Published in final edited form as:

Theor Chem Acc. 2013 October 1; 132(10): 1388–. doi:10.1007/s00214-013-1388-y.

Modeling helical proteins using residual dipolar couplings, sparse long-range distance constraints and a simple residue-based force field

Becky L. Eggimann,

Department of Chemistry, Chemical Theory Center, University of Minnesota, 207 Pleasant St. SE, Minneapolis, MN 55455, USA

Department of Chemistry, Wheaton College, 501 College Ave., Wheaton, IL 60187, USA

Vitaly V. Vostrikov,

Molecular Biology and Biophysics, University of Minnesota, 321 Church St. SE, Minneapolis, MN 55455, USA

Gianluigi Veglia, and

Department of Chemistry, Chemical Theory Center, University of Minnesota, 207 Pleasant St. SE, Minneapolis, MN 55455, USA

Molecular Biology and Biophysics, University of Minnesota, 321 Church St. SE, Minneapolis, MN 55455, USA

J. Ilja Siepmann

Department of Chemistry, Chemical Theory Center, University of Minnesota, 207 Pleasant St. SE, Minneapolis, MN 55455, USA

Abstract

We present a fast and simple protocol to obtain moderate-resolution backbone structures of helical proteins. This approach utilizes a combination of sparse backbone NMR data (residual dipolar couplings and paramagnetic relaxation enhancements) or EPR data with a residue-based force field and Monte Carlo/simulated annealing protocol to explore the folding energy landscape of helical proteins. By using only backbone NMR data, which are relatively easy to collect and analyze, and strategically placed spin relaxation probes, we show that it is possible to obtain protein structures with correct helical topology and backbone RMS deviations well below 4 Å. This approach offers promising alternatives for the structural determination of proteins in which nuclear Overhauser effect data are difficult or impossible to assign and produces initial models that will speed up the high-resolution structure determination by NMR spectroscopy.

Keywords

Structural determination; Helical proteins; NMR; Residual dipolar couplings; Dipolar waves; Paramagnetic relaxation enhancements; Simulated annealing

1 Introduction

Knowledge of three-dimensional structure is an essential step in discerning the biological function of proteins. As the number of available gene sequences continues to rise, there is greater need for methods designed to rapidly determine the protein structures. It has been speculated that in the long-term, most protein structures will be determined by structural models (from comparative modeling studies or structure prediction methods), rather than by experiment [1]. Indeed, the goal of the structural genomics initiative is to experimentally determine only those structures that represent unique folds, thereby creating a database representing approximately 90 % of all existing folds [2, 3]. With such a database in place, the vast majority of protein structures would be solvable by modeling techniques. A best-case scenario for the rapid determination of protein structures in the short-term would combine the strengths of current techniques in both computation and experiment. Carefully chosen experimental constraints can greatly reduce the amount of conformational space that must be sampled by computational search algorithms, making them both more efficient and more accurate [4]. In exchange, an accurate intermolecular force field can reduce the amount of experimental data that must be collected before a protein structure of sufficient accuracy can be determined.

An important consideration for any use of de novo structure prediction methods is the relatively low accuracy of predicted structures. Despite better sampling schemes and improved algorithms seen in the more recent Critical Assessment of Techniques for Protein Structure Prediction experiments [5, 6], prediction methods are still hindered by both a limited understanding of the protein folding process (which is necessary for the development of accurate force fields) and an incomplete fold database. Even with the best available algorithms, the correct overall topology can be obtained only about half of the time [7], with typical RMS deviations (RMSDs) between target and calculated structures falling in the 4–8 Å range [1].

In an effort to reduce the extensive conformational space that must be sampled by prediction algorithms, several methods have been developed to incorporate distance constraints as a means of biasing the possible conformations toward the native protein structure [4, 8–17]. Most of these methods are able to predict correct tertiary folds to a backbone C α RMSD of 3–5 Å given knowledge of the secondary structure and relatively few distances per secondary structural element. In a recent study, using TOUCHSTONEX, 1365 proteins representing the existing folds in the Protein Data Bank (PDB) were simulated using as few as one distance constraint for every eight residues [8]. The majority of structures (>80 %) were found with an RMSD < 6.5 Å. When considering the difficulty of the problem, the prediction results are really quite impressive, but when compared to experimental techniques, which typically achieve RMSDs on the order of 1 Å, there remains significant room for further improvement. Experimental techniques, in contrast, suffer from the burden of labor-intensive procedures and can only be considered high-throughput in favorable cases, i.e., when proteins form well-diffracting crystals or give high-resolution (well-dispersed) NMR spectra. Still, the high accuracy of experimental techniques recommends a closer consideration of the types of data that are available for use as constraints in prediction programs.

NMR spectroscopy is a powerful technique for the determination of the high-resolution structures of proteins under nearly physiological conditions. The basis for the structural determination is typically NOE-based, and for additional degrees of refinement, alternative approaches are employed. One of the popular methods, providing site-specific information on the orientation of the peptide planes, is measurement of residual dipolar couplings

(RDCs) [18–22]. RDCs depend on the orientation of specific protein bonds with respect to the magnetic field axis frame as follows [21]:

$$D^{mn}(\theta, \phi) = D_a^{mn} \left[3 \cos^2 \theta_{mn} - 1 + \frac{3}{2} R \sin^2 \theta_{mn} \cos 2\phi_{mn} \right] \quad (1)$$

where m and n represent the atoms of the bond vector for which RDCs are measured, θ_{mn} and ϕ_{mn} define the orientation of the mn bond vector with respect to the magnetic field axis frame, and D_a and R are the axial and rhombic components of the tensor describing the induced alignment of the protein. Several constant values, including the gyromagnetic ratios of atoms m and n and their inter-nuclear distance, comprise the D_a value.

While RDCs have added important new information, they have not generally been sufficient for structural determination in their own right. This is due in part to intrinsic degeneracies associated with the second rank tensor in Eq. 1 [23–25]. The small protein ubiquitin is a notable exception, in which several independent methods have been used to determine a high-resolution structure using exclusively RDCs [26–30]. Alternatively, RDCs have been used for rigid body docking of multiple proteins or of proteins with ligands [31, 32]. In practice, where limited data sets are the norm, complementary approaches are required, such as homology modeling [33] or the experimental acquisition of distance information. Distances derived from nuclear Overhauser effects (NOEs) are the primary source of these constraints [34, 35]. NOEs are generally considered a local constraint with an upper distance range of approximately 5–6 Å. They can be very powerful when constraining non-sequential residues, but less helpful when this does not occur. Furthermore, inter-domain NOEs in highly crowded spectra of helical proteins are difficult to assign. Finally, long-range NOEs are not routinely attainable for membrane proteins, large proteins or protein complexes, where faster relaxation times cause NOE cross-peaks to broaden and disappear.

Alternative distance constraints are known to be available from NMR and EPR experiments with site-directed spin labels [23, 36, 37] and recently gained popularity in providing long-distance restraints to the NMR structures [38–40]. Paramagnetic spin labels, such as (1-oxyl-1,1,5,5-tetramethylpyrroline-3-methyl) methanethiosulfonate (MTSSL), can be introduced at various locations along the protein backbone by site-specific mutagenesis and then adding the MTSSL probe onto the newly created cysteine side chain [36, 37]. The paramagnetic broadening of resonances observed in the heteronuclear correlation experiments such as [^1H – ^{15}N]-HSQC can be converted into distances between the spin label nitroxide and protein amide hydrogens [36, 37]. Such paramagnetic relaxation enhancement (PRE) arises from the interaction of the unpaired electron on the MTSSL probe with the amide hydrogens of each residue along the protein backbone proximal to the spin label. In EPR experiments, distance information is available when two or more spin labels are introduced along the protein chain. The electron–electron dipolar interactions between spin labels result in a distance-dependent broadening of the EPR spectrum that can be used to calculate distances between different parts of the protein [41]. The clear advantage of distance measurements from paramagnetic spin labels is their long-range character. In contrast to NOEs, PRE and EPR spin label distances range from 8 to as far as 35 Å. This means distances constraining parts of the protein far away in the sequence can be obtained more often with spin label experiments than with NOEs, though they will be longer-ranged and less restrictive.

There are different ways to incorporate RDCs into structural determination procedures, either as a standalone restraint or in combination with other methods [42]. Tjandra et al. [43] implemented RDCs as harmonic restraints into simulated annealing procedures to refine previously determined NOE-based protein structures. Prestegard et al. [44] exploited RDCs

to obtain the relative orientations of secondary structural domains in large proteins, where no long-range NOEs were available. Finally, Clore et al. [45] used RDCs to dock two proteins in large heterologous complexes.

