Journal of Biomolecular NMR, **21:** 335–347, 2001.
KLUWER/ESCOM

Protein backbone structure determination using only residual dipolar couplings from one ordering medium

Michael Andrec, Peicheng Du & Ronald M. Levy*
Department of Chemistry, Wright-Rieman Laboratories, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, NJ 08854-8087, U.S.A.

Received 23 July 2001; Accepted 19 September 2001

Key words: database, overlap, protein fragments, proteomics, structural genomics, tree search

Abstract

Residual dipolar couplings provide significant structural information for proteins in the solution state, which makes them attractive for the rapid determination of protein folds. Unfortunately, dipolar couplings contain inherent structural ambiguities which make them difficult to use in the absence of additional information. In this paper, we describe an approach to the construction of protein backbone folds using experimental dipolar couplings based on a bounded tree search through a structural database. We filter out false positives via an overlap similarity measure that insists that protein fragments assigned to overlapping regions of the sequence must have self-consistent structures. This allows us to determine a backbone fold (including the correct Cα-Cβ bond orientations) using only residual dipolar coupling data obtained from one ordering medium. We demonstrate the applicability of the method using experimental data for ubiquitin.

Introduction

There exists a need for methods that would allow the more rapid determination of protein structure using NMR than can currently be attained, both from the viewpoint of traditional structural biology, as well as from the 'proteomics' and 'structural genomics' perspective. One of the principle rate-limiting steps in NMR structure determination is the sequential assignment of sidechain proton resonances and the assignment of NOESY crosspeaks to particular sidechain resonances. These steps are considerably more difficult and time-consuming than the sequential assignment of chemical shifts along the peptide backbone, for which relatively robust automated methods already exist (Moseley and Montelione, 1999). Therefore, it would be desirable to have a method for obtaining reliable structural information based on the smallest possible additional data collection beyond that needed for the backbone resonance assignments.

*To whom correspondence should be addressed. E-mail: ronlevy@lutece.rutgers.edu

Residual dipolar couplings are of particular interest for this purpose in that they require relatively little data collection time and provide considerable structural information through their dependence on the orientation of an internuclear vector relative to an order frame (Prestegard et al., 1999). The development of a variety of orienting media (such as lipid bicelles and filamentous phage) (Tjandra and Bax, 1997; Hansen et al., 1998; Clore et al., 1998) have increased the practicality of such measurements in recent years, and the use of residual dipolar couplings as a supplement to NOEs and scalar couplings in the refinement of high-resolution NMR solution structures is becoming more common (Tjandra, 1999).

Previous approaches to protein structure determination using residual dipolar couplings in the absence of NOEs have included fold recognition (Annila et al., 1999; Meiler et al., 2000), the searching of a database of protein fragments (Delaglio et al., 2000) to determine the fold, protein structural motif recognition (Andrec et al., 2001), and the direct fitting of peptide group orientations (Hus et al., 2001). All of these methods are limited to different degrees by the ori-

entational ambiguities arising from the many-to-one relationship between internuclear vector orientation and dipolar coupling. One way to reduce the impact of this ambiguity is by the use of dipolar couplings measured using at least two ordering media with different order tensor orientations (Ramirez and Bax, 1998; Al-Hashimi et al., 2000; Hus et al., 2001). Alternatively, one could accept a certain number of false positives and use a post-processing procedure to distinguish them from the true positives. We describe here an alternative method for the construction of a protein backbone fold using a protein fragment database which is similar in spirit to previous approaches (Delaglio et al., 2000; Bowers et al., 2000), but which makes use of a different kind of post-processing procedure.

Our approach differs from previous methods in that we filter out false positives by insisting that protein fragments assigned to overlapping regions of the sequence must have self-consistent structures. In brief, for each N_w-residue window of dipolar coupling data, we choose 15 protein fragments which best satisfy the residual dipolar couplings. Due to orientational ambiguities, these 15 hits in general contain both true positives (fragments which are similar in structure to that which generated the data) and false positives (fragments which are significantly different in structure from that which generated the data). We filter the hits by insisting that overlapping fragments be structurally similar: e.g. a selected hit for data window 1 (residues 1 through 7 with $N_w = 7$) must be structurally similar to the selected hit for data window 2 (residues 2 through 8) over the overlapping region from residues 2 through 7. Finding the best fragment for each data window leads to a combinatorial optimization problem that grows exponentially in the number of windows. However, we show that a bounded tree search algorithm allows the efficient search for optimal selections over a block of up to twenty windows, and that this is sufficient to define the backbone atoms of ubiquitin to a Cα RMSD of 2.8 Å after refinement with respect to the dipolar coupling data from only one orienting medium. Furthermore, the sidechain Cα-Cβ bond orientations are also correctly defined. This will make it much easier to construct all-atom models using the newest generation of sidechain conformation prediction algorithms.

Theory and methods

A residual dipolar coupling associated with a given internuclear vector is related to the orientation of that vector relative to an order tensor and is given by

$$D = D_a[(3\cos^2\theta - 1) + 3/2\, R\, \cos 2\phi\, \sin^2\theta], \quad (1)$$

where D_a is a constant which depends on the internuclear distance and the gyromagnetic ratios of the spins involved, R ($0 \leq R \leq 2/3$) is a measure of the asymmetry of the order tensor, and θ and φ are spherical angles which relate the internuclear vector to the principle axis system (PAS) of the order tensor (Prestegard et al., 1999). Alternatively, one can also rewrite Equation 1 in the form

$$D = (x\ y\ z) \begin{pmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{xy} & D_{yy} & D_{yz} \\ D_{xz} & D_{yz} & D_{zz} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad (2)$$

where D_{ij} are the elements of a symmetric and traceless matrix proportional to the Saupe order tensor (Saupe, 1968; Losonczi et al., 1999) in an arbitrary molecular frame defined by the direction cosines x, y, and z. It is clear from Equation 2 that the inversion of any internuclear vector through the origin leaves the dipolar coupling unchanged. Also, it is clear that the relationship between (θ, φ) or (x, y, z) and D is many-to-one, since there exist manifolds of (θ, φ) or (x, y, z) points which give rise to the same dipolar coupling, e.g. circles of constant θ in the case of R = 0. In the absence of additional data, for example from dipolar couplings obtained using different ordering media with different PAS orientations (Ramirez and Bax, 1998; Al-Hashimi et al., 2000), these degeneracies lead to orientational ambiguities which can give rise to false positive hits when searching a database or to structural ambiguity when constructing a structure *de novo* (Hus et al., 2001).

