the three vectors will be denoted, respectively, by $d_{\beta\gamma}$, $d_{\beta\delta}$ and $d_{\beta\epsilon}$. From Proposition 1 we can easily prove the following (Fig. 2):

Corollary 1 *Given the vector from the atom* C_β *to a sidechain proton and the vector* $\mathbf{v}_{N\beta}$ *from the backbone atom H_N to the atom* C_β*, or the vector* $\mathbf{v}_{\alpha\beta}$ *from the atom* H_α *to the atom* C_β*, the vector* $\mathbf{v}_{NS}$ *from the atom H_N to a sidechain proton (e.g.* H_δ*), or the vector* $\mathbf{v}_{\alpha S}$ *from the atom* H_α *to a sidechain proton can be computed by*

$$\mathbf{v}_{NS} = \mathbf{v}_{N\beta} + \mathbf{v}_{\beta\delta}, \qquad \mathbf{v}_{\alpha S} = \mathbf{v}_{\alpha\beta} + \mathbf{v}_{\beta\delta} \tag{3}$$

The length of these vectors, $r_N = |\mathbf{v}_{NS}|$ *and* $r_\alpha = |\mathbf{v}_{\alpha S}|$, *can be computed by the cosine law where* r_N^2 *and* r_α^2 *are, respectively, trigonometric polynomial functions of the* $\chi_1, \chi_2, \ldots$ *angles.*

In our algorithm the above two vectors, $\mathbf{v}_{N\beta}$, and $\mathbf{v}_{\alpha\beta}$, are computed directly from the backbone structure determined by the algorithm in [30, 29]. Please note that it is not required that the two vectors, $\mathbf{v}_{N\beta}$ and $\mathbf{v}_{\alpha\beta}$, are between intra-residue atoms. The proofs of Proposition 1 and Corollary 1 are provided in Appendix B. According to Eq. (2), the preference for sidechain dihedral angles determines the preference for the vectors, $\mathbf{v}_{\beta\gamma}$, $\mathbf{v}_{\beta\delta}$ and $\mathbf{v}_{\beta\epsilon}$. Thus, we mined the PDB to determine their distributions (Table 1). For example, Fig. 1B and 1C show, respectively, the distribution for the distance $d_{\beta\delta}$ between C$_\beta$ and H$_{\delta 1}$ or H$_{\delta 2}$ of a methylene group, and the distribution for the distance $d_{\beta\epsilon}$ between C$_\beta$ and H$_{\epsilon 1}$ or H$_{\epsilon 2}$ of a methylene group. We have two maxima for $d_{\beta\delta}$, and four major maxima for $d_{\beta\epsilon}$s (Fig. 1B, 1C and Table 1). For each individual maximum the distribution is rather tight and can be fit with a normal distribution. The set of corresponding χ angles that give these maxima can be found by searching through the discrete values for the relevant χ angles as presented in a typical rotamer library [23, 8]. For example, the set of rotamer angles for the sidechain of Arg residue has 31 elements where each element is 4-tuple, $(\chi_1, \chi_2, \chi_3, \chi_4)$. Among the 31 elements, 23 4-tuples have the intra-residue distance $d_{\beta\delta} = 2.70$ (Table 1), while the other 8 4-tuples have $d_{\beta\delta} = 3.35$. The set of χ angles found above can then be used to compute the set of the vectors from the atom $\mathbf{C}_\beta$ to a sidechain proton by Eq. (2). We denote such a finite set of vectors, $\mathbf{v}_{NS}$ or $\mathbf{v}_{\alpha S}$, by $\mathbf{S}_{sc}$ and denote the minimum and maximum length of the vectors in the set $\mathbf{S}_{sc}$ by, respectively, $r_{N,MIN}$ and $r_{N,MAX}$, and $r_{\alpha,MIN}$ and $r_{\alpha,MAX}$.

4.3 The triangle relationship for filtering incorrect assignments for sidechain NOEs

As shown in Corollary 1, the three vectors $\mathbf{v}_{NS}, \mathbf{v}_{N\beta}$ and $\mathbf{v}_{\beta\delta}$ form three sides of a triangle, where $\mathbf{v}_{N\beta}$ is computed from the backbone structure and $\mathbf{v}_{\beta\delta}$ is mined from the PDB rotamers. There are a finite set of vectors $\mathbf{v}_{\beta\delta}$ for each sidechain proton. Consequently, there are a finite set of vectors $\mathbf{v}_{NS}$. These triangle relationships (Eq. 3) are used in our algorithm to filter the assignment of NOE peaks involving sidechain protons. The algorithm first computes

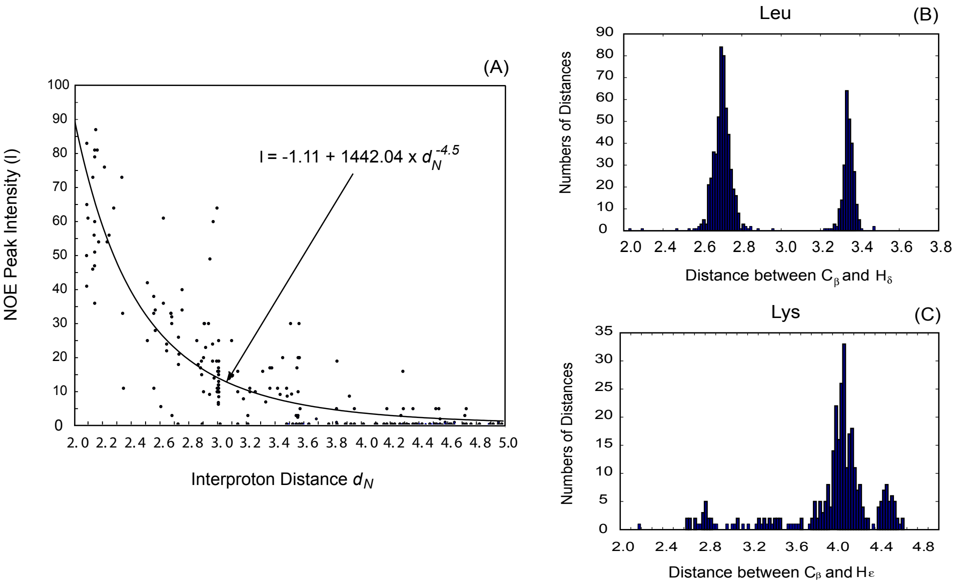

Figure 1. **(A) NOE calibration curve**. The x-axis is the inter-proton distance computed from the backbone structure. The y-axis is the backbone experimental H_N-H_N and H_N-H_α NOE peak intensity, picked from the 3D ^{15}N-edited NOESY spectrum. 155 NOE restraints were used in the fitting. **(B) and (C), distributions of distance between C_β nucleus and sidechain protons**. The x-axis is the internuclear distance between C_β and H_δ of leucine (B), or between C_β and H_ϵ of lysine (C).