An alternative approach is to analyze RDCs in terms of *dipolar waves*. This method was introduced by Opella and coworkers for the analysis of the relative orientation of secondary structural domains with periodic nature, such as helices and β -sheets [46–48]. In fact, RDCs originating from the alignment of regular secondary structures oscillate according to their intrinsic periodicity. By analyzing these oscillations as a function of the protein residue number, it is possible to identify the regular secondary structural elements and at the same time, to determine their relative orientations [46–48]. By exploiting the orientational dependence of the dipolar waves and using an exhaustive search algorithm, we were able to determine the topology of a small helical membrane protein, phospholamban, weakly aligned in stressed gels [49]. Also, Wang and coworkers using the peptide pixel approach demonstrated that it is possible to incorporate the intrinsic periodicity of RDCs into annealing protocols to derive the backbone structures of proteins supplemented with either sparse NOEs or ϕ/ψ angles [50–52]. More recently, Blackledge and coworkers have used dipolar waves to trace transient helical elements of partially folded proteins [53]. To date, however, the utility of dipolar waves for structural determination is still under investigation.

In this paper, we explore the possibility of folding of helical proteins using dipolar waves in concert with long-range distances from paramagnetic centers and a simple residue-based force field [54]. The secondary structural elements (helices) are treated as rigid bodies, and a hybrid Monte Carlo/simulated annealing algorithm is used to search the protein conformational landscape associated with packing and orientation of the helical units. This method requires low-level experimental input and short calculation times. We have critically evaluated both the applicability and the limitation of this approach for several different helical proteins with synthetic data (i.e., the pseudo-experimental data are generated from knowledge of the folded structure), including orientational and long-range distance constraints as well as combinations of both.

2 Methods

2.1 Protein representation

Each protein is represented using only the backbone atoms (C^α , C' , N), the amide hydrogen and the center of mass of the side chain. Side chain centers of mass are used as the interaction site for the force field, while the carbon atoms are needed to define the position of the backbone. Nitrogen and amide hydrogen atoms are included for direct evaluation of N–H RDCs. Bond angles and distances, including the side chain center of mass, are determined by reading the atom coordinates from the PDB file so that the helix retains the experimental (PDB) conformation. We note here that several tests were also performed using model helices built with the ideal helical structure (i.e., generic structures generated without knowledge of the experimental 3D structure for a specific protein); these tests resulted in the same trends observed while using PDB helix conformations, and data for multiple simulation methods are included in the Online Resource (Fig. 1; Tables 1, 2).

There are several methods available for determining the secondary structure of proteins, including an evaluation of NMR chemical shifts [55], the use of prediction methods [56, 57] or extraction of this information directly from RDCs [47]. The orientational dependence of RDCs is such that the intrinsic periodicity of secondary structural elements is encoded in the fluctuation of dipolar couplings with residue number [46–48]. Portions of the protein sequence that can be fit with a sine wave of specific periodicity (3.6 for α -helices and ≈ 2.0 for β -sheets) indicate the locations of the respective secondary structural types. For the

purposes of this work, target structures are limited to small helical proteins. The secondary structure is assumed to have been previously determined by one of these methods. Only helical segments of the protein are treated explicitly. Unstructured loop regions define an additional distance constraint such that consecutive helix ends are forced to be within a distance of less than $3.8 \text{ \AA}$ times the number of residues in the loop. Beyond that, loop regions are neglected; they do not add steric hindrance nor favor any interhelical orientation. As a further simplification, each helix is treated as a rigid body ignoring internal degrees of freedom and conformational changes. The force field that is used with this simplified protein model consists of three terms: a favorable contact energy between residue pairs and energy penalties for orientational and distance constraints. First, the ways that constraint data are incorporated is discussed, and then a complete description of the force field is given.

2.2 Dipolar wave constraints

The use of RDCs as constraint data in a dipolar wave framework requires that the helix orientation first be clearly defined. To this end, the orientation of each helix is described by three Euler angles, as shown in Fig. 1, using a “ZYZ” rotation convention and treating counterclockwise rotations as positive. The central axis of the helix is determined according to the method proposed by Kahn [58]. The relationship of helix orientation, defined by the central axis, with residual dipolar couplings has been presented previously for the spherical [21] and Cartesian [48] coordinate systems. The intrinsic degeneracy of RDCs manifests within dipolar waves as four degenerate possibilities for the Euler angles that describe the position of a helix: (θ, φ, ρ) , $(\theta, \varphi + \pi, \rho)$, $(\pi - \theta, \pi - \varphi, \rho + \pi)$, $(\pi - \theta, 2\pi - \varphi, \rho + \pi)$. A simple exhaustive search over possible angles will readily locate the four degenerate orientations [49]. Lifting this degeneracy requires that another set of RDCs obtained from a sufficiently different alignment (different D_a and R values signify a different overall orientation in the magnetic field) be used together with the first set.

To generate a list of dipolar coupling constraints for use in the simulations, Eq. 1 (which depends only on the orientation of N–H vectors with respect to the principal axis frame) is used to back-calculate a set of pseudo-experimental dipolar couplings from the positions of the nitrogen and amide hydrogen atoms listed in the PDB file. Values for the alignment tensor magnitude (D_a) and shape (R) are calculated using the Simulation of Sterically Induced Alignment tensor (SSIA) program [59] for the first RDC set. Tensor values for additional RDC sets, shown in Table 1, are chosen to differ in both rhombicity (R) and overall alignment (described by a set of alignment tensor Euler angles, α , β and γ , that describe the orientation of the whole protein with respect to the principle axis frame). The difference between successive alignment tensors is chosen to give low correlations between RDCs calculated for each set, making respective sets sufficiently different so as to eliminate the characteristic four-fold degeneracy.

2.3 Long-range distance constraints (PREs and DEER)

Two types of distance constraints are being considered: PREs and DEER constraints (i.e., dipolar distances between EPR spin labels). For distances involving spin labels, specific sites for the introduction of the nitroxide group were chosen by looking for positions least likely to disrupt the native structure. Because residues that are exposed to solvent can incorporate the nitroxide group with less structural disruption than residues buried within the protein core, we used a Kyte–Doolittle hydrophathy plot [60] to select hydrophilic residues as the sites for model spin label placement. We also selected residues that were close to helix ends when possible, reasoning that these would be the most restrictive distance constraints. Figure 2 shows an example of the positions chosen for PRE constraints in the protein with PDB code 1I6Z. From an NMR perspective, the degree to which spin labels change the native protein structure can be assessed by comparing chemical shift changes before and

after the spin labels have been introduced. Battiste and Wagner [36] found that MTSSL spin labels introduced at solvent-accessible sites on the protein eIF4E produced no major chemical shift differences. Furthermore, it has been noted that introducing MTSSL spin labels into the solvent-accessible helical segments of the protein does not generally perturb the protein structure [36, 61]. From a simulation perspective, the interaction energy from the force field can be compared for the same protein with and without spin labels. If the replacement of the native side chain with a spin label has not made significant changes to the protein energy, then it is unlikely that the structure has changed. For each of the six proteins studied here, the changes in interaction energy were found to be small when the spin labels were placed on hydrophilic residues (see “Results and discussion” section).

The actual position of the highly flexible spin label was modeled as an average structure in a manner similar to that reported previously [41]. The average distance from the C^α atom to the nitrogen atom of the MTSSL nitroxide was set to 6.7 Å, with the C^α–N vector passing through the center of mass of the residue side chain. This specific location was chosen for simplicity, but a distribution of orientations could be modeled by introducing another degree of freedom. A list of distance constraints for each protein was generated from the PDB file by replacing the side chain centers of mass of the appropriate hydrophilic residues with the position of the nitroxide nitrogen. EPR constraints were taken from the distances between the nitrogen atoms of one spin label to the nitrogen atoms of all other spin labels on different helices. PRE constraints were taken from the distances from each MTSSL nitrogen to the amide hydrogens of all residues on the other helices. It should be noted that this computational study assumes an ensemble of identically labeled proteins, whereas experimental situations may result in incomplete labeling. Finally, only distances that fell between the limits of 8 and 25 Å were kept. Distances as large as 35 Å have been observed, but the accuracy of the longest distances (>25 Å) is less certain and these were not included here [43, 62].

2.4 Force field

The primary challenge of this study is to quickly calculate the relative energy of several protein conformations given an array of possible constraints. This requires a force field that is both simple and moderately accurate, one that can account for solvent effects and other important contributions to the stability of a protein structure implicitly. A simple square-well potential function was chosen with well-depth energies derived from a statistical study of more than 1,100 proteins in the Protein Data Bank [63]. In the original work, Miyazawa and Jernigan [63] (MJ) considered the average number and type of contacts between different amino acid pairs. This was then related to a statistical (free) energy value (based on a Boltzmann inversion of probabilities) for specific pair interactions that naturally incorporates solvent effects. Use of the MJ contact energies has also been noted to be successful in previous implementations of simple force fields for predicting protein folds [64]. In view of this success, a simple force field based on the contact energies was used to describe the energy of interaction between two interhelical residues (Eq. 2):

$$E_{\text{int}}^{ij} = \begin{cases} \infty & \text{if } r_{ij} < 4.15 \\ \varepsilon_{ij} & \text{if } 4.15 \leq r_{ij} \leq 6.50 \\ 0 & \text{if } r_{ij} > 6.50 \end{cases} \quad (2)$$

where r_{ij} is the distance between the side chain center of mass of residue i on one helix and the side chain center of mass of residue j on a different helix, and ε_{ij} is the contact energy for the interaction between specific residue types i and j as described in the original work [63]. The total interaction energy for each protein structure is the sum of all contact energies for all interhelical residue pairs. Each helix is treated as a rigid body with PDB conformations

allowing the energy from intrahelical residue pairs to be ignored. A shortcoming of this simple square-well force field is that it is always energetically favorable to form additional contacts which in certain cases may lead to overly compact structures.

