Published in final edited form as:
J Biomol NMR. 2011 August ; 50(4): 357–369. doi:10.1007/s10858-011-9521-5.

Backbone resonance assignment and order tensor estimation using residual dipolar couplings

Paul Shealy,
Department of Computer Science and Engineering, University of South Carolina, 315 Main Street, Columbia, SC 29208, USA

Yizhou Liu,
Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30603, USA

Mikhail Simin, and
Department of Computer Science and Engineering, University of South Carolina, 315 Main Street, Columbia, SC 29208, USA

Homayoun Valafar
Department of Computer Science and Engineering, University of South Carolina, 315 Main Street, Columbia, SC 29208, USA

Homayoun Valafar: homayoun@cse.sc.edu

Abstract

An NMR investigation of proteins with known X-ray structures is of interest in a number of endeavors. Performing these studies through nuclear magnetic resonance (NMR) requires the costly step of resonance assignment. The prevalent assignment strategy does not make use of existing structural information and requires uniform isotope labeling. Here we present a rapid and cost-effective method of assigning NMR data to an existing structure—either an X-ray or computationally modeled structure. The presented method, Exhaustively Permuted Assignment of RDCs (EPAR), utilizes unassigned residual dipolar coupling (RDC) data that can easily be obtained by NMR spectroscopy. The algorithm uses only the backbone N–H RDCs from multiple alignment media along with the amino acid type of the RDCs. It is inspired by previous work from Zweckstetter and provides several extensions. We present results on 13 synthetic and experimental datasets from 8 different structures, including two homodimers. Using just two alignment media, EPAR achieves an average assignment accuracy greater than 80%. With three media, the average accuracy is higher than 94%. The algorithm also outputs a prediction of the assignment accuracy, which has a correlation of 0.77 to the true accuracy. This prediction score can be used to establish the needed confidence in assignment accuracy.

Keywords

Assignment; Residual dipolar coupling; Refinement; Protein; Structure; RDC; NMR

Correspondence to: Homayoun Valafar, homayoun@cse.sc.edu.

Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy is a unique and powerful technique for investigating molecular structures, dynamics, and interactions in relevant physiological conditions. A prerequisite to these studies is the assignment of the observed resonance frequencies to the corresponding atoms in the molecule. The modern assignment strategy makes use of a set of H–C–N triple resonance experiments that correlate the amide group of a given residue to carbons or protons of the preceding residue and those of its own (Leopold et al. 1994). This strategy is made possible by the development of ^{15}N and ^{13}C isotopic labeling techniques for bacterially expressed proteins. However, there are certain situations where this assignment strategy fails. For example, missing resonances due to exchange broadening or closely spaced proline residues can hinder unambiguous sequential assignment. Missing resonances are also a problem for large molecules, where considerable signal dampening may be encountered in the triple-resonance experiments. Furthermore, some proteins can fold incorrectly or fail to acquire the essential posttranslational modifications when expressed in bacteria. These proteins must be expressed in eukaryotic hosts, where uniform ^{15}N or ^{13}C labeling is unavailable or extremely costly. However, selectively labeling of a specific type of amino acids is achievable at manageable costs. An immediate consequence of selective labeling is that the traditional sequential assignment strategy, which requires continuous isotope labeling, fails.

For systems like those mentioned above, an alternative assignment strategy is needed. It is worth noting that the traditional assignment strategy was designed for de novo structure determination and is therefore indifferent to whether or not a structure is known for the target of interest. This means that 93% (60972 total protein structures and 64946 X-ray structures) of the contents of the Protein Data Bank (PDB; http://www.pdb.org) (Berman et al. 2000) is not available for immediate use in NMR assignment. The motivation behind this work is to apply existing structural knowledge to challenging assignment problems. Several computational methods have been developed for structure based chemical shift prediction. The traditional NOESY experiment is also useful if distinct proton–proton distance patterns are expected for different labeled sites. However, for deuterated proteins, it can be difficult to obtain sufficient NOEs to make this distinction. A more general distance-based assignment approach may come from strategically introducing an electron spin label that causes differential relaxation enhancements on the labeled sites. This approach is labor-intensive if multiple spin labels are needed to distinguish the isotope-labeled sites, since NMR experiments must be conducted one spin label at a time. Residual dipolar couplings (RDC) are another NMR measurable that can be useful for structure based assignment. RDCs pose global angular constraints that put restrictions on the relative orientations of different inter-atomic vectors (Bax and Grishaev 2005; Prestegard et al. 2000). These types of NMR data are complementary in nature, and existing works have combined their use for assignment with a known structure. Hus et al. (2002) formulate a weighted matching problem with N–H, C′–N, and C′–C^{a} RDCs, C^{a} and C$^{\beta}$ chemical shifts, and Nuclear Overhauser Effect values (NOEs). Langmead et al. (Langmead and Donald 2004; Langmead et al. 2004) use RDCs, NOEs, and chemical shifts in the context of the nuclear vector replacement algorithm. Jung and Zweckstetter (Jung and Zweckstetter 2004) develop the

MARS algorithm, which relies on N–H, C′–C^{α}, and N–C′ RDCs and C′ and C^{α} chemical shifts. It is worth mentioning that a successful combination of these NMR data in an assignment problem relies on robust interpretation of each individual type of data. Therefore any improvement in utility of any one of the data type can make a broad impact on other assignment approaches. Here we explore the possibility of using only backbone N–H RDCs in structure-based assignment. This work is partially motivated by the fact that for large proteins, perdeuterated proteins, and proteins produced in eukaryotic cells, ^{13}C chemical shifts or distinctive NOESY signals may not be available and therefore higher reliance on the more easily measured N–H RDC is expected.

Resonance assignment for proteins with a known X-ray structure allows a number of studies that are useful in a broader context. For example, resonance assignment for a set of residues makes it possible to examine molecular interactions through chemical shift perturbation or paramagnetic relaxation enhancement (PRE), which are not easily predicted from the structure alone. Another example is confirming that the solution-state conformation matches that of a known X-ray structure. For example, the X-ray structure 1HNG (Jones et al. 1992) exhibits more than 99% sequence identity to the NMR structure 1A64 (Murray et al. 1998), but the two structures have 20.9 Å of structural difference measured over the backbone atoms. The structural difference arises from two different folds that the sequence may adopt, one as a monomer, and the other as a metastable dimer. The method described in this paper allows structural validation in the presence of RDCs.

In this work we present *EPAR* (Exhaustively Permuted Assignment of RDCs), an algorithm for assigning NMR data to an already characterized structure using only RDC data acquired in multiple alignment media. The presented method extends the previous work reported by Zweckstetter (2003) and utilizes RDCs, the type of amino acid from which the RDC data originate, and a candidate structure. Acquisition of RDC data for large proteins in multiple alignment media is becoming more routinely accessible due to recent advances in spectroscopic and alignment methods (Gronenborn and Clore 1996; Ou et al. 2001; Prestegard et al. 2000). Developments in specific amino acid labeling make it possible to classify resonances according to their amino acid types. EPAR uses only backbone RDCs (such as $^{15}N–^{1}H$), making it applicable to cases in which carbon labeling or carbon related RDCs are unavailable. The utility is evaluated through experimental (when available) and synthetic data for eight monomeric and homodimeric protein structures. Our results show that *EPAR* can achieve resonance assignment and order tensor estimation simultaneously based on a structure model. A number of methods exist that estimate order tensors (or some of their components) using unassigned RDCs in the absence of structural information, either from one alignment (Clore et al. 1998) or multiple alignments (Mukhopadhyay et al. 2008; Miao et al. 2008) or in the presence of a structure (Zweckstetter 2003). Although proper estimation of alignment tensors is an intermediate step in assignment of RDC data to a given structure, there exist additional challenges. Previous work (Zweckstetter 2003) has illustrated some of these challenges. EPAR incorporates additional features that overcome these previously reported barriers that stand in the way of assignment of RDC data to an existing structure. EPAR is available for download on the web at http://ifestos.cse.sc.edu,

while future plans are to integrate EPAR into the REDCAT (Valafar and Prestegard 2004) software package.