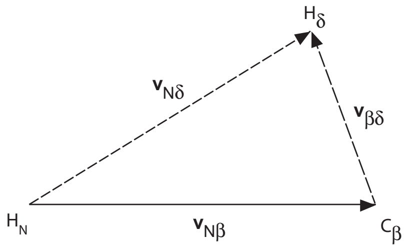

Figure 2. **The triangle relationship**. The vector $\mathbf{v}_{N\beta}$ is computed from the backbone structure, $\mathbf{v}_{\beta\delta}$ is a rotamer ensemble-based intra-residue vector mined from the PDB, and the computed length $d_{N\delta}$ of $\mathbf{v}_{N\delta}$ is compared with the NOE distance computed from Eq. (1) to filter ambiguous NOE assignments.

Amino Acid	H_γ (mean, variance)	H_δ (mean, variance)$^{(a)}$	H_ϵ (mean, variance)$^{(b)}$
Arg	2.06, 0.03	2.70, 0.10; 3.35, 0.07	
Ile	2.06, 0.03	2.75, 0.05; 3.37, 0.04	
Leu	2.06, 0.02	2.69, 0.05; 3.34, 0.03	
Lys	2.05, 0.02	2.70, 0.11; 3.32, 0.06	4.09, 0.11; 2.75, 0.08; 3.32, 0.06; 4.46, 0.06
Phe,Tyr, His		2.67, 0.05	4.55, 0.05

Table 1. **Intra-residue distance between the C_β nuclei and the sidechain protons.** These distances are extracted from 23 ultra-high resolution ($\leq$1.0Å) X-ray structures with proton coordinates. Their PDB IDs are 3AL1, 1BXO, 1CEX, 1C75, 1GDQ, 1G66, 1GDN, 1GCI, 2FDN, 1HJ9, 1IXH, 1GQV, 1IC6, 2ERL, 1HJ8, 1JFB, 1EJG, 1RB9, 3PYP, 1FY5, 1GVK, 1KQP and 1LQT. (a) there are 2 maxima in the distributions. (b) there are 2 or 4 maxima in the distributions. All the units are in Å.

the NOE distances, d_{N} and d_{α}, using Eq. (1) from, respectively, the experimental NOE peaks picked from the 3D ^{15}N NOESY spectrum and 3D ^{13}C NOESY spectrum. The algorithm tests d_{N} and d_{α} to see whether they satisfy the constraints:

$$\begin{aligned} r_{N,MIN} - N_r \leq d_{N} \leq r_{N,MAX} + N_r \\ r_{\alpha,MIN} - \alpha_r \leq d_{\alpha} \leq r_{\alpha,MAX} + \alpha_r \end{aligned} \tag{4}$$

where N_r and α_r are, respectively, the NOE distance error bounds for 3D ^{15}N-edited NOESY peaks and 3D ^{13}C-edited NOESY peaks. In addition, 3.0 Å is added to the NOE distances computed from the peaks of methyl protons, and 1.0 Å is added to the NOE distances computed from the peaks of methylene protons without stereo-specific assignment. The assignment of *sidechain NOEs* proceeds as follows:

1. Sort the list of all ^{1}H chemical shifts.
2. For every NOE peak:
   (a) Search through the sorted chemical shift list to select all the protons (^{1}H) that have chemical shifts in the interval $c_b \pm \delta_c$, where δ_c is the error in ^{1}H chemical shift and insert them into the set $\mathbf{S}_b$. Set $\mathbf{S}_0 \leftarrow \mathbf{S}_b$.
   (b) For every element in $\mathbf{S}_b$, if the computed d_{N} or d_{α} violates Eq. (4) then it is deleted from $\mathbf{S}_b$. The remaining element or elements are selected as the assignment for the NOE peak.
   (c) If $\mathbf{S}_b$ becomes empty after deleting all the incorrect elements,
      i. Then, for every element of the original $\mathbf{S}_0$ compute the difference d, where $d = \min(|d_{N} - r_{N,MIN} + N_r|, |r_{N,MAX} + N_r - d_{N}|)$ or $d = \min(|d_{\alpha} - r_{\alpha,MIN} + \alpha_r|, |r_{\alpha,MAX} + \alpha_r - d_{\alpha}|)$. If $d \leq 2$ Å, insert the element into a new set $\mathbf{T}_b$.
      ii. If $\mathbf{T}_b$ is not empty, select the element with the minimum d as the NOE assignment.
      iii. If $\mathbf{T}_b$ is empty, discard the NOE peak and label it as a noise peak.

5 Results and Discussion

To test our algorithm on real NMR data one must in general record or obtain spectra from an ^{15}N- or ^{13}C-labeled protein sample. We were able to process a suite of real protein NMR spectra including two distinct NOESY spectra for our tests. We now describe how our algorithm performed on these data. We first describe the completely automated assignment, by our algorithm, of the 3D ^{15}N-edited NOESY and ^{13}C-edited NOESY spectra of ubiquitin, followed by a comparison of the complete structure we computed using our automatically-assigned NOE peak lists vs. the X-ray structure [28]. Appendix C contains details of the algorithm for the computation of loops and turns using NH and CH RDCs in a single medium, and Appendix D describes the processing of NMR data starting from raw free-induction decay (FID) data.

5.1 NOE restraint assignment

First, a peak-picking program in the package NMRVIEW [17] was used to automatically pick all the NOE peaks of both the 3D ^{15}N-edited and ^{13}C-edited NOESY spectra after the spectra were properly referenced to the 2D ^{1}H,^{15} N HSQC and 3D HCCH-TOCSY spectra. Next, weak but well-resolved peaks were added manually, which accounts for about $\frac{1}{4}$ of the total peaks. Finally, duplicate peaks were merged. A peak is defined as a *duplicate* of another peak if $\delta_{F1} \leq 0.01$ ppm, $\delta_{F2} \leq 0.03$ ppm and $\delta_{F3} \leq 0.05$ ppm, where $\mathbf{F}1$ corresponds to the direct-detected dimension (amide protons for ^{15}N NOESY or aliphatic protons for ^{13}C NOESY), $\mathbf{F}2$ to the indirect ^{1}H, and $\mathbf{F}3$ to the ^{15}N or ^{13}C dimension. For backbone H$_{N}$-H$_{N}$ peaks, the peak with stronger intensity is selected from the two symmetric peaks. The error δ_{C} for ^{1}H chemical shift in the indirectly-detected dimension ($\mathbf{F}2$) of both the ^{15}N and ^{13}C NOESY spectra is set to be 0.04 ppm. The result of the automated assignment is summarized in Table 2. Out of the 1153 NOE peaks picked from the 3D ^{15}N-edited NOESY spectrum, 1083 originate from backbone amide protons, the remaining 50 are from the sidechain amide protons. Only 420 NOE peaks originating from the backbone H$_{\alpha}$ can be picked from the ^{13}C-edited NOESY spectrum. We were able to assign 1053 NOE peaks from the 1083 peaks picked from ^{15}N-edited NOESY spectrum and 393 peaks from the 420 peaks picked from ^{13}C-edited NOESY spectrum. We further divide the assigned NOE distance restraints into two classes: sequential NOEs and medium/long-range NOEs (Table 2). The number of assigned NOE distance restraints is larger than the number of assigned peaks since it is possible that more than one NOE restraint could be assigned to a single peak. It takes less than *one second* on a 2.4 GHz single-processor Linux workstation for our algorithm to compute the assignments.

