

Published in final edited form as:

*J Biomol NMR*. 2009 November ; 45(3): 265–281. doi:10.1007/s10858-009-9366-3.

# High-Resolution Protein Structure Determination Starting with a Global Fold Calculated from Exact Solutions to the RDC Equations<sup>3,4</sup>

Jiayang Zeng<sup>1</sup>, Jeffrey Boyles<sup>2</sup>, Chittaranjan Tripathy<sup>1</sup>, Lincong Wang<sup>1,†</sup>, Anthony Yan<sup>1</sup>, Pei Zhou<sup>2,\*</sup>, and Bruce Randall Donald<sup>1,2,\*</sup>

<sup>1</sup>Department of Computer Science, Duke University, Durham NC 27708

<sup>2</sup>Department of Biochemistry, Duke University Medical Center, Durham NC 27708

## Abstract

We present a novel structure determination approach that exploits the global orientational restraints from RDCs to resolve ambiguous NOE assignments. Unlike traditional approaches that bootstrap the initial fold from ambiguous NOE assignments, we start by using RDCs to compute accurate secondary structure element (SSE) backbones at the beginning of structure calculation. Our structure determination package, called RDC-PANDA (RDC-based SSE P<sub>A</sub>cking with NOEs for Structure D<sub>E</sub>termination and NOE A<sub>S</sub>signment), consists of three modules: (1) RDC-EXACT; (2) PACKER; and (3) HANA (H<sub>A</sub>usdorff-based NOE A<sub>S</sub>signment). RDC-EXACT computes the global optimal solution of backbone dihedral angles for each secondary structure element by exactly solving a system of quartic RDC equations derived by Wang and Donald (2004a,b), and systematically searching over the roots, each of which is a backbone dihedral  $\phi$ - or  $\psi$ -angle consistent with the RDC data. Using a small number of unambiguous inter-SSE NOEs extracted using only chemical shift information, PACKER performs a systematic search for the core structure, including all SSE backbone conformations. HANA uses a Hausdorff-based scoring function to measure the similarity between the experimental spectra and the back-computed NOE pattern for each side-chain from a statistically-diverse rotamer library, and drives the selection of optimal position-specific rotamers for filtering ambiguous NOE assignments. Finally, a local minimization approach is used to compute the loops and refine side-chain conformations by fixing the core structure as a rigid body while allowing movement of loops and side-chains. RDC-PANDA was applied to NMR data for the FF Domain 2 of human transcription elongation factor CA150 (RNA polymerase II C-terminal domain interacting protein), human ubiquitin, the ubiquitin-binding zinc finger domain of the human Y-family DNA polymerase Eta (pol  $\eta$  UBZ), and the human Set2-Rpb1 interacting domain (hSRI). These results demonstrated the efficiency and accuracy of our algorithm, and show that RDC-PANDA can be successfully applied for high-resolution protein structure determination using only a limited set of NMR data by first computing RDC-defined backbones.

<sup>3</sup>This work is supported by the following grants from National Institutes of Health: R01 GM-65982 to B.R.D. and R01 GM-079376 to P.Z.

<sup>4</sup>The methodology developed in this paper has been applied to compute the ensemble of structures for FF2. The atomic coordinates have been deposited into the Protein Data Bank (PDB ID: 2KIQ).

\*Corresponding authors: Bruce Randall Donald, brd+jbnmr09@cs.duke.edu, tel: 919-660-6583, Fax: 919-660-6519; Pei Zhou, peizhou@biochem.duke.edu, tel: 919-668-6409, Fax: 919-684-8885.

<sup>†</sup>Present address: Medicinal Chemistry, Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield CT, 06877.

**Availability** The RDC-PANDA software is available by contacting the authors, and is distributed open-source under the GNU Lesser General Public License (Gnu, 2002).

## Keywords

Nuclear magnetic resonance (NMR); nuclear Overhauser effect (NOE) assignment; residual dipolar coupling (RDC); structure determination; packing

## Introduction

One of the main bottlenecks in protein NMR structure determination lies in the acquisition and interpretation of a sufficient number of accurate distance restraints from nuclear Overhauser effect (NOE) data, which is often obstructed by assignment ambiguities due to chemical shift degeneracy. Traditional NMR structure determination approaches (Güntert 2003, Mumenthaler et al. 1997, Gronwald et al. 2002, Huang et al. 2006, Kuszewski et al. 2004) rely on heuristic techniques such as molecular dynamics (MD) and simulated annealing (SA), and may use a variety of data, including chemical shift, scalar couplings, NOEs and residual dipolar couplings (RDCs). In these approaches, however, RDCs are typically not employed in the annealing protocol until the end of structure calculation (i.e. refinement). Moreover, SA/MD based structure determination algorithms are used in these approaches as a subroutine in an iterative NOE assignment protocol. The SA/MD structure determination subroutine must typically be carefully initialized the first time using only reliable NOE assignments. The computed structures are then used to prune ambiguous NOE assignments. The SA/MD subroutine is used in an iterative fashion with new NOE assignments until convergence is reached.

Since NOEs provide merely *local* distance restraints, the correctness of an NOE assignment usually depends on the accuracy of other NOE assignments in its neighborhood. Thus, an error in an incorrect NOE assignment can be *propagated*, and hence influence the assignments of other NOEs. In contrast to NOE restraints, RDCs provide *global* orientational restraints on internuclear vectors, for example, backbone NH and C $\alpha$ -H $\alpha$  (henceforth abbreviated to CH) bond vectors with respect to a global alignment frame (Tolman et al. 1995, Tjandra & Bax 1997). In addition, in solution NMR RDC data can be collected with high precision, and assigned much faster than NOEs. Although several attempts have been made to find self-consistent NOE assignments (Güntert 2003, Mumenthaler et al. 1997, Herrmann et al. 2002, Linge et al. 2003, Gronwald et al. 2002, Huang et al. 2006, Kuszewski et al. 2004), little work has been done on exploiting other constraints such as residual dipolar couplings (RDCs) for resolving ambiguous NOE assignments. *NOAH* (Mumenthaler et al. 1997, Güntert 2003), for example, uses the structure determination package *DYANA* (Herrmann et al. 2002), and starts with an initial set of NOE assignments with putatively one or two possible assignments. *ARIA* (Linge et al. 2003) and *CANDID* (Herrmann et al. 2002) improved on *NOAH* by incorporating better modeling of ambiguous distance constraints. For instance, in both programs, the form of a  $(\sum r^{-6})^{-1/6}$  sum is used for handling ambiguous distances when multiple possible assignments exist for an NOE crosspeak. In *AUTO-STRUCTURE* (Huang et al. 2006), several heuristic rules that simulate the expertise of manual assignment are included to generate an initial fold. In *PASD* (Kuszewski et al. 2004) several strategies were proposed to reduce the chance of invoking the structure calculation into a biased path due to the incorrect initial global fold. None of the above NOE assignment approaches applied the global constraints from RDC data to filter ambiguous NOE assignments.

In traditional NMR structure determination approaches, stochastic techniques such as SA/MD are employed in a tight inner-loop and are invoked several times to filter ambiguous NOE assignments (Güntert 2003). The objective function used in these stochastic techniques models both the empirical molecular mechanics energy and the satisfaction of experimental data for a protein structure. Unfortunately, these stochastic techniques can be trapped into local minima

in the energy landscape of the objective function. Missing the global minimum structure solution during the iterative process can subsequently lead to incorrect NOE assignments. Furthermore, most previous approaches for automated NOE assignment (Herrmann et al. 2002, Linge et al. 2003) heavily depend on an accurate initial fold. However, the acquisition of a sufficient number of initial NOE assignments for computing a reliable starting fold is non-trivial, mainly due to the chemical shift degeneracy. Manual intervention and human expertise are often required in assigning these important initial NOEs in order to obtain a reliable initial fold.

To address above issues, a polynomial time *de novo* algorithm, called  $\text{RDC-EXACT}$ , has been proposed to compute high-resolution backbone structures for secondary structure elements (SSEs) using a minimum amount of residual dipolar coupling (RDC) data (Wang & Donald 2004b,a, Wang et al. 2006). The accurate backbone conformations computed by this algorithm enable us to propose a new strategy for NOE assignment. For example, a novel NOE assignment algorithm (Wang & Donald 2005) was proposed to filter ambiguous NOE assignments based on an ensemble of distance intervals computed using intra-residue vectors mined from a rotamer database, against inter-residue vectors from the backbone structure determined from RDCs. This algorithm uses a *triangle-like inequality* between the intra-residue and inter-residue vectors to prune incorrect assignments for side-chain NOEs. However, the algorithm (Wang & Donald 2005) has the following deficiencies: (a) it does not exploit the diversity of the rotamers in the library, (b) it does not exploit the rotamer patterns observable in NOE spectra, and (c) uncertainty in NOE peak position suggest a probabilistic model with provable properties which the previous algorithm (Wang & Donald 2005) did not capture.