Background and theory

Residual dipolar couplings

NMR has been used for a number of years to aid structure determination and refinement of biological macromolecules. The residual dipolar coupling (RDC) of a vector between two magnetically active atoms with spin ½ nuclei can easily be acquired by Nuclear Magnetic Resonance (NMR) spectroscopy when the molecule in question undergoes partial alignment imposed by an aligning medium (Bax 2003; Prestegard et al. 2000). The RDC interaction can be described by Eq. (1). The 3 × 3 matrix *S* is the *Saupe* order tensor matrix (Saupe and Englert 1963), a traceless and symmetric matrix with five independent variables. *S* describes the strength of alignment and orientation of the anisotropic tumbling of the molecule (Prestegard et al. 2000; Tolman et al. 1995). *V* represents an inter-atomic vector associated with a given RDC, expressed in the same molecular frame. The Jacobi method (Greshenfeld 1998; Press et al. 2002) provides a decomposition of the order tensor matrix into two matrices describing the principle order parameters of S (Prestegard et al. 2000; Valafar and Prestegard 2004) denoted by S' and the Euler rotation matrix *R* as shown in Eq. (2). S' describes the strength of alignment along the axes of the principal alignment frame. *R* relates the principle alignment frame to the molecular frame via rotation angles α, β, and γ.

$$D_{ij}=D_{\max}\vec{v}^{T}\times\begin{bmatrix} s_{xx} & s_{xy} & s_{xz} \\ s_{xy} & s_{yy} & s_{yz} \\ s_{xz} & s_{yz} & s_{zz} \end{bmatrix}\times\vec{v}=D_{\max}\vec{v}^{T}\times S\times\vec{v} \quad (1)$$

$$S=R(\alpha,\beta,\gamma)\times\begin{bmatrix} s'_{xx} & 0 & 0 \\ 0 & s'_{yy} & 0 \\ 0 & 0 & s'_{zz} \end{bmatrix}\times R(\alpha,\beta,\gamma)^{T}=R(\alpha,\beta,\gamma)\cdot S'\cdot R(\alpha,\beta,\gamma)^{T} \quad (2)$$

The RDCs from different interacting pair of nuclei within a protein can be collected into a single equation as shown in Eq. (3) where *x*, *y*, and *z* are the Cartesian coordinates for the vector *V* in Eq. (1).

$$\begin{bmatrix} x_1^2-z_1^2 & y_1^2-z_1^2 & 2x_1y_1 & 2x_1z_1 & 2y_1z_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^2-z_n^2 & y_n^2-z_n^2 & 2x_ny_n & 2x_nz_n & 2y_nz_n \end{bmatrix}\begin{bmatrix} S_{xx} \\ S_{yy} \\ S_{xy} \\ S_{xz} \\ S_{yz} \end{bmatrix}=\begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} \quad (3)$$

When given the vector coordinates and the associated RDC values, a best-fit order tensor can be computed (Losonczi et al. 1999; Valafar and Prestegard 2004) by using the Singular Value Decomposition technique. This order tensor can be used to back-compute a set of RDCs. Throughout this work, the difference between the experimental and back-computed

RDC values, *e* and *c* respectively, is expressed through an average *Q*-factor score (Bax et al. 2001) across all media:

$$\overline{Q}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\frac{\sum_{j=1}^{n}(e_{ij}-c_{ij})^2}{\sum_{j=1}^{n}e_{ij}^2}} \quad (4)$$

In this equation *m* is the number of alignment media and *n* is the number of RDCs in each alignment medium (note that the number of RDCs may vary across different alignment media). The *Q*-score represents an error factor between the vectors' experimental and computed RDC values across all media and is used to evaluate a candidate assignment. When a set of RDCs is tentatively assigned to a set of vectors, the correct assignment will have a low *Q*-score between the assigned and computed RDCs.

Theoretically, when the assigned RDCs for a structure are divided into non-overlapping subsets based on certain criteria such as amino acid types, a separate order tensor can be computed for each subset. These order tensors will be identical for a structure under the assumptions of a rigid molecule and perfectly measured RDCs. When measurement noise is present, the order tensors will be similar to within the level of experimental error in the RDCs. The EPAR algorithm considers all *candidate assignments* (potential assignments that it must keep or discard) within a subset by permutation in a manner similar to previous work (Zweckstetter 2003). EPAR leverages information about the order tensors to identify candidate assignments that are invalid. When several subsets of RDCs report similar order tensors, the order tensor and the assigned RDCs are more likely correct. On the other hand, if a certain subset reports an order tensor that is not comparable to the others, the order tensor from this subset is not further used and an assignment for residues within this subset is made based on a final consensus order tensor.

Pseudoassignments

Because of the inherent degeneracy of RDCs, unambiguous assignment of resonances may not be possible in all instances. For example, two parallel backbone N–H vectors will produce identical RDC interactions in any alignment media. In addition to this universal degeneracy, other more circumstantial cases of degeneracy may occur. These degeneracies are dependent on the order tensor and may result up to eight distinct vector orientations in two alignment media. The addition of RDCs from a third alignment medium is likely to eliminate these degeneracies. Under these pathological conditions, the degenerate vector orientations will yield identical RDC values. When considering the accuracy of EPAR's assignments, it is therefore important to account for the errors present in the RDCs. N–H RDCs from a protein usually fall in a ±20 Hz range, with a measurement error of as large as ±2 Hz. Therefore, the possibility exists that two RDCs can be so similar that they cannot be reliably distinguished from each other to within the experimental error (which may vary from experiment to experiment). To distinguish such cases, here we introduce the concept of *pseudoassignment*. A pseudoassignment is defined as a residue whose RDC values are correctly assigned within error bounds, regardless of whether the resonance assignment is correct. Our definition of pseudoassignment requires matching of RDCs to within the

experimental error within all media. If an assignment results in at least one distinguishable false assignment in any of the alignment media, it is categorized as false assignment and not a pseudoassignment. While there is only one correct assignment for each residue, there can be multiple pseudoassignments. The pseudoassignment accuracy for a protein (the percentage of residues that are pseudoassigned) will always be equal to or higher than the assignment accuracy (the percentage of residues that have correct resonance assignments). It is worth noting that a set of pseudoassignments does not imply fewer total assignment errors than a set with false assignments. For example, swapping two assignments with RDCs that differ by 5 Hz (with 2 Hz of experimental error) produces two false assignments and two errors, while cyclic-permuting three assignments of RDCs within 2 Hz of each other produces no false assignments but three pseudoassignments and three errors. Since complete and unambiguous assignment of resonances based on RDC is inherently limited to pseudoassignments, throughout the remainder of this article the terms assignment and pseudoassignment are used interchangeably unless they are explicitly distinguished from each other. The distinction between pseudoassignment and assignment may be critical to some applications and unimportant to others as discussed in “Discussion and conclusion”.