5.2 Structure computation using automatically-assigned NOE restraints

To test the quality of the 1783 NOE restraints automatically generated and assigned by our algorithm (Table 2), we input these assigned distance restraints for ubiquitin to the structure determination program XPLOR/CNS [4], together with 12 hydrogen bonds. No RDC restraints were used. The ranges for NOE restraints were set rather loosely: 1.80 Å–3.50 Å and 1.80 Å–6.00 Å. The NMR structures were computed using a hybrid distance-geometry and SA protocol [4]. After the first round of computation with XPLOR, there were 163 restraints that had NOE violations larger than 0.50 Å in 50 structures out of a total of 70 computed structures. However, none of the NOE violations was larger than 2.50 Å. After all these 163 restraints were deleted from

Spectrum	# of Peaks	# of Assigned NOEs	# of Sequential NOEs$^{(a)}$	# of Medium$^{(b)}$, long-range NOEs$^{(c)}$
^{15}N-NOESY	1083	1288	822	466
^{13}C-NOESY	420	495	228	167

Table 2. The automatically assigned NOE restraints. (a) NOEs between residue i and $i+1$, (b) NOEs between residue i and $i+j$ where $1 < j < 4$, and (c) NOEs between residue i and $i+j$ where $j \geq 4$.

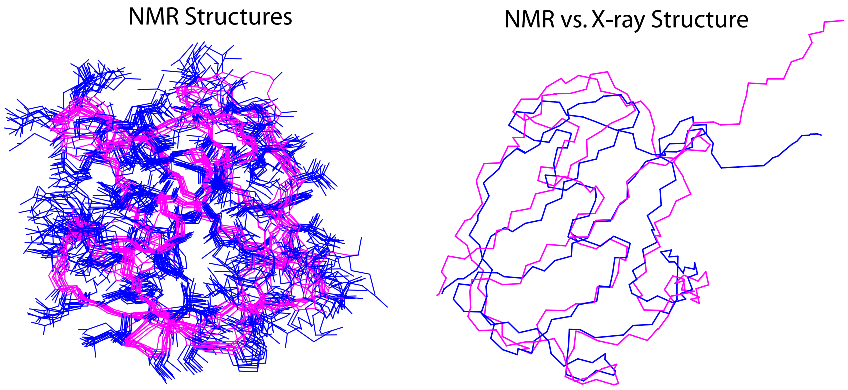

Figure 3. The NMR structures computed from the automatically-assigned NOEs. The left panel shows the 12 best NMR structures with no NOE distance violation larger than 0.50 Å. The sidechains were shown in blue while the backbone in magenta. The right panel is the overlay of the NMR average structure (blue) with the 1.8 Å x-ray structure (magenta) [28].

the NOE list, XPLOR was invoked for the second time to compute the structures using the remaining 1620 NOE restraints and 12 hydrogen bonds. Twelve structures out of 70 total computed structures had no NOE violations larger than 0.50 Å. Thus, our NOE assignment algorithm has an accuracy better than 90%. These 12 best NMR structures (Fig. 3) can be overlayed with a pairwise RMSD of 1.18 ± 0.16 Å for backbone atoms and a pairwise RMSD of 1.84 ± 0.19 Å for all heavy atoms. The accuracy of the NMR structures computed using the automatically-assigned NOEs from our algorithm is in the range of typical high- to medium-resolution NMR structures. The average structure computed from the best 12 structures has a 1.43 Å backbone RMSD and 2.13 Å all heavy-atom RMSD from the 1.8 Å ubiquitin X-ray structure [28].

6 Biological significance

In this paper we have demonstrated that by first computing a complete backbone structure using only NH and CH RDCs in a single medium, a large number of NOE restraints can be assigned automatically using rotamer ensemble-based intra-residue vectors in combination with the backbone structure as an algorithmic filter. These automatically-assigned NOE restraints can then be input to any standard NMR structure determination algorithm to compute a high-resolution complete structure with all the sidechain conformations. Such a complete structure is not only very useful for functional annotation of protein-coding gene sequence but is also the starting point for rational drug design targeting the protein. Furthermore, together with our efficient algorithm for backbone structure computation [30, 29], we obtain a fast protocol for computing complete NMR structures starting from resonance assignment. Our protocol is much more efficient than all the previous approaches, thus it should be important for structural genomics. Compared with previous approaches for NOE automatic assignment (section 3 of the main text) we expect that our algorithm to be much more robust with respect to missing resonance assignments and noise NOE peaks. Our algorithm does not rely on an initial fold computed from only a fraction of assigned NOE peaks. Our initial fold is computed exclusively from RDC data, plus the few unambiguous NOEs that can be identified readily from chemical shifts alone. Our new algorithm should be general and can be applied to any set of NMR spectra (or FIDs) of either proteins or nucleic acids that include two RDCs per residue (resp. nucleotide) and one or more NOE spectra.

7 Algorithmic complexity and practical performance

The algorithm takes $O(n \log n)$ time to sort all the ^{1}H chemical shifts where n is the total number of protons in a protein, which is a constant times of the total number of residues in the protein. The total number of NOE peaks is $O(n^2)$ in the worst-case but $O(n)$ in the average case. It takes at most $O(\log n)$ time to search for all protons which have chemical shifts in the interval, $c_b \pm \delta_c$, for each peak. So the total complexity of NOE assignment is $O(n \log n + n^2 \log n)$, that is $O(n^2 \log n)$ since only one cycle of NOE assignment suffices when a well-defined backbone structure is computed using [30, 29].

In practice, short turns can be computed in less than one second on a 2.4 GHz single-processor Linux workstation (Appendix C). The two loops connecting the helix to the sheet can each be computed in less than 2 minutes. The longest loop (Glu51–Lys63) can be computed in 5 minutes. In practice, it took less than one second to assign 1783 NOE restraints from the NOE peak list picked from both the 3D ^{15}N-edited and ^{13}C-edited NOESY spectra. Compared with all previous heuristic NOE assignment approaches, our NOE assignment algorithm is the first provably-efficient NOE assignment algorithm. As stated in section 3, what makes the previous assignment approaches expensive is their reliance on many cycles of structure determination to filter ambiguous NOE assignments. In contrast, our algorithm uses a rotamer library ensemble in combination with a backbone structure to filter ambiguous NOE assignments. In practice, our algorithm is also much more faster than all the previous approaches. The complete NMR structures using our automatically-assigned NOE restraints can be computed by XPLOR [4] in about 30 minutes. Taken together, our entire protocol, starting from the backbone structure determination [30, 29] using backbone resonance assignment, backbone RDC data, to the final complete solution structures with all the sidechain conformations, can be completed in less than 2.5 hours on a single-processor workstation, that is, at least 4 times faster than any previous approach, which typically require many hours to weeks of computational time.

