Protein Structural Motif Recognition via NMR Residual Dipolar Couplings

Michael Andrec, Peicheng Du, and Ronald M. Levy*

Contribution from the Department of Chemistry, Wright-Rieman Laboratories, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, New Jersey 08854-8087

Received November 16, 2000

Abstract: NMR residual dipolar couplings have great potential to provide rapid structural information for proteins in the solution state. This information even at low resolution may be used to advantage in proteomics projects that seek to annotate large numbers of gene products for entire genomes. In this paper, we describe a novel approach to the structural interpretation of dipolar couplings which is based on structural motif pattern recognition, where a predefined gapless structural template for a motif is used to search a set of residual dipolar couplings for good matches. We demonstrate the applicability of the method using synthetic and experimental data. We also provide an analysis of the statistical power of the method and the effects of order tensor frame orientation, motif size, and structural complexity on motif detection. Finally, we discuss remaining problems that must be overcome before the method can be used routinely to identify protein homologies.

Introduction

Although there has been an explosion of genomic DNA sequence data in the past few years, these data can be fully utilized only if the proteins corresponding to the putative genes can be properly functionally annotated.[1] Unless the sequence homology is relatively high, such annotation cannot be performed reliably in the absence of structural information. This has led to the recent interest in the development of "structural genomics" for the identification of function.[2] Both of the standard methods for high-resolution protein structure determination, X-ray crystallography and NMR spectroscopy, have bottlenecks which make them difficult to apply in a high-throughput manner. One of the principle rate-limiting steps in NMR structure determination is the sequential assignment of side-chain proton resonances and the assignment of NOESY cross-peaks to particular side-chain 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.[3] 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. Several candidate data types exist, including the backbone chemical shifts themselves,[4,5] scalar couplings,[5] cross-correlated relaxation rates,[6] and residual dipolar couplings.[7]

Of these, residual dipolar couplings are of particular interest in that they require relatively little data collection time and provide considerable structural information through their dependence on the orientation of internuclear vectors relative to an order frame.[7] The development of a variety of orienting media (such as lipid bicelles and filamentous phage) has greatly 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 of macromolecules is becoming increasingly routine.[8] The application of residual dipolar couplings to the structural genomics problem is only just beginning. For example, Annila et al.[9] investigated the possibility of fold recognition via dipolar couplings by qualitatively comparing the pattern of residual dipolar coupling as a function of primary sequence between closely related calcium-binding proteins. More recently, Meiler et al.[10] extended this idea by quantitatively fitting structural models of entire proteins to dipolar coupling data using a scaled χ^2 statistic as a quality factor. Delaglio et al.[11] developed a tool for searching the Protein Data Bank (PDB) for seven-residue fragments consistent with local residual dipolar coupling patterns. They showed that an appropriately chosen subset of such fragments that together span the entire data set can be used as a starting point for an optimization procedure to obtain moderate-resolution structural models for the entire protein.

We describe here a method for pattern recognition in residual dipolar couplings targeted at the detection of protein structural motifs ∼20 residues in length using a minimal amount of

* To whom correspondence should be addressed: (phone) (732) 445-3947; (fax) (732) 445-5958; (e-mail) ronlevy@lutece.rutgers.edu.

(1) Eisenstein, E.; Gilliland, G. L.; Herzberg, O.; Moult, J.; Orban, J.; Poljak, R. J.; Banerjei, L.; Richardson, D.; Howard, A. J. *Curr. Opin. Biotechnol.* **2000**, *11*, 25−30.

(2) Skolnick, J.; Fetrow, J. S.; Kolinski, A. *Nature Biotechnol.* **2000**, *18*, 283−287.

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

(4) Wishart, D. S.; Nip, A. M. *Biochem. Cell Biol.* **1998**, *76*, 153−163.

(5) Case, D. A. *Curr. Opin. Struct. Biol.* **2000**, *10*, 197−203.

(6) Reif, B.; Diener, A.; Hennig, M.; Maurer, M.; Griesinger, C. *J. Magn. Reson.* **2000**, *143*, 45−68.

(7) Prestegard, J. H.; Tolman, J. R.; Al-Hashimi, H. M.; Andrec, M. In *Structure Computation and Dynamics in Protein NMR*; Krishna, N. R. Berliner, L. J., Eds.; Plenum Publishers: New York, 1999; Vol. 17, pp 311−355.