Materials and methods

Algorithm

EPAR extends the work reported by Zweckstetter (2003) in a number of ways to achieve a more reliable assignment. In this section we present a detailed description of the EPAR algorithm. We delineate the differences between EPAR and the previous work in “Methodological comparison of EPAR and PALES”. The overall operation of EPAR can be described in two stages: *segmented assignment* and *collective assignment* of RDCs. A flowchart of the EPAR algorithm is shown in Fig. 1. EPAR accepts RDCs from multiple alignment media, the residue type for each RDC, and a candidate structure. It produces as output the assigned RDCs, a set of order tensors, and an *assignment score* indicating the quality of the assignment. Although backbone ^{15}N–^{1}H RDCs is the only vector type used in majority of our analyses, the EPAR algorithm is capable of accepting other types of RDCs. The algorithm also requires the residue type of each RDC to be provided. This can be readily obtained through a number of means, such as selective amino acid labeling (Gronenborn and Clore 1996; Ou et al. 2001), C^{α} and C^{β} chemical shifts (Grzesiek and Bax 1993; Spera and Bax 1991), or ^{1}H chemical shifts (Pons and Delsuc 1999). Requiring the amino acid type from which a given RDC is originated may appear to impose an indirect dependence on the ^{13}C chemical shifts. Although this is true of conventional techniques, new approaches do not impose this requirement. Recent labeling techniques that selectively incorporate ^{15}N-labeled amino acids of a single type can be used for amino acid identification without the need for ^{13}C labeling. These methods have received attention in recent years (Chen et al. 2006; Tong et al. 2008; Whittaker 2007) as a way to provide selective labeling for proteins or environments that are unstable in the presence of high amounts of ^{13}C.

Segmented assignment—The initial strategy employed by EPAR is to partition the problem into computationally manageable *pools* of data. These pools are initially

constructed using the residue type labels, with one pool per amino acid. Each pool contains a set of internuclear vectors as well as a set of RDCs from multiple alignments. RDCs from different alignment media are paired subsequent to data acquisition based on common chemical shifts and are treated accordingly during the analysis. The maximum number of RDCs that can be assigned in a reasonable amount of time is twelve, and so pools with more than twelve entries are treated separately. Pools with fewer than six RDCs are combined to ensure each pool has a sufficient number of RDCs (between 6 and 12) and maximize the assignment success. When investigating small to mid-size proteins, a number of amino acids may be available for combining. Under such conditions, EPAR deploys a strategy in combining pools to maximize the joint information content of the merged pools.

When combining small pools in the early stages of the algorithm, it is advantageous to maximize the likelihood of success during the permutation phase. If the pools to be combined have RDCs that are approximately equal, the algorithm will have difficulty distinguishing between two nearly identical assignments. EPAR chooses the pools to combine by maximizing the separation between RDCs from the two pools. For a pool p, let $s(p)$ be the size (the number of vectors) of the pool and $v_{i,p}$ be the ith vector in the pool. Also let $r_m(v_{i,p})$ denote the RDC for the vector in alignment medium m. The *RDC difference* for two vectors is given by Eq. (5).

$$RDC(v_1, v_2) = \sum_m |r_m(v_1) - r_m(v_2)| \quad (5)$$

The RDC difference over two pools, called the pooling score P, is given by Eq. (6).

$$P(p_i, p_j) = \frac{1}{s(p_i)} \sum_{y=1}^{s(p_i)} \min_{1 \le z \le s(p_j)} RDC(v_{y,i}, v_{z,j}) \quad (6)$$

This equation describes the minimum separation between the RDCs from two pools. EPAR computes the pooling score between all pools and combines the two pools with the maximum pooling score.

EPAR proceeds to identify the best assignment of RDCs for each pool by permuting the pool's RDCs among its vectors in a manner similar to previous work (Zweckstetter 2003). For example, a pool that consists of 6 vectors and 6 RDCs will produce 720 possible permutations of assignments. Each permutation is a candidate assignment for the RDCs. The permutation's fitness to the RDCs is computed using Eq. (4). The permutation with the lowest error for the pool is retained as the best candidate assignment. This process is repeated for each pool (or merged pool), yielding the best candidate assignment for each pool. This may also yield the best estimated order tensor from each pool for each medium. Missing RDCs are given dummy values of 999 (by the REDCAT tradition) and are appropriately treated during the course of the algorithm. Prolines and the *N*-terminal residue are discarded when using N–H RDC data.

Collective assignment—Under ideal conditions, the permutation with the best fitness to the experimental data should constitute the actual assignment. However under practical conditions, the best permutation may be the incorrect assignment. EPAR eliminates this problem by engaging in a second phase of analysis. The general aim is to further improve the assignment accuracy by continually integrating individual pools of RDCs. The final outcome of this phase is one integrated pool of assigned RDCs encompassing the entire protein. More specifically, EPAR compares the distances between order tensors for all pairs of pools using the *M*-score (Mukhopadhyay et al. 2008). The *M*-score compares order tensors based on the difference in the RDCs they produce. This provides an intuitive, sensitive, and reliable measure for comparing order tensors. Pools with *M*-score distances less than 1.5 Hz are considered consistent and are merged into a single pool. The threshold of 1.5 Hz is determined based on the quality of the RDC data in our experiments and can be altered as needed. The pool with the largest number of RDCs after clustering (in number of residues) is denoted as the *converged* pool and is considered to have the most reliable order tensor. The remaining pools are denoted *problematic* pools, and are considered potentially erroneous or represent degenerate order tensor estimation and assignments. The pools that were merged to create the converged pool have an average distance called the *pool convergence score*:

$$C_p = \frac{1}{LN^2} \sum_{i,j \in N} \sum_{l \in L} M(O_{i,l}, O_{j,l}) \quad (7)$$

where N is the number of pools, L is the number of media, $O_{i,j}$ is the order tensor for pool i in medium j, and $M(A, B)$ is the M-score between order tensors A and B. This pool's order tensors are output by the algorithm as the best estimated order tensors for the input RDCs. EPAR then uses this order tensor to back-compute the RDCs for all problematic pools, and RDCs for these pools are assigned using the Hungarian matching algorithm (Kuhn 1955) on the input and back-computed RDCs. This algorithm pairs each input RDC with a back-computed RDC, minimizing the sum of the differences between the paired values. The distance metric used to compare input and back-computed RDCs is the Euclidean distance. Alternatively, the user may specify assignment manually for the problematic pools.

EPAR also outputs an *assignment score*, which shows a 0.77 correlation with respect to the literal assignments. The assignment score is a function of the *Q*-factor between the assigned and back-computed RDCs Eq. (4) and provides the means for assessing reliability of the proposed assignment.

The limit of twelve residues per pool arises from the permutation of the pool's RDCs among its vectors. This process yields a factorial computation time in the maximum pool size n: computing the best assignment for p pools is $O(p \cdot n \cdot n!)$. Matching the pools in the second phase requires $O(p \cdot n^4)$ time using the original Hungarian algorithm implementation. The entire algorithm typically runs in under 5 min on a desktop PC for a protein with 75 residues (75 RDCs) and RDC data from 3 alignment media.

