Protein Fold Recognition Through Application of Residual Dipolar Coupling Data

Y. Qu, J.-T. Guo, V. Olman, and Y. Xu

Pacific Symposium on Biocomputing 9:459-470(2004)

PROTEIN FOLD RECOGNITION THROUGH APPLICATION OF RESIDUAL DIPOLAR COUPLING DATA

Y. QU, J.-T. GUO, V. OLMAN, and Y. XU*
Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602, USA, and Computational Biology Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
(correspondence: xyn@bmb.uga.edu)*

Residual dipolar coupling (RDC) represents one of the most exciting emerging NMR techniques for studying protein structures. However, solving a protein structure using RDC data alone is a highly challenging problem as it often requires that the starting structure model be close to the actual structure of a protein, for the structure calculation procedure to be effective. We report in this paper a computer program, RDC-PROSPECT, for identification of a structural homolog or analog of a target protein in PDB, which best matches the ^{15}N-^{1}H RDC data of the protein recorded in a single ordering medium. The identified structural homolog/analog can then be used as a starting model for RDC-based structure calculation. Since RDC-PROSPECT uses only RDC data and predicted secondary structure information, its performance is virtually independent of sequence similarity between a target protein and its structural homolog/analog, making it applicable to protein targets out of the scope of current protein threading techniques. We have tested RDC-PROSPECT on all ^{15}N-^{1}H RDC data (representing 33 proteins) available in the BMRB database and the literature. The program correctly identified the structural folds for ~ 80% of the target proteins, significantly better than previously reported results, and achieved an average alignment accuracy of 97.9% residues within 4-residue shift. Through a careful algorithmic design, RDC-PROSPECT is at least one order of magnitude faster than previously reported algorithms for principal alignment frame search, making our algorithm fast enough for large-scale applications.

1 Introduction

Since the publication of the seminal work by Tolman et al.[1] and Tjandra & Bax,[2] residual dipolar coupling (RDC) in weak alignment media has gained great popularity in solving protein structures using NMR techniques. RDC provides information about angles of atomic bonds, e.g., N-H bonds, of a protein's amino acids with respect to a specific 3-dimensional (3D) reference frame. Using such information, an NMR structure could, at least theoretically, be solved through molecular dynamics (MD) simulation and energy minimization, under the constraints of the RDC angle information. A key advantage of RDC-based NMR structure solution is that RDC data can be obtained using a small number of NMR experiments and done in a very efficient manner.[3] Potentially, it could also overcome a number of limitations of traditional NOE-based NMR structure determination techniques, e.g., the size limit for a target protein.[4]

Though recognized for its great potential for solving larger proteins faster, direct application of RDC data for protein structure solution remains a highly challenging problem. The problem mainly comes from the well-known four-fold degeneracy

nature of RDC.[5] An RDC value of an N-H bond (for example) does not uniquely define a single orientation of the N-H bond as desired, rather it only restricts the orientation to two symmetric cones, making the search space of feasible structural conformations extremely large. In addition, inclusion of the RDC terms in the NMR energy function for structure calculation has resulted in a highly rippled energy surface with innumerable sharp local minima,[6] making the search problem exceedingly difficult. In the absence of long-range NOE distance information, it is practically intractable to find the global minimum by conventional optimization techniques. However, if the starting model is close to the true structure, convergence will become much easier. Therefore, a great amount of efforts have been made to obtain good starting structures for RDC-based NMR structure calculation.