8 Conclusion

NOE assignment is the bottleneck for high-throughput structure determination by NMR. We have described a provably-efficient, accurate algorithm for automated NOE assignment. In contrast to previous NOE assignment approaches, which rely on many cycles of structure determination to iteratively filter ambiguous NOE assignments, our algorithm uses rotamer ensemble-based internuclear vectors mined from the PDB in combination with a backbone structure computed using only two RDCs per residue, to reduce the ambiguity in assigning NOEs involving sidechain protons, especially aliphatic protons. More than 1,700 NOE restraints were automatically assigned by our algorithm from two distinct, real 3D ^{15}N-edited and ^{13}C-edited NOESY spectra of ubiquitin. An accurate complete structure with all the sidechain conformations has been computed by using these automatically-assigned NOE restraints. Such a complete structure is useful not only for functional annotation of gene sequences, but is also the starting point for drug design. Together with our efficient algorithm for backbone structure computation [30, 29], we have developed a suite of efficient algorithms for computing complete NMR structures of either proteins or nucleic acids starting from resonance assignment. Our protocol is much more efficient than all previous approaches, and thus should be important for structural genomics.

Acknowledgments

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

References

[1] C. Bailey-Kellogg, A. Widge, J. J. Kelley, M. J. Berardi, J. H. Bushweller, and B. R. Donald. The NOESY Jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J. Comput. Biol.*, 7(3–4):537–558, 2000.

[2] C. Bartels, P. Gntert, M. Billeter, and K. Wüthrich. GARANT a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *J. Comp. Chem.*, 18:139–149, 1997.

[3] B. Berger, J. Kleinberg, and F. T. Leighton. Reconstructing a three-dimensional model with arbitrary errors. *Journal of the ACM*, 46(2):212–235, 1999.

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

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

[6] G. Cornilescu, F. Delaglio, and A. Bax. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*, 13:289–302, 1999.

[7] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, Pfeifer. J., and A. Bax. A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, 6:277–293, 1995.

[8] R. L. Jr. Dunbrack and F. E. Cohen. Bayesian statistical analysis of protein sidechain rotamer preferences. *Protein Science*, 6:1661–1681, 1997.

[9] R. R. Ernst, G. Bodenhausen, and A Wokaun. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. Clarendon Press, Oxford., 1987.

[10] A. Grishaev and M. Llinas. Clouds,a protocol for deriving a molecular proton density via NMR. *Proc. Natl. Acad. Sci. USA*, 99:6707–6712, 2002.

[11] 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.

[12] R. Harris. The ubiquitin NMR resource page, BBSRC Bloomsbury center for structural biology. http://www.biochem.ucl.ac.uk/bsm/nmr/ubq/, 2002.

[13] B. Hendrickson. Conditions for unique graph realizations. *SIAM Journal on Computing*, 21:65–84, 1992.

[14] B. Hendrickson. The molecule problem: Exploiting structures in global optimization. *SIAM Journal on Optimization*, 5:835–857, 1995.

[15] 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. *J. Mol. Biol.*, 319:209–227, 2002.

[16] Y. J. Huang, G. V. Swapna, P. K. Rajan, H. Ke, B. Xia, K. Shukla, M. Inouye, and G. T. Montelione. Solution NMR structure of ribosomebinding factor a (RbfA), a coldshock adaptation protein from *escherichia coli*. *J. Mol. Biol.*, 327:521–536, 2003.

[17] B. A. Johnson and R. A. Blevins. NMRView: A computer program for the visualization and analysis of NMR data. *J. Biomol. NMR*, 4:603–614, 1994.

[18] K. Juszewski, C. D. S. 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.*, 126:6258–6273, 2004.

[19] R. H. Kretsinger, R. E. Ison, and S. Hovmoller. Prediction of protein structure. *Methods Enzymol.*, 383:1–27, 2004.

[20] C. J. Langmead and B. R. Donald. An Expectation/Maximization nuclear vector replacement algorithm for automated NMR resonance assignments. *J. Biomol. NMR*, 29(2):111–138, 2004.

[21] G. Lin, Z. Chen, T. Jiang, J. Wen, J. Xu, and Y. Xu. Approximation algorithms for NMR spectral peak assignment. *Theoretical Computer Science*, 299:211–229, 2002.

[22] G. Lin, D. Xu, Z. Chen, T. Jiang, and Y. Xu. A branch and bound algorithm for assignment of protein backbone NMR peaks. In *Proceedings of First IEEE Bioinformatics Conference*, pages 165–174, Stanford University, CA, 2002.

[23] 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.

[24] J. Meiler and D. Baker. Rapid protein fold determination using unassigned NMR data. *Proc. Natl. Acad. Sci. USA*, 100:15404–15409, 2003.

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

[26] J. B. Saxe. Embeddability of weighted graphs in k-space is strongly NP-hard. In *Proceedings of the 17th Allerton Conference on Communications, Control, and Computing*, pages 480–489, 1979.

[27] I. Solomon. Relaxation processes in a system of two spins. *Physical Review*, 99:559–565, 1955.

[28] 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.

[29] 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 *IEEE Computational Systems Bioinformatics Conference*, pages 319–330, 2004.

[30] 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. *J. Biomol. NMR*, 29:223–242, 2004.

[31] L. Wang, A. V. Kurochkin, and E. R. P. Zuiderweg. An iterative fitting procedure for the determination of longitudinal NMR cross–correlation rates. *J. Magn. Reson.*, 144:175–185, 2000.

[32] Y. Xu, D. Xu, D. Kim, V. Olman, J. Razumovskaya, and T. Jiang. Automated assignment of backbone NMR peaks using constrained bipartite matching. In *IEEE Computing in Science and Engineering*, pages 50–62, 2002.

[33] Z. Yang, A. Kolinski, and J. Skolnick. TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophys. J.*, 85(2):1145–1164, 2003.

[34] D. E. Zimmerman, C. A. Kulikowski, W. Feng, M. Tashiro, C-Y. Chien, C. B. Ros, F. J. Moy, R. Powers, and G. T. Montelione. Artificial intelligence methods for automated analysis of protein resonance assignments. *J. Mol. Biol.*, 269:592–610, 1997.