Inclusion of a priori knowledge of the order tensors

A priori knowledge of alignment tensors can be useful in assisting the task of assignment. EPAR is capable of incorporating the axial and rhombic components of anisotropy (D_a/R) from each alignment medium to eliminate implausible assignment of RDCs. Although values of D_a/R can be estimated from the RDC histogram (Varner et al. 1996; Warren and Moore 2001), they should be used with caution. Figure 2a illustrates the order tensor using the min–max approach on the experimental RDCs (from backbone N–H) for the IgG-binding domain of Protein G (PDB 1P7E) and the corresponding powder pattern that has been produced from SVD-based order parameters. This is clearly an example where traditional methods of estimating D_a/R would produce faulty values and could therefore detract from proper assignment of RDCs. Recent work (Miao et al. 2008; Mukhopadhyay et al. 2008) has demonstrated the possibility of accurately estimating relative order tensors in such pathological cases when RDCs are available from multiple alignment media. Figure 2b illustrates the 2D-RDC hull (Mukhopadhyay et al. 2008) corresponding to the order tensors estimated in the absence of assignment or a structure. EPAR can incorporate an order tensor (including the orientational components of the anisotropy) estimated from the unassigned data to further improve its performance. Another source for an a priori order tensor estimate is a small number of assigned RDCs; an order tensor estimate can be obtained by REDCAT (Valafar and Prestegard 2004) and used as an order tensor filter. When a candidate assignment is considered, its order tensor should be reasonably close to the estimated tensor. Any candidate assignment, for which this does not hold, can be discarded.

Filtering assignments based on an estimated order tensor utilizes the *M*-score to compare two order tensors—one from the candidate assignment, the other from the estimate. The *M*-score for two similar order tensors will be lower than the expected value of the experimental error in the RDCs, although a conservative value of 3 Hz is used here. Assignments producing order tensors that are larger than 3 Hz from the provided estimate are discarded. The threshold of 3 Hz is sufficiently generous to allow significant error in the estimated order tensor, yet still allow for meaningful filtering. The average distance between the true and estimated order tensors in a recent study (Mukhopadhyay et al. 2008) was 0.65 Hz; the study used the same method for estimating order tensors that we use in this analysis. This parameter is nonetheless configurable by the user.

RDC assisted assignment of homo-multimeric proteins

EPAR is capable of assigning RDC data acquired from a homo-multimeric protein. We demonstrate results on two homodimers, but other multimeric structures can also be analyzed. EPAR handles these structures by allowing the internuclear vectors from each domain to be specified separately, along with the average observed RDC across all domains. EPAR then computes the average value for the leftmost matrix in Eq. (3) (Bansal et al. 2008) for each set of corresponding vectors across the domains. Because the symmetric axis of the dimer must be collinear with one of the 3 axes of PAF, this step is not required for a perfectly symmetric homo-dimer, as RDCs for both subunits are identical. In this case, the RDCs can be assigned to a single domain. However, for domains that lack perfect symmetry and undergo fast conformational exchange within the NMR measurement time-scale (~50 ms), the observed RDC must be treated as an average of the RDCs from each domain. Due

to various reasons, different pools of amino acids may report different order tensors (Zweckstetter 2003). The problem is further complicated for homo-multimeric proteins because slight differences between the two domains may cause the same pool of amino acids to report different order tensors for different domains. The presented treatment of RDCs eliminates this problem.

Methodological comparison of EPAR and PALES

Previous work by Zweckstetter (2003) has presented an approach that proceeds in a very similar manner to the segmented assignment phase of the EPAR. The strategy by Zweckstetter has been incorporated into the software package PALES (Zweckstetter and Bax 2000). PALES operates in a manner similar to EPAR by dividing the entire set of RDC data into pools of amino acids. PALES also performs an exhaustive permuted assignment of RDCs, in a manner similar to EPAR. Due to the computational complexity of permutated assignment (computational complexity of order $n!$), both approaches have an upper bound maximum on the size of each pool (10 for PALES and 12 for EPAR). Therefore both approaches will produce identical results for pools with five to ten amino acids. One of the main differences between the two algorithms arise in treatment of pools outside of this range.

In application to small proteins, it is likely that some pools of amino acids will contain fewer than five residues and therefore pose a problem for conventional permuted treatment of assignment. PALES has introduced the idea of combining pools of amino acids in order to overcome the problem of data sparcity. Although this is potentially a viable approach to treatment of this ill condition, no formal and automated approach to merging of small pools has been presented. Considering that different merging strategies may improve or decay the quality of assignment (Zweckstetter 2003), an optimal strategy that maximizes the assignment performance becomes prudent. EPAR incorporates a greedy-based optimal merging strategy of smaller pools of amino acids.

Finally, PALES assignment of RDCs is considered complete upon assignment of individual pools. While this approach provides the ideal assignment under the conditions of abundant data and little noise, it may otherwise produce faulty results. EPAR extends the robustness of assignment in the collective-assignment phase. During this phase of EPAR's operation, all pools are gradually merged until assignment of the entire protein is accomplished with one overall consistent estimated order tensor.

The performance of the two approaches has been tested in application to a small protein (Ubiquitin, 76 residues, RDC from 1D3Z) and a large protein (3P76, 271 residues, synthetic RDC data). Selection of these protein structures are inspired by the previous PALES work (Zweckstetter 2003). Data for 3P76 was generated using the order tensors listed in Table 2 with ±1 Hz of uniformly distributed error. As mentioned previously, the performance of both algorithms converge under ideal conditions (clean and abundant data). To illustrate the differences, we have provided results for stressed conditions by either reducing the available RDC data or by assigning the RDC data to a more distantly related structure. It is important to note that despite Ubiquin's reputation for being well-behaved in solution state NMR spectroscopy, it poses a challenging case due to the correlation of RDC data from two

alignment media. Figure 3 presents the correlation plot for the backbone N–H RDCs in two alignment media and the estimated 2D-RDC hull. During the assignment of RDCs using PALES, all pools of amino acids with fewer than five entries were merged with a compatible pool in order to achieve 6 or more RDCs. The merging of pools was repeated between 3 and 5 times and the best results were selected for comparison.

Target proteins

In addition to the two proteins that were used during comparison of EPAR and PALES, we tested EPAR on the eight protein structures listed in Table 1. These structures were selected so that they range in size from 46 to 208 residues and cover a variety of structure types, including all α-helical, all β-strand, $\alpha + \beta$, and dimers. X-ray structures were protonated with Xplor-NIH (Schwieters et al. 2003). We analyzed two dimers. The dimer 2FFG was constructed from the monomer available in the PDB file by rotation of the monomer about the axis of symmetry noted in the file. The dimer 2DWV was provided as a dimer in the PDB file and was unaltered. The two domains of the 2DWV exhibited as much as 0.6 Å of structural difference measured over the backbone atoms.

Simulated and experimental RDC data

We wish to explore the performance of EPAR on a wide variety of protein types. This provides a realistic assessment of EPAR's performance on conditions that might be encountered. Experimental data from multiple alignment media is not available from BMRB (Ulrich et al. 2008) for some structure types, such as an all β-sheet protein, a structure with a large number of residues (>200), or homo-dimeric structures. In these circumstances, we use synthetic data to allow investigation of these structure types. Furthermore, synthetic data provide a method for establishing the fundamental performance of an algorithm in a controlled manner. It is also useful in investigating an algorithm's behavior under non-standard conditions, such as an error level that is higher than normal. We used synthetic data for five structures.

We assigned experimental data obtained from BMRB for three previously reported structures 2KLV (Park et al. 2009), 1RWD (Tian et al. 2001) and 1D3Z (Cornilescu et al. 1998). Data from NMR based structures 1RWD and 1D3Z were assigned to their homologous X-ray structures 1BRF and 1UBQ respectively. These structures exhibited as much as 1.8 Å of structural difference measured over the backbone atoms between the NMR and X-ray structures. 2KLV is a native NMR structure. This is included nonetheless because it illustrates important conditions. 2KLV consists of two α-helices. This is a challenging assignment test because the N–H vectors for each helix are roughly parallel, creating a situation in which the RDCs from each helix vary far less than the RDCs from a typical globular protein. Due to practical circumstances, these proteins did not have a complete set of data. Since EPAR is built on the same computational engine as REDCAT, it possesses the same capabilities as REDCAT, including the ability to accommodate missing data. The percentage of missing data for each protein in each alignment medium is shown in Table 3.