In this paper, we propose a new structure determination framework, called  $\text{RDC-PANDA}$  ( $\text{RDC}$ -based  $\text{SSE}$   $\text{P}$ acking with  $\text{NOEs}$  for  $\text{S}$ tructure  $\text{D}$ etermination and  $\text{NOE}$   $\text{A}$ ssignment), by starting from an accurate global fold computed from RDCs.  $\text{RDC-PANDA}$  consists of three modules: (1)  $\text{RDC-EXACT}$ , which computes orientations and conformations of SSE backbones; (2)  $\text{P}_{\text{ACKER}}$ , which packs SSE backbones using sparse NOE restraints; and (3)  $\text{HANA}$  ( $\text{H}$ Ausdorff-based  $\text{NOE}$   $\text{A}$ ssignment), which uses the SSE backbones to place side-chains and assign NOEs. Unlike previous approaches that randomly sample the conformation space to find solutions that satisfy experimental restraints (Tian et al. 2001, Hus et al. 2001, Andrec et al. 2004),  $\text{RDC-EXACT}$  solves a system of low-degree polynomial equations in closed-form formulated from RDC restraints and protein backbone kinematics, to compute the backbone dihedral angle solutions.  $\text{RDC-EXACT}$  employs a systematic search over the roots of the polynomial system to find the global optimal solution for each secondary structure element.  $\text{P}_{\text{ACKER}}$  first extracts a small number of unambiguous inter-SSE NOEs from the NOESY spectra using only chemical shift information, and then performs a systematic 3-dimensional (3D) grid search over relative translations for the core structures, including all SSE backbone conformations. It considers all possible rotamers and discrete translation points that satisfy these sparse inter-SSE NOEs. The  $\text{HANA}$  module uses a Hausdorff-based scoring function to measure the similarity between the experimental spectra and the back-computed NOE pattern for each rotamer from a statistically-diverse rotamer library (Lovell et al. 2000), and selects optimal position-specific rotamers for filtering ambiguous NOE assignments. Finally, a local minimization approach is used to compute loops, refine side-chain conformations, and eliminate steric clashes among side-chains.  $\text{HANA}$  views the NOE assignment process as a *pattern-recognition* problem, where the objective is to establish a match by explicitly modeling the uncertainty in NOE peak positions, and thereby to choose an ensemble of rotamers with the best match scores between the experimental NOE data and the back-computed NOE pattern. Unlike previous, stochastic algorithms (Güntert 2003, Herrmann et al. 2002, Huang et al. 2006, Linge et al. 2003, Mumenthaler et al. 1997, Kuszewski et al. 2004) for NOE assignment,  $\text{HANA}$  uses the reliable initial fold computed primarily from RDCs, and can hence effectively prune ambiguous NOE assignments. Our strategy for computing the global fold is similar to the hierarchical

approaches in (Hayes-Roth et al. 1986, Chen et al. 1998, Delaglio et al. 2000), which apply the “local-to-global” idea and start with the SSEs to construct the global fold. In these approaches, however, SSEs are either determined by sampling the conformation space to find solutions satisfying the distance restraints (mainly from assigned NOEs) (Hayes-Roth et al. 1986, Chen et al. 1998), or selected from a structure database (Delaglio et al. 2000), while in  $\text{RDC-P}_{\text{ANDA}}$  the global fold is defined from exact solutions to the RDC equations.

We applied  $\text{RDC-P}_{\text{ANDA}}$  to four proteins: the FF Domain 2 of human transcription elongation factor CA150 (RNA polymerase II C-terminal domain interacting protein) (FF2), human ubiquitin, the ubiquitin-binding zinc finger domain of the human Y-family DNA polymerase Eta (pol  $\eta$  UBZ), and the human Set2-Rpb1 interacting domain (hSRI). Our results show that  $\text{RDC-P}_{\text{ANDA}}$  can achieve an accuracy of more than 90% for NOE assignment, and can calculate an ensemble of structures with backbone RMSD of  $0.97 \pm 0.30$  Å and all-heavy-atom RMSD of  $1.74 \pm 0.36$  Å from the reference structures. These results show that  $\text{RDC-P}_{\text{ANDA}}$  can be successfully applied for high-resolution protein structure determination using only a limited set of NMR data by first computing RDC-defined backbones.

## Methods

### Overview

Fig. 1 shows a schematic illustration of the  $\text{RDC-P}_{\text{ANDA}}$  approach (The flow chart of the  $\text{RDC-P}_{\text{ANDA}}$  algorithm is given in SM (Supplementary Material), Section S1 and Fig. S1). Our structure determination approach is divided into two stages. In Stage 1, the conformations and orientations of SSE backbones are computed from the RDC data using the  $\text{RDC-EXACT}$  algorithm (Wang & Donald 2004b, Wang et al. 2006) (Fig. 1 A). Then the SSE backbones are packed using sparse unambiguous inter-SSE NOE restraints, extracted from NOE spectra using only chemical shift information (Fig. 1 B). We call the resulting packed SSE backbones the *core structure*. Next, in the  $\text{HANA}$  module (Fig. 1 C), side-chains are placed on the core structure from the rotamer library (Lovell et al. 2000) by comparing back-computed NOE patterns of rotamers vs. the experimental NOESY spectra using a Hausdorff-based pattern matching technique (Zeng et al. 2008). Then the algorithm iteratively computes the NOE assignments and structures in loop regions (including side-chains) using a *local minimization* approach (Fig. 1 D), in which the core structure is fixed as a rigid body and only the loop regions and side-chains are allowed to move. Details of the local minimization approach are provided in SM Section S7. The local minimization approach allows side-chain flexibility and thus can eliminate the steric clash between rotameric side-chains that could not be resolved by the NOE pattern matching technique. The complete structure (including side-chains) computed by  $\text{HANA}$ , called the *low-resolution structure*, is then used to filter ambiguous NOE assignments. In Stage 2, the NOE assignment table computed by  $\text{HANA}$  in Stage 1 is fed to the structure refinement protocol in  $\text{XPLOR}/\text{XPLOR-NIH}$  to calculate high-resolution structures.

Unlike previous approaches, in which RDCs are only used in final structure refinement, our structure determination applies the global constraints from RDCs at the beginning, to compute an initial accurate global fold for subsequently pruning ambiguous NOE assignments. Moreover, our packing algorithm systematically searches all discrete translation points and outputs all packed structures that satisfy the experimental restraints. In addition, the pattern matching technique drives the selection of optimal rotamers for NOE assignment by statistically exploiting the conformational diversity of the rotamer library, effectively utilizing the accurate backbone conformations solved from RDCs, and efficiently mining the implicit rotamer patterns in the experimental NOE spectra.

## Input Data

The input data to `RDC-PANDA` include: (1) the primary sequence of the protein; (2) the 3D NOE peak list from both  $^{15}\text{N}$ - and  $^{13}\text{C}$ -edited spectra; (3) the resonance assignment list, including both backbone and side-chain resonance assignments; (4) the RDC data, including CH and NH RDCs, and the RDCs of other bond vectors (optional), such as  $\text{C}^\alpha\text{-C}'$  and  $\text{N-C}'$  bond vectors; (5) the TALOS table of dihedral angle ranges from the chemical shift analysis (Cornilescu et al. 1999); (6) the rotamer library (Lovell et al. 2000). Only CH and NH RDCs in one medium are required by `RDC-PANDA` to compute the backbone conformation and orientation, but other additional RDCs such as  $\text{C}^\alpha\text{-C}'$  and  $\text{N-C}'$  RDCs can be also included. This additional data is used to prune the ( $\phi$ ,  $\psi$ ) angles in the `RDC-EXACT` step.

## SSE Backbone Determination from Residual Dipolar Couplings

Residual dipolar couplings (Tjandra & Bax 1997, Tolman et al. 1995) provide global orientational restraints on the internuclear bond vectors, such as NH and CH bond vectors, with respect to a global coordinate frame. Given NH RDCs in two aligning media, the associated NH vector must lie on the intersection of two conic curves (Skrynnikov & Kay 2000, Wedemeyer et al. 2002, Wang & Donald 2004b). In (Wang et al. 2006, Wang & Donald 2004b,a), the authors proposed the first polynomial-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 two RDCs per residue (e.g., assigned NH RDCs in two media or NH and CH RDCs in a single medium). In (Wang & Donald 2004b,a), the authors also showed that given a peptide plane, the orientation of the next peptide plane can have at most 16 possible orientations; a related theorem was shown by (Hus et al. 2008). Given NH and CH RDCs in one medium, the associated NH and CH vectors can be solved from the RDC curve equations (see SM Section S2) and the protein backbone geometry (Wang & Donald 2004a, Wang et al. 2006). For the one-medium case, detailed proofs can be found in (Wang & Donald 2004b,a, Wang et al. 2006), and for completeness we provide a somewhat simpler derivation in SM Section S2. The derivation closely mirrors our new (open-source) software implementation, and the clearer equations therein are easier to interpret and build upon. For a detailed review of this method see (Donald & Martin 2008).

`RDC-EXACT` does not randomly search the conformation space to find solutions consistent with the RDC data. Rather, it formulates the problem such that the structures computed are *exact solutions* of a system of quartic monomial equations derived from the RDC equation. Hence, these roots, and therefore the conformations, are discrete, finite and algebraic. All dihedral angles for each residue are solved exactly from the quartic RDC equations. A depth-first systematic search algorithm is applied to search over all possible roots (conformations) to find an optimal solution for a SSE. The scoring function used in the depth-first systematic search contains RDC RMSD (namely the sum of the squared differences between experimental and back-computed RDCs over all RDCs in the SSE), Ramachandran suitability, van der Waals packing, and other common empirical molecular mechanics energy terms (Wang & Donald 2004b, Wang et al. 2006). As previously described in those references, the computed dihedral angles for a residue are not solely dependent on the local RDC information at that residue. Rather, the algorithm searches over all SSE residues and finds the global minimum of the scoring function for each SSE.

Before applying the `RDC-EXACT` algorithm to solve backbones for the core structure, we first identify the SSE boundaries based on the dihedral angle ranges computed from TALOS (Cornilescu et al. 1999) based on chemical shift information. When the TALOS dihedral intervals for a residue are within the favored or allowed Ramachandran area of  $\alpha$ -helix or  $\beta$ -strand, we consider this residue as part of that SSE. We note that other experimental data such as the J-coupling data from the NMR experiment HNHA (Vuister & Bax 1993) or the NOE patterns from automated



assignment (Bailey-Kellogg et al. 2000) of SSEs can also be used to determine the SSE boundaries.

We have extended the  $\text{RDC-EXACT}$  algorithm (Wang & Donald 2004b, Wang et al. 2006) to incorporate  $\text{C}^\alpha\text{C}'$  and  $\text{NC}'$  RDCs with CH and NH RDCs for the backbone calculation. The alignment tensor is estimated from NH and CH RDCs using the same approach as in (Wang & Donald 2004b, Wang et al. 2006). The  $\text{C}^\alpha\text{C}'$  and  $\text{NC}'$  RDCs are excluded in the alignment tensor computation, because they are in general noisier than NH and CH RDCs. In the new version of  $\text{RDC-EXACT}$ , NH and CH RDCs are first applied to compute and enumerate conformations as solutions to the polynomial systems derived from the RDC equations. Then the remaining  $\text{C}^\alpha\text{C}'$  and  $\text{NC}'$  RDCs are used to prune solutions for which the RMSD between back-computed and experimental RDCs is larger than thresholds 5.0 Hz (after being scaled to NH RDC). In addition, the outlier  $(\phi, \psi)$  angle solutions are pruned by intersecting the  $\text{TALOS}$  ranges with the Ramachandran regions. At this stage, every residue (except glycine and proline) in the backbone structure is replaced with alanine. Such a scheme is called *alanine replacement*. A *serious steric clash* is defined between atoms  $i$  and  $j$  when the distance between them satisfies  $d_{ij} < (r_i + r_j) - \varepsilon$ , where  $r_i$  and  $r_j$  are van der Waals radii, and  $\varepsilon$  is the overlap threshold between two atoms. Currently we choose  $\varepsilon = 0.5\text{\AA}$  for the steric clash checking. We prune those  $(\phi, \psi)$  solutions that result in serious steric clash after alanine replacement, and ensure that no serious clashes occur in our computed backbones. Later (below), we will replace the alanines with proper side-chain rotamers.