Appendix

Appendix **A** describes the principle of NOE experiments and the source of errors in NOE experiments. In appendix **B** we give proofs of Proposition 1 and Corollary 1 from section 4.2. Appendix **C** presents algorithms for computing loops and turns using two RDCs per residue measured in a single medium and the results of applying these algorithms to real experimental data. Appendix **D** presents the processing of NMR spectra and chemical shift resonance assignment.

A Principle of NOE experiments

Physically, the dipole-dipole interaction between two nuclei gives rise to an NOE peak whose intensity is a function of the internuclear distance between the two nuclei. Under the isolated-two spin approximation [27], the NOE peak intensity or volume (in this paper, an NOE peak means an NOE *cross-peak*, not a diagonal peak) between two nuclei (e.g. protons H_a and H_b) is directly proportional to the cross-relaxation rate

$$\sigma(H_a \leftrightarrow H_b) = \frac{\gamma_{_H}^2 \gamma_{_H}^2 \hbar^2}{4r_{ab}^6}\Big(6J_2(2\omega_{_H}) - J_0(0)\Big) \tag{5}$$

where $\gamma_{_H}$ is the proton gyromagnetic ratio (a physical constant), r_{ab} is the internuclear distance between H_a and H_b, and $J_2(2\omega_{_H})$ and $J_0(0)$ are, respectively, the spectral density functions at frequencies $2\omega_{_H}$ and 0, which describe the motion of the internuclear vector caused by the interaction of the 2-spin system with its environment. However, there exist thousands of protons in a macromolecule such as a protein, which interact with one another through their inherent magnetic dipoles. Thus, the measured intensity of each individual NOE peak is really the result of a network of protons where every pair of protons interact with one another as described by Eq. (5) [31, 9]. In principle, one could solve a system of thousands of differential equations to compute cross-relaxation rates. In practice, this is unrealistic, because the initial conditions can not be well-specified in an NOE experiment. Even if the cross-relaxation rates can be computed, approximately, by a relaxation matrix analysis, it is very difficult to accurately measure the individual spectral density functions in Eq. (5). Consequently, to extract distance restraints from an NOE peak, one usually applies the isolated-two spin approximation and, further ignores the spectral density functions. The impossibility of accurately converting the NOE peak intensity into the internuclear distance restraint (r_{ab} in Eq. 5) makes both the NOE automated assignment problem and the NMR structure determination problem dramatically harder. For an approach which uses previous structures to assign NOEs, the large uncertainty in NOE peak intensity back-computed from a structure will increase the number of possible assignments for each peak. Conversely, the large uncertainty in distance restraints computed from NOE peak intensities makes the conventional NMR structure-determination using NOEs alone NP-hard [26, 3, 13, 14] .

B Intra-residue vectors between the C_β nucleus and sidechain protons

In this section we sketch a proof for Proposition 1 and Corollary 1 stated in section 4.2 of the main text. We prove both for the atoms with sp^3 configuration. Propositions for atoms with sp^2 configuration can be proved similarly.

Proposition 1 *Given a backbone structure and the sidechain dihedral angles χ_1, χ_2, χ_3 and χ_4, let H_{γ_i}, H_{δ_i} and H_{ϵ_i} be particular H_γ, H_δ, H_ϵ protons (i = 1, 2, 3, depending on the amino acid). Then the three vectors from the atom C_β to the three respective sidechain protons, H_{γ_i}, H_{δ_i} and H_{ϵ_i}, can be computed analytically by*

$$\begin{aligned} \mathbf{v}_{\beta\gamma_i} &= \mathbf{p}_1(\chi_1, \chi_2) \\ \mathbf{v}_{\beta\delta_i} &= \mathbf{p}_2(\chi_1, \chi_2, \chi_3) \\ \mathbf{v}_{\beta\epsilon_i} &= \mathbf{p}_3(\chi_1, \chi_2, \chi_3, \chi_4), \end{aligned} \tag{6}$$

where $\mathbf{p}_1$, $\mathbf{p}_2$ and $\mathbf{p}_3$ are vector-valued trigonometric polynomials.
Below, the lengths of these three vectors will be denoted, respectively, by $d_{\beta\gamma}$, $d_{\beta\delta}$ and $d_{\beta\epsilon}$.

Proof. In the following, bold letters denote a vector (column vector) or a matrix. Let the matrix $\mathbf{M}_\beta$ be the rotation matrix between a coordinate frame defined in the peptide plane and a frame defined in the plane specified by the three atoms, N, C_α and C_β, with the +Y axis in the NC_α direction, +Z axis in the plane and +X axis defined by right-handedness. The sp^3 atom C_β can be viewed as the *center* of a tetrahedron with C_α as the *top* vertex, and C_γ, H_{β_1}, H_{β_2} as the *bottom* three vertices, v_1, v_2 and v_3. This tetrahedron is connected to other atoms through the backbone amide atom (N), which we call a *connector* (Fig. 4). The shape and pose of the tetrahedron are completely specified by three plane angles, $\theta_{p_1}, \theta_{p_2}$ and θ_{p_3}, three dihedral angles χ_1, θ_{d_2} and θ_{d_3}, and four lengths, L_T, L_1, L_2 and L_3. The plane angles θ_{p_1}, θ_{p_2} and θ_{p_3} are the angles between the vector from the *top* to the *center* and the three vectors from the *center* to, respectively, the three bottom vertices v_1, v_2 and v_3. The dihedral angle χ_1 is between the plane specified by the three atoms, *connector, top* and *center* and the plane specified by the three atoms, *top*, *center* and v_1. The two dihedral angles θ_{d_2} and θ_{d_3} are, respectively, between the plane specified by the three atoms, *top*, *center* and v_1, and the planes specified by the three atoms, *top*, *center* and v_2 and the three atoms, *top*, *center* and v_3. The four lengths, L_T, L_1, L_2 and L_3 are, respectively, the lengths between the *connector* and the *top*, the *top* and v_1, the *top* and v_2, and the *top* and v_3. Except for the dihedral angle χ_1, all the other angles and all the lengths are constants for the standard sp^3 configuration.