Since Equation 2 is linear in the tensor elements D_{ij}, it is possible to solve for the optimal D_{ij}'s which maximize the agreement between a set of bond vector orientations and the dipolar coupling data using a computationally efficient linear least squares procedure (Losonczi et al., 1999). It is therefore straightforward to fit an order tensor for each N_w-residue stretch of dipolar couplings to all fragments in a protein fragment database, back-calculate the best-fit dipolar couplings, and calculate a χ^2 statistic

$$\chi^2 = \sum_i (D_{calc,i} - D_{obs,i})^2, \quad (3)$$

where $D_{calc,\,i}$ and $D_{obs,\,i}$ are the back-calculated best-fit and experimental residual dipolar coupling for the i-th datum, and the index i runs over all data in the current window. In previous work (Andrec et al., 2001), we have found that the scaled χ^2 statistic

$$Q = \frac{\sum_i (D_{calc,\,i} - D_{obs,\,i})^2}{\sum_i D_{obs,\,i}^2}, \quad (4)$$

also known as a 'Q-factor' (Cornilescu et al., 1998), more clearly reflects the goodness-of-fit of a peptide structure and dipolar couplings than the simple χ^2 statistic when comparing between different data sets. For any given data set, however, Q is proportional to χ^2, and where appropriate we will use the two terms interchangeably. In addition, one can efficiently optimize a protein structure with respect to internal coordinates without an explicit representation of the dipolar coupling PAS by minimizing a projected χ^2 obtained after performing the linear fit of D_{ij} at each point in internal coordinate space (Golub and Pereyra, 1973; Moltke and Grzesiek, 1999).

A database consisting of 191 696 18-residue protein fragments from the SCOP40 (Brenner et al., 1998) was constructed. Data windows of length $N_w = 4$, 7, 9, 11, and 18 residues (including dipolar couplings for N-H_N, Cα-Hα, Cα-C, C-N, and C-H_N internuclear vectors) from the 'charged bicelle' data of Ottiger & Bax (Ottiger and Bax, 1998, Supporting Information Table 2) were fit to the first N_w residues of each fragment in this database, and a χ^2 value calculated. For example, if a window size $N_w = 7$ was chosen, then the dipolar coupling data for residues 1–7 were fit to the first 7 residues of database fragments 1, 2, ..., 191 696 and the χ^2 value was calculated for each. This was then repeated using the data for resides 2–8, 3–9, etc. Database fragments derived from domains having clear structural homology to ubiquitin with a CE z-score (Shindyalov and Bourne, 1998) of greater than 4.0 (including ubiquitin itself) were excluded. For each window, the 15 fragments with the smallest χ^2 values were saved for further filtering.

This filtering was done by selecting a single optimal fragment for each window within a given block of N_b windows k through $k+N_b-1$ such that the sum of the RMSD's for all overlapping regions from neighboring windows is minimized. For example, for the 20-window block of windows 1–20 with $N_w = 7$, we choose one fragment for each window from the 15 fragments selected above in such a way that the RMSD of the last six residues of the fragment for window 1 with the first six residues of the fragment for window 2 plus the RMSD of the last six residues of the fragment for window 2 with the first six residues of the fragment for window 3, etc. is minimized. Since the total number of choices is 15^{N_b}, a naive search of all possible choices is clearly impossible for all but the smallest values of N_b. However, due to the additive nature of our overlap score and the tree structure of the problem, it is usually necessary to examine only a very small fraction of these 15^{N_b} possibilities. To see this, consider the selection procedure as a tree (Figure 1) in which we begin with the empty set (level 0) and add a fragment corresponding to each successive window in the block until we reach the last window (level N_b). The nodes at level N_b then represent all possible 15^{N_b} selections. For each node at level 2 or greater, it is possible to calculate a partial overlap score corresponding to all of the fragments selected up to that point, e.g. for the node at level 3 which is circled in Figure 1, the partial overlap score would be the sum of the RMSD of the last six residues of fragment 1 in window k with the first six residues of fragment 1 in window k+1 and the RMSD of the last six residues of fragment 1 in window k+1 with the first six residues of fragment 2 in window k+2. Suppose that this partial overlap score is greater than the best total overlap score found thus far. Since the overlap score is additive and consists of nonnegative terms, it is guaranteed that all nodes which are decendents of this node will have an overlap score which is less optimal than the best selection found thus far. Therefore, it is not necessary to expand this node in searching the tree, thereby greatly reducing the computational burden. The technical details of the algorithms used are given in the Appendix.

Once we have found the optimal choice of fragment for each window, we construct a structural model by performing rigid body superpositions of the selected fragments. For each window $i = k, k+1, \ldots, k+N_b-2$ we translate and rotate the fragment for window i+1 so as to minimize the Cα RMSD with the fragment for window i over the $N_w - 1$ residues where they overlap (Kabsch, 1978). At this point, each residue position will have atomic coordinates from up to N_w fragments. We construct a consensus structure by taking the arithmetic mean of the atomic coordinates for each of the backbone atoms N, Cα, and C. These mean Cα coordinates are used to assess how close each block came to reproducing the 'true' ubiquitin structure (taken to be the X-ray structure in PDB accession code 1UBQ (Vijay-Kumar et al., 1987)) (Ta-

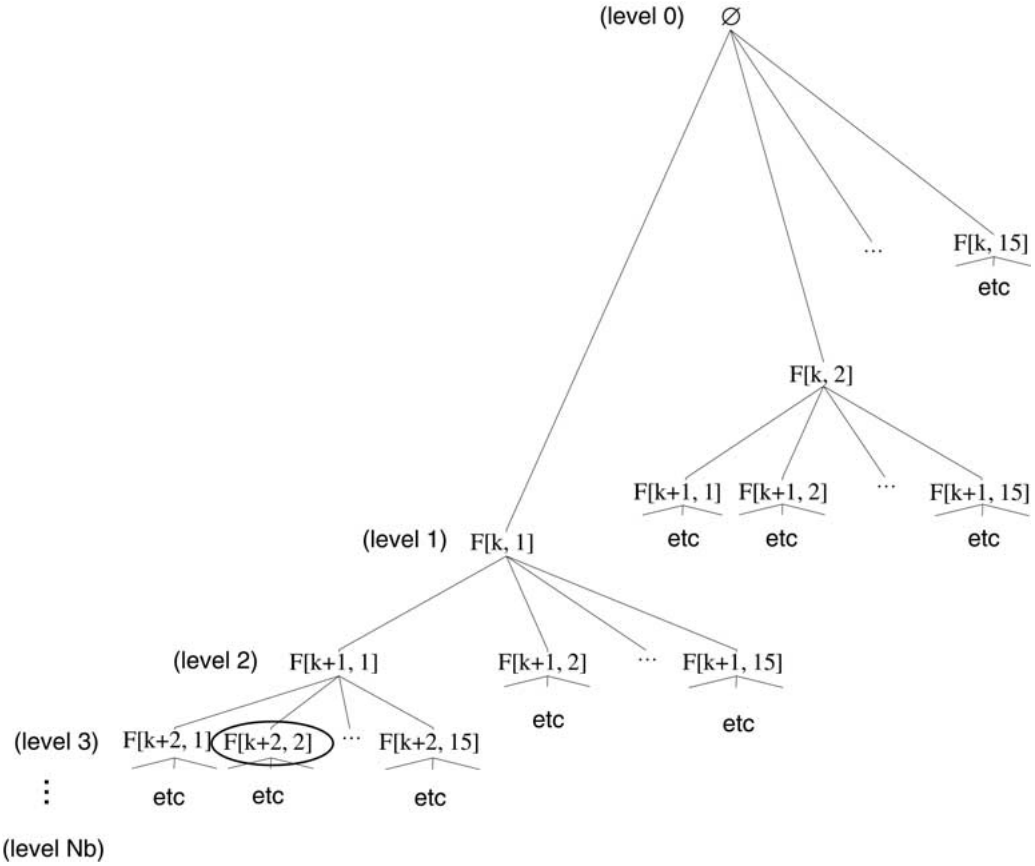

Figure 1. The tree structure of the fragment selection problem. Each node at level L > 0 represents the selection of one of the 15 possible fragments for the L-th window, and F[i, j] represents the fragment with the j-th smallest χ^2 in the database for the i-th window. There are 15^L nodes at each level L ($0 \leq L \leq N_b$), for a total of $(15^{N_b+1} - 1)/14$ nodes. The 15^{N_b} nodes at level N_b represent all possible fragment selections. See Theory and methods and the Appendix for complete discussion.