(8) Tjandra, N. *Structure* **1999**, *7*, R205−R211. Vermeulen, A.; Zhou, H.; Pardi, A. *J. Am. Chem. Soc.* **2000**, *122*, 9638−9647.

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

(10) Meiler, J.; Peti, W.; Griesinger, C. *J. Biomol. NMR* **2000**, *17*, 283−294.

(11) Delaglio, F.; Kontaxis, G.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 2142−2143.

10.1021/ja003979x CCC: $20.00
Published on Web 01/23/2001

residual dipolar coupling data (e.g., amide $N-H_N$ couplings only). Our method, however, is not limited to amide couplings, and data from additional internuclear vectors may be readily incorporated. Critical to our method and in contrast to previous work, we show how the use of a statistical measure appropriate to this pattern recognition problem greatly facilitates the analysis. Such a measure can be formulated in terms of the tail probability (or "*p* value") under a null hypothesis and is similar to measures of statistical significance used in amino acid sequence alignment.[12] The motif size regime used here is smaller than the units previously studied by Annila et al.[9] (>100 residue) and Meiler et al.[10] (30−100 residues), but larger than the 7-residue units used by Delaglio et al.[11] Our choice of this size regime was motivated by the fact that structural motifs of biological interest often are in this size range (e.g., helix−turn−helix DNA-binding[13] and EF-hand Ca^{2+}-binding domains[14]), by our conjecture that suitably chosen gapless motifs of this size range will be more characteristic of given fold families than shorter segments, and by results which indicate that smaller sizes may not provide sufficient statistical signal for motif recognition. We present results for two test systems demonstrating the effectiveness of the methodology. Our first example demonstrates the detection of a helix−turn−helix (HTH) motif in ideal synthetic $N-H_N$ residual dipolar coupling data. In our second example, we demonstrate the detection of a 20-residue template characteristic of the "ubiquitin-related" SCOP superfamily[15] in the experimental ubiquitin $N-H_N$ residual dipolar coupling data obtained by Ottiger and Bax.[16]

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

$$D = D_a[(3\cos^2\theta - 1) + {}^3/_2 R\cos 2\phi \sin^2\theta] \qquad (1a)$$

where D_a is a constant that depends on the internuclear distance and the gyromagnetic ratios of the spins involved, R ($0 \le R \le {}^2/_3$) is a measure of the asymmetry of the order tensor, and θ and ϕ are spherical angles that relate the internuclear vector to the principal axis system (PAS) of the order tensor.[7] Alternatively, one can rewrite eq 1a in the form

$$D = (x\ y\ z)\begin{pmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{xy} & D_{yy} & D_{yz} \\ D_{xz} & D_{yz} & D_{zz} \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} \qquad (1b)$$

where D_{ij} are the elements of a symmetric and traceless matrix proportional to the Saupe order tensor[17,18] in an arbitrary molecular frame defined by the directions cosines x, y, and z of the internuclear vectors relative to that frame. Since eq 1b is linear in the tensor elements D_{ij}, it is possible to solve for the optimal D_{ij}'s that maximize the agreement between a set of bond vector orientations and the dipolar coupling data using a computationally efficient linear least-squares procedure.[18] Specifically, one can find values for five independent D_{ij}'s that minimize the quantity

$$\left\| (\mathbf{M})\begin{pmatrix} D_{yy} \\ D_{zz} \\ D_{xy} \\ D_{xz} \\ D_{yz} \end{pmatrix} - \begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_N \end{pmatrix} \right\| \qquad (2)$$

where D_i ($i = 1, ..., N$) are $N \ge 5$ measured dipolar couplings, and the $N \times 5$ matrix $\mathbf{M}$ is a function of the corresponding direction cosines in the molecular frame. The best-fit solution can be found by constructing the pseudoinverse of $\mathbf{M}$ using singular value decomposition.[18] It is therefore straightforward to fit an order tensor for each ~20-residue stretch of dipolar couplings to a motif template structure, back-calculate the best-fit dipolar couplings, and calculate a χ^2 statistic

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

where $D_{\mathrm{calc},\,i}$ and $D_{\mathrm{obs},\,i}$ are the back-calculated best-fit and experimental residual dipolar coupling data for the ith residue in the template, and the index i runs over all residues in the template for which data are available. Previous work of Meiler et al.[10] aimed at fold recognition using dipolar couplings has made use of a scaled χ^2 statistic (namely, eq 3 divided by the sum of the squares of the observed dipolar couplings) to assess goodness of fit. We show below that while this is an improvement over the simple χ^2 statistic of eq 3, an even more sensitive statistical measure can be formulated.