Synthetic data for five structures was generated in three alignments using REDCAT (Valafar and Prestegard 2004) with an added ±1 Hz of uniformly distributed noise. Typically

observed order tensors (listed in Table 2) are used to compute synthetic data. These order tensors yield a range of RDCs that is comparable to the range observed in experimental RDCs (−19.5 to 12.2 for medium 1, −14.6 to 24.4 for medium 2, −26.8 to 17.0 for medium 3). Synthetic data for each dimer was computed for each monomer, then averaged using REDCAT. This averaging process has two convenient effects. First, this process of averaging produces effective order tensors that satisfy the prerequisite requirements for homo-multimeric proteins (i.e., the axis of symmetry that coincides with a principal axis of the effective order tensor). Second, this averaging process accommodates any structural in-homogeneity between domains of a homo-multimeric protein. This step is critical since structural difference between domains of a homo-multimeric proteins is common. For example, two domains of the dimer 2DWV exhibited 0.59 Å of backbone RMSD, and therefore the observed RDC must be appropriately averaged between the two domains. An assumption is made here that structural exchange between the two subunits is fast in the NMR measurement time scale (~50 ms) so that RDCs are averaged between the two subunits. If this is not the case, inhomogeneous splitting should be observed and can be identified during NMR data analysis. Experimental data for the three structures were downloaded from BMRB. The source of each experimental data set is noted in Table 1. The only data used is ^{15}N–^{1}H RDCs from two or three alignment media. The amino acid type corresponding to each RDC value was determined from the assigned data (Table 3).

EPAR allows incorporation of an estimated order tensor from the unassigned data. We used estimates from λ-maps (Mukhopadhyay et al. 2008) and nD-RDC (Miao et al. 2008) for 10GS and 2DWV to demonstrate the feasibility of this approach. A λ-map yields an estimated order tensor from two unassigned alignment media; because both structures have three alignments available, this analysis was performed separately for each combination of two media. nD-RDC (Miao et al. 2008) was used to estimate an order tensor from all three media simultaneously.

Results

Comparison to PALES

The results from applying PALES and EPAR to four datasets are presented in Table 4. Because PALES is unable to assign some pools from 3P76, the results for EPAR only include those residue types that PALES was able to assign. These are 9% of the total RDCs for the protein.

EPAR outperforms PALES on all four tests. In particular, EPAR significantly outperforms PALES on the focus of this work, which is backbone N–H RDCs from only two alignment media. The collective assignment phase of EPAR is especially useful here. It is able to identify several pools that are problematic and assign them using an estimated order tensor from the more reliable pools. EPAR also significantly outperforms PALES on 3P76. The collective assignment phase is critical here as well. EPAR identified that the pool of glutamines was improperly assigned and matched them using an estimated order tensor.

EPAR results

The results from applying EPAR to eight structures are summarized in Table 5. All structures have RDC data from at least two alignment media and some have experimental data available from three alignment media. The assignment statistics list three values. The *RDCs assigned* column contains two of these values, separated by a slash. The first value is the percentage of residues with correctly assigned RDCs. Because the RDCs are paired across alignments by their chemical shifts, there is no possibility of a residue having the correct RDC assigned in one medium but not in another. The second value is the percentage of correct *pseudoassignments*, which is the percentage of RDCs that were correctly assigned within error bounds across all media (discussed further in “Pseudoassignments”). *Assignment score* reported by EPAR is the confidence score for the RDC assignments and ranges between 0 and 100 (with a higher number indicating a better fitness). For the *two media* columns, when data from three media is available, all values reflect the averages across all combinations of two media.

EPAR reports excellent results on the structures tested. EPAR was able to assign more than 80% of the RDCs correctly for most structures using only N–H RDCs from two media. Furthermore, a majority of the pseudoassignment results are higher than 90%. The pseudoassignment accuracy is, on average, 10% higher than the assignment accuracy. Many of the mis-assigned RDCs are not far from the true value. Remarkably, EPAR was able to assign more than 94% of the RDCs correctly when using three media.

EPAR frequently achieves a pseudoassignment accuracy several percent higher than the assignment accuracy. The difference is as much as 10% higher for several structures. The difference is far higher for two media than three. This is because RDCs of degenerated sizes from two alignment conditions can be resolved from an extra alignment condition. Structures with reported accuracy higher than 90% have an average difference of 2.5% between the accuracy and pseudoassignment accuracy. The majority of assignments using three media fall into this category.

An assessment of the error levels for the data being assigned is insightful into the algorithm’s performance. Table 6 provides detailed information on the RDC errors for each structure. The second column lists the RDC *Q*-score across all alignment media, while the third and fourth columns give the number of RDCs with errors greater than 1 and 2 Hz, respectively. These statistics are for the true assigned RDCs, not the assignments generated by EPAR. Some of the errors greater than 2 Hz are as high as 5 Hz, but the majority lie in the range of 2–3 Hz. Data sets with many high RDC errors, >2 Hz, are more challenging to assign. This is due to the average range of observed N–H RDCs, which varies by alignment and structure but is −23 to 21 for the largest experimental dataset analyzed. An error of 2 Hz spans 5% of this total range; for a dataset with a small range, this could be as high as 10%. A RDC with an error greater than 2 Hz may be better matched to an incorrect vector. This makes it more likely that the permutation with the lowest error still contains mis-assigned RDCs. This is somewhat mitigated by the fact that a vector may have a high error in one medium but not another, leading to an average error across media that is between 1 and 2 Hz.

Predicting the assignment score

EPAR outputs an *assignment score* that may be used as a confidence score that exhibits a reasonably tight correlation (0.77 correlation) to the percentage of correctly assigned RDCs (literal assignments). A higher assignment score indicates that the assignment has a higher degree of reliability. When the true assignments are unknown, as will typically be the case, this prediction provides a valuable feedback about the validity of EPAR's results.

The assignment score is computed as a function of the *Q*-factor between the assigned and back-computed RDCs. A higher assignment score indicates that the assignment has a higher degree of reliability. Linear regression on the *Q*-factor yields the equation for the assignment score:

$$S = -65.81 * Q + 101.03 \quad (8)$$

where *S* is the assignment confidence score and *Q* is the *Q*-factor. Equation (8) can be used to predict the assignment accuracy for an assignment session when the true assignment is unknown. Figure 4 plots all assignment results for all of the tested proteins as a function of the assignment score as well as with Eq. (8). The assignment score has a correlation of 0.77 with the true assignment accuracy. Comparing the predicted and true accuracies for all structures analyzed yielded an average difference of 4.71 and standard deviation of 4.40.

The pool convergence score also has a correlation with the assignment accuracy. Unfortunately, it also has a very high correlation with the *Q*-factor, and so it provides little additional information. Determining the assignment score through linear regression on both the pool convergence score and *Q*-factor yielded a negligible improvement in prediction accuracy.

Dimeric proteins

EPAR achieved a perfect score on the dimer 2FFG for both two and three media. To further examine EPAR's performance on this dimer, we repeated the assignment experiments with 2 Hz of noise to the RDCs. The result was 85% accurate with data available from two alignment media and 100% accurate with data available from three alignment media.