Existing methods for deriving protein structures from RDC data alone mainly fall into two categories: *de novo* fragment assembly methods[7-10] and whole protein structural homology search methods.[11,12] *De novo* methods build protein structures by assembling structural fragments that are consistent with RDC data. These methods typically require a complete or near-complete set of RDC data to be effective, and are often very time-consuming. One example of such methods is the RosettaNMR program,[10] which typically need more than 3 RDC data per residue for its structure calculation to be accurate. As these methods typically attempt to assemble a protein structure in a sequential manner, they often suffer from problems resulting from accumulation and propagation of small errors from each individual fragment. Structural homology search methods generally require fewer RDC data and much less computing time, but are applicable only to proteins with solved homologous structures. Based on theoretical estimates on the total number of unique structural folds in nature and on the low percentage (< 5%) of novel structural folds among all structures submitted to PDB in the past few years,[13] people generally believe that the majority of the unique structural folds in nature are already included in PDB. Hence structural homology search methods are becoming increasingly popular. Annila *et al.*[11] are the first to use assigned RDC to search for structural homologs. Their work demonstrated the feasibility of fold recognition using RDC data alone. Meiler *et al.*[12] developed a program, DipoCoup, for structural homology search using secondary structure alignment. While all the aforementioned methods contain interesting ideas, they have been tested only on a very small set of proteins, in a few cases only on one protein, ubiquitin. Therefore, their true practical usefulness is yet to be determined.

We have recently developed a computer program, RDC-PROSPECT (**RDC**-**PRO**tein **S**tructure **Pr**ed**i**C**t**ion **T**oolkit), for protein fold recognition and protein backbone structure prediction. Currently the program uses only assigned N-H RDC data in a single ordering medium and predicted secondary structure to identify structural homologs or analogs from the PDB database. RDC-PROSPECT identifies a structural fold through finding a structural fold in PDB, which best matches the N-H RDC data, using a dynamic programming approach. Compared with existing methods, RDC-PROSPECT has a number of unique capabilities. Firstly, RDC-PROSPECT requires only a small number of RDC for fold recognition. On our test set consisting of all publicly available N-H RDC data of 33 proteins deposited in

the BMRB database (www.bmrb.wisc.edu) and published in the literature, RDC-PROSPECT achieves an 80% fold recognition rate on an average of 0.7 RDC data per residue. The requirement of fewer RDC data implies smaller number of NMR experiments needed to solve a structure. Secondly, RDC-PROSPECT does not require sequence similarity information for fold recognition, making the program equally applicable to proteins with only remote homologs or structural analogs in the PDB database, which represents a significant challenge to current threading methods. Thirdly, RDC-PROSPECT runs significantly faster than almost all existing RDC-based methods, using a novel search algorithm for the principal alignment frame of the RDC data.

2 Methods

An RDC measures the relative angle of an atomic bond in a residue, with respect to the principal alignment frame[14] of the protein (more rigorously, each rigid portion of the protein structure). The principal alignment frame, represented as an (x, y, z) Cartesian coordinate system, is dependent on the medium where the protein situates and the protein structure itself. In this paper, we consider only the RDC data of N-H bonds, the easiest RDC data to get experimentally. The RDC data measured by NMR experiments for each N-H bond is defined as[15]

$$D = D_a (3\cos^2\_ - 1) + 1.5\ D_r (\sin^2\_ \cos 2\_) \qquad (1)$$

where _ is the angle between the bond and the z-axis of the principal alignment frame (x, y, z), and _ is the angle between the bond's projection in the x-y plane and the x-axis; D_a and D_r represent the axial and rhombic component of the alignment tensor, respectively. Intuitively, D_a and D_r measure the magnitude (intensity) of the alignment. From an NMR experiment, we will get a set of $\{D_i\}$ values without knowing which D_i corresponds to the N-H bond of which residue in a protein and what the principal alignment frame is. Our goal here is to develop a computational procedure to find a protein fold in the PDB database and search for an (x, y, z) Cartesian coordinate system that produces a set of calculated N-H bond RDC values using equation (1), which best match the experimental RDC data. In this paper, we solve a constrained version of this fold recognition problem, assuming that the RDC data are already correctly assigned to individual residues.