Since residual dipolar couplings for fixed-length internuclear vectors depend only on their orientations and contain no translational information,[19] it is useful to have measures of structural similarity based only on vector orientations. One can define an angular equivalent of the RMSD superposition of two structures by recognizing that a direction cosine can be represented as a point on the surface of a three-dimensional unit sphere. The set of bond vectors for two structures to be compared is then two ordered lists of points on the unit sphere, and one can use a standard rigid-body superposition algorithm[20,21] to find the rotation that minimizes the mean-square distance between corresponding "pseudoatoms". The degree of structural similarity can then be described by an angular similarity parameter

$$\langle \cos\theta \rangle = \frac{1}{N}\sum_{i=1}^{N} \mu_i^{(1)} \cdot \mu_i^{(2)} \qquad (4)$$

where $\mu_i^{(j)}$ is the unit vector in the direction of bond vector i in structure j after best-fit superposition. For any given structure, one can also construct a matrix $\mathbf{A}$ with elements $A_{ij} = |\mu_i \cdot \mu_j|$, which can be thought of as the angular counterpart of the commonly used distance matrix (e.g., ref 22) (Figure 1). Like the distance matrix, the angle matrix $\mathbf{A}$ is an internal coordinate representation and is independent of the choice of molecular frame. The use of absolute values is appropriate for purposes of the dipolar coupling problem since dipolar couplings are

(12) Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis: Probabilistic Models for Proteins and Nucleic Acids*; Cambridge University Press: Cambridge, 1998.
(13) Wintjens, R.; Rooman, M. *J. Mol. Biol.* **1996**, *262*, 294−313.
(14) Ikura, M. *Trends Biochem. Sci.* **1996**, *21*, 14−17.
(15) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. *J. Mol. Biol.* **1995**, *247*, 536−540.
(16) Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 12334−12341.
(17) Saupe, A. *Angew. Chem., Int. Ed. Engl.* **1968**, *7*, 97−112.
(18) Losonczi, J. A.; Andrec, M.; Fischer, M. W. F.; Prestegard, J. H. *J. Magn. Reson.* **1999**, *138*, 334−342.

(19) It should be noted that dipolar couplings are also sensitive to internal motions.[7] For purposes of this paper, we assume that these effects are smaller than those due to structural differences between the template and the protein which generated the data.

(20) Kabsch, W. *Acta Crystallogr. A* **1976**, *32*, 922−923.
(21) Kabsch, W. *Acta Crystallogr. A* **1978**, *34*, 827−828.
(22) Holm, L.; Sander, C. *J. Mol. Biol.* **1993**, *233*, 123−138.

