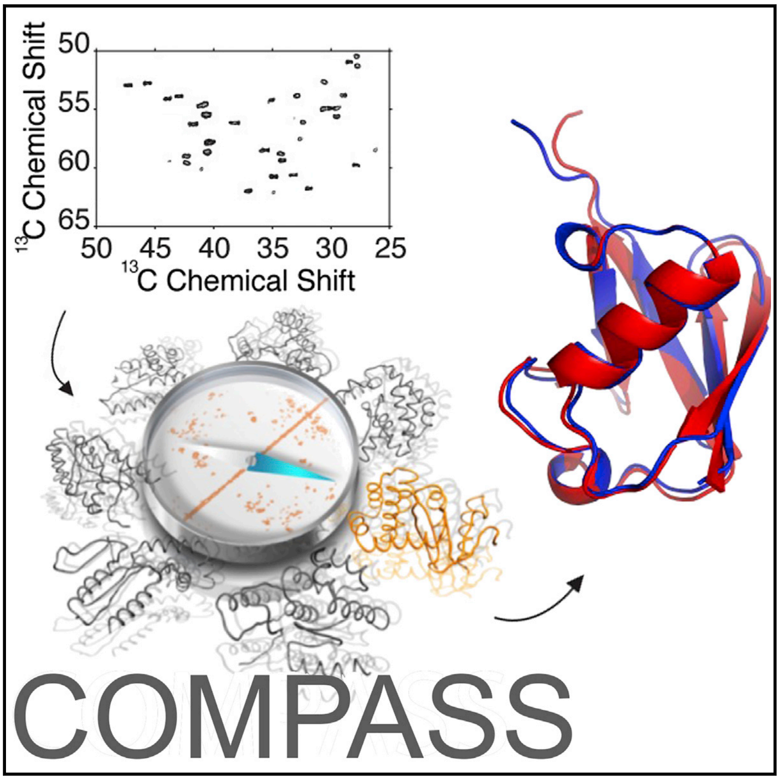
Structure

Resource

Experimental Protein Structure Verification by Scoring with a Single, Unassigned NMR Spectrum

Graphical Abstract



Authors

Joseph M. Courtney, Qing Ye,
Anna E. Nesbitt, ...,
Joseph R. Peterson,
James H. Morrissey, Chad M. Rienstra

Correspondence

rienstra@illinois.edu

In Brief

Courtney et al. have developed an algorithm that scores protein structural models against a previously unanalyzed NMR spectrum. This method, named COMPASS, does not require chemical shift assignments and identifies the correct structure in most cases within 1.5 Å RMSD of the reference structure.

Highlights

- An algorithm to numerically compare NMR spectra and protein structures is developed
- An unassigned ^{13}C-^{13}C NMR spectrum can be used to identify the correct protein fold
- Resonance assignments are not needed to use NMR data in structure development

Courtney et al., 2015, Structure *23*, 1958–1966
October 6, 2015 ©2015 Elsevier Ltd All rights reserved
http://dx.doi.org/10.1016/j.str.2015.07.019

Experimental Protein Structure Verification by Scoring with a Single, Unassigned NMR Spectrum

Joseph M. Courtney,[1] Qing Ye,[1] Anna E. Nesbitt,[1,4] Ming Tang,[1,5] Marcus D. Tuttle,[1] Eric D. Watt,[1,6] Kristin M. Nuzzio,[1] Lindsay J. Sperling,[1,7] Gemma Comellas,[2] Joseph R. Peterson,[1] James H. Morrissey,[3] and Chad M. Rienstra[1,2,3,*]
[1]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[2]Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[3]Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[4]Present address: School of Earth, Society, and Environment, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[5]Present address: Department of Chemistry, College of Staten Island, Staten Island, NY 10314, USA
[6]Present address: National Center for Computational Toxicology, Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA
[7]Present address: Department of Chemistry & Biochemistry, Santa Clara University, Santa Clara, CA 95053, USA
*Correspondence: rienstra@illinois.edu

http://dx.doi.org/10.1016/j.str.2015.07.019

SUMMARY

Standard methods for de novo protein structure determination by nuclear magnetic resonance (NMR) require time-consuming data collection and interpretation efforts. Here we present a qualitatively distinct and novel approach, called Comparative, Objective Measurement of Protein Architectures by Scoring Shifts (COMPASS), which identifies the best structures from a set of structural models by numerical comparison with a single, unassigned 2D ^{13}C-^{13}C NMR spectrum containing backbone and side-chain aliphatic signals. COMPASS does not require resonance assignments. It is particularly well suited for interpretation of magic-angle spinning solid-state NMR spectra, but also applicable to solution NMR spectra. We demonstrate COMPASS with experimental data from four proteins—GB1, ubiquitin, DsbA, and the extracellular domain of human tissue factor—and with reconstructed spectra from 11 additional proteins. For all these proteins, with molecular mass up to 25 kDa, COMPASS distinguished the correct fold, most often within 1.5 Å root-mean-square deviation of the reference structure.

INTRODUCTION

Nuclear magnetic resonance (NMR) is a powerful technique for studying protein structure and dynamics in near-native conditions. Substantial progress has been made in the solution of high-resolution protein structures by solid-state NMR (SSNMR) in the last decade. Structures previously inaccessible by solution NMR and X-ray crystallography, such as fibrils of the HET-s protein and amyloid-β, have been solved at atomic detail, offering insight into important biomedical problems (Wasmer et al., 2008; Lu et al., 2013). SSNMR approaches to solving structures of membrane proteins also have several notable successes (Shahid et al., 2012; Wang et al., 2013a; Park et al., 2012).

However, NMR methods, and SSNMR in particular, still require extensive sample preparation, data collection, and interpretation efforts. Typically, tens of milligrams of ^{13}C,^{15}N-labeled protein and several weeks of instrument time are required to collect the half a dozen or more 3D datasets necessary for the resonance assignments. Additional samples with sparse ^{13}C labeling and weeks of instrument time are needed to obtain a sufficient number of inter-residue distances to determine the fold uniquely (Comellas and Rienstra, 2013). Methods are in development to shorten the lengthy process of data collection, including non-uniform sampling (Paramasivam et al., 2012; Hyberts et al., 2010; Sun et al., 2012), proton detection with fast magic-angle spinning (MAS) (Knight et al., 2011; Zhou et al., 2012; Barbet-Massin et al., 2014), and combinations of these two approaches (Linser et al., 2014). Dynamic nuclear polarization is also a very promising method for accelerating data collection times, yet is usually not compatible with conditions that yield high-resolution spectra (Maly et al., 2008; Wang et al., 2013b; Renault et al., 2012).