2.1 Alignment of RDC data with structural fold

The RDC-based fold recognition problem can be rigorously stated as follows. Let $D = (D_1, \ldots, D_K)$ be a list of assigned experimental N-H RDC data (D_{NH}) of a target protein. Let $D^*(T, F) = (D^*_1, \ldots, D^*_M)$ be the calculated RDC data of a template structure T, assuming the principal alignment frame is F. We want to find an alignment A: i_A(i) between D and $D^*(T, F)$, that minimizes the following function:

$$\__i \left(\__1 |D_i - D^*_{A(i)}| / \_ \ + \ \__2 M (S_i, \ S^*_{A(i)}) \right) + \__j pG_j \qquad (2)$$

where D_i is aligned with $D^*_{A(i)}$, and _ is the standard deviation of the experimental D_{NH}; S_i and $S^*_{A(i)}$ are the predicted secondary structure type of position i of the target protein and the assigned secondary structure of position A(i) of the template structure; M() is a penalty function for secondary structure type match/mismatch, with M() equals –1 for match and 1 for mismatch; pG_j is the total gap penalty for the j-th gap in the alignment, which has the following form $a + L_j b$, with a being the opening gap penalty, b being the elongation gap penalty and L_j being the length of the j-th gap (the number of consecutive skipped elements). $_{-1}$ and $_{-2}$ are two scaling factors, which are empirically determined (using simulated data) as $_{-1} = 1$ and $_{-2} = 1$.

The $D^*(T, F)$ values of the template structure T are calculated using equation (1) for a specified alignment frame F (we will discuss how to systematically search for the correct alignment frame in the next subsection). To estimate D_a and D_r in (1), we use the equations in the histogram method proposed by Clore *et al.*:[16]

$$\begin{aligned} D_{zz} &= 2\, D_a \\ D_{yy} &= -\, D_a\, (1 + 1.5\, D_r/D_a) \end{aligned} \qquad (3)$$

where D_{zz} and D_{yy} are the maximum or the minimum values of the experimental $D_{NH,}$ respectively, with $|D_{zz}| > |D_{yy}|$. _ and _ in equation (1) are calculated for the N-H bond of each residue of the template structure with respect to the specified alignment frame F.

We have used PSIPRED[17] for secondary structure prediction of a target protein sequence. We consider three classes of secondary structures: helix (H), strand (E), and coil (C). In assessing secondary structure matches (using function M()), we consider only PSIPRED predictions with confidence level of at least 8 on the scale of 0-9. For a prediction with confidence level < 8, we assign a special category U (uncertainty) to this position and set $M(S_i, S^*_{A(i)}) = 0$ when $S_i = U$.

The alignment problem also employs a few additional rules as hard constraints, when aligning a list of RDC data with a protein structure. These include (a) if a position in the target protein does not have assigned RDC data, its corresponding alignment score (the D-portion in (2)) will be set to zero; (b) no penalty for gaps in the beginning and the end of a global alignment; (c) no alignment gap is allowed in the middle of an H- or E- secondary structure of the template structure; and (d) we consider alignment scores defined by (2) only for helix and strand regions while for coil regions, we penalize length difference of aligned coils. This is done for the following consideration: homologous proteins are generally more conserved among their corresponding core secondary structures (helices and strands) but not the coil regions. Considering detailed sequence alignment between coil regions often hurts the fold recognition and alignment accuracy, especially when dealing with remote homologs and structural analogs.

We have implemented a simple dynamic programming algorithm for finding the globally optimal solution of this alignment problem under the specified hard constraints. The dynamic programming algorithm consists of a set of recurrences, similar to the Needleman-Wunsch algorithm.[18] At each step of the recurrence calculation, the hard constraints are checked to guarantee no violation of constraints.