### SSE Packing from Sparse NOE Restraints

Once the conformations and orientations of individual SSE backbones have been solved by  $\text{RDC-EXACT}$ , their translations can be determined using a small number of inter-SSE NOE distance restraints (Fowler et al. 2000, Wang & Donald 2004b). Although the NOE restraint provides only an interval bound on the distance between atoms, in general a small number of NOE distance restraints can confine the translation into a bounded conformation space in which all discretized solutions within the parameterized resolution can be enumerated using a systematic grid search.

When packing a pair of SSEs solved from RDCs, we must consider the fourfold orientational ambiguity. Although the orientational ambiguity certainly can be resolved given a sufficient number of independent media, it cannot be resolved based only on RDCs in a single medium. Suppose that we pack all SSEs sequentially, that is, we first assemble the first two SSEs, and then pack their combination with the third SSE, and so on, as previously described (Wang & Donald 2004b). Before packing each pair of SSEs, sparse inter-SSE NOE restraints were extracted using chemical shift information alone (details are provided in SM Section S3). Using these sparse NOEs plus a van der Waals packing score, we can prune the orientational ambiguity (Georgiev et al. 2008, Potluri et al. 2006).

Since most inter-SSE NOEs involve side-chain interactions, we must consider all of the possible side-chain rotamer conformations when using these unambiguous NOE restraints in packing each pair of the SSE backbones. Let  $H_1$  and  $H_2$  denote two SSE backbones to be packed together. Let  $D$  be the set of inter-SSE NOE restraints between  $H_1$  and  $H_2$ , where for  $d_i = (h_{i1}, h_{i2}, \ell_i, u_i) \in D$ ,  $h_{i1}$ ,  $h_{i2}$  are the NOE-interaction protons in  $H_1$  and  $H_2$  respectively, and  $\ell_i$ ,  $u_i$  are the lower and upper bounds of the NOE. Our goal is to find all possible translations  $\mathbf{t} \in \mathbb{R}^3$  between  $H_1$  and  $H_2$ , such that there exists a pair of rotamers in  $H_1$  and  $H_2$  in which the distance between the corresponding protons  $h_{i1}$  and  $h_{i2}$ , denoted by  $d_i$ , satisfies the NOE restraint, namely  $\ell_i \leq d_i \leq u_i$ .

Without loss of generality, we choose the centers of  $H_1$  and  $H_2$ , denoted by  $a_0$  and  $b_0$ , as the representative points for  $H_1$  and  $H_2$  respectively. Then the translation  $\mathbf{t}$  between  $H_1$  and  $H_2$  can

be represented by the translation between points  $a_0$  and  $b_0$ , namely  $\mathbf{t} = a_0 - b_0$ . Let  $a$  be a proton of a residue in a particular rotamer state in  $H_1$  and  $b$  be a proton of a residue in a particular rotamer state in  $H_2$ . Suppose that an NOE  $(a, b, \ell_i, u_i)$  gives the distance restraint between proton  $a$  and  $b$ . Then we have

$$\ell_i \leq \|a - b\| \leq u_i. \quad (1)$$

Let  $\mathbf{t}_a$  be the vector from  $a_0$  to  $a$ , namely  $a = \mathbf{t}_a + a_0$ . Let  $\mathbf{t}_b$  be the vector from  $b_0$  to  $b$ , namely  $b = \mathbf{t}_b + b_0$ . Substituting into Equation (1), we have

$$\ell_i \leq \|\mathbf{t} - (\mathbf{t}_b - \mathbf{t}_a)\| \leq u_i. \quad (2)$$

Equation (2) restricts the translation  $\mathbf{t}$  to a spherical shell with center at  $\mathbf{t}_b - \mathbf{t}_a$  and radii of  $\ell_i$  and  $u_i$ . Let  $q$  be the number of inter-SSE NOE restraints. Given the  $i$ th NOE restraint, let  $\lambda_i$  be the total number of rotamer pairs between two corresponding residues. Let  $A_{ij}$  denote the spherical shell that represents the  $i$ th NOE restraint given the  $j$ th pair of rotamers at corresponding residues, where  $1 \leq i \leq q$  and  $1 \leq j \leq \lambda_i$ . Then the complete space of translation between  $H_1$  and  $H_2$  that satisfies all inter-SSE NOE restraints is represented by

$$\bigcap_{i=1}^q \bigcup_{j=1}^{\lambda_i} A_{ij}, \quad (3)$$

where  $q$  is the total number of NOE restraints, and  $\lambda_i$  is the total number of rotamer pairs between two corresponding residues.

We are interested in finding an ensemble of packed structures (within a parameterized resolution) that satisfy all NOE restraints, rather than just one single maximum-likelihood solution. Below we describe our algorithm for computing an ensemble of packed structures:

1. When packing each pair of SSE backbones, consider all four-fold symmetries of SSE orientations due to the symmetry of the dipolar operator reflected in the RDCs.
2. Apply a 3D grid search over relative translations  $\mathbf{t} \in \mathbb{R}^3$  with a resolution of  $0.2 \text{ \AA}$  to find all discrete translation points in which the set of solutions in Equation (3) is not empty.
3. Prune those packed structures containing steric clashes between atoms from difference SSE fragments.
4. Cluster all packed structures from Step (2) using an agglomerative hierarchical clustering algorithm (Han & Kamber 2006), in which two packed structures are allowed to be in a cluster only if their backbone RMSD is within  $0.4 \text{ \AA}$ . The centroids of all clusters form the final set of representative packed structures.

Our packing algorithm finds all of the discrete translation solutions within a parameterized resolution (i.e.  $0.4 \text{ \AA}$ ) that satisfy SSE orientations determined from RDCs and all inter-SSE NOE restraints. The time complexity analysis for  $P_{\text{ACKER}}$  is given in SM Section S4. In practice,  $P_{\text{ACKER}}$  runs in 30–60 minutes on a 3 GHz single-processor Linux workstation.

After we obtained the set of packed SSEs, which are called the *initial* packed structures, we computed a subset of *well-packed satisfying* (WPS) structures that had both high-quality van der Waals (vdW) score and good NOE satisfaction score (Potluri et al. 2006,2007). We used

a 6-12 Lennard-Jones potential to compute the packing score between SSEs, and a square-well potential (Brünger 1992) to calculate the NOE satisfaction score.

### NOE Assignment Using a Pattern Matching Technique

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 protons that are possibly assigned to an NOE cross peak if the resonances of corresponding atoms fall within a tolerance window around the NOE peak. We use error windows of 0.4 ppm for heavy atoms ( $^{15}\text{N}$  and  $^{13}\text{C}$ ), and 0.04 ppm for protons in the NOE assignment. In the rotamer selection phase (phase 2), we place all the side-chain rotamers of each residue into the backbone and compute all expected NOEs for protons within 6 Å apart. Based on the set of the expected NOEs and the resonance assignment list, we back-compute the expected NOE peak pattern for each rotamer. By matching the back-computed NOE pattern with the experimental NOE spectrum using an extended model of the Hausdorff distance (Huttenlocher & Jaquith 1995), we measure how well a rotamer fits the actual side-chain conformation when interpreted in terms of the NOE data (Fig. 1 C). We then select the top  $k$  rotamers with highest fitness scores at each residue, and obtain a “low-resolution” structure, which typically has approximately 2.0–3.0 Å (all heavy atom) RMSD from the reference structures solved by X-ray or traditional NMR approaches. The low-resolution structure is then used (in phase 3) to filter ambiguous NOE assignments. The details of the NOE assignment algorithm are provided in SM Section S5.

### NOE pattern matching based on the Hausdorff distance measure

The Hausdorff distance measures the closeness between two sets of points by computing the distance from the point in one set that is farthest from any point in the other set, and vice versa. Two sets of points have a small Hausdorff distance if every point in one set is close to some point in the other set. Thus, the Hausdorff distance is suitable for determining the degree of resemblance between two sets of points when they are superimposed on each other. The Hausdorff distance has been widely used in image processing and computer vision, e.g., visual correspondence, pattern recognition, and shape matching (Huttenlocher & Kedem 1992). 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 (Huttenlocher & Kedem 1992) 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 NOE peaks. The Hausdorff distance between  $B$  and  $Y$  can be measured by  $H(B, Y) = \max(h(B, Y), h(Y, B))$ , where  $h(B, Y) = \max_{b \in B} \min_{y \in Y} \|b - y\|$ . The notion of nested Min and Max in the Hausdorff definition identifies the point  $b \in B$  that is farthest from any point in  $Y$ , and measures the distance from point  $b$  to its closest neighbor  $y$  in  $Y$ . Thus, the Hausdorff distance  $H(B, Y)$  describes the discrepancy between the configurations of the two point sets  $B$  and  $Y$ , since it actually measures the distance from a point in  $B$  that is farthest from any point in  $Y$ , and vice versa. A review article on the Hausdorff distance can be found in (Huttenlocher & Kedem 1992).