In addition to challenges associated with data collection, the assignment and interpretation of spectra to yield a structure remain major bottlenecks and can take months of manual data analysis. Although methods are now available to automate the assignment process (Moseley et al., 2010; Güntert 2009; Guerry and Herrmann, 2011; Schmidt et al., 2013), these approaches still require complete sets of 3D data and extensive manual intervention. Once resonance assignments are available, methods such as CS-ROSETTA (Shen et al., 2008) and CHESHIRE (Cavalli et al., 2007; Robustelli et al., 2010) are available to leverage the chemical shift data for structure determination. These approaches have been highly successful; yet still require complete sets of site-specific resonance assignments. Therefore, there remains a compelling need for alternative methods that are faster and more cost-effective, requiring less sample, instrument time, and analysis. Combining NMR with advances in protein structure prediction (both homology modeling and ab initio methods)

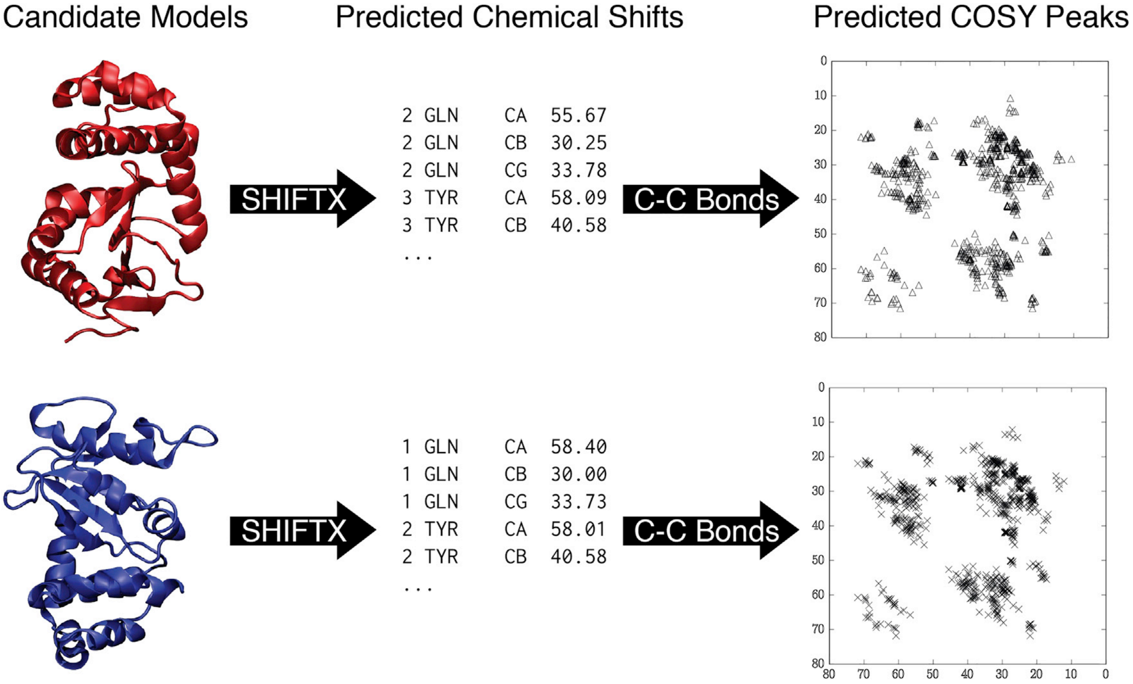


Figure 1. Prediction of ^{13}C-^{13}C Correlation Spectra from Protein Models with SHIFTX
The predicted chemical shifts are paired using a Python function that enumerates all directly bonded carbon pairs in the structure, and the corresponding chemical shifts are stored in a list without any assignment information. COSY, correlation spectroscopy.

offers a potential increase in efficiency (Simons et al., 1997; Eswar et al., 2002; Moult et al., 2014). This approach requires validation by comparing predicted NMR observables from the models with empirical or experimental data. In all prior methods, this has been done using sequence-specific resonance assignments.

Here we present a method, called Comparative, Objective Measurement of Protein Architectures by Scoring Shifts (COMPASS), which aims to extract structural information from NMR spectra by fully leveraging a limited amount of experimental data—one 2D ^{13}C-^{13}C spectrum—to accurately distinguish the correct protein fold from a set of proposed models. This avoids the lengthy structure determination process and requires no manual analysis of spectra. COMPASS solely employs the numerical comparison of predicted spectra from structural models, produced by various methods (e.g., homology modeling, molecular dynamics, ab initio quantum chemistry), with a single, unassigned 2D ^{13}C-^{13}C NMR spectrum, utilizing the dependence of chemical shifts upon protein conformation.

COMPASS leverages the accuracy of ^{13}C chemical shift prediction methods, and in this study we utilize SHIFTX2 (Han et al., 2011). For each protein, we collect a ^{13}C-^{13}C homonuclear correlation spectrum under conditions of scalar or dipolar mixing that yield exclusively one-bond correlations throughout the entire aliphatic region (Chen et al., 2006; Hohwy et al., 1999). Cross-peaks in this spectrum are enumerated and filtered according to a simple heuristic to generate a list of unassigned peaks. Meanwhile, a series of models are generated from the amino acid sequence using either homology or ab initio methods, and the ^{13}C chemical shifts are predicted for each model by SHIFTX2. Due to the simplicity and predictability of single-bond homonuclear correlation spectra, the hypothetical cross-peaks that would result from each model can be predicted (Figure 1). Then, using a scoring method based on the modified Hausdorff distance (Dubuisson and Jain, 1994) (see Figure 10), the models can be ranked according to their consistency with the experimental peak list. In the large majority of cases, the best model identified is consistent with the experimentally solved structure (see Figure 9).

RESULTS AND DISCUSSION

We selected 16 proteins, ranging in molecular mass from 6.6 to 33.6 kDa, to test COMPASS. For all selected proteins, high-quality structures of the monomeric form in the absence of any perturbing ligands are available in the PDB (Bernstein et al., 1977). 2D one-bond ^{13}C-^{13}C correlation spectra under solid-state conditions (MAS) were collected for four of these proteins: GB1, ubiquitin, DsbA, and the extracellular domain of human tissue factor (TF). For GB1, ubiquitin, and DsbA, constant-time, uniform-sign cross-peak correlation spectroscopy (CTUC-COSY) spectra were collected. For TF, we collected an SPC5 spectrum with a short mixing time to observe only one-bond transfers (Hohwy et al., 1999). Other pulse sequences that generate one-bond correlations could also be employed.

Automated Peak Filtering

Peaks were picked using the automated peak picking function of the Sparky NMR data analysis program (Goddard and Kneller,

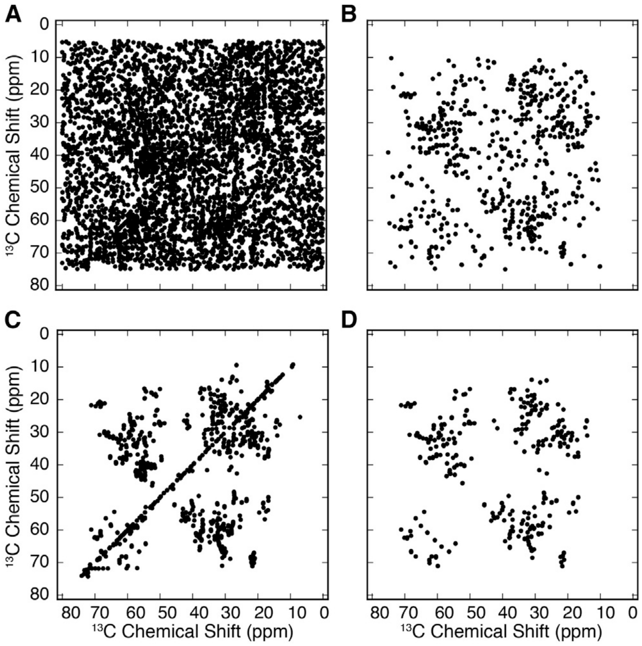


Figure 2. Peak Filtering Procedure

(A) Peaks automatically picked in the Sparky analysis program with a noise floor set at twice the root-mean-square (RMS) noise level.

(B) The same peaks after being filtered to exclude points near the diagonal and peaks without corresponding peaks opposite the diagonal.

(C) Peaks automatically picked with a noise floor set at six times the RMS noise level.

(D) The same data as (C), but filtered as in (B).