2.2 Assessment of prediction confidence

Considering that the alignment scores are not normalized with respect to the lengths and the composition of amino acids, we use Z-score to assess the quality of an alignment. For an RDC alignment problem with a set of experimental RDC data D_{NH} and a template structure T, we calculate the Z-score of the alignment score T_0 as follows. The RDC data with their respective secondary structure types are randomly shuffled multiple times. For each reshuffled RDC list, we calculate the alignment score with the template T. The Z-score of T_0 is defined as

$$Z = (T_a - T_0) / \_a \qquad (4)$$

where T_a and $\_a$ are the average alignment score of the reshuffled RDC lists and their standard deviation. For our current work, we run 500 times of reshuffling (we have also tried significantly larger number of reshuffling but found that 500 gives similar Z-scores to that with higher numbers). Figure 1 shows a plot of Z-score with respect to the fold recognition specificity on our test set of 33 proteins against our template structure database. For example, when Z-score is > 20, the prediction specificity is > 70%.

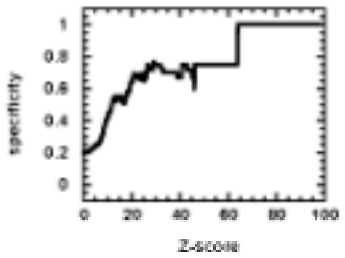

Figure 1. Fold recognition Z-score versus prediction specificity.

2.3 Principal alignment frame search and fold recognition

One of the challenging issues with the RDC-based fold recognition problem is that we do not know the principal alignment frame from the experimental data, which is required for the calculation of RDC values using equation (1). If the 3D structure of the target protein is known, this problem is equivalent to finding the correct rotation, in a fixed 3D Cartesian coordinate system of the structure that gives the (_, _)-angles of its N-H bonds and hence the calculated RDC values, which best match the experimental RDC data. For our fold recognition work, the problem is to find the rotation of a template structure that gives the best match with the experimental data, defined by equations (2) and (4). Note that any rotation of a 3D protein structure (say in PDB format) can be accomplished by a combination of clockwise rotations around x-axis by _ degree and around z-axis by _ degree. More specifically, the new coordinates of a data point [x, y, z], after a (_, _)-rotation, can be calculated as

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = R_z(\_)\, R_x(\_) \begin{bmatrix} x \\ y \\ z \end{bmatrix} \qquad (5)$$

where the two rotation matrices are defined as

$$R_x(\_) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix} \text{ and } R_z(\_) = \begin{bmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

For each given template structure, our fold recognition algorithm will search through all possible (_, _)-rotations. For each (_, _)-rotation, the algorithm employs the alignment algorithm of Section 2.1 to find the optimal alignment between the (assigned) experimental RDC data and the calculated RDC data for the template under this particular rotation. One thing to notice is that the range of both _ and _ is between 0 and 180 degrees as there is no need to consider $180 < \_, \_ \leq 360$ because of the four-fold degeneracy of RDC data.[20]

We have extensively tested and evaluated different increments for _ and _, ranging from 1 degree to 30 degrees. We found that the search surface (made of values of the calculated RDC) over the (_, _)-plane is very smooth, and an increment of 30 degree is adequate for our fold recognition. So we use 30 degrees as default increment value for RDC-PROSPECT. For each template, our algorithm will conduct 36 (6x6) RDC data alignments. The alignment with the optimal alignment score among the 36 alignments is considered the best alignment between the RDC data and the template. For cases we need to get very accurate alignment frame, we use a finer grid for searching the (α,γ) angles, which takes longer search time.

Our overall fold recognition procedure is carried out as follows. For each set of assigned RDC data, we search our template database consisting of all proteins in the SCOP40 database.[19] Currently, SCOP40 (release 1.63 of May 2003) consists of approximately 5,200 protein domains covering 765 folds and 2,164 families. Hydrogen atoms are added to the structure using the program REDUCE.[20] Secondary structure assignment is carried out using the program DSSPcont.[21] For each template, we calculate the Z-score of its best alignment with the experimental RDC data using equation (4). Then all the templates are ranked based on their alignment raw scores.