Additional energy penalties from the various distance and orientational constraints are added in the form of harmonic penalty functions. For the orientational constraints, the calculated dipolar coupling for a given residue is compared to the target dipolar coupling directly (Eq. 3):

$$E_{\text{orient}}^{\text{NH}} = \frac{1}{2} k_{\text{orient}} \left(D_{\text{calc}}^{\text{NH}} - D_{\text{target}}^{\text{NH}} \right)^2 \quad (3)$$

Using ^{15}N - ^1H dipolar couplings, there is, at most, one value per residue, and the total energy penalty from orientational constraints is the sum over all residues with a measurable coupling (i.e., non-proline residues). The force constant, k_{orient} , in this equation is kept at a constant value for all of the six proteins studied. When only one set of RDCs is used, the force constant was set to 1.0 energy units per Hz^2 . For two sets of RDCs, the force constant was scaled to 0.5 energy units per Hz^2 , and for five sets, the force constant was set to 0.2 energy units per Hz^2 . Because the orientational energy penalty is a simple sum over the number of residues, treating the force constant in this manner results in somewhat weaker penalties for shorter helices. Another alternative, though not implemented here, would be to use protein-specific force constants that scale with the length of the helix and/or number of possible constraints.

For the distance constraints, the calculated distances are evaluated with an upper (d_{max}) and lower (d_{min}) bound consistent with the accuracy of the measurement (Eq. 4):

$$E_{\text{dist}}^{ij} = \begin{cases} \frac{k_{\text{dist}}}{2} (d_{ij} - d_{\text{min}})^2, & \text{if } d_{ij} < d_{\text{min}} \\ 0, & \text{if } d_{\text{min}} \leq d_{ij} \leq d_{\text{max}} \\ \frac{k_{\text{dist}}}{2} (d_{ij} - d_{\text{max}})^2, & \text{if } d_{ij} > d_{\text{max}} \end{cases} \quad (4)$$

For PREs, the upper and lower bounds were set to ± 4 Å as suggested by Battiste and Wagner [36]. Distances from EPR spin labels can be obtained with better accuracy than the PREs measured in the NMR [41], so the upper and lower bounds were set to ± 1 Å. Alternate bounds for the EPR spin label distances were also tested, ranging from ± 1 to ± 4 Å. For all tests, final simulated structures were comparable; small increases in average and variance of RMSDs were observed given less restrictive boundary criteria (see Online Resource 2), but these differences were not large enough to affect the overall comparison of different methods. In the case of distance constraints, the total energy penalty is a sum over all the interhelical residue pairs for which a distance could be measured. For distance constraints, the force constant was simply set to 1.0 energy units per $\text{\AA}^2$.

The total protein energy is a sum of the interaction energy and any constraints applied to the system (Eq. 5):

$$E_{\text{total}} = \sum_{p=1}^{M-1} \sum_{q=p+1}^M \sum_{i=1}^{N(p)} \sum_{j=1}^{N(q)} E_{\text{int}}^{ij} + \sum_{p=1}^{M-1} \sum_{q=p+1}^M \sum_{i=1}^{N(p)} \sum_{j=1}^{N(q)} E_{\text{dist}}^{ij} + \sum_{p=1}^M \sum_{i=1}^{N(p)} E_{\text{orient}}^{ij} \quad (5)$$

where M is the number of helices, and N is the number of residues per helix.

2.5 Simulation details

Each set of simulations utilizes several different starting configurations ($M!$, where M is the number of helices) for each protein. The Monte Carlo/simulated annealing simulations are initiated at a high temperature (1,000 in arbitrary simulation units selected via trial and error) and slowly cooled in a standard simulated annealing process to a temperature less than 1 in the same simulation units. The initial temperature was selected to ensure that each system has enough thermal energy to sample phase space adequately and avoid becoming trapped in local minima. The initial helix orientations of each starting configuration are assigned randomly, with initial positions set so that all loop length constraints are satisfied. Various rates of cooling have been attempted for each protein in an effort to find the fastest rate for which most of the starting configurations converge to the same final structure. Tests were done at six annealing rates for all proteins. The final choice of a protocol with 9×10^7 steps (each annealing step reduced the temperature by a factor of 0.9999999) presents a suitable compromise of converged energies across starting configurations and length of time for the simulation. Actual time for each simulation was variable, depending on protein size and number of constraints; times ranged from several hours to a couple of days.

The simulations proceed with a Monte Carlo search over translational and rotational degrees of freedom using the normal Metropolis acceptance rule. A summary of the target proteins can be found in Table 2, along with the respective numbers and types of constraints that are used for each of the different simulations. Each protein was chosen for its helical secondary structure, small size and the ability to compare with the simulation results of Naniias et al. who used a similar force field [64]. Additional details for the protein model, the constraints, the force field and the simulation method can be found in [54].

Simulation results were analyzed for relative contributions to the total interaction energy arising from the force field and each type of constraint, C $^\alpha$ -RMSDs between calculated structure and PDB structure, and average errors in the internal angles between each pair of helices in a given protein (δ_{IA} ; Eq. 6):

$$\delta_{IA} = \frac{1}{M} \sum_{k=1}^M \sqrt{\sum_{i=1}^{N-1} \sum_{j=i+1}^N (a_{ij} - a_{\text{target}})^2} \quad (6)$$

where M is the number of internal angles, N is the number of helices, and a is the value of the internal angle. The internal angles are found by first calculating the helix orientation of individual helices according to the convention shown in Fig. 1 and then determining the dot product of central axis vectors for each helix pair.

3 Results and discussion

3.1 Force field

For every simulation, the correctness of calculated properties depends on the accuracy of the force field and the ability to adequately sample conformational/configurational space. For simulations in this work, absolute accuracy of the computed structures was less important than the relative accuracy, i.e., determining which combination of constraints gives a better answer than the other combinations. Furthermore, the search algorithm was aided to different degrees by the constraints added to the potential function. As an estimate of the range of structures that are possible using the methodology of this work, two test simulations were run for each protein.

In the first test, the orientational degrees of freedom were eliminated. Backbone conformations and orientations of each helix were read from the respective PDB files and

held fixed throughout the simulated annealing Monte Carlo search. This amounted to a simulation using only translation in three dimensions and allowed the force field to locate a preferred packing arrangement of the orientationally constrained PDB helices. This search, mimicking perfectly accurate orientational constraint data, reflects the ability of our simple force field to find correct protein structures in the presence of good constraint information. Results for this first test protocol are shown in Table 3 as “translation only.” For each of the six target proteins, the final RMSDs of the best structure are compared with those found previously [64] and demonstrate that the residue-based force field employed here is similarly accurate despite its simplicity.

In the second test, both orientation and position of the helices were allowed to change. In the Monte Carlo search, this required both translation and rotation moves. Rotation moves were performed for an entire rigid helix about a given axis such that all three Euler angles (Fig. 1) were effectively sampled. In this test (referred to as method 1 for later comparisons), the search algorithm proceeded in the absence of any constraint data and located the helix orientations and packing arrangement preferred by the force field alone. For all but protein 1I6Z, the unconstrained search was unable to find the appropriate orientation of helices. This is shown by the high errors in internal angles (δ_{IA}) listed in Table 3 and the increases in RMSDs compared to the constrained test.