See also Figure S1.

2004). A range of noise floors was tested and an optimal minimum signal-to-noise ratio of 6 was chosen on the basis of testing shown in Figure S1. Peaks were then filtered to retain only those in the aliphatic region (0–80 ppm), at least 0.5 ppm away from the diagonal. The lists were then further filtered to retain only those peaks that were observed on both sides of the diagonal within a cutoff of 0.3 ppm (Figures 2B and 2D). This automated peak picking and filtering heuristic contributes significantly to the noise tolerance of COMPASS, as observed by the exclusion of the majority of the noise peaks even in a spectrum picked with a noise floor of twice the root-mean-square (RMS) noise (Figure 2B).

Evaluation of COMPASS Score

Next, we investigated the relationship between the scores of a group of models and their Cα RMS deviations (RMSDs) measured against the reference structure deposited in the PDB to test the behavior of the COMPASS score on models of differing accuracy. Figure 3 shows plots of the COMPASS score versus Cα RMSD for the four proteins with peak lists obtained directly from 2D spectra. For all four examples, models with lowest scores have low RMSDs. The obverse, however, is not always true. As can be seen, especially for GB1 (Figure 3A), many models with RMSD below 2 Å have scores greater than or equal to those models with RMSD >10 Å. This phenomenon occurs because the scores depend not only on the Cα-Cβ correlations, which report most strongly on secondary structure, but also on cross-peaks involving side-chain carbons, which report more strongly on the local environment (Han et al., 2011). Therefore, models with the correct side-chain conformations will agree best with the NMR data (i.e., exhibit the lowest scores). This behavior gives the COMPASS score a conservative character in that it rejects some models that have good coarse-grain structure but incorrect side-chain packing, while uniformly rejecting models with incorrect folds. Consistent with the score's sensitivity to side-chain conformation, there is a decreased correlation between the score and RMSD at higher RMSD values, since models with extremely different backbone structure but energetically optimized side chains are very unlikely to have conformations that would produce similar side-chain ^{13}C chemical shifts.

Overlays of the reference structure (red) with the model with the lowest score (blue) for each protein are shown in Figures 3E–3H. For all tested proteins, the bundle RMSD acts as a good surrogate for the actual RMSD from the true structure. When the bundle of five lowest-score structures had an acceptably small average pairwise RMSD, the consensus structure also had a low RMSD with respect to the reference structure (Figure 4).

We chose an additional 11 proteins with known structure and complete ^{13}C chemical shift assignments from the Biological Magnetic Resonance DataBank (BMRB) to test the performance of COMPASS on a wider range of structures (Ulrich et al., 2008). In lieu of raw spectra, we reconstructed peak lists from the known assignments using the same algorithm applied to predicting model peak lists. Although the sequence-specific assignments were available for these cases, the assignment information was not carried forward in the calculation.

The COMPASS score performed similarly well for most proteins in the synthetic dataset (Figures 5, 6, and 7). However, for the protein StR65, none of the models predicted by MODELLER had an RMSD below 10 Å. For this dataset, the COMPASS score exhibits the desirable quality that the five structures that agree most closely with the experimental data have an average pairwise RMSD of over 22.4 Å, providing an unambiguous indication that a consensus structure does not exist in the model set (Figures 7D–7F). As expected, if the set of models supplied to COMPASS does not contain any models that are consistent with the experimental data, a consensus structure cannot be identified.

In one case, a model with a low score but a high RMSD was observed. In this calculation on coactosin-like protein, a single model was generated with a Cα RMSD of 13 Å but had a COMPASS score comparable with much better models (Figure 7A). Upon manual inspection of the outlying model, it is clear that the majority of the secondary and tertiary structure elements are correct, but the model corresponds to a protein with two domains dissociated from each other, tethered by an unstructured loop. While this outlier did not perform as expected, its score is still well above that of the consensus, which agrees with the reference structure to within an RMSD of 0.72 Å. Manual inspection or the application of structure validation programs would easily identify this model as incorrect, enabling its removal from the structure pool.

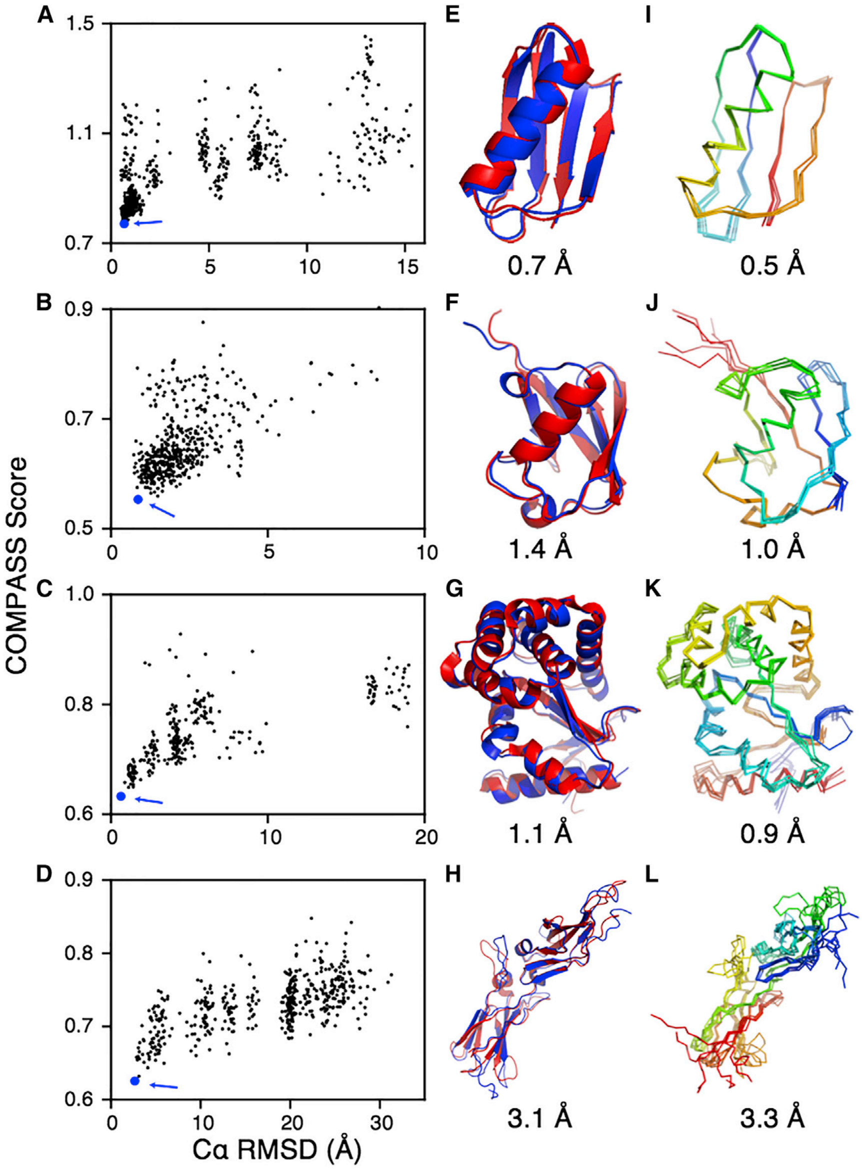


Figure 3. COMPASS Results for Four Proteins with Unassigned NMR Data

(A–D) COMPASS Score versus Cα RMSD from the reference structure for (A) GB1 (PDB: 2LGI), (B) ubiquitin (PDB: 1UBQ), (C) DsbA (PDB: 1FVK), and (D) TF (PDB: 1BOY). The structure with the lowest COMPASS score is shown in blue and indicated with an arrow.