3 Results

We have tested RDC-PROSPECT on all publicly available N-H RDC data deposited in the BMRB database and published in the literature (by July, 2003), which contain 51 sets of RDC data for 33 proteins. The goal of the tests is to evaluate the fold recognition rate using RDC data (plus predicted secondary structure of a target protein) and the accuracy of the alignment with the correct structural folds. Tables 1 and 2 summarize the fold recognition and alignment results on the 33 proteins using 51 sets of RDC data – for some proteins, there are multiple sets of RDC data collected by different labs and/or in different ordering media.

For the fold recognition prediction, we consider a prediction as correct if a member protein from the same family or superfamily of the target protein is ranked in top three among all proteins in SCOP40, otherwise as incorrect. From Table 1, we can see that RDC-PROSPECT correctly identified the structural folds for 41 out of 51 RDC data sets (80.4% success rate), and identified 26 structural folds for 33 target proteins (78.8% success rate). Hence we consider the performance of RDC-PROSPECT as quite successful even under our very conservative definition of correct fold recognition, i.e. ranked among top three out of thousands of possible structures.

It is somewhat unfortunate that there is very little published data by other RDC-based structure prediction programs. Most of them were tested only on one protein, ubiquitin. The only meaningful comparison we can do is with RosettaNMR that was tested on 4 proteins using experimental RDC data, ubiquitin (1d3z), BAF (1cmz), cyanovirin-N (1ci4), and GAIP (2ezx), and 7 proteins using simulated RDC data.[10] Of the 4 proteins with experimental data, RosettaNMR predicted correct structures for 1d3z and 1cmz, and partially (~50%) correct structures for 1ci4 and 2ezm. Our program correctly identified the backbone structures for 1d3z, 1cmz, and 2ezx (the same protein as 1ci4), but did not find the correct structural fold for 2ezm due to inadequate secondary structure information (only 9.9% of the residues have reliable secondary structure prediction by PSIPRED).

From Table 2, we can see that alignment accuracy for the 26 target proteins with correct fold recognition is very high. The percentage of 4-shifts is commonly used for assessing threading alignment accuracy. RDC-PROSPECT achieved an average alignment accuracy of 97.9% residues aligned within 4-residue shifts to their correct positions. None of the other RDC-based structure prediction programs provide this kind of statistics.

Figure 2 shows the predicted structures (right) *versus* the actual structures (left) for four target proteins with < 25% sequence identity with their best structural templates.

4 Discussion

Our results have clearly demonstrated that RDC-based fold recognition, when

Table 1. A summary of fold recognition accuracy

Target No.	PDB code	Length	Data Set No.	template name	template length	Seq. Iden.	Rank	Z-score

1	1ap4	89	1	d2pvba_	107	19.1	1	10.2
			2	d2pvba_	107	19.1	1	10.2
			3	d2pvba_	107	19.1	1	10.5
2	1b4c	92	4	d1ksoa_	93	37.2	2	11.0
3	1brf	53	5	d1rb9__	52	64.8	1	7.0
			6	d1dx8a_	70	32.9	1	5.0
4	1c05	159	7	d1fjgd_	208	45.2	1	12.6
5	1cmz	152	8	d1dk8a_	147	28.8	1	10.1
			9	d1agre_	128	37.5	1	12.3
6	1d3z	76	10	d1bt0a_	73	59.2	1	12.5
			11	d1bt0a_	73	59.2	1	13.7
			12	d1h4ra3	84	20.4	1	14.4
			13	d1h8ca_	82	18.6	1	16.0
			14	d1bt0a_	73	59.2	1	13.4
			15	d1h4ra3	84	20.4	1	15.8
			16	d1h4ra3	84	20.4	1	15.2
			17	d1h4ra3	84	20.4	1	16.9
			18	d1bt0a_	73	59.2	1	14.7
			19	d1h4ra3	84	20.4	1	17.1
			20	d1bt0a_	73	59.2	1	13.1
7	1d8v	263	21	d1hwma_	251	37.0	1	97.5
8	1e8l	129	22	d31zt__	129	100	1	14.1
			23	d31zt__	129	100	1	12.8
9	1f3y	165	24	d1jkna_	165	99.4	1	14.8
10	1i42	89	25	d1i42a_	89	100	1	10.6
11	1j6t,A	148	26	d1a6ja_	150	24.2	1	24.4
12	1j6t,B	85	27	d1opd__	85	97.6	1	23.6
13	1j7o	76	28	d1exra_	146	49.0	1	21.3
14	1j7p	67	29	d1j7qa_	86	18.6	1	9.7
15	1jwe	114	30	d1b79a_	102	88.6	1	16.2
16	1khm	89	31	d1j4wa1	74	26.7	1	45.1
17	1kqv	76	32	d1irja_	85	28.7	1	11.4
18	1l3g	136	33	d1bm8_	99	72.8	1	16.5
19	1lud	162	34	d1ra9__	159	29.8	1	26.6
20	1n7t	103	35	d1mfga_	95	91.3	1	19.4
21	1ny9	90	36	d1ash__	147	18.2	1	12.6
22	2ezx	89	37	d1ci4a_	89	97.8	1	7.4
23	3eza,A1	123	38	d1zyma1	123	100	1	14.4
24	3eza,A2	125	39	d1zyma2	125	100	1	8.6
25	3eza,B	85	40	d1opd__	85	97.6	2	14.8
26	3gb1	56	41	d2igd__	61	82.0	1	13.2