Comparing results for the test simulations to the energy of the target structure as calculated with the potential function employed here demonstrates that the target structure is generally not the true minimum of the force field. The differences in energy between the simulations with fixed and variable orientations suggest a bias. The residue-based force field is a simple sum over all the interhelical contacts in the protein (Eq. 5). As the number of contacts increases (and the protein becomes more closely packed), the energy of the protein becomes more favorable. When the orientations of each helix are allowed to change throughout the simulation, the force field attempts to maximize the contacts between helices. This generally corresponds to the situation where all of the helices are roughly parallel or have approximately the same orientation. This tendency can explain the larger error in internal angles and higher RMSDs for the case where helix orientations are allowed to vary. Not surprisingly, the one strikingly successful (average RMSD = 1.1 Å) protein in this group, 1I6Z, has a native structure with long, nearly parallel helices. The difference in interaction energies between variable and fixed helical orientations is much smaller for 1I6Z than for any of the other proteins in the list. Additionally, the large errors in the internal angles for all but protein 1I6Z indicate that the force field alone struggles to determine both the orientation and packing of helical segments in proteins with non-parallel helix alignments. However, when the orientations are held fixed, the simple force field performs remarkably well. The best RMSDs for each protein in this group are all below 3 Å for the PDB conformations. The implication is that as the constraint data become more accurate and restrictive, the force field will more readily locate a structure that resembles the native target. In this respect, a simple force field is all that is needed to evaluate different types of constraint data that can be used with more sophisticated protein structure prediction methodologies and should be prioritized for experimental measurement. For all subsequent simulations, the interaction energies and corresponding structures are expected to fall somewhere between the target structure and the structure preferred by the force field (“translation and rotation” field in Table 3).

3.2 Orientational constraints

The challenges posed by the orientational degeneracy of RDC measurements and the potential benefit of unique constraint data are well known and should not be overlooked when using RDCs in prediction algorithms. In the general case, RDC data from a single

alignment are not sufficient for determining unique protein orientations and must be supplemented either by an additional set of RDCs from a different alignment, by distance constraints, or by a force field. Combination of ^{15}N – ^1H RDCs with the chemical shifts in the Rosetta [65] algorithm for protein structure prediction is one such approach [4, 66–68]. In these studies, the RDC data were used as a supplement to the force field. The inherent degeneracies were noticeable in the first study as an increase in the range of errors (i.e., some structures were much better, but some were much worse) in the distance matrix for the protein [49]. When combined with additional data, such as NMR chemical shifts, overall results and uncertainties improved relative to the unconstrained case [4]. In the second study, no explicit mention of RDC degeneracies was made, but it is interesting to note that at least one of the six proteins in the study failed to show structural improvements with the inclusion of a limited set of RDC constraints from a single alignment [68].

In the present work, a series of simulations are performed using an increasing amount of RDC data in conjunction with the simple force field described above (methods 2–4). The results, presented in Table 4, demonstrate that the use of data from only one alignment is not adequate for uniquely determining helix orientations. In fact, the addition of a single set of RDCs to the force field (method 2) actually made structures worse compared to the force field alone for four of the proteins tested. This can be seen by the higher RMSDs and larger errors in the internal angles (δ_A). The situation begins to improve with the addition of a second set of RDCs from a unique alignment. With the exception of 1A6S, the errors in the internal angles improved dramatically compared to the single set. RMSDs decreased for all proteins. As the number of unique alignments was increased to five, the results improved further. The RMSDs observed using two and five data sets (methods 3 and 4) improve steadily, with all proteins showing an RMSD of the lowest energy structure below 6 Å for five sets of RDCs. There is also a best structure in each case with an RMSD below 3.5 Å. The differences in total energy between the lowest energy structure and the best RMSD structure also improve going from a single set to five sets. With a single set of RDCs, the average difference in energy between the lowest energy and best RMSD structure is 27 units. When five RDC sets are used, this average energy difference reduces to 10 units.

A further assessment of the quality of constraint data is the convergence of starting configurations. Experimental data that result in few local minima will direct each starting configuration into the global minimum more rapidly than a less restrictive data set. The better constraint sets will be identified by a higher level of convergence of starting configurations to the global minimum. The plots in Fig. 3 compare energy and RMSD for all starting configurations of each simulation, the sweet spot being the lower left-hand corner, corresponding to low energy (high accuracy) and low RMSD (high precision). As discussed above, the single set of RDCs (Fig. 3a) does not have an appreciable effect on the convergence of helix packing. The scatter of starting configurations for each protein was expected to get smaller and moves toward the lower left-hand corner of the plot as the number of RDC sets increased. Nevertheless, in moving from a single set of RDCs to five sets (Fig. 3b), RMSD convergence is not noticeably better, but overall there is an increasing number of structures achieving RMSDs below 5 Å. A closer look at the structures obtained from these multi-set simulations indicates that the improvement in RMSD is not as dramatic as expected due to the persistence of translational errors. In many cases, the orientations of the helices are correct (note especially the decrease in errors for internal angles; Table 4), but the packing has not been optimized.

These results underscore the challenges of using RDC data to constrain conformational search algorithms. A single RDC data set only becomes beneficial when there are enough data from other sources to overcome the intrinsic degeneracy arising from second rank tensor in Eq. 1. Without additional data, RDCs can potentially introduce additional

unwanted uncertainty. Unique helix orientations were found consistently only with the use of two or more data sets.

3.3 Distance constraints

Distance constraints become more effective at limiting the extent of conformational space when they constrain portions of the protein far apart in the sequence to be close in space. NOE distances, due to their short range (on the order of 5 Å), are the most precise distance constraints when they are available, but PRE and EPR distances, due to their longer range, can provide complementary information about more distant parts of the tertiary structure. In many previous studies, as in this one, distance information is often adequate for determining both the orientation and packing of helices within the protein [4, 8–17, 69]. As can be seen from Table 5, in each of the two types of simulations using distance constraints (methods 5 and 6), the RMSDs of the lowest energy structures are all below 7 Å, and for the best RMSD structure, below 5 Å. The one exception is protein 1J7O for simulations using only EPR distances. This protein has fairly loose packing in the target structure and was a consistent challenge for the force field to fit (cf. Eq. 2). For most of the proteins, however, the ability of the force field to find correct structures improves with distance constraints as seen by the smaller differences in energy between the best structure and the lowest energy structure relative to the orientational methods. The average energy differences have decreased from over 25 to 14 units for the PREs and 18 units for the EPR method. Only the simulations using five sets of RDCs had better agreement between the lowest energy structure and the best RMSD structure ($\Delta E_{\text{tot}} < 10$ on average). A comparison of energies and RMSDs for each of the distance constraint simulations shows an improving trend in results going from EPR simulations to PRE simulations. Thus, the range of RMSDs observed for EPR distance constraints (Fig. 3d) is similar to that of five sets of RDC constraints. Conversely, with the PRE data set, the RMSDs for all starting configurations fall below 10 Å (Fig. 3c), but none of the observed structures has an RMSD below 1.5 Å. This is the only simulation type where structures below 1.5 Å were not observed. The simulations using PRE or EPR distances resulted in average RMSDs in the 5–7 Å range. By comparison, the simulations using a single set of RDCs produced average RMSDs in the 8–10 Å range. While the RMSDs for distance methods are generally lower than those seen in simulations with RDCs, the errors in the internal angles do not improve substantially relative to RDC-based methods. The simulations using EPR constraints resulted in the highest orientational errors of any of the distance methods. However, these errors are still somewhat better than a single set of RDCs alone. In the absence of definitive NOE data, experimentalists can make use of the long-range distances available from paramagnetic spin labels, such as the PRE and EPR distances, as a supplement or replacement of NOE distances.

Referring again to Table 5, the results of our simulations indicate that the long-range distances from either PRE or EPR experiments do not converge satisfactorily to the target structure, which we attribute to the placement of the paramagnetic center. In each of the paramagnetic distance simulations, the optimal position of spin labels was placing one label at each of the helix ends. With the further requirement that the spin labels be incorporated at solvent-accessible sites, the ideal placement of spin labels was only achieved for 2ABD. In the case of 1A6S, there were no solvent-accessible sites on the second helix, so all distances for that protein came from helices 1, 3 or 4. When measuring EPR distances, a minimum of two useable distances per helix was imposed. In most cases, this meant placing several spin labels on one helix to account for the lack of hydrophilic sites on the other helices. For PRE data, the distances were measured from each spin label to amide hydrogen atoms on all other helices. Each set of distances from a given spin label has only a single origin. Despite larger numbers of actual distances (see Table 2), PRE distances are less unique and less discriminating than an equal number of NOEs. The number of distances available from EPR

experiments is limited further to those between spin labels. Each spin label introduced requires an additional amino acid mutation, thereby increasing the experimental effort required to obtain enough useable distances.

3.4 Combined distance and orientational constraints

A final set of simulations was performed using both distance and orientational constraints (methods 7 and 8). This set of simulations generated the best overall results (Table 6). Using a combination of distance and orientational information places constraints on both the translational (helix packing) and rotational (helix orientation) degrees of freedom. With these simulations, the energies tend to increase for each protein reflecting the larger number of constraints imposed on the system. Looking more closely at the energy penalties associated with the addition of constraints reveals that the structures obtained using combined constraints are a better match for the target structure. For the combined simulations using EPR and PRE distances, significant improvements are observed. For all of the combined simulations, the RMSD of the lowest energy structure is lower compared to the corresponding methods using either the corresponding distance or orientational constraints. Plots of energy and RMSD for the combined methods (Fig. 3e, f) show trends similar to those observed using only distance constraints. There is increasing convergence going from RDC/EPR simulations to RDC/PRE simulations. These combined simulations are also showing the best agreement between the lowest energy structure and the best RMSD structure. Average differences in the energies are 10 (RDC/PRE) and 18 (RDC/EPR) in the arbitrary units used in the simulations. Errors in internal angles for the combined simulations are also better than the separate distance and orientational methods. The δ_{IA} values for RDC/PRE simulations are generally better than those observed for a single set of RDCs, but not quite as good as those for two and five sets of RDCs.