(E–H) The structure with the lowest COMPASS score (blue) overlaid with the reference structure (red). The Cα RMS deviation (RMSD) is noted.

(I–L) The five lowest scoring structures aligned and overlaid. The average pairwise Cα RMSD is noted.

Application of COMPASS to Solution NMR Data

Although the COMPASS framework was developed to address the problems of spectral overlap and low sensitivity in NMR experiments, it does not rely on any special feature of SSNMR experiments. The performance of COMPASS on solution NMR data was tested by collecting ^{1}H-^{15}N HSQC (heteronuclear single-quantum coherence) and ^{13}C-^{13}C-^{1}H TOCSY (total correlation spectroscopy) spectra for a uniformly ^{13}C,^{15}N-labeled ubiquitin solution. The 3D TOCSY spectrum was projected through the ^{1}H dimension to generate a ^{13}C-^{13}C 2D spectrum.

The results for the HSQC comparison (Figure 8A) do not show a strong relationship between the COMPASS score and the RMSD. We attribute this result to the relative inaccuracy of chemical shift predictions for ^{15}N and ^{1}H amide resonances, due to the stronger dependence on hydrogen bonding and electrostatics, as well as backbone conformation and nearest neighbor residue type. For example, in contrast to the ^{13}Cα predictions which have an RMSD of 0.38 ppm (relative to known chemical shifts for a set of test proteins) (Han et al., 2011), amide ^{15}N predictions have an RMSD of 1.23 ppm, representing a 3-fold larger error over a similar range of chemical shifts (~30 ppm overall, or ~6–10 ppm for a given residue type). Moreover, the amide ^{1}H shifts have an RMSD of 0.24 ppm over a range of ~3 ppm. Thus the relative error in predicting a ^{1}H-^{15}N correlation spectrum is significantly higher than for ^{13}C-^{13}C spectra, leading in the case of ^{1}H-^{15}N to an inability to conclusively identify the best structure among a set, even for the relatively simple case of ubiquitin.

In contrast, the COMPASS scores for the projected ^{13}C-^{13}C-^{1}H TOCSY spectrum demonstrate a clear correlation and sharp convergence at a low RMSD value (Figure 8B), similar to the results observed for the solid-state NMR ^{13}C-^{13}C spectra, confirming that the strength of this method comes from its use of ^{13}C chemical shifts.

Conclusions

We present a new method for objective direct comparison of a modeled protein structure with experimental NMR data. COMPASS greatly reduces the time and effort required to validate a structure with experimental data by circumventing the lengthy process of chemical shift assignment and the collection of large datasets to obtain distance and orientation information required for de novo structure determination. The method is robust with respect to data collection and peak picking protocols, and has good tolerance for noise and artifacts. Here we have demonstrated successful calculations for 15 proteins, four with experimental SSNMR data, one with experimental solution NMR data, and ten with reconstructed spectra from the BioMagResBank chemical shift database.

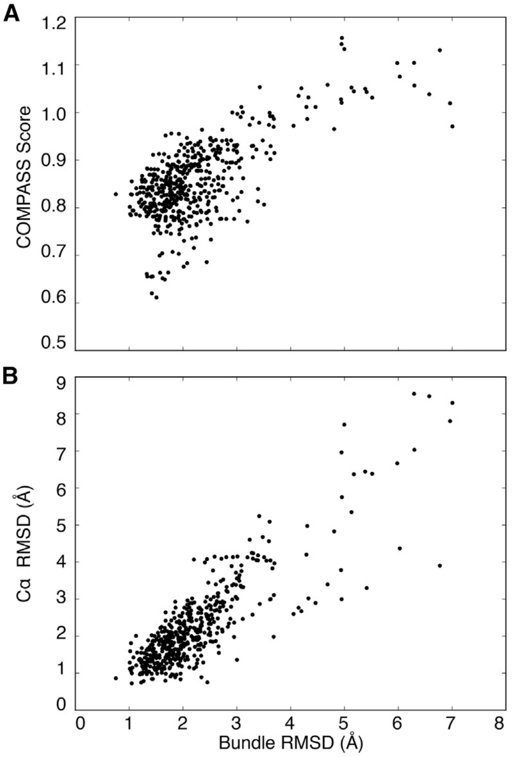


Figure 4. Ordered Bundle RMSD

Models are scored and ordered by the COMPASS scores. The bundle RMSD is the average RMSD of the four models with COMPASS scores closest to its own.

(A) COMPASS score versus bundle RMSD showing the "funneling" toward the origin, indicating a dataset containing a correct consensus structure.

(B) The bundle RMSD is highly correlated with the Cα RMSD to the correct structure, which enables its use as a surrogate when the true structure is unknown.

The COMPASS algorithm exploits the fact that the ^{13}C chemical shift is an exquisitely sensitive reporter on conformation, including not only backbone conformation as evidenced in the secondary chemical shifts (Spera and Bax, 1991), but also the conformation of side chains and packing in the protein core, which give rise to ring current and van der Waals packing effects. COMPASS leverages developments in chemical shift prediction methodology that take these effects into account. Strategies based on empirical models, homology methods, quantum mechanical calculations, and machine learning have progressively improved the accuracy, Here we used SHIFTX2 (Han et al., 2011), which uses a hybrid approach combining a sequence homology module with an ensemble machine-learning method to attain good accuracy for both backbone and side-chain atoms. SHIFTX2 attains prediction accuracy of better than 0.6 ppm for α, β, and carbonyl carbons and better than 1.0 ppm accuracy for most side-chain carbons. This level of prediction accuracy enables us to use the inherent sensitivity of ^{13}C chemical shifts to discern structural information from NMR data at a much earlier stage of analysis, and to quantitatively judge consistency of raw spectra with structural models. The rapid discrimination of valid protein folds by COMPASS may enable rational prioritization of subsequent data collection for structure refinement and acceleration of data analysis. For example, the experimentally consistent folds identified by COMPASS may be used to perform assignments of ambiguous correlations in spectra with long mixing times, reporting on long-range correlations.

As NMR is applied to systems of increasing complexity, manual data analysis becomes unfeasible. We envision potential future improvements including the application of COMPASS to 3D spectra, the use of the COMPASS score directly in model refinement and structure determination, as well as continued improvements in the accuracy of chemical shift prediction. In the current implementation only ^{13}C chemical shifts are used but, to accommodate the inclusion of higher dimensionality data, weighted aggregate scoring functions could be devised to account for differing chemical shift prediction accuracy of different nuclei.