ble 1), and the mean N, Cα, and C are used to calculate backbone dihedral angles ϕ and ψ.

It should be noted that up to this point we have made no use of the 'long range' information inherent in residual dipolar couplings. In order to re-introduce this information, we further refine the backbone dihedral angles determined from the mean coordinates by direct minimization of the projected χ^2 (Golub and Pereyra, 1973; Moltke and Grzesiek, 1999) as a function of the ϕ and ψ angles for the entire protein using an ideal peptide geometry and peptide bond torsion angle ω fixed at 180°. This is done by sequentially minimizing the projected χ^2 as a function of (ψ_1, ϕ_2), (ψ_2, ϕ_3), (ψ_3, ϕ_4), ... (ψ_{75}, ϕ_{76}), (ψ_1, ϕ_2), ..., (ψ_{75}, ϕ_{76}), etc. using the simplex algorithm (Nelder and Mead, 1965) until convergence is reached. We have chosen to iteratively minimize over the two-dimensional subspaces (ψ_i, ϕ_{i+1}) rather than the individual angles because this allows for correlated changes which can lead to rotations of the peptide plane while keeping the Cα trace unchanged (Peticolas and Kurtz, 1980) (so-called 'crankshaft motion' (Fadel et al., 1995) or 'peptide plane rotation' (Parker, 1999)). We expect that this will allow for more efficient convergence. It should be emphasized, however, that at no time do we constrain the minimization to peptide plane rotations, but allow the minimization algorithm to find the nearest local minimum in the entire (ψ_i, ϕ_{i+1}) plane.

Results and discussion

Database search

It is well known that dipolar couplings obtained using only one orienting medium may be insufficient to uniquely determine a structure, because of the nature

Table 1. Results of fragment selection for ubiquitin using $N_w = 7$

Window block	Residues	$D(S_{greedy})$[a] (Å)	$D(S_{optimal})$[a] (Å)	# nodes (efficiency)[b]	CPU time for fragment selection (min)	RMSD of 1UBQ to mean coordinates of fragments[c] (Å)
1–20	1–26	3.57	2.15	6.7×10^8 (5×10^{14})	4905.0	1.56
21–25	21–31	0.35	0.23	499 (6×10^4)	0.6	0.37
26–35	26–41	2.02	0.87	4.4×10^6 (7×10^6)	33.7	2.03
36–50	36–56	2.28	1.53	2.8×10^7 (6×10^{11})	205.2	2.11
51–70	51–76	6.14	2.85	9.2×10^7 (3×10^{16})	674.3	2.14

[a] $D(S_{optimal})$ is the total overlap RMSD score for the optimal fragment selection. $D(S_{greedy})$ is the initial upper bound for the bounded tree search determined using a greedy algorithm (see Appendix).
[b] The number of nodes in the fragment tree of Figure 1 visited during the determination of $S_{optimal}$ using the depth-first bounded tree search (see Appendix). The number in parentheses is the increase in efficiency over naive exhaustive search (i.e. evaluation of all $(15^{N_b+1} - 1)/14$ nodes).
[c] The RMSD of the Cα positions defined by the arithmetic mean of the Cα atoms of the superimposed fragments (see Theory and methods) to the corresponding atoms of the crystal structure 1UBQ.

of the relationship between internuclear vector orientation and the observed coupling (Ramirez and Bax, 1998; Al-Hashimi et al., 2000; Hus et al., 2001). This can be seen quite prominently in Figure 2, where we show the results of the database search for two different window positions using $N_w = 7$. It is clear that for window number 52 (Figure 2a) there are true positives (green ellipse) which have a small χ^2 (or Q-factor) with respect to the dipolar couplings and a small RMSD to the true ubiquitin structure, false positives (red ellipse) which have a small χ^2 with respect to the dipolar couplings but a large RMSD to the true ubiquitin structure, and false negatives (cyan ellipse) which have a poor χ^2 with respect to the dipolar couplings but which nonetheless have a small RMSD to the true ubiquitin structure. The effect of the false positive hits can be quite deleterious, especially when the local structural information for a given residue is binned together irrespectively of what fragment it came from. In such a case, one will only see a lack of clustering and the information will be deemed ambiguous. The degree to which false positives contribute to the low-χ^2 hits varies significantly, however. For example, the low-χ^2 hits for window number 55 (Figure 2b) consist only of true positives.

This variability in the number of false positives for different data windows can be easily seen in Figure 3, where we plot the RMSD to the true ubiquitin structure for each of the 15 smallest χ^2 hits in each data window. It is clear that for $N_w = 7$ (Figure 3a), some regions (e.g. windows 13–18, 31–39, and 47–63) have many more false positives than others (e.g. windows 1–12 and 20–30). Some of this variability is easily explained. For example, windows 20–30 correspond to the alpha-helical region of ubiquitin (residues 23–34). Since helices are common and tend to have relatively little structural variability, there will be many true positive fragments in the database, many of which will have a small χ^2 when fit with data derived from a helical structure. Some of the regions with a large number of false positives correspond to regions in which there is a large amount of missing data, such as windows 31–39, which overlap the very data-poor region of residues 36–39. However, despite the false positives, there is still a sufficient number of true positives for us to expect to be able to construct a reasonably accurate structure. In this respect, it is useful to compare the best hits obtained via χ^2 filtering with the best hit that can be found in a fragment database based on structural similarity (RMSD) alone. The latter is indicated in Figure 3 by the red line, and corresponds to the best achievable fragment in SCOP90. Clearly, the best hits based on the χ^2 filtering do not deviate too far from this ideal result.