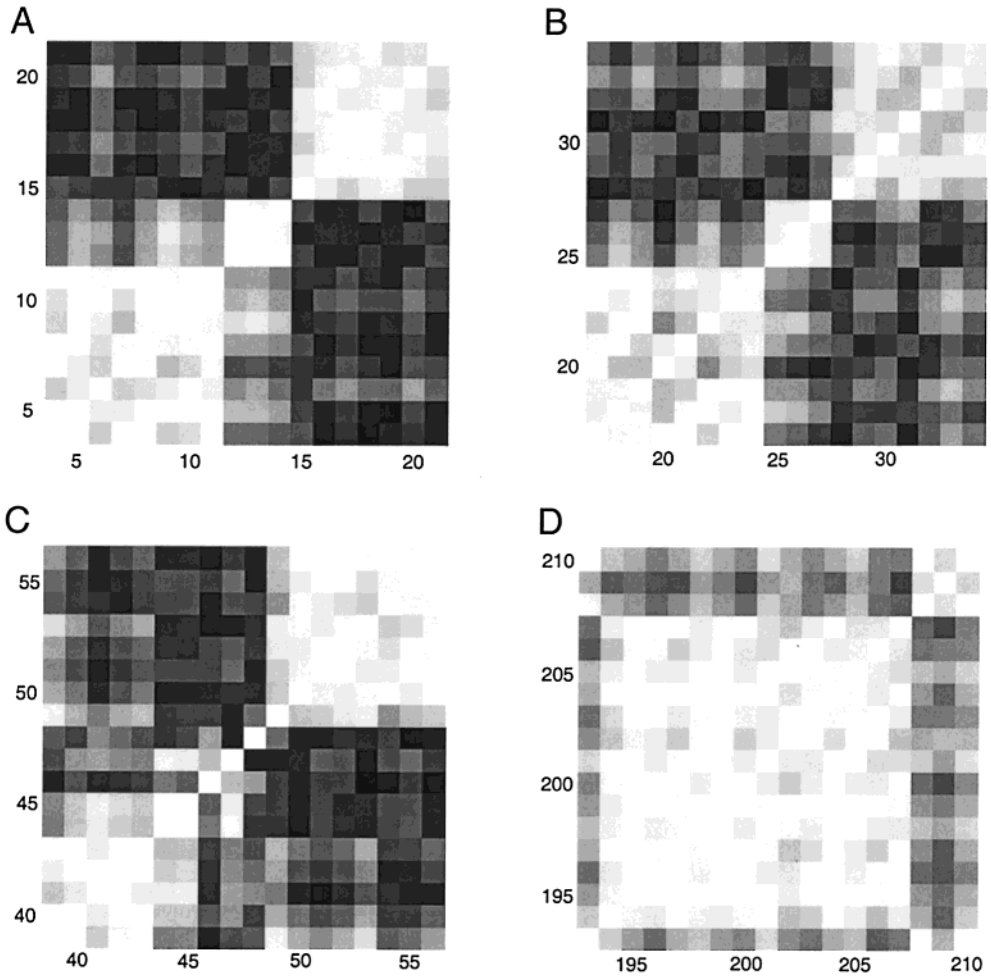

Figure 1. Angle matrices A describing the relative orientations of amide $N-H_N$ bond vectors in (A) the HTH motif 2PUE:A(4−21), (B) the HTH motif 1RPE:R(17−34) used as a motif template, (C) the window 2PUE:A(39−56), which has angular similarity to an HTH and which gives rise to the false positive in Figure 4a, and (D) the window 2PUE:A(193-210), which gives rise to data that fits the 1RPE:R(17−34) template with a small χ^2 (see Figure 2), but which is nonetheless statistically insignificant (see Figure 3). Each element of this matrix corresponds to the absolute value of the dot product between the unit vectors along $N-H_N$ bonds for each pair of residues in the fragment, with (anti)parallel $N-H_N$ bond vector pairs indicated by white squares and perpendicular pairs by black.

invariant with respect to inversion (i.e., substitution of $-x$, $-y$, and $-z$ for x, y, and z in eq 1b leaves D unchanged).

To test our methodology, we generated synthetic amide $N-H_N$ residual dipolar couplings for all nonproline residues in the purine repressor structure 2PUE:A (which contains an HTH motif at residue 4) calculated using eq 1 and tensor magnitudes $D_a = 5.0$ and $R = 0.2$. The PAS of the order tensor was arbitrarily chosen to be the Cartesian axis system of the PDB file. A set of five 18-residue template HTH structures (1B0N:A(17−34), 1LCC:A(6−23), 1LMB:3(33−50), 1NEQ(25−42), and 1RPE:R(17−34)) was also chosen. These are members of the "434 cro" and "lac repressor" HTH families as defined by Wintjens and Rooman.[13] The C_α distance RMSDs between them and the 2PUE:A HTH motif are 0.9, 0.5, 0.5, 1.0, and 0.6 Å, respectively, while the corresponding angular similarity parameters are 0.97, 0.96, 0.98, 0.90, and 0.98, respectively. These fragments were used as structural templates for searching dipolar coupling data for windows that match to a high statistical significance.

To demonstrate the effectiveness of our methodology using experimental data, we made use of the ubiquitin $N-H_N$ residual dipolar coupling data obtained by Ottiger and Bax.[16] As a template, we chose a 20-residue fragment of elongin B 1VCB:A(25−44); the template was identified by searching for the segments of that length in each of five representative members of the "ubiquitin-related" SCOP superfamily[15] (1UBQ, 1VCB:A, 1A5R, 1NDD:A, 1BT0) which minimizes the mean-square deviation between the elements of the corresponding submatrices of the angle matrices **A** in each pair of structures. The C_α distance RMSD for 1VCB:A(25−44) and the corresponding position in 1UBQ (residues 24−43) is 0.6 Å, while the angular similarity parameter is 0.96. All calculations were performed using only amide $N-H_N$ bond vector directions, which were chosen to lie in the plane and along the bisector of the $C_{i-1}-N_i-C\alpha_i$ bond angle. Template positions corresponding to proline residues were assigned fictitious $N-H_N$ bond vector directions calculated in the same manner.