While the combination of MODELLER and SHIFTX works well for the primarily monomeric, globular proteins presented here, the COMPASS algorithm could straightforwardly be

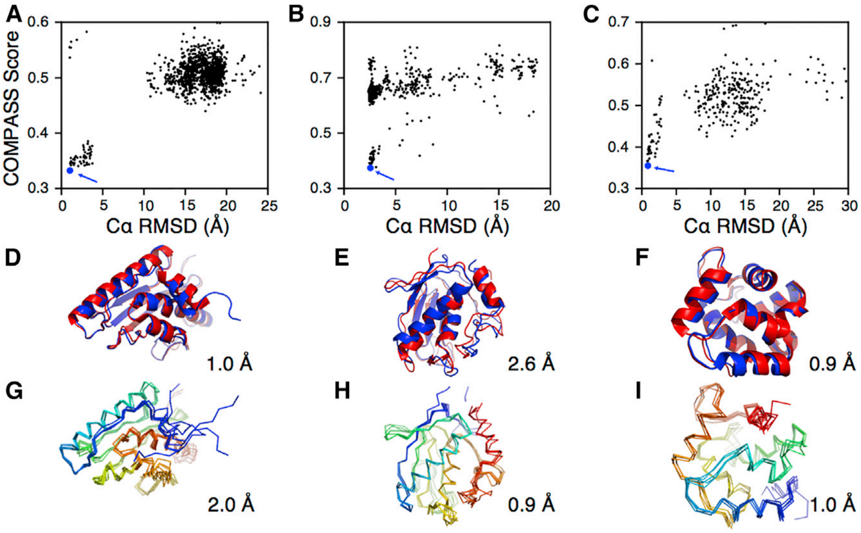


Figure 5. Additional COMPASS Results for Synthetic Peak Lists Constructed from BMRB-Deposited Chemical Shifts

(A–C) COMPASS score versus Cα RMSD from the reference structure for (A) Ufm1-conjugating enzyme 1 (PDB: 2Z6O), (B) macrophage metalloelastase (PDB: 2KRJ), (C) α-parvalbumin (PDB: 1RWY). The structure with the lowest COMPASS score is shown in blue and indicated with an arrow.

(D–F) The structure with the lowest COMPASS score (blue) overlaid with the reference structure (red). The Cα RMSD is noted.

(G–I) The overlay of five structures from each calculation with the lowest COMPASS scores. The average pairwise Cα RMSD is noted.

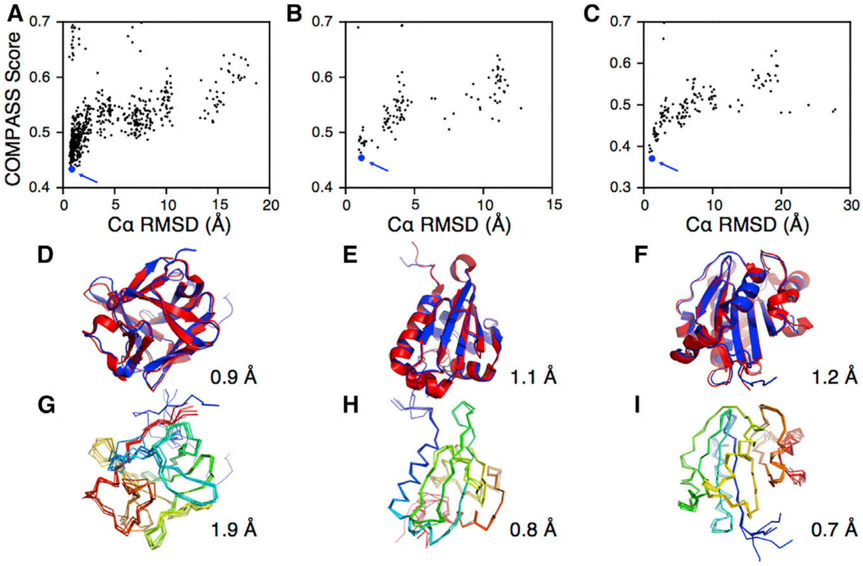


Figure 6. Additional COMPASS Results for Synthetic Peak Lists Constructed from BMRB-Deposited Chemical Shifts

(A–C) COMPASS score versus Cα RMSD from the reference structure for (A) Basic fibroblast growth factor (PDB: 1BFG), (B) sterol carrier protein 2 (PDB: 1C44), and (C) integrin α-L (PDB: 1XUO). The structure with the lowest COMPASS score is shown in blue and indicated with an arrow.

(D–F) The structure with the lowest COMPASS score (blue) overlaid with the reference structure (red). The Cα RMSD is noted.

(G–I) The overlay of five structures from each calculation with the lowest COMPASS scores. The average pairwise Cα RMSD is noted.

See also Figures S2 and S3.

extended to more specialized areas by using integrative structure prediction approaches for multimeric assemblies (Sali et al., 2015) and utilizing molecular dynamics averaged chemical shift predictions for dynamic loops (Robustelli et al., 2012). In addition, our assignment-free approach can be used to replace many chemical shift similarity-based potentials for structure refinement, and possibly in methods utilizing chemical shifts to develop models of structural ensembles (Kannan et al., 2014).

The continual progression in the quality of model prediction methods and chemical shift prediction algorithms will benefit COMPASS because of its modular approach. By leveraging these increasingly accurate predictions combined with the simple automated analysis of COMPASS, previously inaccessible systems will become feasible. These advances may be particularly significant to address categories of proteins, such as membrane proteins and fibrils, which have historically been very challenging.

EXPERIMENTAL PROCEDURES

The COMPASS framework can be applied to any combination of model-generation method and chemical shift prediction algorithm. In this study, models were prepared using the MODELLER protein structure-modeling program, using a standard protocol (Eswar et al., 2002), and subsequently relaxed using the ab initio relaxation function in the Rosetta software package to ensure low-energy side-chain conformations (Simons et al., 1997). SHIFTX2 was used to predict chemical shifts due to its speed and its applicability to both backbone and side-chain carbons.

To simulate the 2D spectra, a Python program enumerates all adjacent ^{13}C pairs, assembles the corresponding predicted chemical shifts into pairs, and records them in a list (Figure 9). The simulated peak list for each model is

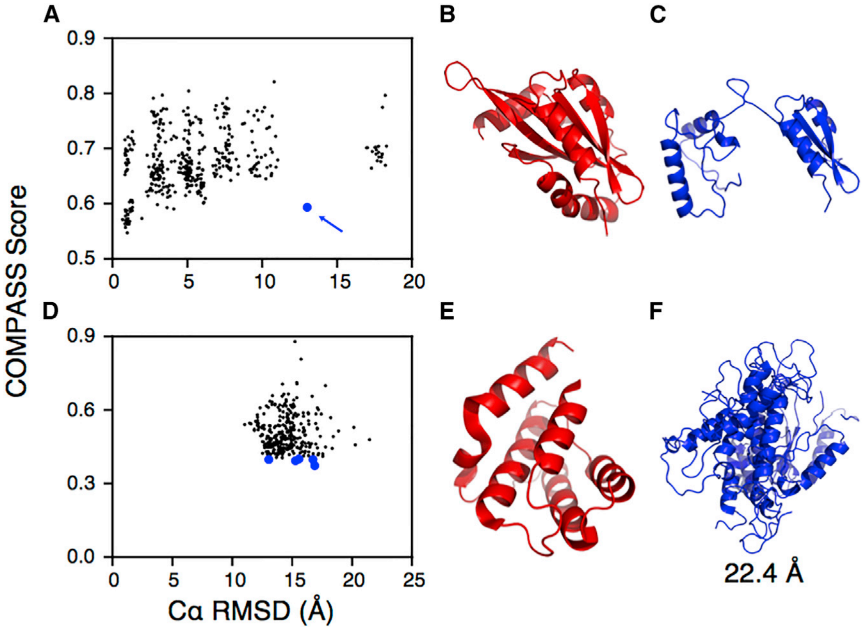


Figure 7. Behavior of the COMPASS Scoring Method when Applied to Incorrect Models

(A–C) Coactosin-like protein (A) COMPASS score versus Cα RMSD from PDB: 1T3Y. Point with anomalously low score is blue and noted with an arrow. (B) Structure from PDB: 1T3Y. (C) Structure of outlier model showing split structure.

(D–F) NorthEast Structural Genomics consortium target STR65 (D) COMPASS score versus Cα RMSD from PDB: 2ES9. Points with five lowest COMPASS scores are denoted by large blue dots. (E) Structure from 2ES9. (F) Aligned overlay of five lowest COMPASS score structures. Cα RMSD is noted.