Intuitively, we expect that there is an optimal fragment size N_w, since on the one hand it should be

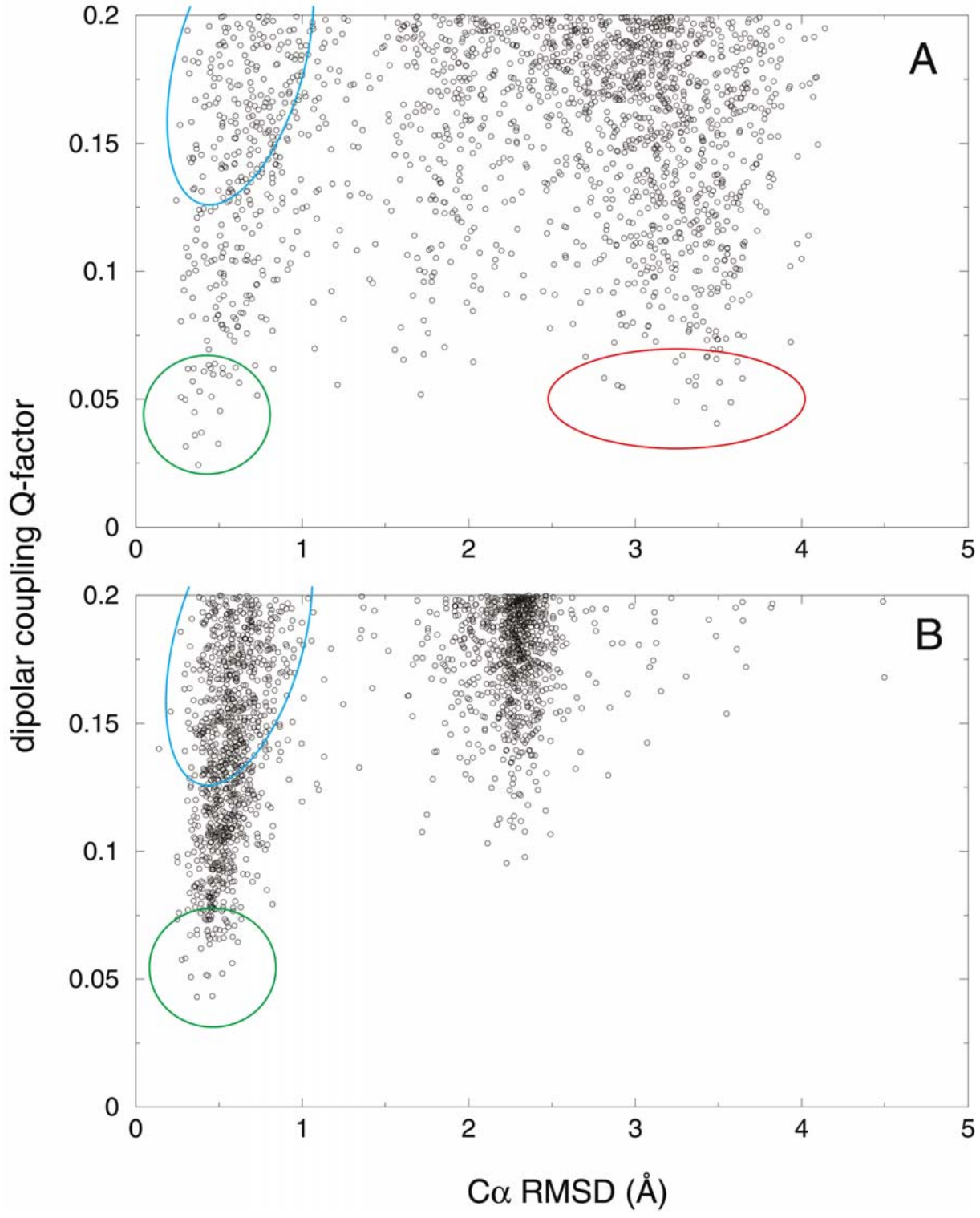

Figure 2. The results of the fragment database search for the seven-residue ubiquitin data windows 52 (a) and 55 (b), corresponding to residues 52–58, and 55–61, respectively. Each small circle represents one fragment in the database, and its position is given by the Q-factor for the fit to the dipolar coupling data and that fragment's Cα RMSD to the corresponding residues of the 1UBQ crystal structure. Only those fragments with a Q-factor less than or equal to 0.2 are shown. The green ellipses denote fragments which are true positives, the cyan ellipses denote false negative fragments, and the red ellipse denotes false positive fragments.