Results and Discussion

Structural Motif Recognition via *p* Values. One way in which we can attempt to recognize a structural motif in residual dipolar coupling data is to find the best-fit D_{ij} parameters for the N-residue template and each N-residue window in the data and calculate the resulting χ^2. The result of such a calculation for the purine repressor data and the 1RPE:R(17−34) template is shown as the solid curve in Figure 2. Although the correct motif location does give rise to a lower χ^2 value, there are also many other regions that are not structurally related but give comparably small χ^2 values (e.g., the 18-residue window beginning at residue 193). We can significantly improve on this

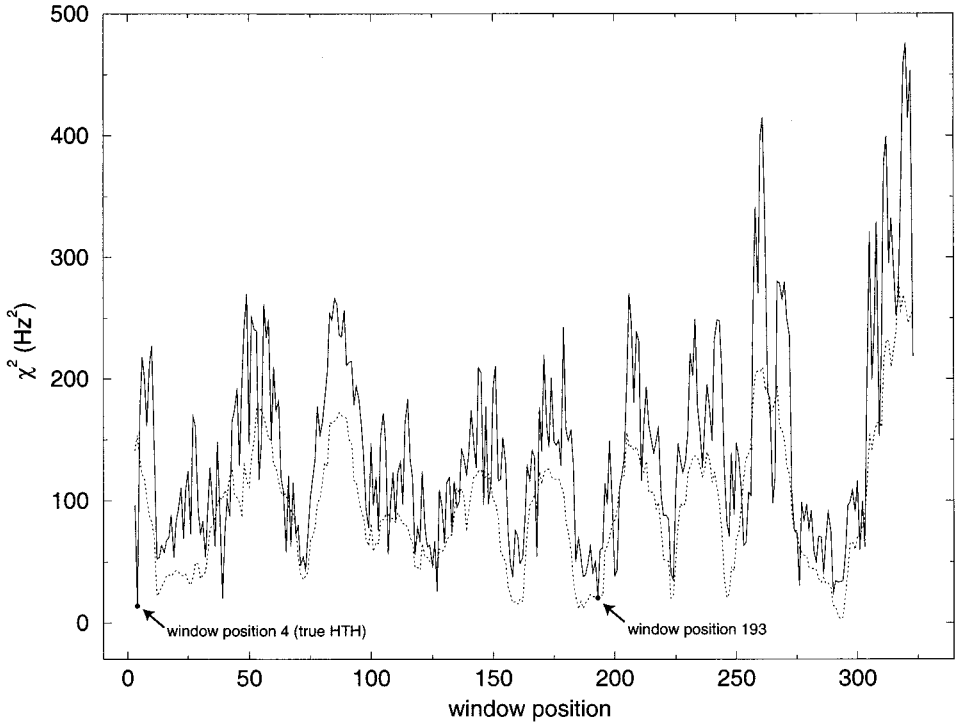

Figure 2. Plot of the χ^2 goodness of fit (solid line) of each 18-residue segment of the synthetic residual dipolar coupling data for 2PUE:A to the HTH motif template 1RPE:R(17−34) used to search the dipolar coupling data. The dotted line represents the mean minus 1.5 times the standard deviation of the distribution of χ^2 for the fits of the data and 1500 18-residue-long peptide fragments drawn with a uniform probability from the PDB coordinates of all domains in the SCOP40 database.[24]

result by recognizing that χ^2 alone does not constitute a measure of statistical significance but, in general, must be compared to its distribution under an appropriate null hypothesis (just as the significance of a Smith−Waterman sequence alignment can be determined by comparing a raw score to an extreme value distribution[12]). For some problems (e.g., when the null hypothesis is that the data were generated by a known linear model plus Gaussian noise), this is not necessary, since the distribution of the χ^2 statistic scaled by the variance of the noise depends only on the number of degrees of freedom and not on the data values themselves.[23] In such a case, one can rank order and compare the goodness of fit of different data sets of the same length by comparing their χ^2 values directly, without having to compare them to the distribution of χ^2 under the null hypothesis.