2DWV was determined experimentally as a homodimer and is not perfectly symmetric. The backbone RMSD between the two domains over the entire protein (49 residues) is 0.59 Å. Residues 10–39 have an RMSD of 0.29 Å. The *N*-terminal and *C*-terminal fragments are much higher; residues 1–9 have an RMSD of 0.62 Å, while residues 40–49 have an RMSD of 0.86 Å. This degree of structural difference may correspond to RDC differences of as much as 5 Hz between the two sister RDCs (RDCs from the same residue but different domains). EPAR is nonetheless able to assign 2DWV well, with assignment accuracies of 100% for both two and three media. The dynamic averaging feature is invaluable in correctly handling this structure.

Discussion and conclusion

Study of RDCs presents some unexpected challenges. For example, it is a common expectation to reduce the task of RDC assignment to estimation of order tensors. There are several challenges that stand in the way of equating alignment tensor estimation problem with RDC assignment problem. For example, within the context of our presented work and previous work (Zweckstetter 2003), the estimated alignment tensors vary widely between pools, necessitating methods to consolidate discrepancies reported by each pool of RDCs. Second, due to a number of degenerate conditions, it is possible to obtain a reasonable estimate of order tensors with a false assignment of RDC data. These limitations have served as impediments that have stood in the way of assignment of RDC data.

Here we have presented an algorithm for resonance assignment using only RDCs. The algorithm uses $^{15}N–^{1}H$ RDCs from multiple alignments and is based on a known structure. The results presented in Table 5 demonstrate the strong potential of this method. The results for two media illustrate that this data set is sufficient for analyses that require reasonably high accuracy; adding a third medium results in accuracy rates that are frequently above 95%. Our algorithm provides an assignment score to evaluate the performance of the program by establishing a confidence in the assigned RDCs. In summary, EPAR can be considered as a viable alternative approach to assignment of NMR resonances to existing structures that are within 1.8 Å of the actual structure and produce the needed RDC data with less than 10% missing data. Assignment of RDC data to lower quality structures may fail, with the confidence score confirming the failed attempt. Structures with higher levels of missing data may have more pseudo-assignments.

We utilize both synthetic and experimental data in our analysis of EPAR. Synthetic data allows examination of a broader range of candidate structures, including an all β-sheet structure and two dimers. The algorithm's performance is not significantly different when using experimental data. 2KLV, with experimental data, has one of the best scores for two alignment media. Assignment of RDC data from 1D3Z to its homologous 1UBQ structure does have a pseudoassignment accuracy of 75%, but its assignment score is one of the lowest among all results for two media. This case represents a scenario that the structure model is inaccurate potentially due to improper placement of the backbone HN atoms. While fewer experimental data sets include a third alignment, the experimental data sets have results that are comparable to the results using synthetic data.

We present separate assignment results for pseudoassignments. The distinction between correct assignments and correct pseudoassignments is important for applications that rely only on RDCs, as opposed to assigning RDCs in order to use related chemical shifts, NOEs, or PREs. For example, the impact of pseudoassignments in a study that aims to determine a protein's solution-state structure by refining an x-ray structure using only RDCs may be negligible since by definition, pseudoassignment produces an alternative assignment of RDCs that do not deviate beyond the experimental error range. Any alteration of RDCs within the experimental error is not expected to have significant impact on RDC-based structure refinement or determination protocols. This is not true for applications that use other sources of data, such as NOEs or PREs. In these cases, one must use caution when

using assignment information from EPAR. The utility of EPAR results is highly compatible with newly emerging structure determination/refinement protocols that are entirely based on RDCs (Ulmer et al. 2003).

EPAR allows incorporation of an estimated order tensor from the unassigned RDCs. An estimate constrains the search space of possible order tensors. This increases the likelihood that the best assignment from each pool will be close to the true assignment, which in turn increases the likelihood that the algorithm will converge to the correct assignment. An estimated order tensor from the unassigned RDCs may come from a variety of sources, some of which are designed specifically for multiple alignment media. As these methods become increasingly sophisticated and more accurate, incorporating an estimated order tensor will become commonplace.

The dimers 2FFG and 2DWV are the structures with the best results of all, achieving a perfect accuracy of 100% with both two and three media. This may be due in part, to the averaging process, which reduces the effective noise by $1/\sqrt{2}$. Increasing the noise to 2 Hz for 2FFG also yielded comparable results, with an accuracy of 85% for two media and 100% for three media. A dimer is known to have an order tensor in which a tensor axis lies on the molecular axis of symmetry (Al-Hashimi et al. 2000). Incorporating this knowledge into the algorithm would likely further increase the accuracy score under conditions of high noise level.

Finally, RDC-only methods have well-known limitations. In particular, using only N–H values makes it impossible to distinguish between values from parallel vectors, leading to pseudoassignments. EPAR can be readily combined with other sources of NMR information such as chemical shifts, NOE, PRE or other sources or RDCs for further improvement in assignment reliability. For example, RDCs of similar values are fundamentally difficult to distinguish with a RDC-only method, but may be resolved by different distances of the underlying vectors to an electron spin label in PRE studies. To further distinguish a pseudoassignment from the unique correct assignment, other types of NMR data such as chemical shifts, NOE and PRE are called for. Additional RDC types, such as C^{a}–H^{a} or C′–N, could also prove useful in reducing the number of pseudoassignments if carbons can be matched to their bonded nitrogen atoms by HNCO or HNCA experiments. Future plans include incorporating all of these types of information.

Acknowledgments

This work has been funded by NSF Grant number MCB-0644195. The authors are grateful to the Rothberg fellowship at USC for support to PGS.

References

Al-Hashimi HM, Bolon PJ, Prestegard JH. Molecular symmetry as an aid to geometry determination in ligand protein complexes. J Magn Reson. 2000; 142:153–158. [PubMed: 10617446]

Bansal S, Miao X, Adams MWW, et al. Rapid classification of protein structure models using unassigned backbone RDCs and probability density profile analysis (PDPA). J Magn Reson. 2008; 192:60–68. [PubMed: 18321742]

Bax A. Weak alignment offers new NMR opportunities to study protein structure and dynamics. Protein Sci. 2003; 12:1–16. [PubMed: 12493823]

Bax A, Grishaev A. Weak alignment NMR: a hawk-eyed view of biomolecular structure. Curr Opin Struct Biol. 2005; 15:563–570. [PubMed: 16140525]

Bax A, Kontaxis G, Tjandra N. Dipolar couplings in macromolecular structure determination. Methods Enzymol. 2001; 339:127–174. [PubMed: 11462810]

Berman HM, Westbrook J, Feng Z, et al. The protein data bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]

Chen C, Cheng C, Chen Y, et al. Preparation of amino-acid-type selective isotope labeling of protein expressed in *Pichia pastoris*. Proteins. 2006; 62:279–287. [PubMed: 16283643]

Clore GM, Gronenborn AM, Bax A. A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information. J Magn Reson. 1998; 133:216–221. [PubMed: 9654491]

Cornilescu G, Marquardt JL, Ottiger M, et al. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. Journal of American Chemical Society. 1998; 120:6836–6837.

Greshenfeld, NA. The nature of mathematical modeling. Cambridge University Press; Cambridge: 1998.