We apply an extended model of Hausdorff distance (Huttenlocher & Kedem 1992, Huttenlocher & Jaquith 1995) to measure the match between the back-computed NOE pattern and experimental NOE spectrum. Below, we assume 3D NOE spectra without loss of generality. Given the back-computed NOE pattern  $B$  with  $m$  peaks, and the set of NOE peaks  $Y$  with  $w$  peaks, the  $\tau$ -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  $\tau$ th is the  $\tau$ -th largest of  $m$  values. We call  $s = \tau/m$  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 in chemical shift. The similarity score for a rotamer given  $\delta$  can be computed using a scheme similar to that of (Huttenlocher & Jaquith 1995):

$$s = \frac{|B \cap Y_\delta|}{|B|}, \quad (4)$$

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

Let  $(a_1, a_2, a_3, d)$  represent a *distance restraint* back-computed from a structure, where  $a_1$  and  $a_3$  are the involved protons in the structure,  $a_2$  is the heavy atom covalently bound to the proton  $a_1$ , and  $d$  is the distance between protons  $a_1$  and  $a_3$ . Let  $(p_1, p_2, p_3, I_p)$  denote an *experimental NOE peak* from a 3D NOESY spectrum, where  $p_1$  and  $p_3$  are frequencies of a pair of (unassigned) interacting protons,  $p_2$  is the frequency of the heavy atom covalently bound to the first proton, and  $I_p$  is the intensity of the cross peak. Let  $b_i = (\omega(a_1), \omega(a_2), \omega(a_3), I(d))$  denote the *back-computed NOE peak* for a distance restraint  $(a_1, a_2, a_3, d)$  back-computed from a structure, where  $\omega(a_j)$  is the assigned chemical shift of atom  $a_j$ ,  $1 \leq j \leq 3$ , and  $I(d)$  is the back-computed peak intensity of distance  $d$ . The equation  $I(d) = kd^{-6}$  is used to back-compute the peak intensity  $I(d)$  from the distance  $d$ , where the constant  $k$  is calculated using the same strategy as in (Mumenthaler et al. 1997). The *likelihood* for a back-computed peak  $b_i = (\omega(a_1), \omega(a_2), \omega(a_3), I(d))$  in the NOE pattern  $B$  to match an experimental NOE peak within the distance  $\delta$  in  $Y_\delta$  can be defined as

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

where  $(p_1, p_2, p_3, I_p)$  is the experimental NOE peak matched to the back-computed NOE peak  $(\omega(a_1), \omega(a_2), \omega(a_3), I(d))$  under the Hausdorff distance measure, and  $\mathcal{N}(|x - \mu|, \pi)$  is the probability of observing the difference  $|x - \mu|$  in a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ . We note that the normal distribution and other similar distribution families have been widely used to model the noise in the NMR data, e.g., see (Rieping et al. 2005) and (Langmead & Donald 2004).

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

$$s = \frac{1}{m} \sum_{i=1}^m \mathcal{N}_i(b_i). \quad (5)$$

When back-computing the NOE pattern for each rotamer, HANA also considers the stereospecific assignment ambiguity for  $H^\alpha$  in glycine, all  $\beta$ -methylene protons, and methyl groups in leucine and valine. HANA back-computes all possible NOE patterns resulting from the different possible stereospecific assignments for all protons in a residue, and chooses the stereospecific assignment with the best match score for each rotamer when compared vs. the experimental data.

We provide the detailed pseudocode for computing the similarity score and for <sub>HANA</sub> in SM Section S5. The time complexity analysis (see SM Section S6) shows that <sub>HANA</sub> runs in polynomial time. In practice, <sub>HANA</sub> runs in 1–2 minutes on a 3 GHz single-processor Linux workstation.

### Incorporation of Rotamer Probabilities into the Scoring Function

Different side-chain rotamers for each residue occur at different probabilities (Lovell et al. 2000). To consider the occurrence rates of different rotamers, the following inference is applied to extend our scoring function in Equation (5). Let  $X_j$  be the boolean proposition that the  $j$ -th rotamer is selected. Let  $\Pr(D|X_j)$  be the likelihood function that quantifies the likelihood of observing data  $D$  given the  $j$ -th rotamer. Then  $\Pr(D|X_j)$  is equivalent to the similarity score in Equation (5), that is,

$$\Pr(D|X_j) = \frac{1}{m} \sum_{i=1}^m \mathcal{N}_i(b_i). \quad (6)$$

By Bayes' Theorem, the posterior probability,  $\Pr(X_j|D)$  is given by

$$\Pr(X_j|D) \propto \Pr(D|X_j) \cdot \Pr(X_j) \propto \left( \frac{1}{m} \sum_{i=1}^m \mathcal{N}_i(b_i) \right) \Pr(X_j). \quad (7)$$

Equation (7) is used as the scoring function for computing the ensemble of rotamers that best fit the experimental data.

## Results

We tested <sub>RDC-PANDA</sub> on four proteins: the FF Domain 2 of human transcription elongation factor CA150 (RNA polymerase II C-terminal domain interacting protein) (FF2), the ubiquitin-binding zinc finger domain of the human Y-family DNA polymerase Eta (pol  $\eta$  UBZ), human ubiquitin, and the human Set2-Rpb1 interacting domain (hSRI). The lengths (number of amino acid residues) of these proteins are 62 for FF2, 39 for pol  $\eta$  UBZ, 76 for ubiquitin, and 112 for hSRI.

All NMR data except the RDC data of ubiquitin were recorded and collected using Varian 600 and 800 MHz spectrometers at Duke University. The NMR spectra were processed using the program <sub>NMRPIPE</sub> (Delaglio et al. 1995). All NMR peaks were picked by the programs <sub>NMRVIEW</sub> (Johnson & Blevins 1994) or <sub>XEASY/CARA</sub> (Bartels et al. 1995), followed by manual editing. Backbone assignments were obtained from the set of triple NMR experiments HNCA, HN(CO)CA, HN(CA)CB, HN(COCA)CB, and HNCO, combined with the HSQC spectra using the program <sub>PACES</sub> (Coggins & Zhou 2003), followed by manual checking. The side-chain resonances were assigned from the HA(CA)NH, HA(CACO), HCCH-TOCSY, and HC(CCO)NH-TOCSY spectra. The NOE cross peaks were picked from three-dimensional <sup>15</sup>N- and <sup>13</sup>C-edited NOESY-HSQC spectra. In addition, we removed the diagonal cross peaks and water artifacts from the picked NOE peak list. The NH and CH RDC data of FF2, pol  $\eta$  UBZ and hSRI were measured from a 2D <sup>1</sup>H-<sup>15</sup>N IPAP experiment (Ottiger et al. 1998) and a modified (HACACO)NH experiment (Ball et al. 2006) respectively. The C $^{\alpha}$ C' and NC' RDC data of FF2 were measured from a set of HNCO-based experiments (Permi et al. 2000). The CH and NH RDC data of ubiquitin were obtained from the Protein Data Bank (PDB ID: 1D3Z).

The solution structures of FF2, pol  $\eta$  UBZ, hSRI and ubiquitin have been solved using conventional NMR structure determination approaches (PDB ID of ubiquitin NMR reference structure: 1D3Z; PDB ID of FF2 NMR reference structure: 2E71; PDB ID of pol  $\eta$  UBZ NMR reference structure: 2I5O; PDB ID of hSRI NMR reference structure: 2A7O). In addition, the X-ray structure of human ubiquitin (PDB ID: 1UBQ) was available. We used these previously-solved NMR or X-ray structures as the reference structures for evaluating the structure calculation results from  $\text{RDC-P}_{\text{ANDA}}$ .

### Evaluation of SSE Backbones Determined from RDCs

$\text{RDC-EXACT}$  computed accurate backbones that are similar to the reference structures and satisfy the experimental RDC data. As shown in Table 1, each of the computed SSEs has a backbone RMSD 0.2–1.3 Å from the reference structure, and the back-computed RDCs have a RMSD 0.2–2.3 Hz for NH bond vectors and 0.3–2.0 Hz for CH bond vectors vs. experimental RDCs.

To examine the individual RDC deviation at each residue, we plotted the back-computed vs. experimental RDCs (Fig. 2 and SM Fig. S2). The plots of back-computed vs. experimental RDCs fit the diagonal line well for all four proteins, indicating good agreement between the computed RDCs and the experimental data. The back-computed RDCs of CH and NH bond vectors fit the experimental RDCs better than the  $C^{\alpha}C'$  and  $NC'$  RDCs, likely because  $C^{\alpha}C'$  and  $NC'$  RDCs contain more noise than CH and NH RDCs. Most back-computed RDCs are within deviation 0–2.0 Hz from the experimental RDCs. The maximum deviations (about 3–6 Hz) between back-computed and experimental RDCs occur in the short  $\beta$ -strands with length less than 3 residues, including residues K48–L50 in ubiquitin, and residues Q9–P11 and L18–P20 in pol  $\eta$  UBZ, which indicates that  $\text{RDC-EXACT}$  has difficulty in computing accurate orientations for these short structure fragments. When only a small number of RDCs are available for the short fragments, it is difficult for  $\text{RDC-EXACT}$  to calculate the orientations of fragments in the principal order frame (POF) accurately from the RDC equations. Moreover, short  $\alpha$ -helices or  $\beta$ -strands often contain proline as their terminal residues and hence do not contain the observed NH RDC in the proline residue. In addition, when a modified (HACACO)NH experiment (Ball et al. 2006) is used to measure RDCs, the CH RDC in the residue proceeding proline is also missing. This situation makes it more difficult to solve the backbone conformation. For example, in pol  $\eta$  UBZ only two NH RDCs and two CH RDCs are available for each  $\beta$ -strand (namely residues Q9–P11 and L18–P20) (since each strand contains one proline), which make it harder to calculate the accurate orientations and conformations for the  $\beta$ -sheet in pol  $\eta$  UBZ. The paucity of NH RDCs in the short  $\beta$ -strands might explain why the overall RDC RMSD of pol  $\eta$  UBZ is slightly larger than other three proteins (Table 1). However, as we will demonstrate below, such structure deviations in the short  $\beta$ -strands are manageable and did not significantly degrade our NOE assignment and structure calculation results.

### Quality of SSE Packing

When packing each pair of SSE backbones, all three other ambiguous symmetric orientations were pruned by the sparse unambiguous inter-SSE NOE restraints (extracted using chemical shift information alone) and van der Waals packing score. About 3–12 unambiguous NOE assignments were extracted from NOE spectra given only chemical shift information (Table 2). About 3–10% percent of the initial packed structures were chosen as WPS structures that have NOE satisfaction score less than 0.5 Å and packing score less than –10 kcal/mol (Fig. 4 and SM Fig. S3).

The ensemble of the initial packed structures have average backbone RMSD 1.2–1.7 Å to the mean structure, while the ensemble of the WPS structures have average backbone RMSD 0.6–1.4 Å to the mean structure (Table 2). The mean structure has backbone RMSD 0.9–1.9 Å for the initial packed structures and backbone RMSD 0.8–1.8 Å for WPS structures to the reference

structure, which indicates that the computation of WPS structures yields an ensemble of packed structures closer to the reference structure by improving the RMSD between the computed structures and the reference structure. Moreover, the selection of WPS structures increases the percentage of the structures that are within RMSD 2.0 Å to the reference structure in the chosen ensemble (Fig. 4 and SM Fig. S3, columns 2–3).