The RDC/EPR simulations are more interesting. In this case, the orientational error is as small as that observed for five sets of RDCs, with the exception of 1DV5. These simulations illustrate both the potential benefits and difficulties arising from RDC information. Inclusion of RDCs resulted in significant improvement in the orientational errors compared to the distance simulation where EPR constraints resulted in the highest δ_{IA} values. For the protein 1DV5, however, despite the low energy penalties, the error in the internal angles is nearly 80°. The structure obtained is one of the degenerate orientations that satisfy the orientational constraints while at the same time not generating any significant distance penalty. In this structure, the second helix is rotated 180° relative to the correct orientation. With comparatively few distances between helices, this situation is more likely to occur with EPR constraints than with any of the others. When this protein is excluded, both the RMSDs of the lowest energy structure and the best RMSDs for all other proteins are comparable to those observed using five sets of RDCs.

For each of the combined constraint methods considered here, the addition of distance data to a single set of RDCs is able to overcome the degeneracy problems seen in the high orientational errors and RMSDs using only orientational constraints. These combined simulations utilize the strengths of each type of constraint to yield optimal results. Of course, better convergence and complete relief of RDC degeneracy can be achieved using short- and long-range NOEs. Indeed, when we ran our simulations using dipolar waves in conjunction with NOE constraints randomly distributed along the protein sequences, we consistently obtained RMSDs lower than 1 Å (data not shown). However, we would like to underscore that NOE assignment is a very time-consuming process, while both PREs and EPR distances can be readily available (easy to measure and analyze). Note that in some cases, obtaining samples for PRE measurements can be challenging. Nonetheless, once the

samples are prepared, the NMR experiments for these samples are relatively simple to set up and analyze.

3.5 Summary of results

In Figs. 4 and 5, the simulation results for RMSDs and errors in the internal angles are compared for each of the eight methods. In Fig. 4, results are presented for the structure with the best RMSD and the structure with the lowest energy for each method averaged over all six proteins in the test set. Figure 5 presents the degree of convergence of each method. For the simulations using a single set of RDC constraints (method 2) or only distance constraints (methods 5–6), we found high orientational errors and poor convergence. Simulations using constraint data that include multiple RDC data sets from unique alignments or RDC data paired with sparse distance constraints significantly reduce the observed errors in internal angles and result in better predictions of helix orientation. RDC and EPR constraints generated the best RMSDs, whereas RDC and PRE constraints and 5 RDC data sets produced the best internal angles.

In addition to the results discussed so far, additional tests were also performed using ideal helix structures and another set of simulations using larger distance boundaries for the EPR constraints (see Online Resources). Tests with ideal helix conformations show the same general trends as simulations with PDB conformations, particularly with respect to the relative orientations and packing of the helices. For a given set of force constants, however, the magnitude of E_{orient} is larger for the ideal helix structures, which could easily be corrected by scaling the relative force constants. Increasing the EPR distance boundaries does not lead to any significant increase in the RMSD, but E_{dist} decreases slightly when the boundaries are looser.

Ribbon diagrams of the lowest energy structures for the three best methods are shown in Fig. 6. All of the structures shown have RMSDs to the target of less than 6 Å, with the exception of 1DV5 with RDC/EPR constraints noted earlier. Of the remaining structures, only three of the proteins (1G2H, 1A6S and 2ABD) simulated with five sets of RDC constraints have RMSDs above 4.0 Å. All the other structures for all of the other methods have a lowest energy structure with an RMSD of 3.8 Å or less. Orientational errors also persist for 1G2H and 1A6S. For 1G2H, the higher errors are related to the short length of the helices. As described in the Methods Section, the force constants can lead to weaker constraint penalties for shorter helices; this could be overcome by optimizing force constants for each protein. 2ABD has the correct orientations, but suffers from persistent translational errors. The ribbon diagram reveals that helices 2 and 3 are not correctly packed despite having the correct orientations.

In summation, the best performing methods were found to be methods 4 (5 sets of RDCs), 7 (RDC + EPR) and 8 (RDC + PRE). These methods produced, on average, the best RMSDs and lowest orientational errors. In general, the introduction of PRE data is preferable, as it still provides the site-specific information, manifested in the quenching of NMR lines. The exception here is the small globular tightly packed proteins, for which the PRE data would quench the majority of the NMR signals. In this case, there is a preference toward the EPR restraints, as seen for the 1A6S protein (Table 5). In the absence of the spin-labeled data, similar precision and accuracy can be obtained with multiple RDC sets, provided a number of sufficiently different alignments can be achieved. Note that introducing other types of RDC data (e.g., C_{α} – C') does not reduce the degeneracy of the system. Our simulations show that it is possible to have relatively low RMSDs while retaining a fair amount of orientational error and suggest that both RMSD and orientational error should be taken into account when assessing the quality of calculated structures. The success of the method involving multiple RDC data sets further underscores the importance of orientational input

when determining protein structure, but multiple unique RDC alignments pose a significant experimental challenge. Our analysis further confirms that the combined use of both long-range distance and orientational constraints is a better approach.

3.6 Comparison with other methods and perspectives

In the past few years, concomitant with the structural biology initiative, there has been a tremendous effort devoted to advance computational methods for structural predictions with and without auxiliary NMR data [4, 23, 69–71]. In particular, the incorporation of RDCs into prediction methods using fragment- or residue-based approaches has improved dramatically the convergence of the prediction programs. In our approach, we advocate the use of dipolar waves (characteristic patterns arising from the RDC data of periodic structures) for both secondary structure and three-dimensional structure determination. The approach presented in this paper builds on our previous work [49] using dipolar waves and an exhaustive search algorithm to generate a set of geometric solutions compatible with the overall protein folds. Here, we have coupled this approach with a minimal force field and Monte Carlo/simulated annealing sampling method to determine the tertiary fold of proteins. As with the approach proposed by Prestegard and coworkers [72], this method does not rely on sequence content and utilized sparse and easily obtainable NMR/EPR data. Unlike the method proposed by Blackledge and coworkers [73], our approach is compatible with the use of dipolar waves derived from only one aligning medium (methods 7 and 8). While our approach embraces a philosophy similar to that of Wang and coworkers, we do not make use of short range distances or ϕ/ψ angles [50]. Rather, we treat the helical segments as rigid units to build the protein three-dimensional structure. Our aim is to obtain a fast fold at moderate-resolution (below 4 Å) that can constitute excellent starting points for high-resolution structure determination. As with the method proposed by Wang and coworkers, our structures mainly contain information regarding protein backbones and side chains need to be reconstructed using more realistic force fields [50]. Specifically, these low-resolution structures can be used to design new experiments (i.e., labeling sites, NMR and EPR experiments) to derive further constraints to be implemented in a more comprehensive manner. Fully atomistic force fields can be used in conjunction with simulated annealing approaches to improve convergence and resolution of both backbone and side chains [74, 75].

Our long-term goal, however, is to establish a robust framework to apply these methods for the structural determination of helical membrane proteins utilizing orientational constraints, where NOE data can be difficult to obtain. In fact, membrane protein backbone structural topologies are obtainable from anisotropic solid-state NMR data (dipolar couplings and chemical shift anisotropies) derived from protein reconstituted in mechanically or magnetically aligned lipid bilayers. For these samples, dipolar and chemical shift waves have been instrumental for backbone resonance assignments, topology and structure, simultaneously. The confinement of the transmembrane domains within the ~40 Å of the membrane bilayer will significantly restrict the conformational landscape (e.g., a higher propensity for parallel helix alignment), making our strategy (but with a force field reflecting the lipid environment) more advantageous for membrane protein structure determination.

4 Conclusions

Using six helical proteins for validation, we demonstrate that the combination of minimal experimental data and a simple residue-based force field is able to determine both orientation and packing of several small helical proteins. Specifically, the combinations of EPR or PRE and orientational data derived from RDCs (i.e., dipolar waves) are very promising. Although the structural models we have obtained are at moderate-resolution, they

have the correct topology of all helical elements. In the future, our computational approach may aid in the design of new experiments to obtain optimal long-distance information (e.g., predict the location of spin labels that can best discriminate between degenerate candidate structures using only RDCs) and, hence, enables one to find the correct solution for the three-dimensional protein structure in a more expedited manner than the classical experimental methods for structural determination.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Dr. A. Mascioni for helpful discussion. Financial supports from the National Science Foundation (CTS-0553911 and CBET-0756641) to J.I.S., the National Institutes of Health (GM64742 and GM072701) to G.V. and postdoctoral fellowship from American Heart Association (13POST14670054) to V.V. are gratefully acknowledged. B.L.E. would like to thank the NSF-IGERT program for fellowship support. Part of the computer resources for this project was provided by the Minnesota Supercomputing Institute.