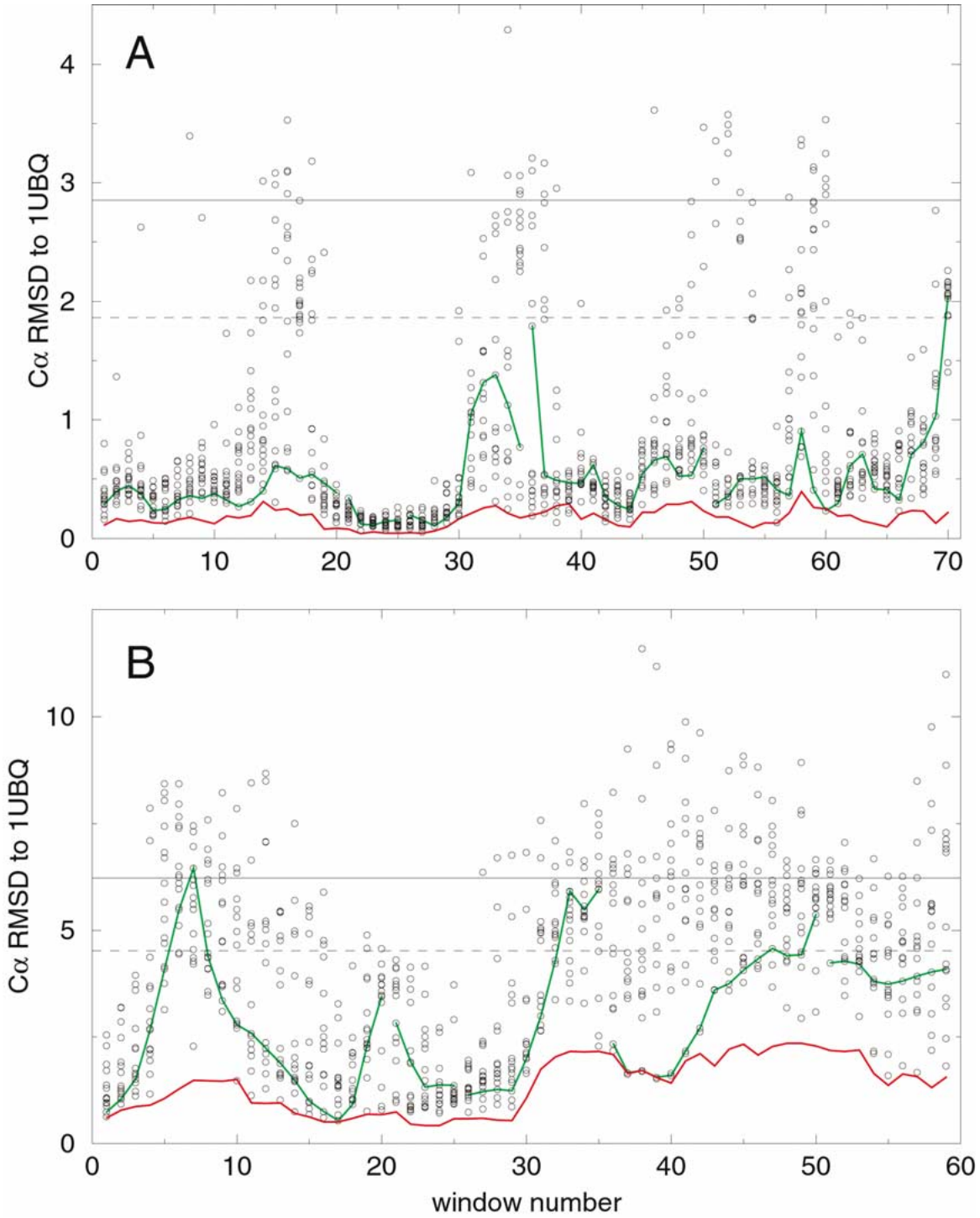

Figure 3. A summary of the fragment database search results for all data windows in ubiquitin using a window size $N_w = 7$ (a) and $N_w = 18$ (b). The small circles represent (for each window) the 15 fragments in the database with the smallest Q-factor, and the ordinate is the fragment's Cα RMSD to the corresponding residues of the 1UBQ crystal structure. The solid and dashed horizontal lines represent the mean and the mean minus one standard deviation of the Cα RMSD's of randomly chosen pairs of N_w-residue fragments. The red lines represents the RMSD of the fragment in a protein fragment database constructed using SCOP90 (Brenner et al., 1998) with the smallest RMSD to the corresponding residues in 1UBQ. Since the SCOP40-based fragment database is not a strict subset of the SCOP90-based fragment database, one can have fragments in the former that have smaller RMSD's to 1UBQ than the best fragment in the latter. The green lines represent the fragment selections with the optimal overlap score (see Theory and methods) for the five window blocks 1–20, 21–25, 26–35, 36–50, and 51–70 (Table 1).

easier to find a good match in the database when N_w is small, while the ability to filter out false positives increases with increasing N_w. The former is confirmed by comparing Figures 3a and 3b for $N_w = 7$ and 18, respectively. Since RMSD values increase with the length of the fragment irrespective of structural similarity, it is important to properly define the scale with respect to which one makes comparisons. One simple way in which this can be done is by comparing an RMSD value to the distribution of RMSD values for randomly selected pairs of protein fragments of the same length. These distributions for fragments of size 7 and 18 residues are summarized in Figure 3 by the solid and dashed horizontal lines, which represent the mean and the mean minus one standard deviation of the distribution. Not only is the false positive rate substantially greater for $N_w = 18$, but the best hits are also of poorer quality. This can be seen by noting that the best hits for $N_w = 7$ are approximately 2.5 standard deviations from the mean RMSD of randomly selected pairs, while for $N_w = 18$ there are large regions (particularly windows 32–53) for which the best hits are only 1.5 standard deviations from the mean. Results for $N_w = 9$ and 11 confirm this trend (data not shown). Therefore, searching the database with fragments of shorter length yields better hits than with longer fragments, however, very short fragments provide very little 'resolving power' to discriminate false positives from true positives, as shown below.

Fragment selection

In our method, we select one hit for each window which maximizes the overlap RMSD with the neighboring windows, as described in the Theory and Methods section above. We show here that this procedure does in fact greatly reduce the impact of the false positives. In the case of $N_w = 7$ (Figure 3a), we performed the selection procedure independently for the five non-overlapping blocks of windows 1–20, 21–25, 26–35, 36–50, and 51–70. The results for each block is shown in Table 1, and the hits selected are indicated by the green line in Figure 3. In general, the selections made using the overlap RMSD criterion are quite good, though usually not the best in terms of the RMSD of the fragment to 1UBQ. The only substantial deviation is in the area of windows 31–36, which could be another consequence of the low data density for residues 36–38 mentioned above. It should be noted that previous approaches have also had problems in this region of ubiquitin (Hus et al., 2001). The analogous results for $N_w = 18$ (Figure 3b) show similar difficulty in this region, but are in general substantially worse due to the overall poorer hit quality and greater false positive rate.

The amount of CPU time required to find the optimal fragment selection was quite reasonable; furthermore, these CPU times represent a very substantial speedup (by up to 16 orders of magnitude) over a naive exhaustive search. For example, for window block 36–50 ($N_b = 15$) an exhaustive search would have required the examination of $(15^{16} - 1)/14 \approx 5 \times 10^{17}$ nodes in the tree structure of Figure 1. By using the bounded tree-search algorithm described in the Appendix, only 3×10^7 nodes needed to be examined, and the algorithm reached level N_b only 12 times. Nonetheless, the CPU time still appears to scale roughly as $e^{0.5N_b}$, making block sizes greater than 20 prohibitively expensive. There are situations in which the algorithm used here is very inefficient, especially when there are many adjacent windows early in a block all of whose fragments give very small overlap RMSDs with each other. This can happen in the case of alpha helices, for example. It can also be seen in the unusually long running time for the first window block (1–20) (Table 1). Such pathologies could be identified by examining the distributions of structural similarity of the fragments in each window and the pairwise overlap RMSDs between neighboring windows. Regions which will cause an inefficient search can then be isolated within their own smaller blocks, or redundant fragments can be removed after appropriate clustering. If necessary, alternative approaches based on Monte Carlo or genetic/evolutionary algorithms could be developed to deal with larger block sizes and/or local inefficiency.

The local structure corresponding to each block is generally quite good (Table 1). For example, the selections for the block corresponding to windows 1–20 (residues 1–26) gives a remarkably tight bundle with a 1.6 Å Cα RMSD to the corresponding region of 1UBQ (Figure 4). The overall structure determined using $N_w = 7$ before refinement is topologically correct but of somewhat poorer quality than the fit in the individual blocks: the RMSD of the mean Cα positions to 1UBQ is 5.94 Å. This is not surprising, however, as absolutely no 'long range' information has been used in the construction of this initial model. In fact, our result is similar to that obtained previously with a different algorithm which included database searching with respect to chemical shifts as well as dipolar