Fig. 3 shows the ensemble of WPS structure, and the overlay between the mean structure we computed vs. the reference structure for ubiquitin, hSRI and pol  $\eta$  UBZ. The ensemble of WPS structures (Fig. 3) fall in a cluster of structures with reasonably small coordinate variance. The small deviation from the mean structure to the reference structure indicates that our packing algorithm and computation of WPS structures are able to calculate sufficiently accurate core structures to support our subsequent NOE assignment and structure calculation (see below).

To check the NOE assignment performance resulting from the packed structures, we investigated the percentage of the SSE NOE assignments computed by HANA that agreed with the reference structure. We say an NOE assignment is *compatible* if it is consistent with the reference structure (namely the distance between the assigned pair of NOE protons in the reference structure satisfies the NOE restraint calibrated from peak intensity), otherwise we call it an *incompatible* NOE assignment. Then the percentage of compatible SSE NOE assignments is defined as the fraction of SSE NOE assignments (over all SSE NOE assignments) computed by HANA. As shown in Table 2, HANA computes more than 90% compatible NOE assignments for SSE regions, after placing the side-chains onto the packed backbones. Together these results show that our packing algorithm can assemble SSE backbones and compute sufficiently accurate core structures (within backbone RMSD 0.8–1.8 Å to the reference structure) such that they can then be effectively used for filtering ambiguous NOE assignments.

### Evaluation of Rotamers Computed by HANA

To assess the accuracy of rotamers selection from HANA, we studied the details of computed rotamers in HANA's low-resolution structure of ubiquitin. Our test on ubiquitin shows that HANA selects more than 70% of rotamers that are consistent with either the X-ray or NMR reference structure (SM Fig. S4, Fig. S5 and Fig. S7).

Since most aromatic side-chains are hydrophobic and located in the core of protein, the selection of correct aromatic rotamers plays an important role in filtering ambiguous assignment of long-range NOEs, and thus is crucial in calculating the accurate global fold. For ubiquitin, HANA chooses all correct aromatic rotamers (i.e., consistent with either the X-ray or NMR reference structure) (SM Fig. S4). Fig. 5 illustrates the structure overlay between aromatic rotamers computed by HANA and side-chains from X-ray and NMR reference structures, which verifies that our algorithm can choose the correct aromatic rotamers that are close to both X-ray and NMR reference structures.

Further tests show that HANA computes all accurate leucine rotamers that are consistent with either X-ray or NMR reference structures (SM Fig. S6). In addition, the number of consistent rotamers does not vary significantly with the backbone resolution (the variance is less than 10% of the consistent rotamers), which indicates that our rotamer selection algorithm is not sensitive to small variations in the backbone conformation (SM Fig. S7).

We classified all residues into two categories: *class I* residues, including valine, isoleucine, leucine, methionine, phenylalanine, tryptophan and cysteine; and *class II* residues, including remaining residues except alanine and glycine. Note that class I residues are generally more hydrophobic than class II residues. Alanine and glycine are excluded since they have only one conformation. Our investigation shows that HANA chose 90.0% correct rotamers (i.e., consistent

with either the X-ray or NMR reference structure) for class I residues, and 65.1% correct rotamers for class II residues. The reason that our algorithm finds fewer correct rotamers for class II residues than for class I residues is because more class II residues are located on the protein surface. A further investigation shows that about three quarters of class II residues in ubiquitin exhibit solvent-accessibility over 20 %, while more than 90 % of class I residues are buried inside the protein with solvent-accessibility below 20 %. It can be easily rationalized that more NOEs are generally observed for class I residues buried in the core of the protein than for those class II residues on the surface. Thus, class II residues generally have fewer observable NOE restraints to constrain the correct rotamers than class I residues do. On the other hand, as we will demonstrate in the next subsection, the wrong NOE assignments caused by these incorrect rotamers in class II residues are manageable and do not degrade the quality of our final structure ensemble.

### Evaluation of Calculated Structures

The final NOE assignment table computed by *HANA* was fed into the structure calculation software XPLOR/XPLOR NIH together with other experimental restraints including dihedral angle and hydrogen bond distance constraints. We used the same hydrogen bond and dihedral angle constraints as in computing the NMR reference structures, hence our structure calculation test should be fair to demonstrate the accuracy of our NOE assignment results. Compared with the calculation of the NMR reference structures, in which RDCs were incorporated along with NOE restraints only during the final structure calculation, here we only used RDCs to compute the initial backbone fold. From an algorithmic point of view, our final structure determination using NOEs but not RDCs can be considered a good control test of the quality of the NOE assignment. The structure calculation was performed in two rounds. For FF2, the atomic coordinates of the ensemble of top 20 structures with the lowest energies, using the NOE assignments computed by our algorithm, have been deposited into the Protein Data Bank (PDB ID: 2KIQ).

Table 3 summarizes the results on NOE assignment and final structure calculation for proteins ubiquitin, FF2, hSRI and pol  $\eta$  UBZ. The ensemble of computed lowest-energy structures for all three proteins converged to a cluster of structures with small coordinate deviations (Fig. 6). The average RMSD to the mean coordinates is within 0.31–0.63 Å for backbone and 0.61–1.24 Å for all heavy atoms. The long loops including residues 51–64 for ubiquitin, and residues 35–50 for hSRI in our structure ensembles exhibited slightly more disorder than other regions (Fig. 6, Column 1). Our calculated structures have only small deviations from idealized structure geometry, and the Ramachandran plots show that more than 90 % of backbone dihedral angles are in favored regions (Table 3), which indicates the good quality of our computed structures. The comparisons of our structures with the reference structures either from X-ray crystallography or traditional NMR approach show that our structures agree well with the reference structures (Fig. 6). For all four proteins, the mean structure of final top 20 structures with lowest energies has a backbone RMSD less than 1.3 Å and an all-heavy-atom RMSD less than 2.1 Å from the reference structure. This result indicates that our NOE assignment can provide a sufficient number of accurate distance restraints for the structure determination.

### Discussion

We have compared our Hausdorff-based algorithm with other metrics, such as a Bayesian metric (i.e., essentially changing lines 9/10 of SM Algorithm 1 to multiplication) and an RMSD-based metric. In the Bayesian metric, when more experimental peaks are observed around a back-computed peak (within the error window), the likelihood of this back-computed NOE peak is weakened rather than been enhanced. In our Hausdorff-based metric, only the



closest experimental peak within the error window is matched to the corresponding back-computed peak. It will therefore be more robust than the Bayesian metric, and less affected by noisy peaks. The RMSD-based metric can be biased when experimental peaks are missing (i.e., when no experimental peaks are observed within the error window of the back-computed NOE peak). This is because when experimental peaks are missing, the RMSD-based metric will find a closest experimental peak that incorrectly matches to the back-computed NOE peak. Tests on our ubiquitin data show that the Hausdorff-based measurement performs 6–12% better than the other two metrics, in terms of percentage of consistent rotamers and compatible NOE assignments. We believe that the Hausdorff-based metric is in general more robust to the noisy and missing peaks, which are common in the NMR data. Further discussions of the differences between our algorithm and previous approaches can be found SM Sections S2 and S5.

We note that recently a similar pattern-matching technique has been independently proposed in *ASCAN* (Fiorito et al. 2008) to compare the back-computed NOE pattern with the experimental NOE data for the side-chain resonance assignment. *ASCAN* (Fiorito et al. 2008) computes the initial fold from a subset of NOE assignments based on given backbone resonance assignments and a subset of highly confident side-chain resonance assignments, and then uses an iterative strategy to refine the side-chain assignment, NOE assignment, and structure calculation. Compared to *ASCAN* (Fiorito et al. 2008), our approach starts with an RDC-defined backbone and performs a systematic search for the rotamer selection, and thus is potentially a more robust approach for the structure determination. The Hausdorff-based pattern matching technique for NOE assignment, which we introduced in Zeng et al. (2008), also allows us to efficiently measure the similarity between the back-computed NOE patterns and the experimental spectra in the presence of noise and experimental uncertainty.

## Limitations and Extensions

Our approach assumes that dynamics can be neglected, although it has shown in recent studies that modest dynamics averaging can be tolerated, albeit with reduced accuracy in the calculation of the bond vector orientations (Ruan et al. 2008). When order parameters  $S^2$  are measured for the same bond vectors as the RDCs (e.g. using relaxation experiments), we can neglect the dynamics within the time scale of the dynamics measurements. Thus, we can heuristically assume that when  $S^2$  is sufficiently uniform (i.e. the core of the protein is largely rigid), then the dynamic averaging due to  $S$  in the RDC measurement is safe to tolerate for the structure determination.

Although our current implementation of *HANA* uses 3D NOE spectra, *HANA* is general and can be easily extended to higher-dimensional (e.g., 4D) NOE data (Coggins & Zhou 2008). In addition, it would be interesting to extend the current version of *HANA* for NOE assignment with missing resonance assignments. In principle, *HANA* can be extended to assign the NOEs with a partially assigned resonance list, as long as the back-computed NOE patterns with missing peaks can still support the identification of the accurate rotamers.

Our structure determination starts with the high-resolution core structure computed from RDCs. The loop regions are computed by a local minimization approach, which does not incorporate the RDC data into the structure calculation. Thus, the loop regions are less accurate than the SSE regions in our final structures (see Table 3 and Fig. 6). Currently we are developing efficient algorithms for computing the loop conformations that satisfy both NOE and RDC data.

## Conclusion

We have developed a novel algorithm that exploits the global RDC restraints for filtering ambiguous NOE assignments. We provided a novel approach for computing the initial structure

template for NOE assignment by exactly solving backbones from RDCs and systematically choosing rotamers based on NOE pattern matching. Our automated structure calculation framework extends previous work (Wang & Donald 2004b, Wang et al. 2006) on backbone calculation from RDCs to high-resolution structure determination, and introduces a systematic packing algorithm that finds the ensemble of all packed structures and considers all possible side-chain conformations consistent with all inter-SSE NOE restraints. We presented a new automated NOE assignment algorithm that simultaneously exploits the accurate high-resolution backbone computation from RDC data (Wang & Donald 2004b, Wang et al. 2006), the statistical diversity of rotamers from a rotamer library (Lovell et al. 2000), and the robust Hausdorff measure (Huttenlocher & Jaquith 1995) for comparing the back-computed NOE patterns vs. the experimental NOE spectra to choose accurate rotamers. Our algorithm computed the NOE assignments with high accuracy. Tests on NMR data for four proteins yielded accurate NOE assignment and structure calculation results, and suggest that  $\text{RDC-P}_{\text{ANDA}}$  can play an important role in high-throughput structure determination.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

We thank Dr. A. Yershova, Mr. J. Martin, Prof. J. Richardson, Prof. D. Richardson, Dr. S. Apaydin 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  $\eta$  UBZ NMR data. We thank Mr. D. Keedy, Prof. J. Richardson, Prof. D. Richardson and other members in the Richardson lab for helpful comments on our computed FF2 structures.