Gronenborn AM, Clore GM. Rapid screening for structural integrity of expressed proteins by heteronuclear NMR spectroscopy. Protein Sci. 1996; 5:174–177. [PubMed: 8771212]

Grzesiek S, Bax A. Amino acid type determination in the sequential assignment procedure of uniformly 13C/15 N-enriched proteins. J Biomol NMR. 1993; 3:185–204. [PubMed: 8477186]

Hus J, Prompers J, Bruschweiler R. Assignment strategy for proteins with known structure. J Magn Reson. 2002; 157:119–123. [PubMed: 12202140]

Jones EY, Davis SJ, Williams AF, et al. Crystal structure at 2.8 Å resolution of a soluble form of the cell adhesion molecule CD2. Nature. 1992; 360:232–239. [PubMed: 1279440]

Jung Y, Zweckstetter M. Backbone assignment of proteins with known structure using residual dipolar couplings. J Biomol NMR. 2004; V30:25–35. [PubMed: 15452432]

Kuhn HW. The Hungarian method for the assignment problem. Naval Res Logist Q. 1955; 2:83–97.

Langmead CJ, Donald BR. An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments. J Biomol NMR. 2004; 29:111–138. [PubMed: 15014227]

Langmead CJ, Yan A, Lilien R, et al. A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. J Comput Biol. 2004; 11:277–298. [PubMed: 15285893]

Leopold M, Urbauer J, Wand A. Resonance assignment strategies for the analysis of NMR spectra of proteins. Mol Biotechnol. 1994; 2:61–93. [PubMed: 7866869]

Losonczi JA, Andrec M, Fischer MWF, Prestegard JH. Order matrix analysis of residual dipolar couplings using singular value decomposition. J Magn Reson. 1999; 138:334–342. [PubMed: 10341140]

Miao X, Mukhopadhyay R, Valafar H. Estimation of relative order tensors, and reconstruction of vectors in space using unassigned RDC data and its application. J Magn Reson. 2008; 194:202–211. [PubMed: 18692422]

Mukhopadhyay, R.; Shealy, P.; Valafar, H. Protein fold family recognition from unassigned residual dipolar coupling data. 2008. p. 633-638.

Murray AJ, Head JG, Barker JJ, et al. Engineering an intertwined form of CD2 for stability and assembly. Nat Struct Biol. 1998; 5:778–782. [PubMed: 9731771]

Ou HD, Lai HC, Serber Z, et al. Efficient identification of amino acid types for fast protein backbone assignments. J Biomol NMR. 2001; 21:269–273. [PubMed: 11775743]

Park SH, Son WS, Mukhopadhyay R, et al. Phage-induced alignment of membrane proteins enables the measurement and structural analysis of residual dipolar couplings with dipolar waves and lambda-maps. J Am Chem Soc. 2009; 131:14140–14141. [PubMed: 19761238]

Pons JL, Delsuc MA. RESCUE: an artificial neural network tool for the NMR spectral assignment of proteins. J Biomol NMR. 1999; 15:15–26. [PubMed: 10549132]

Press, W.; Teukolsky, SA.; Vetterling, WT.; Flannery, BP. Numerical recipes in C: the art of scientific computing. Cambridge University Press; Cambridge: 2002.

Prestegard JH, al-Hashimi HM, Tolman JR. NMR structures of biomolecules using field oriented media and residual dipolar couplings. Q Rev Biophys. 2000; 33:371–424. [PubMed: 11233409]

Saupe A, Englert G. High-resolution nuclear magnetic resonance spectra of orientated molecules. Phys Rev Lett. 1963; 11:462–464.

Schwieters CD, Kuszewski JJ, Tjandra N, et al. The Xplor-NIH NMR molecular structure determination package. J Magn Reson. 2003; 160:65–73. [PubMed: 12565051]

Spera S, Bax A. Empirical correlation between protein backbone conformation and C.alpha. and C.beta. 13C nuclear magnetic resonance chemical shifts. J Am Chem Soc. 1991; 113:5490–5492.

Tian F, Valafar H, Prestegard JH. A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. J Am Chem Soc. 2001; 123:11791–11796. [PubMed: 11716736]

Tolman JR, Flanagan JM, Kennedy MA, et al. Nuclear magnetic dipole interactions in field-oriented proteins—information for structure determination in solution. Proc Natl Acad Sci USA. 1995; 92:9279–9283. [PubMed: 7568117]

Tong K, Yamamoto M, Tanaka T. A simple method for amino acid selective isotope labeling of recombinant proteins in *E. coli*. J Biomol NMR. 2008; 42:59–67. [PubMed: 18762866]

Ulmer TS, Ramirez BE, Delaglio F, et al. Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. J Am Chem Soc. 2003; 125:9179–9191. [PubMed: 15369375]

Ulrich EL, Akutsu H, Doreleijers JF, et al. BioMagResBank. Nucleic Acids Res. 2008; 36:D402–D408. [PubMed: 17984079]

Valafar H, Prestegard J. REDCAT: a residual dipolar coupling analysis tool. J Magn Reson. 2004; 167:228–241. [PubMed: 15040978]

Varner S, Vold R, Hoatson G. An efficient method for calculating powder patterns. J Magn Reson. 1996; 123:72–80.

Warren JJ, Moore PB. A maximum likelihood method for determining D(a)(PQ) and R for sets of dipolar coupling data. J Magn Reson. 2001; 149:271–275. [PubMed: 11318629]

Whittaker J. Selective isotopic labeling of recombinant proteins using amino acid auxotroph strains. Methods Mol Biol. 2007; 389:175–187. [PubMed: 17951643]

Zweckstetter M. Determination of molecular alignment tensors without backbone resonance assignment: aid to rapid analysis of protein–protein interactions. J Biomol NMR. 2003; 27:41–56. [PubMed: 12878840]

Zweckstetter M, Bax A. Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. J Am Chem Soc. 2000; 122:3791–3792.