References

1. Baker D, Sali A. Protein structure prediction and structural genomics. *Science*. 2001; 294:93–96. [PubMed: 11588250]
2. Chandonia JM, Brenner SE. The impact of structural genomics: expectations and outcomes. *Science*. 2006; 311:347–351. [PubMed: 16424331]
3. Service R. Structural biology. Structural genomics, round 2. *Science*. 2005; 307:1554–1558. [PubMed: 15761136]
4. Rohl CA, Baker D. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J Am Chem Soc*. 2002; 124:2723–2729. [PubMed: 11890823]
5. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round 6. *Proteins*. 2005; 61(Suppl 7):3–7. [PubMed: 16187341]
6. Kryshchuk A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. *Proteins*. 2005; 61(Suppl 7):225–236. [PubMed: 16187365]
7. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. *Proc Natl Acad Sci USA*. 2003; 100:12105–12110. [PubMed: 14528006]
8. Li W, Zhang Y, Skolnick J. Application of sparse NMR restraints to large-scale protein structure prediction. *Biophys J*. 2004; 87:1241–1248. [PubMed: 15298926]
9. Li W, Zhang Y, Kihara D, Huang YJ, Zheng D, Montelione GT, Kolinski A, Skolnick J. TOUCHSTONE: protein structure prediction with sparse NMR data. *Proteins*. 2003; 53:290–306. [PubMed: 14517980]
10. Bowers PM, Strauss CE, Baker D. De novo protein structure determination using sparse NMR data. *J Biomol NMR*. 2000; 18:311–318. [PubMed: 11200525]
11. Standley DM, Eylich VA, Felts AK, Friesner RA, McDermott AE. A branch and bound algorithm for protein structure refinement from sparse NMR data sets. *J Mol Biol*. 1999; 285:1691–1710. [PubMed: 9917406]
12. Debe DA, Carlson MJ, Sadanobu J, Chan SI, Goddard WA III. Protein fold determination from sparse distance restraints: the restrained generic protein direct Monte Carlo method. *J Phys Chem B*. 1999; 103:3001–3008.
13. Kolinski A, Skolnick J. Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. *Proteins*. 1998; 32:475–494. [PubMed: 9726417]
14. Skolnick J, Kolinski A, Ortiz A. MONSTER: a method for folding globular proteins with a small number of distance restraints. *J Mol Biol*. 1997; 265:217–241. [PubMed: 9020984]
15. Aszodi A, Gradwell MJ, Taylor WR. Global fold determination from a small number of distance restraints. *J Mol Biol*. 1995; 251:308–326. [PubMed: 7643405]

16. Connolly PJ, Stern AS, Hoch JC. Estimating protein fold from incomplete and approximate NMR data. *J Am Chem Soc.* 1994; 116:2675–2676.
17. Smith-Brown MJ, Kominos D, Levy RM. Global folding of proteins using a limited number of distance constraints. *Protein Eng.* 1993; 6:605–614. [PubMed: 7694274]
18. Prestegard JH, al-Hashimi HM, Tolman TR. NMR structures of biomolecules using field oriented media and residual dipolar couplings. *Q Rev Biophys.* 2000; 33:371–424. [PubMed: 11233409]
19. Prestegard JH. New techniques in structural NMR— anisotropic interactions. *Nat Struct Biol.* 1998; 5(Suppl):517–522. [PubMed: 9665182]
20. 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]
21. Bax A, Kontaxis G, Tjandra N. Dipolar couplings in macromolecular structure determination. *Methods Enzymol.* 2001; 339:127–174. [PubMed: 11462810]
22. Chen K, Tjandra N. The use of residual dipolar coupling in studying proteins by NMR. *Top Curr Chem.* 2012; 326:47–67. [PubMed: 21952837]
23. Clore GM, Schwieters CD. Theoretical and computational advances in biomolecular NMR spectroscopy. *Curr Opin Struct Biol.* 2002; 12:146–153. [PubMed: 11959490]
24. Ramirez BE, Bax A. Modulation of the alignment tensor of macromolecules dissolved in a dilute liquid crystalline medium. *J Am Chem Soc.* 1998; 120:9106–9107.
25. Hus JC, Salmon L, Bouvignies G, Lotze J, Blackledge M, Bruschweiler 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]
26. 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.
27. 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]
28. Andrec M, Du P, Levy RM. Protein backbone structure determination using only residual dipolar couplings from one ordering medium. *J Biomol NMR.* 2001; 21:335–347. [PubMed: 11824753]
29. Prestegard JH, Mayer KL, Valafar H, Benison GC. Determination of protein backbone structures from residual dipolar couplings. *Methods Enzymol.* 2005; 394:175–209. [PubMed: 15808221]
30. Ruan K, Tolman JR. Composite alignment media for the measurement of independent sets of NMR residual dipolar couplings. *J Am Chem Soc.* 2005; 127:15032–15033. [PubMed: 16248635]
31. Jain NU. Use of residual dipolar couplings in structural analysis of protein-ligand complexes by solution NMR spectroscopy. *Methods Mol Biol.* 2009; 544:231–252. [PubMed: 19488703]
32. Berlin K, O'Leary DP, Fushman D. Structural assembly of molecular complexes based on residual dipolar couplings. *J Am Chem Soc.* 2010; 132:8961–8972. [PubMed: 20550109]
33. Qu Y, Guo JT, Olman V, Xu Y. Protein structure prediction using sparse dipolar coupling data. *Nucleic Acids Res.* 2004; 32:551–561. [PubMed: 14744980]
34. Wuthrich, K. NMR of proteins and nucleic acids. Wiley; New York: 1986.
35. Kumar A, Ernst RR, Wuthrich K. A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton–proton cross-relaxation networks in biological macromolecules. *Biochem Biophys Res Commun.* 1980; 95:1–6. [PubMed: 7417242]
36. Battiste JL, Wagner G. Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear Overhauser effect data. *Biochemistry.* 2000; 39:5355–5365. [PubMed: 10820006]
37. Gaponenko V, Howarth JW, Columbus L, Gasmi-Seabrook G, Yuan J, Hubbell WL, Rosevear PR. Protein global fold determination using site-directed spin and isotope labeling. *Protein Sci.* 2000; 9:302–309. [PubMed: 10716182]
38. Zhuang T, Jap BK, Sanders CR. Solution NMR approaches for establishing specificity of weak heterodimerization of membrane proteins. *J Am Chem Soc.* 2011; 133:20571–20580. [PubMed: 22084929]