## Abbreviations used

NMR, nuclear magnetic resonance  
 ppm, parts per million  
 RMSD, root mean square deviation  
 HSQC, heteronuclear single quantum coherence spectroscopy  
 NOE, nuclear Overhauser effect  
 NOESY, nuclear Overhauser and exchange spectroscopy  
 RDC, residual dipolar coupling  
 PDB, Protein Data Bank  
 pol  $\eta$  UBZ, ubiquitin-binding zinc finger domain of the human Y-family DNA polymerase Eta  
 CH, C $^{\alpha}$ -H $^{\alpha}$   
 hSRI, human Set2-Rpb1 interacting domain  
 FF2, FF Domain 2 of human transcription elongation factor CA150 (RNA polymerase II C-terminal domain interacting protein)  
 POF, principal order frame  
 SA, simulated annealing  
 MD, molecular dynamics  
 SSE, secondary structure element  
 C', carbonyl carbon  
 WPS, well-packed satisfying  
 vdW, van der Waals  
 SM, Supplementary Material.

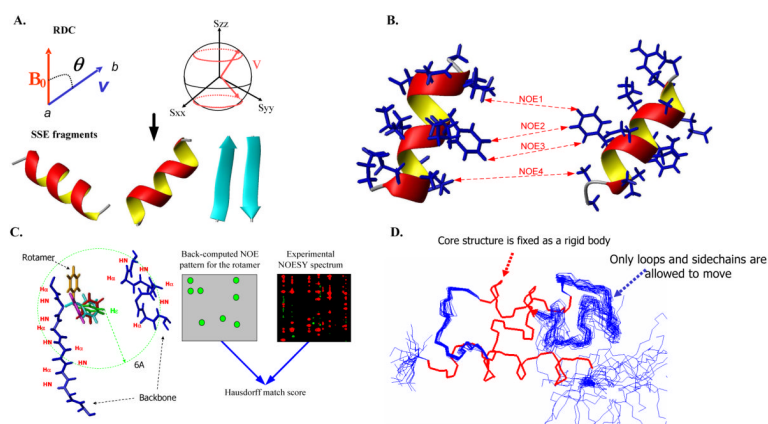
## References

- Andrec M, Du P, Levy RM. 'Protein backbone structure determination using only residual dipolar couplings from one ordering medium'. *Journal of Biomolecular NMR* 2004;21:335–347. [PubMed: 11824753]
- Bailey-Kellogg, C.; Widge, A.; Kelley, JJ.; Berardi, MJ.; Bushweller, JH.; Donald, BR. The noesy jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned nmr data. *Proceedings of the fourth annual international conference on Research in computational molecular biology*; 2000. p. 33–44. PMID: 11108478
- Ball G, Meenan N, Bromek K, Smith BO, Bella J, Uhrin D. 'Measurement of one-bond  $^{13}\text{C}^{\alpha}$ - $^1\text{H}^{\alpha}$  residual dipolar coupling constants in proteins by selective manipulation of  $\text{C}^{\alpha}\text{H}^{\alpha}$  spins'. *Journal of Magnetic Resonance* 2006;180:127–136. [PubMed: 16495100]
- Bartels C, Xia T, Billeter M, Güntert P, Wüthrich K. 'The program XEASY for computer-supported NMR spectral analysis of biological macromolecules'. *Journal of Biomolecular NMR* 1995;6:1–10.
- Bomar MG, Pai M, Tzeng S, Li S, Zhou P. 'Structure of the ubiquitin-binding zinc finger domain of human DNA Y-polymerase  $\eta$ '. *EMBO reports* 2007;8:247–251. [PubMed: 17304240]
- Brünger AT. 'X-PLOR, Version 3.1: a system for X-ray crystallography and NMR'. *Journal of the American Chemical Society*. 1992
- Chen CC, Singh JP, Altman RB. 'The hierarchical organization of molecular structure computations'. *J. Comp. Biol* 1998;5:409–422.
- Coggins BE, Zhou P. 'PACES: Protein sequential assignment by computer-assisted exhaustive search'. *Journal of Biomolecular NMR* 2003;26:93–111. [PubMed: 12766406]
- Coggins BE, Zhou P. 'High resolution 4-D spectroscopy with sparse concentric shell sampling and FFT-CLEAN'. *Journal of Biomolecular NMR* 2008;42:225C239. [PubMed: 18853260]
- Cornilescu G, Delaglio F, Bax A. 'Protein backbone angle restraints from searching a database for chemical shift and sequence homology'. *Journal of Biomolecular NMR* 1999;13:289–302. [PubMed: 10212987]
- Cornilescu G, Marquardt JL, Ottiger M, Bax A. 'Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase'. *Journal of the American Chemical Society* 1998;120:6836–6837.
- Davis IW, Murray LW, Richardson JS, Richardson DC. 'MolProbity: structure validation and all-atom contact analysis for nucleic acids and their complexes'. *Nucleic Acids Research* 2004;32:W615–W619. [PubMed: 15215462]
- Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. 'NMRPipe: a multidimensional spectral processing system based on UNIX pipes'. *Jour. Biomolecular NMR* 1995;6:277–293.
- Delaglio F, Kontaxis G, Bax A. 'Protein structure determination using molecular fragment replacement and NMR dipolar couplings'. *J. Am. Chem. Soc* 2000;122:2142–2143.
- Donald BR, Martin J. 'Automated NMR assignment and protein structure determination using sparse dipolar coupling constraints'. *Progress in NMR Spectroscopy*. 2008[Epub ahead of Print] doi: 10.1016/j.pnmrs.2008.12.001
- Fiorito F, Herrmann T, Damberger F, Wüthrich K. 'Automated amino acid side-chain NMR assignment of proteins using (13)C- and (15)N-resolved 3D [(1)H, (1)H]-NOESY'. *J Biomol NMR* 2008;42:23–33. [PubMed: 18709333]
- Fowler CA, Tian F, Al-Hashimi HM, Prestegard JH. 'Rapid determination of protein folds using residual dipolar couplings'. *Journal of Molecular Biology* 2000;304:447–460. [PubMed: 11090286]
- Georgiev I, Lilien RH, Donald BR. '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* 2008;29:1527–1542. [PubMed: 18293294]
- Gronwald W, Moussa S, Elsner R, Jung A, Ganslmeier B, Trenner J, Kremer W, Neidig K-P, Kalbitzer HR. 'Automated assignment of NOESY NMR spectra using a knowledge based method (KNOWNOE)'. *Journal of Biomolecular NMR* 2002;23:271–287. [PubMed: 12398348]
- Güntert P. 'Automated NMR Protein Structure Determination'. *Progress in Nuclear Magnetic Resonance Spectroscopy* 2003;43:105–125.

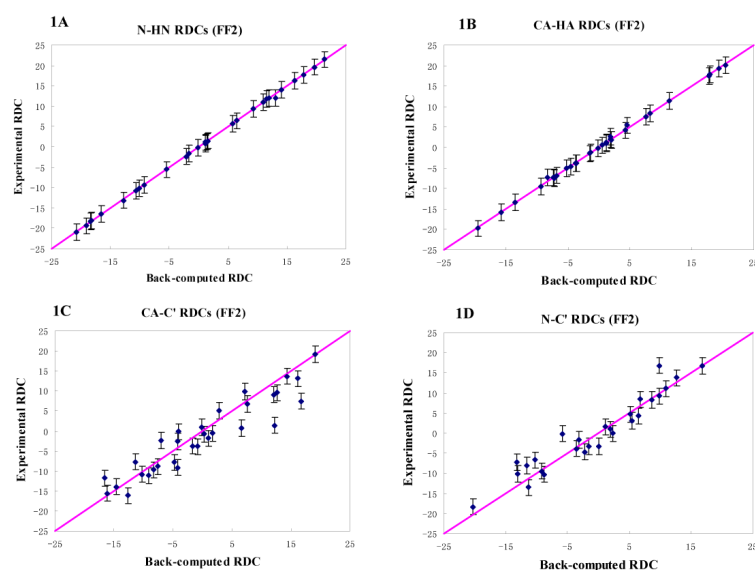
21. Han, J.; Kamber, M. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers; 2006.
22. Hayes-Roth, B.; Buchanan, B.; Lichtarge, O.; Hewitt, M.; Altman, R.; Brinkley, J.; Cornelius, C.; Duncan, B.; Jardetzky, O. PROTEAN: Deriving Protein Structure from Constraints. *Proceedings of the fifth national conference on artificial intelligence*; 1986. p. 904-908.
23. Herrmann T, Güntert P, Wüthrich K. 'Protein NMR Structure Determination with Automated NOE Assignment Using the New Software CANDID and the Torsion Angle Dynamics Algorithm DYANA'. *Journal of Molecular Biology* 2002;319(1):209–227. [PubMed: 12051947]
24. Huang YJ, Tejero R, Powers R, Montelione GT. 'A topology-constrained distance network algorithm for protein structure determination from NOESY data'. *Proteins: Structure Function and Bioinformatics* 2006;62(3):587–603.
25. Hus JC, Marion D, Blackledge M. 'Determination of protein backbone structure using only residual dipolar couplings'. *J. Am. Chem. Soc* 2001;123:1541–1542. [PubMed: 11456746]
26. Hus JC, Salmon L, Bouvignies G, Lotze J, Blackledge M, Brüschweiler R. '16-Fold Degeneracy of Peptide Plane Orientations from Residual Dipolar Couplings: Analytical Treatment and Implications for Protein Structure Determination'. *J. Am. Chem. Soc* 2008;130:15927–15937. [PubMed: 18959402]
27. Huttenlocher, DP.; Jaquith, EW. Computing visual correspondence: Incorporating the probability of a false match. *Proceedings of the Fifth International Conference on Computer Vision (ICCV 95)*; 1995. p. 515-522.
28. Huttenlocher, DP.; Kadem, K. Distance Metrics for Comparing Shapes in the Plane. In: Donald, BR.; Kapur, D.; Mundy, J., editors. *Symbolic and Numerical Computation for Artificial Intelligence*. Academic press; 1992. p. 201-219.
29. Johnson BA, Blevins RA. 'NMRView: a computer program for the visualization and analysis of NMR data'. *Jour. Biomolecular NMR* 1994;4:603–614.
30. Kuszewski J, Schwieters CD, Garrett DS, Byrd RA, Tjandra N, Clore GM. 'Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments'. *J. Am. Chem. Soc* 2004;126(20):6258–6273. [PubMed: 15149223]
31. Langmead C, Donald B. 'An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments'. *J. Biomol. NMR* 2004;29(2):111–138. [PubMed: 15014227]
32. Li M, Phatnani HP, Guan Z, Sage H, Greenleaf AL, Zhou P. '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* 2005;102:17636–17641.
33. Linge JP, Habeck M, Rieping W, Nilges M. 'ARIA: Automated NOE assignment and NMR structure calculation'. *Bioinformatics* 2003;19(2):315–316. [PubMed: 12538267]
34. Lovell SC, Word JM, Richardson JS, Richardson DC. 'The Penultimate Rotamer Library'. *Proteins: Structure Function and Genetics* 2000;40:389–408.
35. Mumenthaler C, Güntert P, Braun W, Wüthrich K. 'Automated combined assignment of NOESY spectra and three-dimensional protein structure determination'. *Journal of Biomolecular NMR* 1997;10(4):351–362. [PubMed: 9460241]
36. Ottiger M, Delaglio F, Bax A. 'Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra'. *J. Magn. Reson* 1998;138:373–378. [PubMed: 9571116]
37. Permi P, Rosevear PR, Annala A. 'A set of HNCO-based experiments for measurement of residual dipolar couplings in  $^{15}\text{N}$ ,  $^{13}\text{C}$ , ( $^2\text{H}$ )-labeled proteins'. *Journal of Biomolecular NMR* 2000;17:43–54. [PubMed: 10909865]
38. Potluri S, Yan AK, Chou JJ, Donald BR, Bailey-Kellogg C. 'Structure Determination of Symmetric Homo-oligomers by a Complete Search of Symmetry Configuration Space using NMR Restraints and van der Waals Packing'. *Proteins* 2006;65:203–219. [PubMed: 16897780]
39. Potluri S, Yan AK, Donald BR, Bailey-Kellogg C. 'A complete algorithm to resolve ambiguity for intersubunit NOE assignment in structure determination of symmetric homo-oligomers'. *Protein Science* 2007;16:69–81. [PubMed: 17192589]
40. Rieping W, Habeck M, Nilges M. 'Inferential Structure Determination'. *Science* 2005;309:303–306. [PubMed: 16002620]