Given χ_1 and the position vector of the atom C_β (*center*), $\mathbf{C}_\beta$, we can compute the vectors of the three bottom atoms, $\mathbf{C}_\gamma, \mathbf{H}_{\beta_1}$ and $\mathbf{H}_{\beta_2}$ and the rotation matrix, $\mathbf{M}_\gamma$, to the plane specified by the three atoms C_α, C_β and C_γ (with the +Y axis in the $\mathrm{C}_\alpha\mathrm{C}_\beta$ direction and +Z axis in the plane) by

$$\begin{aligned} \mathbf{C}_\gamma &= \mathbf{C}_\beta + \mathbf{M}_\gamma(0, L_1, 0)^T \\ \mathbf{H}_{\beta_1} &= \mathbf{C}_\beta + \mathbf{M}_\beta \mathbf{R}^{-1}{}_y(\chi_1 + \theta_{d_2})\mathbf{R}^{-1}{}_x(\theta_{p_2})(0, L_2, 0)^T \\ \mathbf{H}_{\beta_2} &= \mathbf{C}_\beta + \mathbf{M}_\beta \mathbf{R}^{-1}{}_y(\chi_1 + \theta_{d_3})\mathbf{R}^{-1}{}_x(\theta_{p_3})(0, L_3, 0)^T \\ \mathbf{M}_\gamma &= \mathbf{M}_\beta \mathbf{R}^{-1}{}_y(\chi_1)\mathbf{R}^{-1}{}_x(\theta_{p_1}), \end{aligned} \tag{7}$$

where

$$\mathbf{R}_x(\theta) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}$$

$$\mathbf{R}_y(\theta) = \begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix}$$

$$\mathbf{R}_z(\theta) = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and $(0, L_1, 0)^T$ etc. denotes a column vector. Similarly, we can view the sp^3 atom C_γ as the *center* of a new tetrahedron with C_β as the *top* vertex, and C_δ, $\mathrm{H}_{\gamma 1}$, $\mathrm{H}_{\gamma 2}$ as the *bottom* three vertices. This tetrahedron is connected to other atoms through the backbone atom C_α. Thus, given χ_2 and the

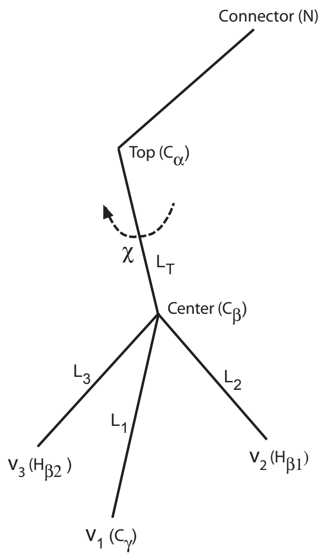

Figure 4. The sp^3 configuration. An sp^3 configuration with the atom C_β as the *center*, amide atom N as the *connector*, atom C_α as the *top* and the three atoms, C_γ, H_{β_1}, H_{β_2}, as the *bottom* three vertices, v_1, v_2 and v_3. L_T, L_1, L_2 and L_3 are lengths.

vector of the atom C_γ (*center*), $\mathbf{C}_\gamma$, we have the same equations, *mutatis mutandis*, to compute the vectors of the three bottom atoms, $\mathbf{C}_\delta$, $\mathbf{H}_{\gamma 1}$ and $\mathbf{H}_{\gamma 2}$ and the rotation matrix, $\mathbf{M}_\delta$, to the plane specified by the three atoms, C_β, C_γ and C_δ (with the +Y axis in the $\mathrm{C}_\beta\mathrm{C}_\gamma$ direction and +Z axis in the plane) by

$$\begin{aligned}
\mathbf{C}_\delta &= \mathbf{C}_\gamma + \mathbf{M}_\gamma \mathbf{R}^{-1}{}_y(\chi_2)\mathbf{R}^{-1}{}_x(\theta_{p_1})(0, L_1, 0)^T \\
\mathbf{H}_{\gamma_1} &= \mathbf{C}_\gamma + \mathbf{M}_\gamma \mathbf{R}^{-1}{}_y(\chi_2 + \theta_{d_2})\mathbf{R}^{-1}{}_x(\theta_{p_2})(0, L_2, 0)^T \\
\mathbf{H}_{\gamma_2} &= \mathbf{C}_\gamma + \mathbf{M}_\gamma \mathbf{R}^{-1}{}_y(\chi_2 + \theta_{d_3})\mathbf{R}^{-1}{}_x(\theta_{p_3})(0, L_3, 0)^T \\
\mathbf{M}_\delta &= \mathbf{M}_\gamma \mathbf{R}^{-1}{}_y(\chi_2)\mathbf{R}^{-1}{}_x(\theta_{p_1}). \qquad (8)
\end{aligned}$$

We have derived similar equations for the sp^3 atom C_δ with C_γ as the *top* vertex, and C_ϵ, H_{δ_1}, H_{δ_2} as the *bottom* three vertices and C_β as the *connector*. Thus, given χ_3 and the vector for the atom C_δ (*center*), $\mathbf{C}_\delta$, we can compute the vectors of the three bottom atoms, $\mathbf{C}_\epsilon$, $\mathbf{H}_{\delta_1}$ and $\mathbf{H}_{\delta_2}$ and the rotation matrix, $\mathbf{M}_\epsilon$, to the plane specified by the three atoms, C_γ, C_δ and C_ϵ (with the +Y axis in the $\mathrm{C}_\gamma\mathrm{C}_\delta$ direction and +Z axis in the plane) by

$$\begin{aligned}
\mathbf{C}_\epsilon &= \mathbf{C}_\delta + \mathbf{M}_\delta \mathbf{R}^{-1}{}_y(\chi_3)\mathbf{R}^{-1}{}_x(\theta_{p_1})(0, L_1, 0)^T \\
\mathbf{H}_{\delta_1} &= \mathbf{C}_\delta + \mathbf{M}_\delta \mathbf{R}^{-1}{}_y(\chi_3 + \theta_{d_2})\mathbf{R}^{-1}{}_x(\theta_{p_2})(0, L_2, 0)^T \\
\mathbf{H}_{\delta_2} &= \mathbf{C}_\delta + \mathbf{M}_\delta \mathbf{R}^{-1}{}_y(\chi_3 + \theta_{d_3})\mathbf{R}^{-1}{}_x(\theta_{p_3})(0, L_3, 0)^T \\
\mathbf{M}_\epsilon &= \mathbf{M}_\delta \mathbf{R}^{-1}{}_y(\chi_3)\mathbf{R}^{-1}{}_x(\theta_{p_1}). \qquad (9)
\end{aligned}$$

Substituting the expressions for $\mathbf{C}_\gamma$ and $\mathbf{M}_\gamma$ of Eq. (7) into Eq. (8) we can compute the two vectors, $\mathbf{v}_{\beta\gamma_1}$ and $\mathbf{v}_{\beta\gamma_2}$, from the atom C_β to the two H_γ atoms,

$$\begin{aligned}
\mathbf{v}_{\beta\gamma_1} &= \mathbf{M}_\beta \mathbf{R}^{-1}{}_y(\chi_1)\mathbf{R}^{-1}{}_x(\theta_{p_1})(0, L_1, 0)^T \\
&\quad + \mathbf{M}_\beta \mathbf{R}^{-1}{}_y(\chi_1)\mathbf{R}^{-1}{}_x(\theta_{p_1})\mathbf{R}^{-1}{}_y(\chi_2 + \theta_{d_2}) \\
&\quad \mathbf{R}^{-1}{}_x(\theta_{p_2})(0, L_2, 0)^T \\
\mathbf{v}_{\beta\gamma_2} &= \mathbf{M}_\beta \mathbf{R}^{-1}{}_y(\chi_1)\mathbf{R}^{-1}{}_x(\theta_{p_1})(0, L_1, 0)^T \\
&\quad + \mathbf{M}_\beta \mathbf{R}^{-1}{}_y(\chi_1)\mathbf{R}^{-1}{}_x(\theta_{p_1})\mathbf{R}^{-1}{}_y(\chi_2 + \theta_{d_3}) \\
&\quad \mathbf{R}^{-1}{}_x(\theta_{p_3})(0, L_3, 0)^T. \qquad (10)
\end{aligned}$$