Only the highest ranked correct template is listed for each protein. The first two columns represent the target id in our test and in PDB code. The third column represents the sequence length of the target. The fourth column represents the id of the RDC data set for each protein, some of which have multiple data sets. The fifth and sixth columns are the correct template id in SCOP code and its sequence length. The seventh column represents the sequence identity between a target protein and its correct template. The eighth column shows the rank of the top template among all SCOP40 proteins while the ninth shows the corresponding Z-score. No correct templates are identified in top three templates for proteins 27-33 (including 1d2b, 1ghh, 1o8r, 1qn1, 2ezm, 2gat, 4gat).

Table 2. Summary of alignment accuracy

Shift	0-shift	1-shift	2-shift	3-shift	4-shift
Accuracy (%)	63.1	90.1	95.3	96.8	97.9

x-shift represents the percentage of residues that are within x residues to its correct alignment positions.

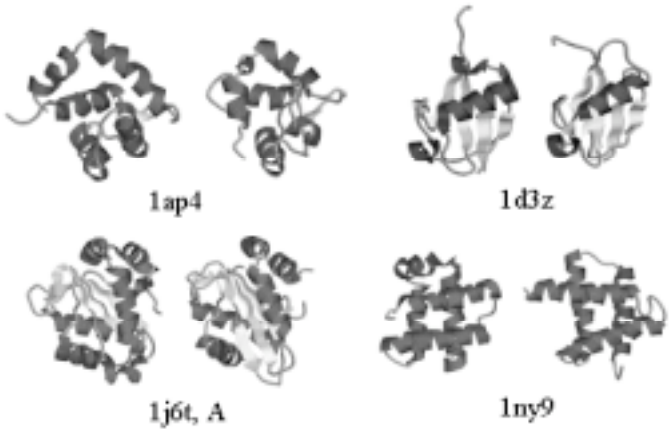

Figure 2. Actual (left) and predicted structure (right) on four target proteins with < 25% sequence identity with their best structural folds in SCOP40.

coupled with predicted secondary structure, is highly effective and robust for identification of native-like structural folds and prediction of its backbone structure. Our test examples cover a wide range of prediction scenarios. The test proteins span over 5 SCOP classes and more than 20 SCOP fold families with varying sequence lengths. Their N-H RDC data coverage ranges from 43.4% to 95.5%, and their predicted secondary structure ranges from 9.9% to 76.3% (for the remaining residues, their predictions are "uncertain" and hence not used). We now discuss some key advantages and unsolved issues of RDC-PROSPECT along with some future developments.