39. Skasko M, Wang Y, Tian Y, Tokarev A, Munguia J, Ruiz A, Stephens EB, Opella SJ, Guatelli J. HIV-1 Vpu protein antagonizes innate restriction factor BST-2 via lipid-embedded helix-helix interactions. *J Biol Chem*. 2012; 287:58–67. [PubMed: 22072710]
40. Sengupta I, Nadaud PS, Helmus JJ, Schwieters CD, Jaroniec CP. Protein fold determined by paramagnetic magic-angle spinning solid-state NMR spectroscopy. *Nat Chem*. 2012; 4:410–417. [PubMed: 22522262]
41. Rabenstein MD, Shin YK. Determination of the distance between two spin labels attached to a macromolecule. *Proc Natl Acad Sci USA*. 1995; 92:8239–8243. [PubMed: 7667275]
42. Duncan B, Buchanan BG, Hayes-Roth B, Lichtarge O, Altman R, Brinkley J, Hewett M, Cornelius C, Jardetzky O. PROTEAN: a new method for deriving solutions structures of proteins. *Bull Magn Reson*. 1986; 8:111–119.
43. Tjandra N, Omichinski JG, Gronenborn AM, Clore GM, Bax A. Use of dipolar ^1H - ^{15}N and ^1H - ^{13}C couplings in the structure determination of magnetically oriented macromolecules in solution. *Nat Struct Biol*. 1997; 4:732–738. [PubMed: 9303001]
44. Fischer MW, Losonczi JA, Weaver JL, Prestegard JH. Domain orientation and dynamics in multidomain proteins from residual dipolar couplings. *Biochemistry*. 1999; 38:9013–9022. [PubMed: 10413474]
45. Clore GM. Accurate and rapid docking of protein-protein complexes on the basis of intermolecular nuclear Overhauser enhancement data and dipolar couplings by rigid body minimization. *Proc Natl Acad Sci USA*. 2000; 97:9021–9025. [PubMed: 10922057]
46. Mesleh MF, Veglia G, DeSilva TM, Marassi FM, Opella SJ. Dipolar waves as NMR maps of protein structure. *J Am Chem Soc*. 2002; 124:4206–4207. [PubMed: 11960438]
47. Mesleh MF, Lee S, Veglia G, Thiriot DS, Marassi FM, Opella SJ. Dipolar waves map the structure and topology of helices in membrane proteins. *J Am Chem Soc*. 2003; 125:8928–8935. [PubMed: 12862490]
48. Mascioni A, Veglia G. Theoretical analysis of residual dipolar coupling patterns in regular secondary structures of proteins. *J Am Chem Soc*. 2003; 125:12520–12526. [PubMed: 14531696]
49. Mascioni A, Eggimann BL, Veglia G. Determination of helical membrane protein topology using residual dipolar couplings and exhaustive search algorithm: application to phospholamban. *Chem Phys Lipids*. 2004; 132:133–144. [PubMed: 15530454]
50. Wang J, Walsh JD, Kuszewski J, Wang YX. Periodicity, planarity, and pixel (3P): a program using the intrinsic residual dipolar coupling periodicity-to-peptide plane correlation and phi/psi angles to derive protein backbone structures. *J Magn Reson*. 2007; 189:90–103. [PubMed: 17892961]
51. Walsh JD, Kuszewski J, Wang YX. Determining a helical protein structure using peptide pixels. *J Magn Reson*. 2005; 177:155–159. [PubMed: 16084744]
52. Walsh JD, Wang YX. Periodicity, planarity, residual dipolar coupling, and structures. *J Magn Reson*. 2005; 174:152–162. [PubMed: 15809182]
53. Jensen MR, Blackledge M. On the origin of NMR dipolar waves in transient helical elements of partially folded proteins. *J Am Chem Soc*. 2008; 130:11266–11267. [PubMed: 18665596]
54. Eggimann, BL. Monte Carlo simulations for NMR-based structure determination, aqueous interfacial properties, and pressure-volume curves (Ph. D. thesis). Department of Chemistry, University of Minnesota; 2006.
55. Wishart DS, Sykes BD, Richards FM. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J Mol Biol*. 1991; 222:311–333. [PubMed: 1960729]
56. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*. 2000; 40:502–511. [PubMed: 10861942]
57. Chen J, Won HS, Im W, Dyson HJ, Brooks CL III. Generation of native-like protein structures from limited NMR data, modern force fields and advanced conformational sampling. *J Biomol NMR*. 2005; 31:59–64. [PubMed: 15692739]
58. Kahn PC. Defining the axis of a helix. *Comput Chem*. 1989; 13:185–189.
59. Zweckstetter M, Bax A. Evaluation of uncertainty in alignment tensors obtained from dipolar couplings. *J Biomol NMR*. 2002; 23:127–137. [PubMed: 12153038]
60. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. 1982; 157:105–132. [PubMed: 7108955]

61. Hubbell WL, Cafiso DS, Altenbach C. Identifying conformational changes with site-directed spin labeling. *Nat Struct Biol.* 2000; 7:735–739. [PubMed: 10966640]
62. Clore GM, Iwahara J. Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. *Chem Rev.* 2009; 109:4108–4139. [PubMed: 19522502]
63. Miyazawa S, Jernigan RL. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol.* 1996; 256:623–644. [PubMed: 8604144]
64. Nanas M, Chinchio M, Pillardy J, Ripoll DR, Scheraga HA. Packing helices in proteins by global optimization of a potential energy function. *Proc Natl Acad Sci USA.* 2003; 100:1706–1710. [PubMed: 12571353]
65. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D. Free modeling with Rosetta in CASP6. *Proteins.* 2005; 61(Suppl 7):128–134. [PubMed: 16187354]
66. Meiler J, Baker D. Rapid protein fold determination using unassigned NMR data. *Proc Natl Acad Sci USA.* 2003; 100:15404–15409. [PubMed: 14668443]
67. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A. Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA.* 2008; 105:4685–4690. [PubMed: 18326625]
68. Haliloglu T, Kolinski A, Skolnick J. Use of residual dipolar couplings as restraints in ab initio protein structure prediction. *Biopolymers.* 2003; 70:548–562. [PubMed: 14648765]
69. Hirst SJ, Alexander N, McHaourab HS, Meiler J. RosettaEPR: an integrated tool for protein structure determination from sparse EPR data. *J Struct Biol.* 2011; 173:506–514. [PubMed: 21029778]
70. Bax A. Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci.* 2003; 12:1–16. [PubMed: 12493823]
71. Blackledge M. Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. *Prog Nucl Magn Reson Spectrosc.* 2005; 46:23–61.
72. Mayer KL, Qu Y, Bansal S, LeBlond PD, Jenney FE Jr, Brereton PS, Adams MW, Xu Y, Prestegard JH. Structure determination of a new protein from backbone-centered NMR data and NMR-assisted structure prediction. *Proteins.* 2006; 65:480–489. [PubMed: 16927360]
73. Beraud S, Bersch B, Brutscher B, Gans P, Barras F, Blackledge M. Direct structure determination using residual dipolar couplings: reaction-site conformation of methionine sulfoxide reductase in solution. *J Am Chem Soc.* 2002; 124:13709–13715. [PubMed: 12431100]
74. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The Xplor-NIH NMR molecular structure determination package. *J Magn Reson.* 2003; 160:66–74.
75. Schwieters CD, Kuszewski JJ, Clore GM. Using Xplor-NIH for NMR molecular structure determination. *Progr NMR Spectrosc.* 2006; 48:47–62.

Fig. 1.

Angle definitions. The position of a helix is described by three angles as depicted in the *left* and *center figures*. The angles are the result of a “*ZZZ*” Euler angle convention treating counterclockwise rotations as positive. θ is the angle between the central axis of the helix and the z -axis of the reference frame starting from a zero point position where the central axis and z -axis are collinear. ϕ is the angle between the projection of the central axis and the positive x -axis of the reference frame. ρ is defined as the rotation about the central axis referenced to a zero point that occurs when the central axis is collinear with the z -axis, as shown in the *center figure*, and the first N–H bond vector in the helix lies along the positive x -axis. The figure at the *right* shows the angles defining the position of an N–H bond vector with respect to the central axis of the helix. β is the polar angle from the positive z' -axis of the helix axis frame, and α is the azimuthal angle in the $x'y'$ -plane measured from the positive x' -axis

Fig. 2.

a Kyte–Doolittle hydropathy plot for 1I6Z that was used to determine solvent-accessible sites for the positioning of site-directed spin labels. **b** Ribbon structure of 1I6Z showing the location of MTSSL spin labels along the protein backbone

Fig. 3.

Comparing energy and RMSD for RDC, PRE and EPR constraint methods. **a** Force field plus a single set of RDCs (method 2); **b** force field plus five sets of RDCs (method 4); **c** force field plus PRE distance constraints (method 5); **d** force field plus EPR distance constraints (method 6); **e** force field plus PRE constraints and a single set of RDCs (method 7); **f** force field plus EPR constraints and a single set of RDCs (method 8). In each case, the six proteins are represented by *circles* (1I6Z), *squares* (1G2H), *diamonds* (1DV5), *up-triangles* (1J7O), *left-triangles* (1A6S) and *down-triangles* (2ABD). For each protein, data points represent an independent starting configuration

Fig. 4.

Average RMSD (black bars) and δ_{IA} (gray bars) for each simulation type. Averages are taken over the six proteins studied. **a** Structures with the lowest RMSD; **b** structures giving the lowest energy for each protein. The simulations are as follows: the force field alone (1), the force field plus a single set of RDC constraints (2), the force field plus two sets of RDC constraints (3), the force field plus five sets of RDC constraints (4), the force field plus PRE distance constraints (5), the force field plus EPR constraints (6), the force field plus RDC/PRE constraints (7), the force field plus RDC/EPR constraints (8)

Fig. 5. Percentage of total structures with δ_{LA} (a) and RMSD (b) values above the following cutoffs: 25° or 6 Å (black bars), 10° or 4 Å (light gray bars), and 5° or 3 Å (dark gray bars). The methods are the same as described in the caption of Fig. 4

Fig. 6. Ribbon diagrams of the lowest energy structure found for each protein in the best three methods. Diagrams are colored, so the first helix is *red*, the second helix is *blue*, the third helix is *green*, and the fourth helix is *yellow*. Unstructured loop regions are shown in *gray* only for the target structure

Table 1

Alignment tensors for multiple RDC data sets

				1I6Z	1G2H	1DV5	1J7O	1A6S	2ABD
$D_a=$				36.5	−6.4	−6.8	11.0	5.8	−8.9
Set	α	β	γ	Rhombicity (R)					
1	0	0	0	0.281	0.552	0.328	0.154	0.063	0.517
2	$\pi/4$	0	0	0.056	0.276	0.656	0.077	0.126	0.259
3	$\pi/2$	$\pi/4$	0	0.422	0.138	0.164	0.308	0.252	0.121
4	$3\pi/4$	$\pi/2$	$\pi/4$	0.633	0.069	0.082	0.462	0.378	0.035
5	π	$3\pi/4$	$\pi/2$	0.112	0.652	0.041	0.616	0.630	0.611