Substituting Eq. (7) into Eq. (8), then Eq. (8) into Eq. (9), we can compute the two vectors, $\mathbf{v}_{\beta\delta_1}$ and $\mathbf{v}_{\beta\delta_2}$, from the atom C_β to the two H_δ atoms,

$$\begin{aligned}
\mathbf{v}_{\beta\delta_1} &= \mathbf{M}_\beta \mathbf{R}^{-1}{}_y(\chi_1)\mathbf{R}^{-1}{}_x(\theta_{p_1})(0, L_1, 0)^T \\
&\quad + \mathbf{M}_\beta \mathbf{R}^{-1}{}_y(\chi_1)\mathbf{R}^{-1}{}_x(\theta_{p_1})\mathbf{R}^{-1}{}_y(\chi_2) \\
&\quad \mathbf{R}^{-1}{}_x(\theta_{p_1})(0, L_1, 0)^T \\
&\quad + \mathbf{M}_\beta \mathbf{R}^{-1}{}_y(\chi_1)\mathbf{R}^{-1}{}_x(\theta_{p_1})\mathbf{R}^{-1}{}_y(\chi_2) \\
&\quad \mathbf{R}^{-1}{}_x(\theta_{p_1})\mathbf{R}^{-1}{}_y(\chi_3 + \theta_{d_2})\mathbf{R}^{-1}{}_x(\theta_{p_2})(0, L_2, 0)^T \\
\mathbf{v}_{\beta\delta_2} &= \mathbf{M}_\beta \mathbf{R}^{-1}{}_y(\chi_1)\mathbf{R}^{-1}{}_x(\theta_{p_1})(0, L_1, 0)^T \\
&\quad + \mathbf{M}_\beta \mathbf{R}^{-1}{}_y(\chi_1)\mathbf{R}^{-1}{}_x(\theta_{p_1})\mathbf{R}^{-1}{}_y(\chi_2) \\
&\quad \mathbf{R}^{-1}{}_x(\theta_{p_1})(0, L_1, 0)^T \\
&\quad + \mathbf{M}_\beta \mathbf{R}^{-1}{}_y(\chi_1)\mathbf{R}^{-1}{}_x(\theta_{p_1})\mathbf{R}^{-1}{}_y(\chi_2) \\
&\quad \mathbf{R}^{-1}{}_x(\theta_{p_1})\mathbf{R}^{-1}{}_y(\chi_3 + \theta_{d_3})\mathbf{R}^{-1}{}_x(\theta_{p_3})(0, L_3, 0)^T. \\
&\qquad (11)
\end{aligned}$$

Eq. (10) and Eq. (11) show that the functions $\mathbf{p}_1$ and $\mathbf{p}_2$ are trigonometric polynomials in the χ_1, χ_2, χ_3 angles. Similar trigonometric polynomials of the χ_1, χ_2, χ_3 and χ_4 angles can be derived for the two vectors, $\mathbf{v}_{\beta\epsilon_1}$ and $\mathbf{v}_{\beta\epsilon_2}$, that is, the function $\mathbf{p}_3$ is also a trigonometric polynomial in the $\chi_1, \chi_2, \chi_3, \chi_4$ angles. ▌

As a corollary of the above proposition we can easily prove that

Corollary 1 *Given the vector from the atom* C_β *to a sidechain proton and the vector* $\mathbf{v}_{_{N\beta}}$ *from the backbone atom* $H_{_N}$ *to the atom* C_β*, or the vector* $\mathbf{v}_{_{\alpha\beta}}$ *from the atom* H_α *to the atom* C_β*, the vector* $\mathbf{v}_{_{NS}}$ *from the atom* $H_{_N}$ *to a sidechain proton (e.g.* H_δ*), or the vector* $\mathbf{v}_{_{\alpha S}}$ *from the atom* H_α *to a sidechain proton can be computed by*

$$\mathbf{v}_{_{NS}} = \mathbf{v}_{_{N\beta}} + \mathbf{v}_{_{\beta\delta}}, \qquad \mathbf{v}_{_{\alpha S}} = \mathbf{v}_{_{\alpha\beta}} + \mathbf{v}_{_{\beta\delta}}. \qquad (12)$$

The length of these vectors, $r_{_N} = |\mathbf{v}_{_{NS}}|$ *and* $r_{_\alpha} = |\mathbf{v}_{_{\alpha S}}|$*, can be computed by the cosine law where* $r^2_{_N}$ *and* $r^2_{_\alpha}$ *are, respectively, trigonometric polynomial functions of the* $\chi_1, \chi_2, \ldots$ *angles.*

Proof. The rotation matrix $\mathbf{M}_\beta$ defined in the proof of Proposition 1 can be computed from the backbone dihedral angle ϕ,

$$\mathbf{M}_\beta = \mathbf{R}^{-1}{}_x(\theta_1)\mathbf{R}^{-1}{}_y(\phi)\mathbf{R}^{-1}{}_y(\theta_{\beta_1})\mathbf{R}^{-1}{}_x(\theta_{\beta_2}), \quad (13)$$

where the angles θ_1, β_1 and β_2 are known from standard protein backbone geometry [30]. Given the protein backbone structure, the vector $\mathbf{v}_{_{N\beta}}$ from the backbone atom H_N or the vector $\mathbf{v}_{_{\alpha\beta}}$ from the H_α to the atom C_β is therefore known. As shown in Proposition 1, the vectors $\mathbf{v}_{_{\beta\gamma}}$, $\mathbf{v}_{_{\beta\delta}}$ and $\mathbf{v}_{_{\beta\epsilon}}$ are trigonometric polynomials in the χ angles. Thus, the vector from the atom H_N or H_α to a sidechain proton can be computed as the addition of two vectors such as $\mathbf{v}_{_{NS}} = \mathbf{v}_{_{N\beta}} + \mathbf{v}_{_{\beta\delta}}$ (Fig. 2 of the main text). ▌

C Algorithm for the computation of loops and turns

We employ the backbone structure-determination algorithm in [30, 29] as a first stage, namely, given the RDC data, the algorithms in [30, 29] are called as a subroutine to compute an accurate backbone structure. This backbone structure, in turn, is then used to boot-strap the automated assignment of NOEs: the assignment proceeds by *filtering* the experimentally-measured NOEs based on consistency with the structure. For the purposes of computing an initial structure for automated NOE assignment, the method by which we compute loops and turns is somewhat different from the algorithms in [30, 29]. In order to fully understand our automated assignment algorithm, these details are given here.