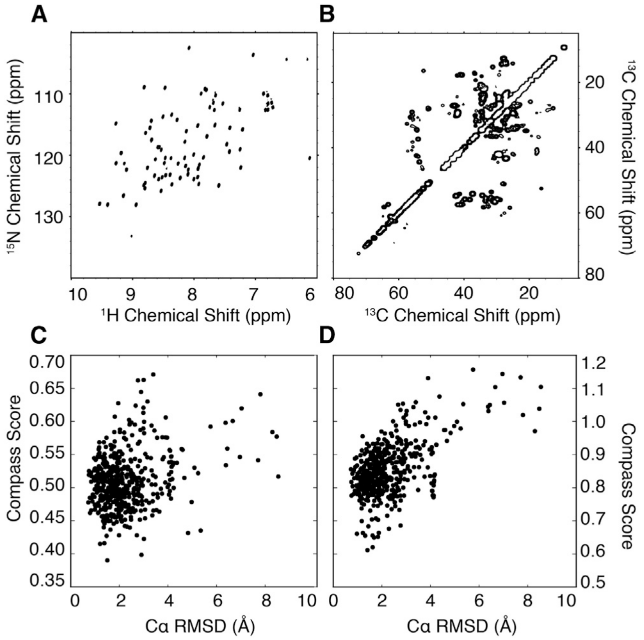


Figure 8. COMPASS Applied to Solution NMR Data of Ubiquitin

(A) SOFAST ^{1}H-^{15}N HSQC of ubiquitin.

(B) F3-projection of ^{13}C-^{13}C-^{1}H TOCSY of ubiquitin.

(C) COMPASS score versus Cα RMSD for ubiquitin using peaks from HSQC. Difficulty in predicting amide proton and nitrogen shifts makes it unsuited for use with the COMPASS algorithm.

(D) COMPASS score versus Cα RMSD for ubiquitin using peaks from TOCSY spectrum projection. Just as in SSNMR data, the COMPASS score based on ^{13}C-^{13}C correlations has a strong relationship with Cα RMSD, allowing its use in the determination of experimentally consistent data.

then compared with the experimental peak list using the COMPASS score, which is based on the modified Hausdorff distance. Hausdorff distances are a popular family of metrics in computational image analysis, and have found applications both in structure comparison and NOESY (nuclear Overhauser effect spectroscopy) peak matching (Zeng et al., 2008; Kozin and Svergun, 2001).

The COMPASS score is defined by Equations 1 and 2.

$$d(a,B)=\min_{b\in B}\|a-b\|, \qquad \text{(Equation 1)}$$

$$d_{COMPASS}(A,B)=\frac{1}{N_A}\sum_{a\in A}d(a,B). \qquad \text{(Equation 2)}$$

Equation 1 defines the distance between a point *a* and a point set *B* as the distance from point *a* to the closest point in set *B*. The COMPASS score is then defined in Equation 2 as the average of these minimum distances for every point in set *A*. This definition makes the COMPASS score directional, meaning that switching sets *A* and *B* gives different results. While this diverges from typical Hausdorff distances, it emphasizes the importance of the points in set A (chosen as the experimental peak set) over the points in set B (the predicted peaks). This way, every experimental peak is used in the calculation of the score but if the peak sets are very different, many of the predicted peaks (set B) may be ignored; for example, some regions of a protein may yield lower signal intensities experimentally.

The COMPASS score for each model is computed by matching each experimental peak with the nearest predicted peak in the model peak list, and calculating the average minimum distance for these pairings (Figure 10). The COMPASS score is therefore smaller for models that predict peak patterns similar to the experimental spectrum. In the limit of identical peak patterns, it would be identically zero. By weighting each experimental peak equally, the COMPASS score naturally addresses overlap and missing peaks in experimental spectra. If a peak is missing from the experimental spectrum, nearby peaks in the predicted spectrum are not matched and thus do not contribute to the overall score. Similarly, noise signals are deemphasized by the averaging procedure. Significant outliers that have no near matches in any model peak list contribute a similar magnitude to the scores of all models, manifesting as a nearly constant offset of all resulting scores.

Sample Preparation

The expression, purification, and crystallization of isotopically labeled recombinant ubiquitin was previously reported (Igumenova et al., 2004). The β1-immunoglobulin binding domain of protein G (GB1) was expressed and purified as previously reported (Franks et al., 2005). DsbA was expressed and purified according to the method of Sperling et al. (2010). Soluble TF

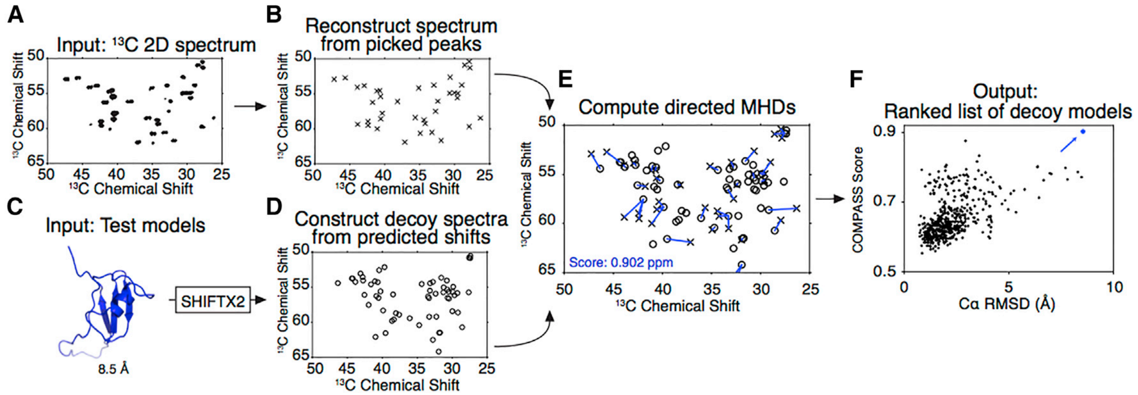


Figure 9. Flow Chart of the COMPASS Algorithm

(A) The algorithm takes as input a ^{13}C-^{13}C correlation spectrum. A selected region for a spectrum of ubiquitin is shown.

(B) The peaks are enumerated and stored as a list of unassigned chemical shift pairs.

(C) A collection of test models is produced. The model shown was generated by MODELLER and has a Cα RMSD of 8.5 Å with respect to the reference structure, PDB: 1UBQ.

(D) The chemical shifts for each model are predicted by SHIFTX2, and a list of peaks that would occur in a ^{13}C-^{13}C correlation spectrum is generated.

(E) The experimental and model peak lists are compared using the COMPASS score. Blue lines indicate the minimum distances described in the text.

(F) In this example the COMPASS score from the experimental peak list to the model is 0.902 ppm (point indicated with blue arrow), a relatively high value. The models are then ranked in the order of their computed COMPASS score.