4.1 Efficient algorithm for alignment tensor orientation search

If we use N to represent the number of rotation angles we have to search along each axis, previous similar algorithms[9,22] all require N^3 combinations of rotations while our algorithm requires only N^2, saving at least one order of magnitude of search time and making our program much faster than other similar programs.

4.2 Combination of RDC data and predicted secondary structure for fold recognition

We found that predicted secondary structure, though not perfect, complements the RDC data for fold recognition. While RDC data are good for identification of global structural environment, secondary structure is good for finding the local structural environment (e.g., in a helix or in a strand). Our test data have shown that without either one of the two types of data, RDC-PROSPECT's performance drops significantly. In this work, we used predicted secondary structures based on protein sequence information only. Actually, secondary structures could be derived more

accurately using experimental data, like chemical shifts data. The only reason we did not use chemical shifts is that only 10 out of 33 proteins have such data available in the BMRB database. Using chemical shifts data will improve the performance of the program. For example, the otherwise missed correct template for the protein 2ezm can be identified when chemical shifts based secondary structure prediction is used.

4.3 Why some protein structures cannot be correctly predicted?

For 7 out of 33 target proteins, RDC-PROSPECT did not place the correct structural folds in the top three templates. We have done a detailed analysis on the failed predictions and found that the failures can be attributed to two classes of reasons.

a. proteins composed mainly of coils: this group includes 1o8r, 1qn1, 2gat, 4gat (6gat). As discussed in Section 2, RDC-PROSPECT considers only coil length conservation but does not conduct detailed alignment for coil region. When a protein is mainly composed of coils, RDC-PROSPECT does not perform well. Work is currently under way to improve on such cases.

b. others: we found that various other reasons could also contribute to the failure of our RDC-based fold recognition. The reasons range from inaccurate estimation of D_a and D_r, to incorrect prediction of secondary structures, to errors in the measured RDC data.

In this work, we have used raw RDC data without treatment of the data for contributions from internal dynamics. Our results suggest this is feasible in practice. As Rohl and Baker discussed,[10] internal dynamics likely contribute to the observed RDC to a greater content in flexible loops. Our method doesn't perform alignment in the coil region, so this greatly alleviates the effect of dynamics that could potentially harm the alignment.

4.4 Comparisons with DipoCoup

DipoCoup is a popular program to perform 3D structure homology search using RDC and pseudo-contact shifts together with secondary structure information. A basic problem with DipoCoup is that it does not use gap penalty in alignment, thus its applicability is significantly limited. In contrast, RDC-PROSPECT allows the flexibility of having gaps inside or outside secondary structures. Moreover, DipoCoup uses secondary structure fragment as alignment unit, while RDC-PROSPECT conducts alignments at residual level, making it more flexible and robust. This also allows us to use sparse secondary structure information, which DipoCoup could not handle.

4.5 Assignment of RDC data

Like other RDC-based structure prediction programs, RDC-PROSPECT assumes that the RDC data have been assigned to individual residues. This should not limit its applications, as sequential assignments of NMR data (RDC data included), unlike NOE data assignments, are generally solvable using existing programs. A recently

published work by Coggins & Zhou[23] has achieved ~80% assignments without any error for 27 test proteins using their PACES program. Assignments at such level are adequate for RDC-PROSPECT to perform well for most proteins. We have previously published an algorithm/software[24] for sequential assignments of NMR data using chemical shifts data. We are in the process of merging the two programs to do fold recognition using unassigned RDC data.

In conclusion, our method has convincingly testified the capability of fast and accurate protein fold recognition through combining sparse RDC data and threading technology. An important feature of our RDC-based homology search method is that it does not use sequence information for alignment. Our program provides a good complimentary and crosscheck tool to the conventional threading methods. It is especially attractive for the low sequence identity situations that the conventional structure prediction methods generally do not perform reliably. As we continue to work on this project, we will (a) use chemical shifts data for more reliable prediction of secondary structures, (b) include other types of RDC data, such as C-H RDC, which can be easily added into the framework of RDC-PROSPECT, and (c) include traditional statistics-based threading energy terms, such as pair-wise interaction potentials, in our RDC-based fold recognition method, as in our threading program PROSPECT.[25] We expect that RDC-PROSPECT will prove to be useful in structural genomics projects for high-throughput structure determinations, due to the efficient and effective application of RDC-PROSPECT to fit sparse RDC data with solved structures from a minimum number of NMR experiments.