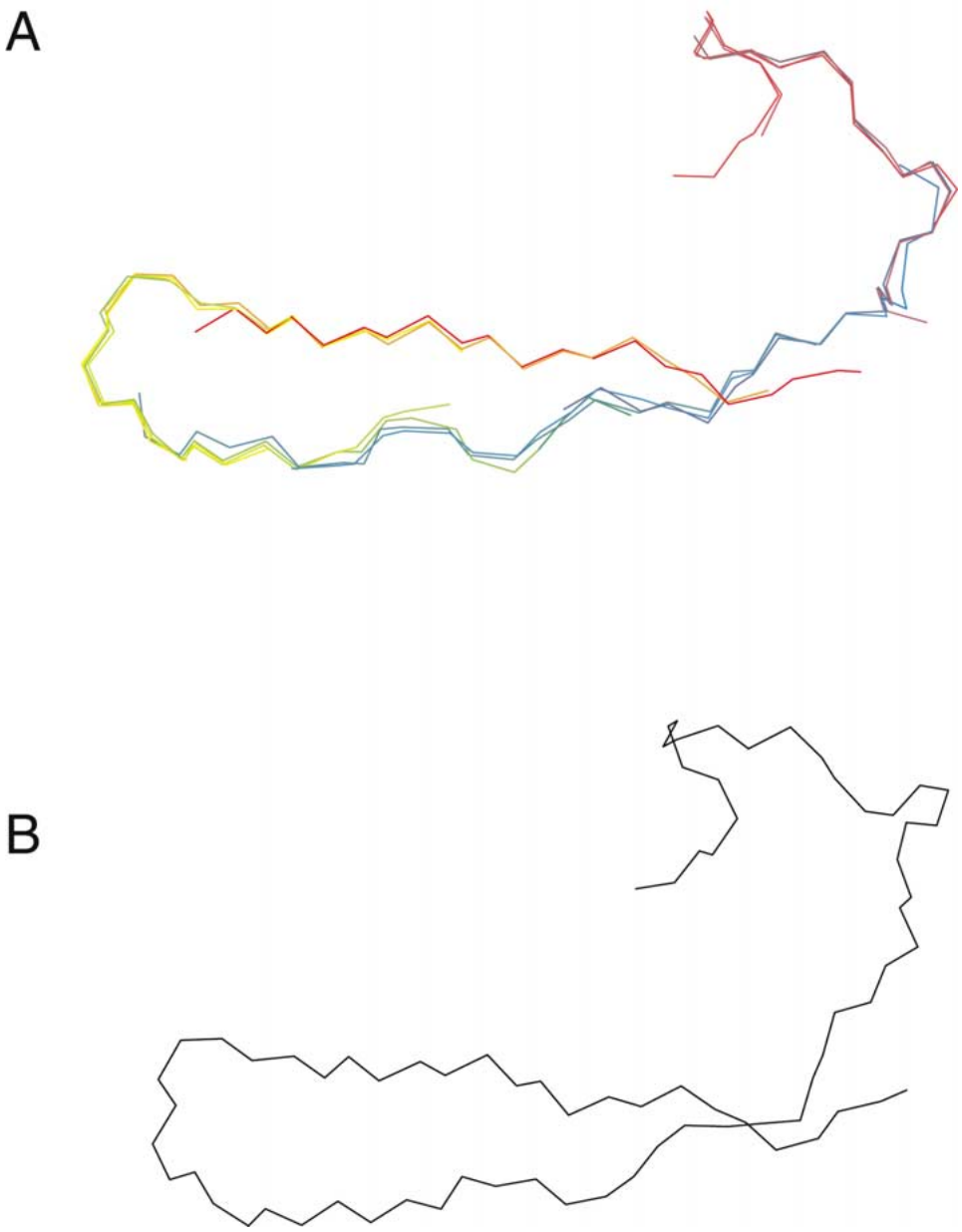

Figure 4. (a) The result of the superposition of the selected fragments for the window block 1–20 (residues 1–26) as described in Theory and methods. The 20 fragments chosen are 1SLU:A(69–75), 1BTN(75–81), 1BTN(76–82), 1BTN(77–83), 1TSS(35–41), 1AGQ:A(115–121), 1AGQ:A(116–122), 1DNP:A(67–73), 1DNP:A(68–74), 1RIE(84–90), 1RIE(85–91), 1AGQ:A(44–50), 1AGQ:A(45–51), 1AGQ:A(46–52), 1ALO(15–21), 1ALO(16–22), 1ALO(17–23), 1EXN:A(90–96), 1EXN:A(91–97), and 1HVD(116–122), respectively. Each fragment is shown in a different color. (b) The corresponding region of 1UBQ. The Ca RMSD between it and the mean Cα positions of the fragments shown in (a) is 1.56 Å (Table 1).

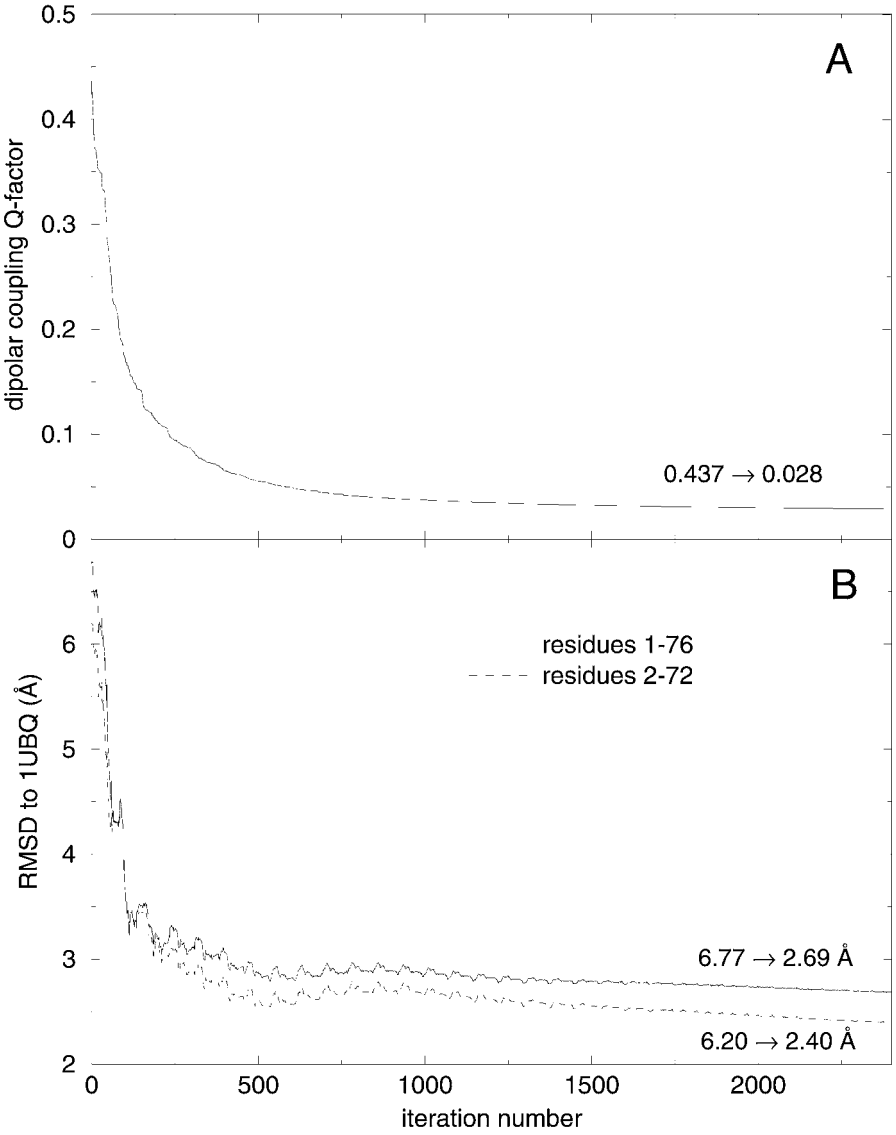

Figure 5. Results of the refinement of the model constructed from the fragment selections shown in Table 1 as described in Theory and methods. (a) shows the convergence monitoring in terms of decrease in dipolar coupling χ^2 (Q-factor). (b) Shows the RMSD of the refined structural model relative to the 1UBQ crystal structure for the complete protein as well as the protein without the unstructured N- and C-terminal regions.