Euler angles α , β and γ are given in radians. D_a and R are given in Hz

Table 2

Protein information and numbers of available constraints for each type

PDB code	<i>M</i>	<i>N</i>	<i>N</i> _{hel}	<u>Number of constraints</u>		
				RDC	PRE	EPR
1I6Z	3	130	106	105	424 (6)	10 (6)
1G2H	3	61	34	34	93 (4)	13 (6)
1DV5	3	80	40	39	81 (3)	6 (6)
1J7O	4	76	51	50	228 (6)	9 (8)
1A6S	4	87	57	56	162 (4)	11 (7)
2ABD	4	86	59	59	354 (8)	14 (8)

M, *N* and *N*_{hel} are the number of helices, total number of residues and number of residues in helices, respectively. The number of spin labels used for PRE and EPR distances is shown in parenthesis beside the number of constraints

Table 3

Comparison of test simulations for the simple, residue-based force field in the absence of experimental data

PDB code	Target E_{int}	Lowest energy			Best RMSD		Average		Ref. [43] RMSD
		E_{int}	δ_{IA}	RMSD	RMSD	ΔE_{tot}	RMSD	σ_{RMSD}^2	
Translation only									
1I6Z	−305.6	−336.8	0.0	0.4	0.4	0.0	0.4	0.0	2.5
1G2H	−91.3	−122.3	0.0	2.4	2.4	0.0	3.5	3.4	3.4
1DV5	−107.5	−128.1	0.0	0.9	0.9	0.0	1.9	0.4	2.2
1J7O	−91.6	−264.9	0.0	3.0	3.0	0.0	3.6	1.1	-
1A6S	−229.3	−250.9	0.0	2.7	2.7	0.0	6.4	4.2	4.4
2ABD	−199.6	−225.3	0.0	5.3	0.4	8.1	6.0	7.2	6.9
Translation and rotation (method 1)									
1I6Z	−305.6	−376.7	3.1	1.5	0.8	1.6	1.1	0.1	2.5
1G2H	−91.3	−249.2	54.3	5.9	5.9	0.0	7.5	0.7	3.4
1DV5	−107.5	−244.4	59.6	8.2	6.5	20.4	8.0	0.9	2.2
1J7O	−91.6	−419.9	60.3	10.3	7.5	28.0	9.3	1.2	-
1A6S	−229.3	−442.4	48.4	9.7	7.7	31.0	9.5	0.6	4.4
2ABD	−199.6	−392.6	49.2	10.5	6.5	30.6	9.3	1.5	6.9

Results for the lowest energy structure in each set of simulations are compared to the structure giving the best RMSD and to the average values. Column headings are defined as follows: E_{int} refers to the interaction energy (Eq. 5); δ_{IA} is the average error in internal angles (Eq. 6) in degrees;

RMSD is the C $^{\alpha}$ RMS deviation of the calculated structure from the target (PDB) structure; “Best RMSD” is the best observed RMSD over all starting configurations; ΔE_{tot} is the difference in the total energy between the lowest energy structure and the best RMSD structure; the average

RMSD is the average over all starting configurations; and lastly, σ_{RMSD}^2 is the variance of RMSDs. RMSDs and σ_{RMSD}^2 are given in units of Å, while energies are reported in arbitrary units

Table 4

Summary of structural data for orientational constraint simulations

PDB code	Lowest energy				Best RMSD		Average	
	E_{int}	E_{orient}	δ_{IA}	RMSD	RMSD	ΔE_{tot}	RMSD	σ_{RMSD}^2
<i>RDC 1 set (method 2)</i>								
1I6Z	-365.8	4.7	0.7	0.7	0.6	0.3	0.7	0.0
1G2H	-201.0	26.9	31.4	7.1	5.7	9.1	6.9	0.5
1DV5	-194.4	7.9	76.8	6.8	6.4	33.8	8.5	2.2
1J7O	-324.0	14.0	42.6	10.1	7.6	25.6	10.0	1.3
1A6S	-385.5	26.3	39.3	9.5	7.9	31.0	9.6	0.6
2ABD	-315.4	28.6	57.1	9.7	0.8	60.5	10.2	18.6
<i>RDC 2 sets (method 3)</i>								
1I6Z	-365.8	4.7	0.6	0.6	0.6	0.0	0.6	0.0
1G2H	-170.3	12.4	2.5	4.3	3.0	19.0	6.1	3.3
1DV5	-183.2	13.7	3.9	8.3	1.7	17.4	4.9	7.0
1J7O	-305.2	14.9	2.8	3.3	3.2	0.5	8.0	6.4
1A6S	-359.9	23.8	30.8	9.3	5.2	76.0	8.9	1.6
2ABD	-261.6	6.0	3.3	5.6	0.5	27.7	5.7	10.8
<i>RDC 5 sets (method 4)</i>								
1I6Z	-360.8	3.1	0.5	0.6	0.6	0.0	0.6	0.0
1G2H	-174.6	16.5	2.4	4.2	1.9	13.1	4.6	5.4
1DV5	-153.8	4.3	3.0	2.1	1.8	3.7	2.2	0.1
1J7O	-302.3	15.9	1.9	3.3	3.2	1.2	4.4	2.6
1A6S	-355.2	51.3	4.3	5.1	2.7	11.8	6.7	3.8
2ABD	-274.6	18.2	2.1	5.7	0.5	31.8	5.2	19.9

E_{orient} is the energy penalty for violating the orientational constraints; other column headings are defined in the caption for Table 3

Table 5

Summary of structural data for distance constraint simulations

PDB code	Lowest energy				Best RMSD		Average	
	E_{int}	E_{dist}	δ_{IA}	RMSD	RMSD	ΔE_{tot}	RMSD	σ_{RMSD}^2
<i>PRE (method 5)</i>								
1I6Z	-365.2	43.0	2.5	1.9	1.9	0.0	7.7	6.6
1G2H	-203.7	5.1	55.7	6.0	3.9	9.3	4.8	0.7
1DV5	-208.3	8.2	24.7	3.2	3.1	8.8	5.1	4.7
1J7O	-321.9	14.1	32.4	6.5	3.1	7.3	4.7	1.4
1A6S	-359.2	11.2	10.3	4.4	3.4	33.8	5.7	4.5
2ABD	-321.6	11.3	8.5	2.2	1.7	22.1	3.2	5.0
<i>EPR (method 6)</i>								
1I6Z	-372.7	0.8	3.1	1.5	0.8	5.9	1.0	0.1
1G2H	-232.7	14.8	36.3	7.1	4.4	13.3	5.5	1.1
1DV5	-214.3	9.5	41.4	6.5	4.9	4.5	5.7	0.6
1J7O	-392.2	26.7	60.1	8.9	6.3	8.9	8.5	3.0
1A6S	-343.9	14.2	34.0	6.6	3.5	8.6	6.9	2.8
2ABD	-381.1	14.7	19.6	4.6	3.2	64.0	6.1	6.2

E_{dist} is the energy penalty for violating the distance constraints (Eq. 4); other column headings are defined in the caption for Table 3

Table 6

Summary of structural data for combined distance and orientational constraint simulations

PDB code	Lowest energy					Best RMSD		Average	
	E_{int}	E_{orient}	E_{dist}	δ_{IA}	RMSD	RMSD	ΔE_{tot}	RMSD	k18
<i>RDC and PRE (method 7)</i>									
1I6Z	−381.3	11.2	3.9	0.8	0.8	0.8	0.0	8.0	10.7
1G2H	−166.6	11.0	2.6	6.8	2.3	2.1	2.9	3.3	2.6
1DV5	−189.2	8.4	7.6	12.5	3.1	1.2	23.7	4.9	8.9
1J7O	−274.1	4.3	7.6	4.5	2.6	2.3	11.9	5.0	4.6
1A6S	−295.5	7.3	1.6	16.5	2.6	2.1	13.6	7.2	6.9
2ABD	−255.6	3.4	0.9	4.7	1.7	1.0	6.6	1.3	0.1
<i>RDC and EPR (method 8)</i>									
1I6Z	−366.5	5.3	0.0	0.7	0.7	0.6	2.2	4.0	23.1
1G2H	−188.8	10.0	7.3	3.3	2.3	2.3	0.0	3.7	3.8
1DV5	−205.0	11.4	1.9	79.7	7.1	1.2	43.6	4.8	6.4
1J7O	−304.3	13.2	26.4	3.5	3.7	3.2	46.6	7.7	3.6
1A6S	−268.5	1.7	3.8	3.9	1.5	1.5	0.0	7.8	5.0
2ABD	−261.9	5.0	13.7	3.2	3.0	1.1	13.2	3.3	4.0

Column headings are defined in the captions for Tables 3 and 4