Acknowledgments

This work was funded in part by the Structural Biology Program of the Office of Health and Environmental Research, U.S. Department of Energy, under Contract No. DE-AC05-000R22725 managed by UT-Battelle, LLC. We thank Drs. Nitin Jain, Dong Xu and Dongsup Kim for helpful discussions.

References

1. J.R. Tolman, J.M. Flanagan, M.A. Kennedy, and J.H. Prestegard, *Proc. Natl. Acad. Sci. U. S. A.* **92**, 9279 (1995)
2. N. Tjandra and A. Bax, *Science* **278**, 1111 (1997)
3. J.R. Tolman, *Curr. Opin. Struct. Biol.* **11**, 532 (2001)

4. J.H. Prestegard, *Nat. Struct. Biol.* 5 Suppl, 517 (1998)
5. J.H. Prestegard, H.M. Al Hashimi, and J.R.Tolman, *Q. Rev. Biophys.* **33**, 371 (2000)
6. A. Bax, *Protein Sci.* **12**, 1 (2003)
7. F. Delaglio, G. Kontaxis, and A. Bax, *J. Am. Chem. Soc.* **122**, 2142 (2000)
8. J.C. Hus, D. Marion, and M. Blackledge, *J. Am. Chem. Soc.* **123**, 1541 (2001)
9. F. Tian, H. Valafar, and J.H. Prestegard, *J. Am. Chem. Soc.* **123**, 11791 (2001)
10. C.A. Rohl and D. Baker, *J. Am. Chem. Soc.* **124**, 2723 (2002)
11. A. Annila, H. Aitio, E. Thulin, and T. Drakenberg, *J. Biomol. NMR* **14**, 223 (1999)
12. J. Meiler, W. Peti, and C. Griesinger, *J. Biomol. NMR* **17**, 283 (2000)
13. D. Lee, A. Grant, D. Buchan, C. Orengo, *Curr. Opin. Struct. Biol.* **13**, 359 (2003)
14. J.A. Losonczi, M. Andrec, M.W. Fischer, and J.H. Prestegard, *J. Magn Reson.* **138**, 334 (1999)
15. G.M. Clore, A.M. Gronenborn, and N. Tjandra, *J. Magn Reson.* **131**, 159 (1998)
16. G.M. Clore, A.M. Gronenborn, and A. Bax, *J. Magn Reson.* **133**, 216 (1998)
17. D.T. Jones, *J. Mol. Biol.* **292**, 195 (1999)
18. S.B. Needleman, C.D. Wunsch, *J. Mol. Biol.* **48**, 443 (1970)
19. A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia, *J. Mol. Biol.* **247**, 536 (1995)
20. J.M. Word, S.C. Lovell, J.S. Richardson, and D.C. Richardson, *J. Mol. Biol.* **285**, 1735 (1999)
21. P. Carter, C.A. Andersen, and B. Rost, *Nucleic Acids Res.* **31**, 3293 (2003)
22. J.C. Hus, J.J. Prompers, and R. Bruschweiler, *J. Magn Reson.* **157**, 119 (2002)
23. B.E. Coggins and P. Zhou, *J. Biomol. NMR* **26**, 93 (2003)
24. Y. Xu, D. Xu, D. Kim, V. Olman, J. Razumovskaya, and T. Jiang, *IEEE Computing in Science & Engineering* **4**, 50 (2002)
25. Y. Xu and D. Xu. *Proteins* **40**, 343 (2000)