couplings (6.28 Å)* (Delaglio et al., 2000, Supporting Information Table 1). Construction of a regularized structural model in internal coordinate (Ramachandran ϕ/ψ) space is not trivial, however. For example, the 5.94 Å for the RMSD of the mean Cα positions increases to 6.77 Å for a model constructed using ideal peptide geometry, $\omega = 180°$, and ϕ and ψ angles determined from the mean Cα, C, and N positions. The effect is even more dramatic when we constructed, as a test, a model using the fragment in each window with the smallest RMSD with respect to the corresponding window of 1UBQ (i.e., the lowest circles in Figure 3a). While the RMSD of the mean Cα positions to 1UBQ for the resulting structure is only 3.0 Å, the full backbone internal coordinate representation is 6.0 Å away from 1UBQ. Although we have found that dihedral angles determined using the mean Cα, C, and N positions result in a better consensus of the 'fragment bundle'

*Delagio et al. (2000) report backbone RMSD rather than Cα RMSD. We find that the Cα RMSDs that we report are very similar or identical to the backbone RMSDs for the same structures, therefore the numbers are directly comparable.

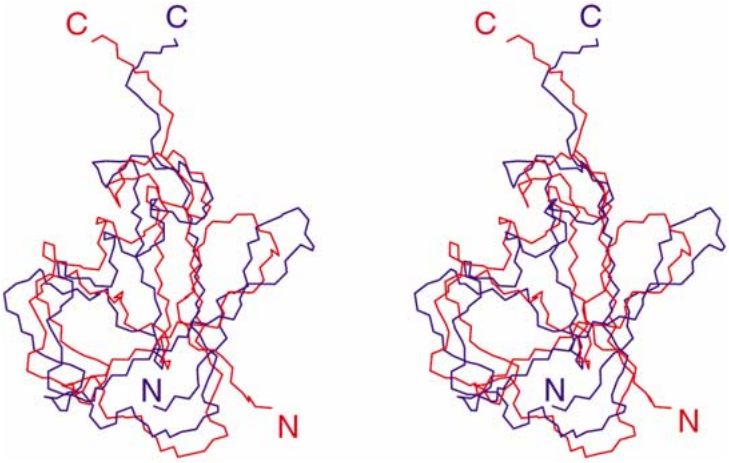

Figure 6. Stereo diagram of the backbone traces of the Cα superposition of the structural model generated using the methodology described in this paper (red) compared with the crystal structure (blue). The N- and C-termini are indicated.

than those obtained using other methods, such as the median dihedral angle, the method used here is clearly far from optimal.

The quality of the initial structural model varies quite strongly with the window size used: the RMSDs to 1UBQ for residues 1–76 for the structures determined in a similar manner with $N_w = 4$, 7, 9, 11, and 18 are 19.19, 6.77, 10.04, 14.32, and 14.73 Å, respectively. The poorer results for larger window sizes most likely arise from the greater difficulty of finding good database hits, while the poor result for $N_w = 4$ can be attributed to the lower information content of the 3 residue overlap as opposed to the 6 (or more) residue overlap afforded by larger window sizes, which degrades both the filtering of false positives as well as the construction of the overlap-based initial models (e.g. Figure 4a). Therefore, it appears that the ideal window size (at least for this example) is in the neighborhood of 7 residues, and is in agreement with that used by previous workers (Delaglio et al., 2000).

Refinement

In order to re-introduce long range dipolar coupling information into the structure, the backbone torsion angles were adjusted so as to minimize the overall χ^2 as described in Theory and methods above using the structure obtained from the $N_w = 7$ fragment search as a starting point. The convergence was monitored by observing the decrease in the overall χ^2 (Figure 5a), and was achieved after approximately 2000 iterations, or approximately 25 cycles through the entire protein. The resulting decrease in RMSD relative to 1UBQ is shown in Figure 5b: the overall RMSD decreased from 6.77 to 2.69 Å, while the RMSD for the core of the protein (neglecting the unstructured N and C termini) decreased from 6.20 to 2.40 Å. The resulting final model is shown in Figure 6 along with the best-fit superposition to the crystal structure 1UBQ. While the final model does contain some mis-orientation of secondary structural elements (e.g. the two N-terminal β-sheets), it is still very good. The previous work of Delaglio et al. (2000, Supporting Information Table 1) obtained a slightly lower backbone RMSD of 1.65 Å, however it made use of additional data in the form of chemical shifts. Not only is the backbone fold correct, but the peptide planes and Cα-Cβ bond vectors also have the correct relative orientations. This can be seen by comparing the RMSD of the final model to 1UBQ using the Ca atoms only (2.69 Å) and using the Cα and Cβ atoms (2.78 Å). Alternatively, one can perform a structural superposition of our model and the 1UBQ structure using Cα and Cβ atoms and calculate the angle between the direction cosines corresponding to the Cα-Cβ bond vectors for each residue. For more than half of the non-glycine residues (53%), the resulting angle is less than 15°, while 74% have an angle of less than 23°. Our model is accurate enough to identify the backbone fold (for purposes of structural genomics), and to use as a starting structure for further refinement using molecular modeling with or without additional NMR data. The fact that the side-chain Cα-Cβ directions are accurately defined raises the possibility of constructing all-atom models using modern side-chain conformation prediction algorithms and

molecular mechanics refinement (Xiang and Honig, 2001; M. Jacobs, personal communication).

Conclusions

We have shown that a protein fragment database search approach using overlap RMSD as a filtering tool is an efficient way of generating a backbone fold from residual dipolar coupling data. Our approach is similar in spirit to the earlier pioneering work of Delaglio et al. (2000); it differs in the use of a fragment filter based on the structural consistency between different windows rather than on consistency at the level of a single residue (Delaglio et al., 2000) or single peptide plane (Hus et al., 2001). Our filtering methodology is not specific to residual dipolar couplings, but could be used to filter any collection of fragments generated based on chemical shift, amino acid sequence, or other criteria. The necessary computer time is not large by the standards of NMR structure determination, and the resulting structural model is of sufficient quality to allow for backbone fold identification and further refinement using molecular modeling.

There is a great deal of information contained in the residual dipolar couplings, not all of which is used by the methods described here, and the overall reliability and robustness of our procedure could be improved by incorporating this information. For example, we used the dipolar couplings to assemble the list of candidate fragments at each window position, but this information was not used at all in the fragment filtering and the construction of the initial model. Clearly, one could use the dipolar couplings to assist in this, for example, by insisting on self-consistency between the rotation matrix associated with the overlap RMSD superposition and the PAS orientations derived from the fits to the dipolar couplings. We are currently investigating these possibilities, as well as developing methods for the further refinement of backbone and all-atom structural models using molecular modeling approaches.

Acknowledgement

This research was supported by the National Institutes of Health (NRSA Fellowship GM19856-02 to MA and grant GM-30580 to RML).