For the protein structural motif detection problem, the situation is quite different. Here, we consider the data to match a motif template if the χ^2 for a fit to the 18-residue template is much smaller than might be expected for a fit to a protein fragment chosen at random. To represent this null hypothesis, we constructed a database of direction cosines from 1500 and 20 000 18-residue fragments randomly chosen from the SCOP40 database[24] and calculated the χ^2 of the fit of each 18-residue-long data window to each structure in the 1500-fragment database, resulting in a distribution of χ^2 values for each data window. The dashed line in Figure 2 is the mean minus 1.5 times the standard deviation of that distribution as a function of window position, and the histograms in Figure 3 show the shape of these distributions for two windows (residues 4−21 and residues 193-210). Unlike in the familiar "χ^2 test" described above, the null hypothesis appropriate to the motif recognition problem results in distributions of χ^2 that depend not only on the number of dipolar couplings but also on their values. For example, the dipolar couplings of residues 193−210 have less information content, in the sense that they are much more likely to be fit well by a randomly chosen protein fragment, than the data for residues 4−21. Therefore, the statistical significance of the fits of the HTH template (as shown by the arrows) is quite different for the two segments, even though both χ^2 values are similarly small. There is a large variability in the information content of different stretches of dipolar couplings, and this

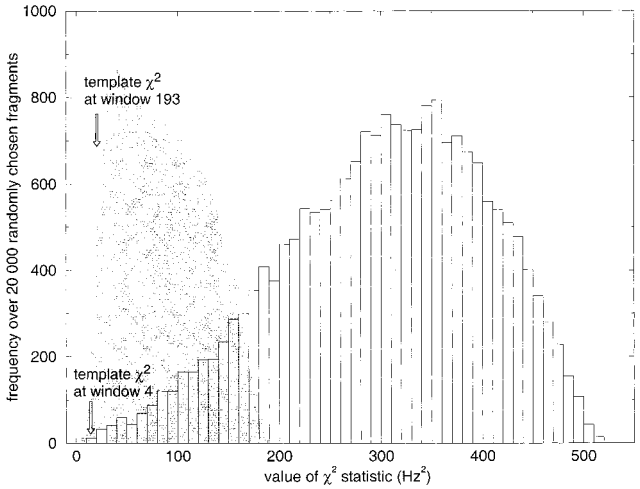

Figure 3. Comparison of the distribution of the χ^2 statistic for the fit of two 18-residue segments of the synthetic purine repressor residual dipolar couplings (residues 4−21 in open boxes and residues 193−210 in shaded boxes) to the 20 000 random 18-residue-long fragments described in Figure 2. The χ^2 values of the 1RPE:R template to the dipolar couplings are shown by the arrows (open and shaded for 4−21 and 193−210, respectively).

(23) Press: W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C: The Art of Scientific Computing*; Cambridge University Press: Cambridge, 1992.

(24) Brenner, S. E.; Chothia, C.; Hubbard, T. J. P. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 6073−6078.