41. Ruan K, Briggman KB, Tolman JR. 'De novo determination of internuclear vector orientations from residual dipolar couplings measured in three independent alignment media'. *Journal of Biomolecular NMR* 2008;41:61–76. [PubMed: 18478335]
42. Skrynnikov NR, Kay LE. 'Assessment of molecular structure using frame-independent orientational restraints derived from residual dipolar couplings'. *J. Biomol. NMR* 2000;18(3):239–252. [PubMed: 11142514]
43. Tian F, Valafar H, Prestegard JH. 'A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones'. *J Am Chem Soc* 2001;123:11791–11796. [PubMed: 11716736]
44. Tjandra N, Bax A. 'Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium'. *Science* 1997;278:1111–1114. [PubMed: 9353189]
45. Tolman JR, Flanagan JM, Kennedy MA, Prestegard JH. 'Nuclear magnetic dipole interactions in field-oriented proteins: Information for structure determination in solution'. *Proc. Natl. Acad. Sci. USA* 1995;92:9279–9283. [PubMed: 7568117]
46. Vijay-Kumar S, Bugg CE, Cook WJ. 'Structure of ubiquitin refined at 1.8 Å resolution'. *Journal of Molecular Biology* 1987;194:531–44. [PubMed: 3041007]
47. Vuister GW, Bax A. 'Quantitative J correlation: A new approach for measuring homonuclear three-bond  $J(\text{H}^N\text{H}^\alpha)$  coupling constants in  $^{15}\text{N}$ -enriched proteins'. *J. Am. Chem. Soc* 1993;115:7772–7777.
48. Wang, L.; Donald, BR. Analysis of a Systematic Search-Based Algorithm for Determining Protein Backbone Structure from a Minimal Number of Residual Dipolar Couplings. *Proceedings of The IEEE Computational Systems Bioinformatics Conference (CSB)*; Stanford CA. August, 2004; 2004a. p. 319-330.PMID: 16448025
49. Wang L, Donald BR. '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* 2004b;29(3):223–242.
50. Wang, L.; Donald, BR. '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; 2005. p. 189-202.PMID: 16447976
51. Wang L, Mettu R, Donald BR. 'A Polynomial-Time Algorithm for De Novo Protein Backbone Structure Determination from NMR Data'. *Journal of Computational Biology* 2006;13(7):1276–1288.
52. Wedemeyer WJ, Rohl CA, Scheraga HA. 'Exact solutions for chemical bond orientations from residual dipolar couplings'. *J. Biomol. NMR* 2002;22:137–151. [PubMed: 11883775]
53. Zeng, J.; Tripathy, C.; Zhou, P.; Donald, BR. A Hausdorff-Based NOE Assignment Algorithm Using Protein Backbone Determined from Residual Dipolar Couplings and Rotamer Patterns. *Proceedings of the 7th Annual International Conference on Computational Systems Bioinformatics*, Stanford CA; 2008. p. 169-181.ISBN 1752-7791. PMID: 19122773



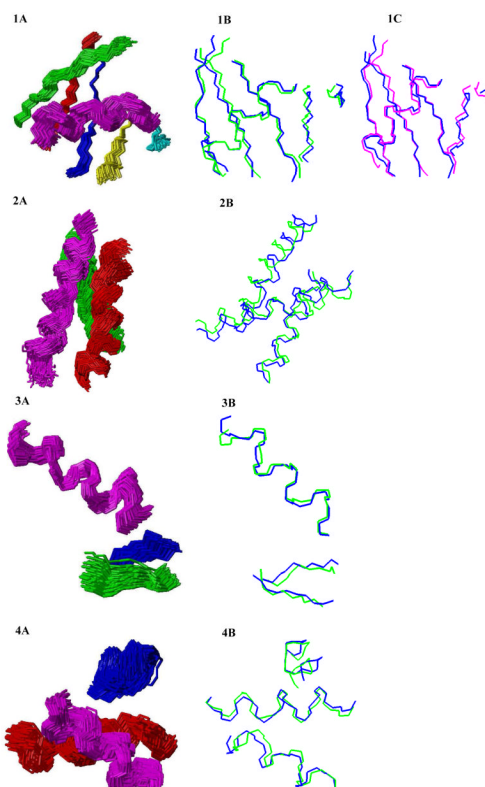


**Figure 1.** Schematic illustration of  $RDC-PANDA$ . Panel A:  $RDC-EXACT$ . Panel B:  $PACKER$ . Panel C:  $HANA$ . Panel D: the local minimization approach for loops.

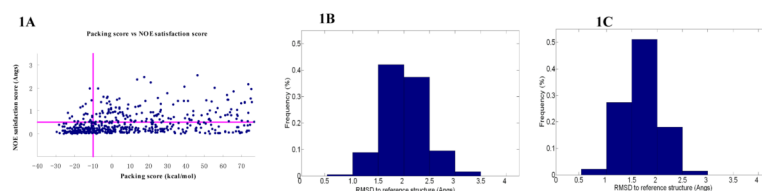


**Figure 2.**

Back-computed vs. experimental RDCs. Panels 1A, 1B, 1C and 1D: CH, NH,  $C^{\alpha}C'$  and NC' RDCs for FF2. All RDCs are scaled to the NH RDCs; a window of 2.0 Hz is shown as the error bars for the experimental RDCs. The plots of back-computed vs. experimental RDCs for ubiquitin, hSRI and pol  $\eta$  UBZ are shown in SM Fig. S2.

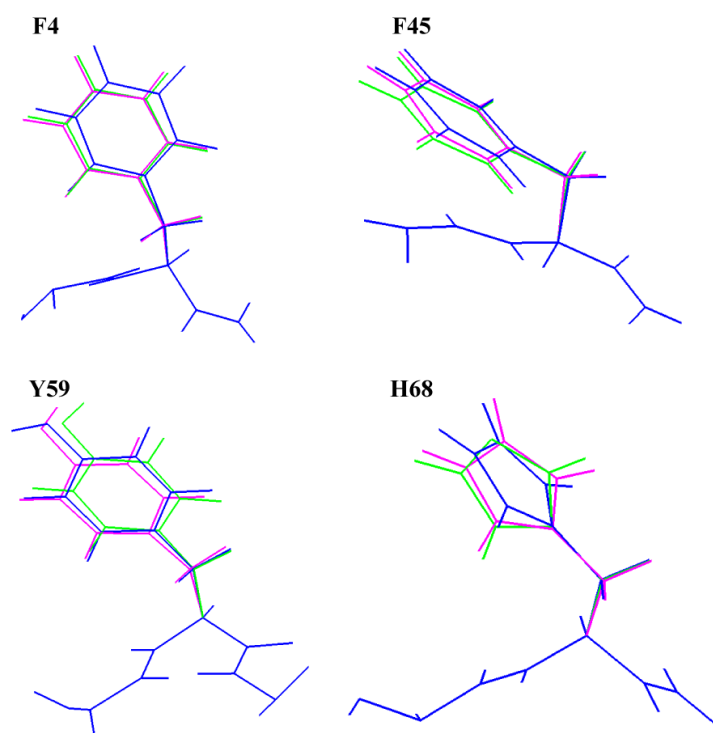


**Figure 3.**  
The SSE backbones of core structures computed by  $P_{\text{ACKER}}$ . Column 1: ensemble of WPS structures. Column 2: structure overlay of the mean WPS structure (blue) vs. the NMR reference structure (green). Column 3: structure overlay of the mean WPS structure (blue) vs. the X-ray structure (magenta).