Appendix

In this Appendix we provide the technical details of the algorithms used to select the optimal fragment in each data window based on overlap RMSD. Let F[i,j] denote the fragment with the j-th smallest χ^2 value for data window i (i.e. residues i through i+N_w − 1). We define the overlap RMSD OL(F[i, m], F[i+1, n]) to be the Cα RMSD of the last N_w − 1 residues of F[i, m] with the first N_w − 1 residues of F[i+1, n]. Consider a block of N_b windows k through k+N_b − 1. We define a selection S to be N_b fragments such that each S[i] (i = k, k+1, ..., k+N_b − 1) is one of F[i,j] (j = 1, 2, ..., 15), and the overlap score for the selection S to be the sum of the individual overlap RMSDs for the entire block:

$$D(S) = \sum_{i=k}^{k+N_b-2} OL(F[i, S[i]], F[i+1, S[i+1]]).$$

Given a block of fragments F[i, j] (i = k, k+1, ..., k+N_b − 1) (j = 1, 2, ..., 15), we wish to find the selection S which minimizes D(S).

One can quickly obtain a reasonable (but in general suboptimal) selection S_{greedy} using the following greedy algorithm:

(1) Choose S_{greedy}[k] and S_{greedy}[k+1] such that

$$OL(F[k,S_{greedy}[k]], F[k+1, S_{greedy}[k+1]]) = \min_{\substack{i=1,\ldots,15 \\ j=1,\ldots,15}} OL\ (F[k, i], F[k+1,j])$$

(2) For each i = k+2, k+3, ..., k+N_b − 1 choose S_{greedy}[i] such that

$$OL(F[i\text{-}1, S_{greedy}[i\text{-}1], F[i, S_{greedy}[i]]) = \min_{j=1,\ldots,15} OL(F[i-1, S_{greedy}[i\text{-}1]], F[i, j])$$

While this does not lead to the best selection in general, the overlap score $D(S_{greedy})$ is usually not too much larger than the optimal overlap score (Table 1), and it can be used as an initial upper bound in a subsequent bounded tree search.

Consider the construction of a selection S in which we begin with the empty set (level 0) and add a fragment corresponding to each successive window in the block until we reach the last window (level N_b). This can be visualized as a tree structure as shown in Figure 1. For each node at level 2 or greater, it is possible to calculate a partial overlap score corresponding to all of the fragments selected up to that point, e.g., for the node F[k+2, 2] at level 3 circled in Figure 1, the partial overlap score would be the sum of the two overlaps

OL(F[k, 1], F[k+1, 1]) and OL(F[k+1, 1], F[k+2, 2]). Suppose that this partial overlap score is greater than the greedy overlap score $D(S_{greedy})$. Since the overlap score is additive and consists of nonnegative terms, it is guaranteed that all nodes which are decendents of this node will have an overlap score which is less optimal than the greedy selection, and therefore it is not necessary to expand this node in searching the tree. Furthermore, when searching the tree in a depth-first manner (Smith, 1989), if the algorithm reaches level N_b with an overlap score less than the upper bound (initially $D(S_{greedy})$), the upper bound can be set to this new value. The overall algorithm can be summarized as follows:

(1) Let bestscore = $D(S_{greedy})$.

(2) Traverse the tree in Figure 1 recursively in a depth-first manner.

  (2a) For each node at level ≥ 2 calculate the partial overlap score for that node.

  (2b) If the partial overlap score > bestscore, ignore all decendents of this node.

  (2c) If level = N_b and the overlap score $\leq$ bestscore, then output the selection corresponding to this node and let bestscore = overlap score.

(3) The final selection output is the optimal selection $S_{optimal}$ (ignoring ties).

References

Al-Hashimi, H.M., Valafar, H., Terrell, M., Zartler, E.R., Eidsness, M.K. and Prestegard, J.H. (2000) *J. Magn. Reson.*, **143**, 402–406.

Andrec, M., Du, P. and Levy, R.M. (2001) *J. Am. Chem. Soc.*, **123**, 1222–1229.

Annila, A., Aitio, H., Thulin, E. and Drakenberg, T. (1999) *J. Biomol. NMR*, **14**, 223–230.

Bowers, P.M., Strauss, C.E.M. and Baker, D. (2000) *J. Biomol. NMR*, **18**, 311–318.

Brenner, S.E., Chothia, C. and Hubbard, T.J.P. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 6073–6078.

Clore, G.M., Starich, M.R. and Gronenborn, A.M. (1998) *J. Am. Chem. Soc.*, **120**, 10571–10572.

Cornilescu, G., Marquardt, J.L., Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 6836–6837.

Delaglio, F., Kontaxis, G. and Bax, A. (2000) *J. Am. Chem. Soc.*, **122**, 2142–2143.

Fadel, A.R., Jin, D.Q., Montelione, G.T. and Levy, R.M. (1995) *J. Biomol. NMR*, **6**, 221–226.

Golub, G.H. and Pereyra, V. (1973) *SIAM J. Numer. Anal.*, **10**, 413–432.

Hansen, M.R., Mueller, L. and Pardi, A. (1998) *Nat. Struct. Biol.*, **5**, 1065–1074.

Hus, J.-C., Marion, D. and Blackledge, M. (2001) *J. Am. Chem. Soc.*, **123**, 1541–1542.

Kabsch, W. (1978) *Acta Cryst.* **A34**, 827–828.

Losonczi, J.A., Andrec, M., Fischer, M.W.F. and Prestegard, J.H. (1999) *J. Magn. Reson.*, **138**, 334–342.

Meiler, J., Peti, W. and Griesinger, C. (2000) *J. Biomol. NMR*, **17**, 283–294.

Moltke, S. and Grzesiek, S. (1999) *J. Biomol. NMR*, **15**, 77–82.

Moseley, H.N.B. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.

Nelder, J.A. and Mead, R. (1965) *Comp. J.*, **7**, 308–313.

Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 12334–12341.

Parker, J.M.R. (1999) *J. Comp. Chem.*, **20**, 947–955.

Peticolas, W.L. and Kurtz, B. (1980) *Biopolymers*, **19**, 1153–1166.

Prestegard, J.H., Tolman, J.R., Al-Hashimi, H.M. and Andrec, M. (1999) In *Structure Computation and Dynamics in Protein NMR*, Krishna, N.R. and Berliner, L.J. (Eds.), Plenum Publishers, New York, NY, pp. 311–355.

Ramirez, B.E. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 9106–9107.

Saupe, A. (1968) *Angew. Chem. Internat. Ed.*, **7**, 97–112.

Shindyalov, I.N. and Bourne, P.E. (1998) *Prot. Eng.*, **11**, 739–747.

Smith, J.D. (1989) *Design and Analysis of Algorithms*, PWS-Kent Publishing, Boston.

Tjandra, N. (1999) *Structure*, **7**, R205–R211.

Vijay-Kumar, S., Bugg, C.E. and Cook, W.J. (1987) *J. Mol. Biol.*, **194**, 531–544.

Xiang, Z. and Honig, B. (2001) *J. Mol. Biol.*, **311**, 421–430.