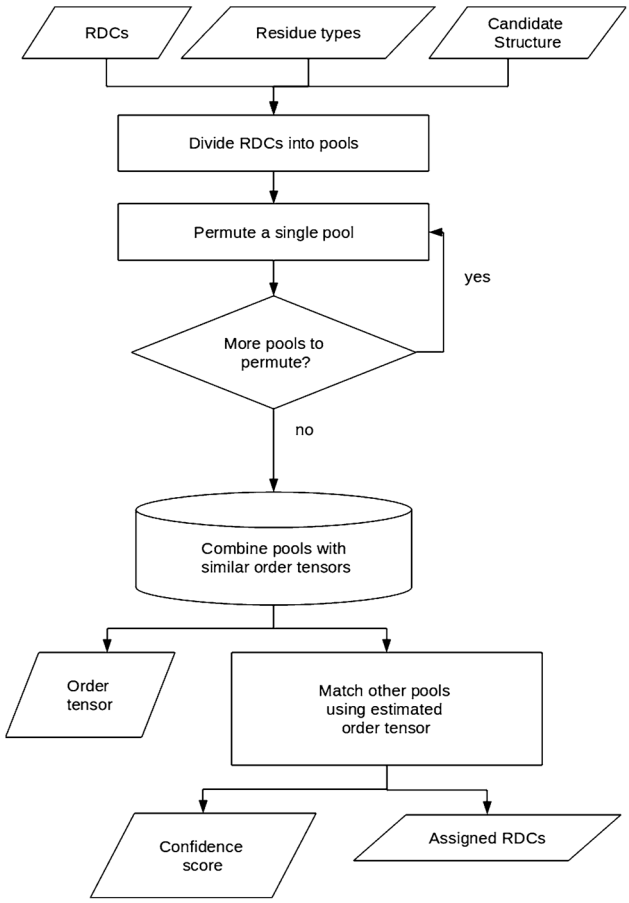

Fig. 1.
A flowchart of the EPAR algorithm

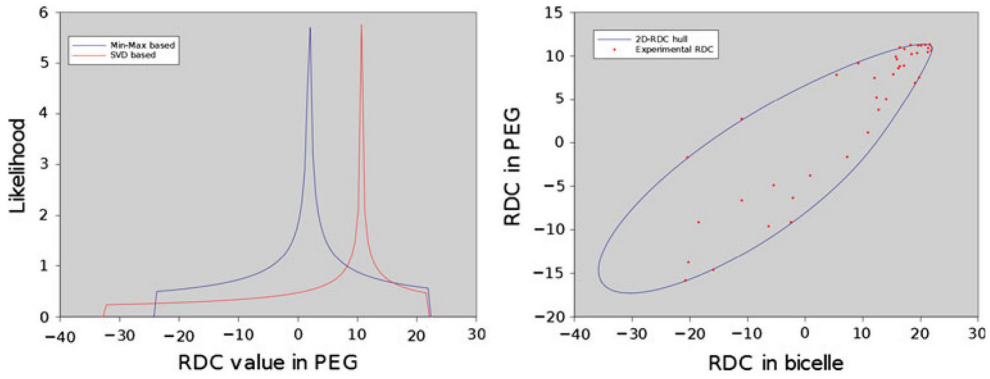

Fig. 2.

Order tensor estimates for the protein IgG-binding domain of protein G from **a** Min–Max and SVD methods, and **b** 2D-RDC method

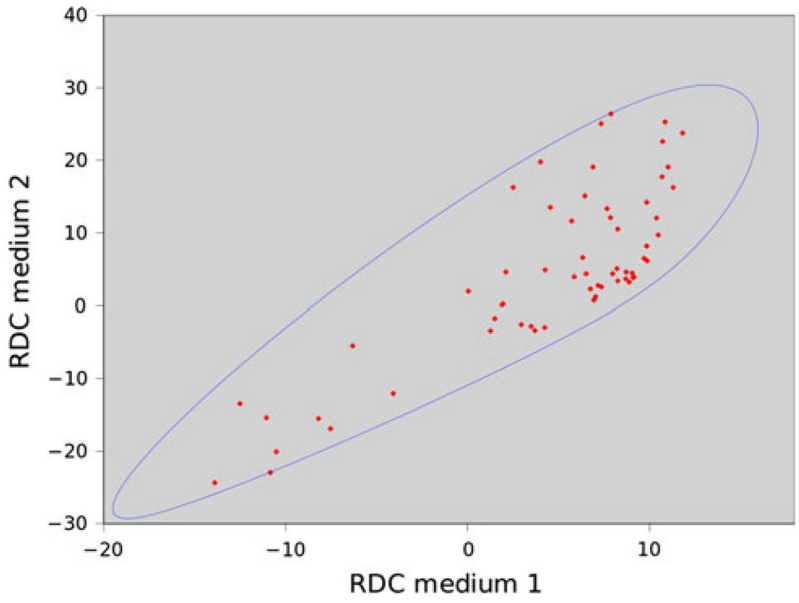

Fig. 3.

Correlation plot of RDC data for Ubiquitin in two alignment media, demonstrating linearity of the two data sets

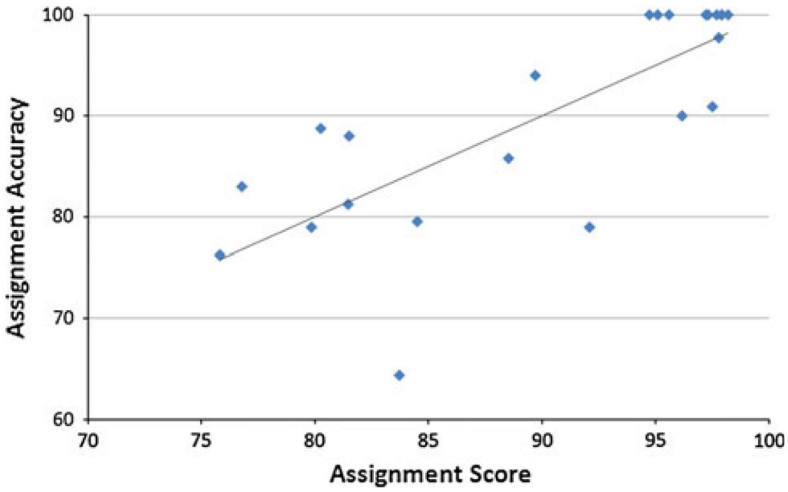

Fig. 4.

The assignment accuracy versus assignment score for all structures tested. The data includes points for both two and three media

Table 1

Structures tested with EPAR

PDB ID	Structure type	Number of residues	Data type	Number of alignments available
1A1Z	α	91	Synthetic	3
2GAL	β	135	Synthetic	3
2KLV	α	46	Experimental (Park et al. 2009)	2
10GS	α	208	Synthetic	3
1BRF	$\alpha + \beta$	53	Experimental (Tian et al. 2001)	2
2FFG	$\alpha + \beta$ dimer	87	Synthetic	3
2DWV	β dimer	49	Synthetic	3
1UBQ	$\alpha + \beta$	76	Experimental (Cornilescu et al. 1998)	2

Table 2

Order tensors used to compute synthetic RDCs for 1A1Z, 2GAL, 10GS, 2FFG, and 2DWV

	α	β	γ	S_{xx}	S_{yy}	S_{zz}
Medium I	0	0	0	3e–4	5e–4	−8e–4
Medium II	40	50	60	−4e–4	−6e–4	10e–4
Medium III	−70	−60	30	4e–4	7e–4	−11e–4

Table 3

Percentage of missing data for proteins 2KLV, 1RWD and 1D3Z in each alignment media

	Medium 1 (%)	Medium 2 (%)
2KLV	2	7
1RWD	0	7
1D3Z	5	9

Table 4

A comparison of results from PALES and EPAR as applied to four datasets

Structure	Data	Number of alignment media	Accuracy	
			PALES (%)	EPAR (%)
1UBQ	N–H, N–C	1	67	70
1UBQ	N–H	2	41	64
1AAR	N–H, N–C, CA–C	1	71	76
3P76	N–H, N–C	1	67	92

Table 5

A summary of EPAR results from a variety of structures. *RDCs assigned* has two values. The first is the percent of correctly assigned RDCs. The second is the percent of correctly assigned RDCs, taking into account pseudoassignments (i.e., all assignments within the error bounds)

PDB ID	Two media		Three media	
	RDCs assigned	Assignment score (%)	RDCs assigned	Assignment score (%)
1A1Z	82%/92%	79	100%/100%	97
2GAL	87%/94%	84	94%/97%	90
2KLV	91%/93%	97	N/A	N/A
10GS	81%/92%	87	98%/99%	98
1BRF	80%/90%	85	N/A	N/A
2FFG	100%/100%	98	100%/100%	98
2DWV	100%/100%	96	100%/100%	96
1UBQ	64%/75%	84	N/A	N/A

Table 6

RDC fitness for the true assigned RDCs

PDB ID	*Q*-factor score	Errors >1 Hz	Errors >2 Hz
1A1Z	0.0578	13	0
2GAL	0.0642	8	0
2KLV	0.0375	0	0
10GS	0.0478	16	0
1BRF	0.1973	38	14
2FFG	0.0472	1	0
2DWV	0.0825	13	5
1UBQ	0.1659	51	22