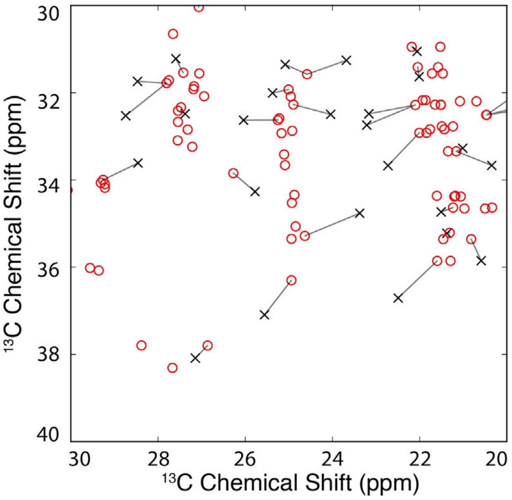


Figure 10. COMPASS Score Calculation

The COMPASS score is calculated by matching every experimental peak (black x) to the closest test peak (red circle) and calculating the average of the distances between them (gray line). A selected region from a comparison between a ubiquitin COSY spectrum and a poorly matching model is shown.

was expressed and purified as described by Boettcher et al. (2010) and crystallized by precipitation in 1.6 M ammonium sulfate with 200 mM NaCl and 100 mM HEPES buffer (pH 7.5) at 4°C as previously reported (Boys et al., 1993). Samples were packed into 3.2-mm thin-walled NMR rotors.

NMR Spectroscopy

The ^{13}C-^{13}C 2D CTUC-COSY spectrum of GB1 has been previously reported (Franks et al., 2005). The CTUC-COSY spectrum of ubiquitin was collected on a 750-MHz Varian VNMRS spectrometer (^{1}H frequency) with an HCN Balun MAS probe. The MAS rate was 16.666 kHz and the variable air temperature was set to -10°C. SPINAL decoupling (85 kHz) was employed during acquisition. The refocusing delay was 4.2 ms. The spectrum was processed with 20-Hz net line broadening in each dimension.

The CTUC-COSY spectrum of DsbA was collected on a 500-MHz Infinity Plus spectrometer (^{1}H frequency) spinning at 22.222 kHz at variable air temperature set point of -10°C. 85 kHz of ^{1}H SPINAL decoupling was employed during acquisition. 30-Hz net line broadening was applied in each dimension.

The ^{13}C-^{13}C 2D SPC5 spectrum of TF was collected on a 750-MHz Varian VNMRS spectrometer (^{1}H frequency) with an HCN BioMAS probe. The MAS rate was 12.500 kHz and the variable air temperature was set to 10°C. The SPINAL ^{1}H decoupling was employed at 80 kHz during the acquisition. The spectrum was processed with 20-Hz net line broadening in each dimension.

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures and can be found with this article online at http://dx.doi.org/10.1016/j.str.2015.07.019.

AUTHOR CONTRIBUTIONS

J.M.C., A.E.N., and C.M.R. conceived of the project and experiments. J.M.C., Q.Y., J.R.P., and A.E.N. designed and implemented algorithms. J.M.C., M.T., M.D.T., E.D.W., K.M.N., L.J.S., and G.C. executed NMR experiments. M.T., E.D.W., K.M.N., L.J.S., and J.H.M. prepared samples. J.M.C. and C.M.R. wrote the manuscript. All authors read and corrected the manuscript.

ACKNOWLEDGMENTS

This work was supported by NIH R01-GM073770 (to C.M.R.), R01-HL103999 (to J.H.M. and C.M.R.), and R21-107905 (to C.M.R.), and NIH S10RR025037 (to C.M.R). J.M.C. and K.M.N. were recipients of National Science Foundation Graduate Research Fellowships. A.E.N. was a recipient of an NIH Ruth L. Kirschstein National Research Service Award (F32 GM095344), and E.D.W. was an American Heart Association Postdoctoral Fellow. The authors thank Dr. Ying Li for expressing and purifying uniformly ^{13}C,^{15}N-labeled ubiquitin, and Dr. Deborah A. Berthold for preparing the uniformly ^{13}C,^{15}N-labeled WT-DsbA sample.

Received: May 8, 2015
Revised: July 22, 2015
Accepted: July 23, 2015
Published: September 10, 2015

REFERENCES

Barbet-Massin, E., Pell, A.J., Retel, J.S., Andreas, L.B., Jaudzems, K., Franks, W.T., Nieuwkoop, A.J., Hiller, M., Higman, V., Guerry, P., et al. (2014). Rapid proton-detected NMR assignment for proteins with fast magic angle spinning. J. Am. Chem. Soc. *136*, 12489–12497.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. *112*, 535–542.

Boettcher, J.M., Clay, M.C., LaHood, B.J., Morrissey, J.H., and Rienstra, C.M. (2010). Backbone ^{1}H, ^{13}C and ^{15}N resonance assignments of the extracellular domain of tissue factor. Biomol. NMR Assign. *4*, 183–185.

Boys, C.W.G., Miller, A., Harlos, K., Martin, D.M.A., Tuddenham, E.G.D., and O'Brien, D.P. (1993). Crystallization and preliminary X-ray analysis of human tissue factor extracellular domain. J. Mol. Biol. *234*, 1263–1265.

Cavalli, A., Salvatella, X., Dobson, C.M., and Vendruscolo, M. (2007). Protein structure determination from NMR chemical shifts. Proc. Natl. Acad. Sci. USA *104*, 9615–9620.

Chen, L., Olsen, R.A., Elliott, D.W., Boettcher, J.M., Zhou, D.H., Rienstra, C.M., and Mueller, L.J. (2006). High resolution (13)C-detected solid-state NMR spectroscopy of a deuterated protein. J. Am. Chem. Soc. *128*, 9992–9993.

Comellas, G., and Rienstra, C.M. (2013). Protein structure determination by magic-angle spinning solid-state NMR, and insights into the formation, structure, and stability of amyloid fibrils. Annu. Rev. Biophys. *42*, 515–536.

Dubuisson, M.P., and Jain, A.K. (1994). A modified Hausdorff distance for object matching. In Proceedings of the 12th International Conference on Pattern Recognition, *Vol. 1*, P. Storms, ed. (IEEE), pp. 566–568.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M., Pieper, U., and Sali, A. (2002). Protein structure modeling with MODELLER. Curr. Prot. Bioinform. *15*, 1–30.

Franks, W.T., Zhou, D.H., Wylie, B.J., Money, B.G., Graesser, D.T., Frericks, H.L., Sahota, G., and Rienstra, C.M. (2005). Magic-angle spinning solid-state NMR spectroscopy of the β1 immunoglobulin binding domain of protein G (GB1): ^{15}N and ^{13}C chemical shift assignments and conformational analysis. J. Am. Chem. Soc. *127*, 12291–12305.

Goddard, T.D., and Kneller, D.G. (2004). SPARKY 3 (University of California, San Francisco).

Guerry, P., and Herrmann, T. (2011). Advances in automated NMR protein structure determination. Q. Rev. Biophys. *44*, 257–309.

Güntert, P. (2009). Automated structure determination from NMR spectra. Eur. Biophys. J. *38*, 129–143.

Han, B., Liu, Y., Ginzinger, S.W., and Wishart, D.S. (2011). SHIFTX2: significantly improved protein chemical shift prediction. J. Biomol. NMR *50*, 43–57.

Hohwy, M., Rienstra, C.M., Jaroniec, C.P., and Griffin, R.G. (1999). Fivefold symmetric homonuclear dipolar recoupling in rotating solids: application to double quantum spectroscopy. J. Chem. Phys. *110*, 7983.

Hyberts, S.G., Takeuchi, K., and Wagner, G. (2010). Poisson-gap sampling and forward maximum entropy reconstruction for enhancing the resolution and sensitivity of protein NMR data. J. Am. Chem. Soc. *132*, 2145–2147.

Igumenova, T.I., McDermott, A.E., Zilm, K.W., Martin, R.W., Paulson, E.K., and Wand, A.J. (2004). Assignments of carbon NMR resonances for microcrystalline ubiquitin. J. Am. Chem. Soc. *126*, 6720–6727.