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

Proposition 2 *Given the orientation of peptide planes* i *and* $i+2$ *in a principal order frame (POF) of RDCs and the backbone dihedral angle* ϕ_i*, the backbone dihedral angles* ψ_i*,* ϕ_{i+1} *and* ψ_{i+1} *can be computed exactly and in constant time.*

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

For ease of exposition, we use the following example to illustrate our algorithm, which is itself completely general. Suppose a helix is linked to a sheet by two loops, loop1 (connecting the C-terminal of strand A of the sheet to the N-terminal of the helix) and loop2 (connecting the C-terminal of the helix to the N-terminal of strand B of the sheet).

1. Start with the C-terminal of strand A, use a depth-first search (DFS) as described in [30] to compute all the loop1 conformations consistent with the orientation of the first peptide plane of the N-terminal of the helix (Proposition **2**) and without steric clash with other backbone atoms.
2. Translate the helix by overlaying the N (amide) atom of the first residue of the helix with the position for the same N atom which is newly-computed in step 1.
3. Compute the distance, $d_{_{CC}}$ between the C_α atom of the C-terminal residue of the newly positioned helix and the C_α atom of the N-terminal residue of strand B and prune the loop1 conformation if $d_{_{CC}} > d_{_{MAX}}(i, i+j)$ (see below).
4. Start with the C-terminal of the newly-positioned helix, use a DFS as described in [30] to compute all the loop2 conformations consistent with the *fixed* first peptide plane of the N-terminal of strand B (Proposition **3**) and without steric clash with other backbone atoms.
5. Compute the distance $d_{_{CC}}$ between the newly-computed C_α atom of loop2 and the C_α atom of the N-terminal residue of strand B and prune the loop2 conformation if $d_{_{CC}} > d_{_{MAX}}(i, i+j)$ (see below).

The distance, $d_{_{MAX}}(i, i+j)$, is the maximum possible distance between the C_α atoms of the two residues i and $i+j$. Given the standard protein backbone geometry [30, p. 234] and the offset j, $d_{_{MAX}}(i, i+j)$ is a constant. From all the computed conformations for loop1 and loop2 we select a best conformation that has the smallest RMSDs between the experimental RDCs of the residues in loop1 and loop2, and the corresponding RDCs back-computed from the loop conformations using the alignment tensor for the backbone structure.

We have successfully applied our algorithm to compute the turns and loops for the protein human ubiquitin using NH and CH RDCs in a single medium. Two short turns,

NMR spectrum	Information Content
2D ^{1}H, ^{15}N HSQC	Correlation of intra-residue amide proton ^{1}H and nitrogen ^{15}N
3D HNCA	Correlation of intra-residue and inter-residue ^{1}H, ^{15}N, ^{13}C$_\alpha$
3D HNCO	Correlation of intra-residue ^{1}H, ^{15}N, ^{13}C'
3D HNCACB	Correlation of intra-residue and inter-residue ^{1}H, ^{15}N, ^{13}C$_\alpha$, ^{13}C$_\beta$
3D HN(CO)CA	Correlation of inter-residue ^{1}H, ^{15}N, ^{13}C$_\alpha$
3D HN(CA)CO	Correlation of inter-residue ^{1}H, ^{15}N, ^{13}C'
3D HCCH-TOCSY	Correlation of intra-residue sidechain ^{13}C and the bonded aliphatic protons (^{1}H)

Table 3. **NMR spectra used for resonance assignment.** Two sets (7ms and 18ms mixing time, respectively) of 3D HCCH-TOCSY spectra were used for sidechain assignment, while all the other spectra were used for backbone resonance assignment.

Leu8–Gly10 and Gly47–Lys48, could be computed without using any experimental RDCs, since they are less than 3 residues long (Proposition **3**). The two loops, Glu18–Thr22 and Gly35–Glu41, connecting the helix (Leu23–Glu34) to the single sheet (consisting of five strands), can also be computed using only NH and CH RDCs in a single medium. The conformations of these two loops determine the relative position between the helix and sheet. The most difficult problem is the computation of the long loop, Glu51–Lys63, connecting two β-strands in the sheet. Two long-range backbone NOE distances, H$_N$(Thr22)↔H$_N$(Thr55) and H$_N$(Ile23)↔H$_\alpha$(Leu56), automatically-assigned based on chemical shift alone (Section 4.1) are required for improving the accuracy of the conformation. The complete backbone structure computed by our algorithm has a 1.45 Å backbone RMSD from the corresponding X-ray backbone structure (PDB ID 1UBQ) [28].

D Processing of NMR data and Resonance Assignment

All the raw NMR FIDs and the acquisition parameters for ubiquitin were obtained from the Driscoll lab [12]. The NMR data used for resonance assignment have been summarized in Table 3. All the spectra were processed with the program NMRPIPE [7]. We briefly describe the processing of NMR data for the resonance assignment. The 2D ^{1}H, ^{15}N HSQC spectrum was processed by applying a Gaussian window function ^{1}H (F2) dimension and a shifted sine-bell function in the ^{15}N (F1) dimension. The HSQC data set was zero filled to 2048 × 512 real points. The triple resonance spectra (3D HNCA, 3D HNCO, 3D HNCACB, 3D HN(CO)CA, HN(CA)CO) used for backbone assignments were processed by applying a sine-bell or a shifted sine-bell function in each dimension. In general, the data sets were zero-filled to 2048 × 256 × 128 real points. The two 3D HCCHH-TOCSY spectra of 7ms and 18ms mixing time, used for sidechain resonance assignment, were processed by applying a Gaussian window function to the ^{1}H (F3) dimension and a shifted sine-bell function in the F1 and F2 dimensions. The two HCCH-TOCSY data sets were zero-filled to 1024 × 512 × 256 real points. Both the 3D ^{15}N-edited NOESY spectrum and 3D ^{13}C-edited NOESY were processed by applying a Gaussian window function ^{1}H (F3) dimension and a shifted sine-bell function in the F1 (^{1}H) and F2 (^{15}N or ^{13}C) dimensions. The two NOESY data sets were zero-filled to 2048 × 512 × 128 real points.

In general, the peaks were picked by an automatic peak-picking program [17] and subsequently edited manually. The backbone and sidechain resonance assignments followed the manual or semi-automatic procedure described in [5, chapter 8]. Except for the three residues, Met1, Asn24 and Gly53, we were able to assign the backbone resonances of all the remaining residues from Gln2 to Leu71, and assign more than 98% of the sidechain resonances (excluding the above three residues). The C-terminal five residues of ubiquitin are very flexible in solution, so none of their resonances have been assigned.