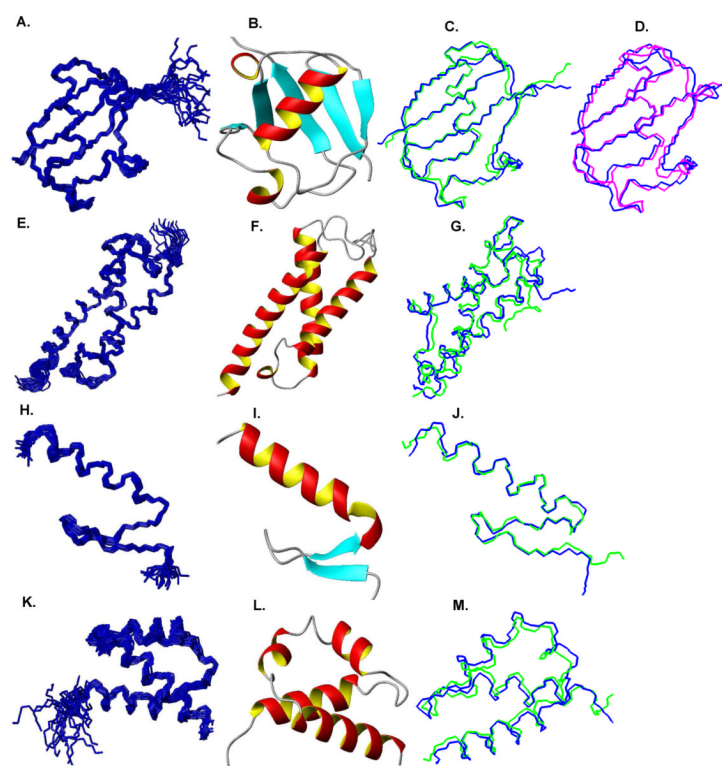
**Figure 4.**

Evaluation of packed structures computed by P<sub>ACKER</sub>. Here we show results for FF2. The results for ubiquitin, hSRI and pol  $\eta$  UBZ are shown in SM Fig. S3. Panel 1A: NOE satisfaction score vs. packing score for all structures in the ensemble (structures with vdW energies larger than 80 and NOE score larger than 10 were truncated from the plot). Panel 1B: histogram of backbone RMSD to the reference structure for all packed structures. Panel 1C: histogram of backbone RMSD to the reference structures for WPS structures. The magenta lines show the cutoffs of NOE satisfaction score (horizontal) and packing score (vertical) for computing the WPS structures.



**Figure 5.** Comparison of aromatic rotamers vs. side-chain conformations in the reference structures. Blue: rotamers computed by HANA. Magenta: X-ray side-chains. Green: side-chains in the NMR reference structure.





**Figure 6.**

The NMR structures of ubiquitin, hSRI, pol  $\eta$  UBZ and FF2 computed using our automatically-assigned NOEs. Panels A, B, C and D: the structures of ubiquitin. Panels E, F and G: the structures of hSRI. Panels H, I and J: the structures of pol  $\eta$  UBZ. Panels K, L and M: the structures of FF2. Panels A, E, H and K: the ensemble of 20 best NMR structures with the minimum energies. Panels B, F, I and L: ribbon view of the mean structures. Panel D: backbone overlay of the mean structures (blue) of ubiquitin vs. its X-ray reference structures (Vijay-Kumar et al. 1987) (magenta). Panels C, G, J and M: backbone overlay of the mean structures (blue) vs. corresponding NMR reference structures (green) (PDB ID of ubiquitin (Cornilescu et al. 1998): 1D3Z; PDB ID of FF2: 2E71; PDB ID of hSRI (Li et al. 2005): 2A7O; PDB ID of pol  $\eta$  UBZ (Bomar et al. 2007): 2I5O).

**Table 1**

Results on SSE backbones computed by  $\text{RDC-EXACT}$ . The backbone RMSD is reported in the format of SSE: RMSD. The RMSD between back-computed and experimental RDCs is computed using the equation  $\text{RMSD}_x = \sqrt{\frac{1}{m} \sum_{i=1}^m (r_{x,i}^b - r_{x,i}^e)^2}$ , where  $x$  indicates the RDC type, such as CH, NH, NC' or C $^\alpha$  C' RDC, and  $m$  is the number of RDCs,  $r_{x,i}^e$  is the experimental RDC, and  $r_{x,i}^b$  is the corresponding back-computed RDC. All RDCs have been scaled to the NH RDC.

Proteins	Backbone RMSD (Å) vs. NMR structure	Backbone RMSD (Å) vs. X-ray structure	RDC RMSD (Hz)
ubiquitin	$\alpha_1$ (E24-E34): 0.38; $\beta_{1,1}$ (Q2-T7): 0.30; $\beta_{1,2}$ (K11-V17): 0.30; $\beta_{1,3}$ (Q41-F45): 0.32; $\beta_{1,4}$ (K48-L50): 0.61; $\beta_{1,5}$ (S65-V70): 0.33.	$\alpha_1$ (E24-E34): 0.39; $\beta_{1,1}$ (Q2-T7): 0.38; $\beta_{1,2}$ (K11-V17): 0.32; $\beta_{1,3}$ (Q41-F45): 0.35; $\beta_{1,4}$ (K48-L50): 0.60; $\beta_{1,5}$ (S65-V70): 0.34.	CH: 1.05; NH: 1.42.
hSRI	$\alpha_1$ (A15-L34): 1.24; $\alpha_2$ (E51-C72): 0.69; $\alpha_3$ (E82-Q97): 0.44.	N/A	CH: 1.31; NH: 1.01.
pol $\eta$ UBZ	$\alpha_1$ (M24-E35): 0.64; $\beta_{1,1}$ (Q9-P11): 0.50; $\beta_{1,2}$ (L18-P20): 0.44.	N/A	CH: 1.98; NH: 2.27;
FF2	$\alpha_1$ (A8-E18): 0.66; $\alpha_2$ (F27-K33): 0.22; $\alpha_3$ (D48-A60): 0.38.	N/A	CH: 0.31; NH: 0.23; C $^\alpha$ C': 3.78; NC': 2.65.

Summary of SSE packing results (computed by P<sub>ACKER</sub>). For ubiquitin and pol  $\eta$  UBZ, hydrogen bonds were used to assemble  $\beta$ -strands into  $\beta$ -sheets (Wang & Donald 2004b). The percentage of compatible SSE NOE assignments is defined as the fraction of compatible SSE NOE assignments over all SSE NOE assignments computed by HANA.

Proteins	# of unambiguous NOEs used for packing	Total # of packed structures	Total # of WPS structures	Average backbone RMSD to mean structure (all, WPS) (Å)	Backbone RMSD between mean and reference structure (all, WPS) (Å)	Percentage of compatible SSE NOE assignments (%)
ubiquitin	3	1658	117	1.21, 0.69	To NMR structure: 0.96, 0.86 To X-ray structure: 1.00, 0.87	91.1
hSRI	12	2386	110	1.66, 1.32	To NMR structure: 1.84, 1.79	90.5
pol $\eta$ UBZ	7	1645	54	1.63, 0.79	To NMR structure: 1.12, 0.97	93.6
FF2	9	1472	139	1.36, 1.08	To NMR structure: 1.50, 1.31	91.1

**Table 3**

Results on  $\text{RDC-PANDA}$ 's NOE assignments and structures calculated from those assignments using  $\text{XPLOR}$ . A: Summary of NOE restraints from assignments computed by  $\text{RDC-PANDA}$ . B: Summary of structures calculated by  $\text{XPLOR}$  using the NOE assignments in (A). MolProbity (Davis et al. 2004) was used to assess the quality of our structures. The ordered regions are residues 2-70 for ubiquitin, residues 15-102 for hSRI, residues 9-35 for pol  $\eta$  UBZ and residues 8-60 for FF2. The MolProbity scores are reported only based on the ordered regions. We chose an ensemble of 20 structures with the lowest energies out of 50 final structures computed by  $\text{XPLOR/XPLOR-NH}$ . Compared with traditional NMR approaches, our NOE assignment table generally has more intraresidue NOEs. This is because in traditional NMR approaches using  $\text{D}_2\text{O}$  (Herrmann et al. 2002), redundant NOE restraints that could never be violated were removed, while the NOE table computed from  $\text{H}_2\text{O}$  included all these NOE restraints.

**A: Summary of NOE assignments computed by  $\text{RDC-PANDA}$** 

	ubiquitin	hSRI	pol $\eta$ UBZ	FF2
1. NOE restraints computed by $\text{RDC-PANDA}$	1331	3327	960	1070
Intraresidue	570	1193	429	422
Sequential ( $ i - j  = 1$ )	353	612	244	243
Medium-range ( $ i - j  \leq 4$ )	167	832	188	259
Long-range ( $ i - j  \geq 5$ )	242	690	99	146

**B: Summary of structures calculated from  $\text{RDC-PANDA}$ 's NOE assignments using  $\text{XPLOR}$** 

	ubiquitin	hSRI	pol $\eta$ UBZ	FF2
2. NOE violations ( $> 0.5 \text{ \AA}$ )	0	0	0	0
3. Other experimental restraints				
Hydrogen bonds	27	80	30	46
Dihedral angle restraints	61	178	74	96
4. Average RMSD to mean coordinates				
SSE region (backbone, heavy) ( $\text{\AA}$ )	0.28, 0.69	0.39, 0.85	0.29, 0.58	0.53, 1.11
Ordered region (backbone, heavy) ( $\text{\AA}$ )	0.36, 0.73	0.44, 0.86	0.31, 0.61	0.63, 1.24
5. Ramachandran plot				
Favored (%)	90.5	92.3	94.2	90.8
Allowed (%)	100	100	100	100
6. Deviation from idealized geometry				
Bonds ( $\text{\AA}$ )	$0.0013 \pm 0.0001$	$0.0013 \pm 0.0003$	$0.0014 \pm 0.0002$	$0.0009 \pm 0.0002$
Angles ( $^\circ$ )	$0.41 \pm 0.01$	$0.48 \pm 0.03$	$0.51 \pm 0.02$	$0.36 \pm 0.02$
Impropers ( $^\circ$ )	$0.22 \pm 0.02$	$0.35 \pm 0.02$	$0.35 \pm 0.02$	$0.22 \pm 0.01$
7. RMSD to X-ray structure				
SSE region (backbone, heavy) ( $\text{\AA}$ )	0.82, 1.49	N/A	N/A	N/A
Ordered region (backbone, heavy) ( $\text{\AA}$ )	0.96, 1.54	N/A	N/A	N/A
8. RMSD to NMR reference structure				
SSE region (backbone, heavy) ( $\text{\AA}$ )	0.75, 1.35	1.09, 2.09	0.47, 1.22	0.76, 1.57
Ordered region (backbone, heavy) ( $\text{\AA}$ )	0.85, 1.38	1.21, 2.09	0.54, 1.50	1.26, 1.97