Kannan, A., Camilloni, C., Sahakyan, A.B., Cavalli, A., and Vendruscolo, M. (2014). A conformational ensemble derived using NMR methyl chemical shifts reveals a mechanical clamping transition that gates the binding of the HU protein to DNA. J. Am. Chem. Soc. *136*, 2204–2207.

Knight, M.J., Webber, A.L., Pell, A.J., Guerry, P., Barbet-Massin, E., Bertini, I., Felli, I.C., Gonnelli, L., Pierattelli, R., Emsley, L., et al. (2011). Fast resonance assignment and fold determination of human superoxide dismutase by high-resolution proton-detected solid-state MAS NMR spectroscopy. Angew. Chem. Int. Ed. Engl. *50*, 11697–11701.

Kozin, M.B., and Svergun, D.I. (2001). Automated matching of high- and low-resolution structural models. J. Appl. Crystallogr. *34*, 33–41.

Linser, R., Bardiaux, B., Andreas, L.B., Hyberts, S.G., Morris, V.K., Pintacuda, G., Sunde, M., Kwan, A.H., and Wagner, G. (2014). Solid-state NMR structure determination from diagonal-compensated proton-proton restraints. J. Am. Chem. Soc. *136*, 11002–11010.

Lu, J.X., Qiang, W., Yau, W.M., Schwieters, C.D., Meredith, S.C., and Tycko, R. (2013). Molecular structure of β-amyloid fibrils in Alzheimer's disease brain tissue. Cell *154*, 1257–1268.

Maly, T., Debelouchina, G.T., Bajaj, V.S., Hu, K.-N., Joo, C.-G., Mak-Jurkauskas, M.L., Sirigiri, J.R., van der Wel, P.C.A., Herzfeld, J., Temkin, R.J., and Griffin, R.G. (2008). Dynamic nuclear polarization at high magnetic fields. J. Chem. Phys. *128*, 052211, 1–19.

Moseley, H.N.B., Sperling, L.J., and Rienstra, C.M. (2010). Automated protein resonance assignments of magic angle spinning solid-state NMR spectra of B1 immunoglobulin binding domain of protein G (GB1). J. Biomol. NMR *48*, 123–128.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP)—round X. Proteins *82* (*Suppl 2*), 1–6.

Paramasivam, S., Suiter, C.L., Hou, G., Sun, S., Palmer, M., Hoch, J.C., Rovnyak, D., and Polenova, T. (2012). Enhanced sensitivity by nonuniform sampling enables multidimensional MAS NMR spectroscopy of protein assemblies. J. Phys. Chem. B *116*, 7416–7427.

Park, S.H., Das, B.B., Casagrande, F., Tian, Y., Nothnagel, H.J., Chu, M., Kiefer, H., Maier, K., De Angelis, A.A., Marassi, F.M., and Opella, S.J. (2012). Structure of the chemokine receptor CXCR1 in phospholipid bilayers. Nature *491*, 779–783.

Renault, M., Pawsey, S., Bos, M.P., Koers, E.J., Nand, D., Tommassen-van Boxtel, R., Rosay, M., Tommassen, J., Maas, W.E., and Baldus, M. (2012). Solid-state NMR spectroscopy on cellular preparations enhanced by dynamic nuclear polarization. Angew. Chem. Int. Ed. Engl. *51*, 2998–3001.

Robustelli, P., Kohlhoff, K., Cavalli, A., and Vendruscolo, M. (2010). Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. Structure *18*, 923–933.

Robustelli, P., Stafford, K.A., and Palmer, A.G., 3rd. (2012). Interpreting protein structural dynamics from NMR chemical shifts. J. Am. Chem. Soc. *134*, 6365–6374.

Sali, A., Berman, H.M., Schwede, T., Trewhella, J., Kleywegt, G., Burley, S.K., Markley, J., Nakamura, H., Adams, P., Bonvin, A.M.J.J., et al. (2015). Outcome of the first wwPDB Hybrid/Integrative Methods Task Force Workshop. Structure *23*, 1156–1167.

Schmidt, E., Gath, J., Habenstein, B., Ravotti, F., Székely, K., Huber, M., Buchner, L., Böckmann, A., Meier, B.H., and Güntert, P. (2013). Automated solid-state NMR resonance assignment of protein microcrystals and amyloids. J. Biomol. NMR *56*, 243–254.

Shahid, S.A., Bardiaux, B., Franks, W.T., Krabben, L., Habeck, M., van Rossum, B.J., and Linke, D. (2012). Membrane-protein structure determination by solid-state NMR spectroscopy of microcrystals. Nat. Methods *9*, 1212–1217.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K., Lemak, A., et al. (2008). Consistent blind protein structure generation from NMR chemical shift data. Proc. Natl. Acad. Sci. USA *105*, 4685–4690.

Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. *268*, 209–225.

Spera, S., and Bax, A. (1991). Measurement of NH-CaH coupling constants in staphylococcal nuclease by two-dimensional NMR and comparison with X-ray crystallographic results. J. Am. Chem. Soc. *113*, 5490–5492.

Sperling, L.J., Berthold, D.A., Sasser, T.L., Jeisy-Scott, V., and Rienstra, C.M. (2010). Assignment strategies for large proteins by magic-angle spinning NMR: the 21-kDa disulfide-bond-forming enzyme DsbA. J. Mol. Biol. *399*, 268–282.

Sun, S., Yan, S., Guo, C., Li, M., Hoch, J.C., Williams, J.C., and Polenova, T. (2012). A timesaving strategy for MAS NMR spectroscopy by combining non-uniform sampling and paramagnetic relaxation assisted condensed data collection. J. Phys. Chem. B *116*, 13585–13596.

Ulrich, E.L., Akutsu, H., Doreleijers, J.F., Harano, Y., Ioannidis, Y.E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., et al. (2008). BioMagResBank. Nucleic Acids Res. *36*, 402–408.

Wang, S., Munro, R.A., Shi, L., Kawamura, I., Okitsu, T., Wada, A., Kim, S.-Y., Jung, K.-H., Brown, L.S., and Ladizhansky, V. (2013a). Solid-state NMR spectroscopy structure determination of a lipid-embedded heptahelical membrane protein. Nat. Methods *10*, 1007–1012.

Wang, T., Park, Y.B., Caporini, M.A., Rosay, M., Zhong, L., Cosgrove, D.J., and Hong, M. (2013b). Sensitivity-enhanced solid-state NMR detection of expansin's target in plant cell walls. Proc. Natl. Acad. Sci. USA *110*, 16444–16449.

Wasmer, C., Lange, A., Van Melckebeke, H., Siemer, A.B., Riek, R., and Meier, B.H. (2008). Amyloid fibrils of the HET-s(218–289) prion form a beta-solenoid with a triangular hydrophobic core. Science *319*, 1523–1526.

Zeng, J., Tripathy, C., Zhou, P., and Donald, B.R. (2008). A Hausdorff based NOE assignment algorithm using protein backbone determined from residual dipolar couplings and rotamer patterns. Comput. Sys. Bioinform. Conf. *2008*, 169–181.

Zhou, D.H., Nieuwkoop, A.J., Berthold, D.A., Comellas, G., Sperling, L.J., Tang, M., Shah, G.J., Brea, E.J., Lemkau, L.R., and Rienstra, C.M. (2012). Solid-state NMR analysis of membrane proteins and protein aggregates by proton detected spectroscopy. J. Biomol. NMR *54*, 291–305.