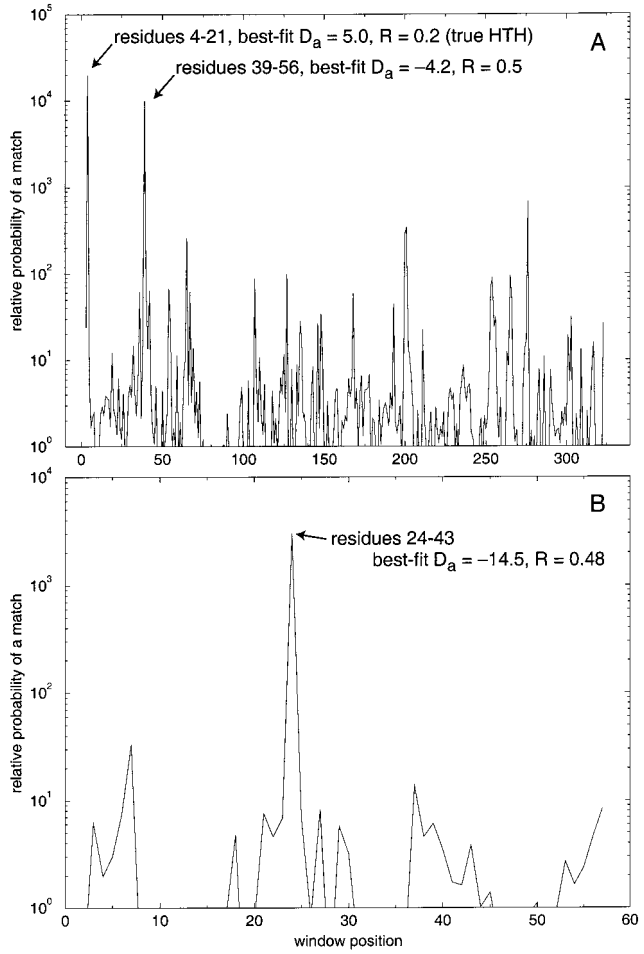

Figure 4. Plots of the relative probability of a match $= (1 - p)/p$ (in log scale) as a function of the position of the data window for (A) the synthetic purine repressor data ($D_a = 5$, $R = 0.2$) to the 1RPE:R HTH template and (B) the experimental "charged bicelle" data for ubiquitin obtained by Ottiger and Bax.[16] The analogous results for the "uncharged bicelle" data are qualitatively identical and are not shown. Estimates of the *p* values were initially performed using a 1500-fragment database generated from SCOP40 as described in the text; window positions resulting in $p < 0.01$ were re-estimated using a 20 000-fragment database constructed in the same manner. The χ^2 of the template at position 4 for the purine repressor was smaller than any of the 20 000 fragments in the database, and the plotted value represents a lower bound. The best-fit D_a and *R* values for each hit with relative probability of a match greater than 1000 are indicated.

variability must be taken into account to properly interpret the χ^2 values of the fit to the template.

The statistical significance of the fit of a given segment of dipolar couplings to a motif template can be quantitatively expressed as a *p* value, which is the probability that a χ^2 as small or smaller than that observed for the template could have been obtained by chance from a population of random protein structures or, equivalently, as the odds against the null hypothesis, which can be thought of as the relative probability that a given data window is a match to the template (Figure 4). The purine repressor data (Figure 4A) shows two strong "hits" to the 1RPE:R(17−34) template. One of these corresponds to the true HTH signal at position 4, and the best-fit D_a and *R* values agree well with those used to generate the data. We see no other hits with significance within 1 order of magnitude of the true HTH, except for position 39, which is still at least a factor of 2 smaller than the match to the true HTH at window position 4. In the ubiquitin example (Figure 4B), we obtain only one very strong hit, which is in the correct location and has best-fit tensor magnitudes comparable to those estimated by Ottiger and Bax.[16] No other hits within 2 orders of magnitude of the true positive are observed.

Table 1. Rigid-Body Superposition of $N-H_N$ Direction Cosines for 1RPE:R(17−34) and Three 18-Residue Windows of 2PUE:A

position in window	cosines of the angles between corresponding $N-H_N$ directions[a]		
	2PUE:A(4−21)	2PUE:A(39−56)	2PUE:A(193-210)
1	0.949	0.917	0.670
2	0.989	0.906	0.931
3	0.998	0.871	0.701
4	0.993	0.832	0.587
5	0.996	0.911	0.983
6	0.972	0.597	0.636
7	0.971	0.691	0.604
8	0.993	−0.263	0.915
9	0.984	0.411	0.764
10	0.984	0.968	0.745
11	0.962	0.103	−0.894
12	0.997	0.802	0.427
13	0.987	0.990	0.379
14	0.937	0.847	0.445
15	0.953	0.733	0.360
16	0.966	0.785	0.620
17	0.985	0.876	0.710
18	0.988	0.989	−0.269
$\langle\cos\theta\rangle$	0.978	0.720	0.517

[a] Positions corresponding to proline residues were assigned fictitious $N-H_N$ bond vector directions based on the $C_{i-1}-N_i-C\alpha_i$ coordinates as described in Theory and Methods.

The hit at position 39 in the purine repressor example (Figure 4A) demonstrates some of the structural ambiguities inherent in the use of only $N-H_N$ residual dipolar couplings. The structure of 2PUE:A(39−56) consists of two helical fragments joined by an extended coil and is quite different from an HTH in terms of backbone conformation (C_α RMSD ≈ 7 Å). However, if we consider only the relative orientations of the amide $N-H_N$ bonds, this structure is similar to that of 1RPE:R(17−34), as can be seen by comparing the angle matrices in Figure 1B and C or the angular similarity parameters of residues 4−21 and 39−56 relative to 1RPE:R(17−34) (Table 1). For the fit of 1RPE:R(17−34) to 2PUE:A(4−21) $\langle\cos\theta\rangle = 0.978$, while for the corresponding fit to 2PUE:A(39−56) $\langle\cos\theta\rangle = 0.720$. The decreased angular similarity parameter reflects the very different amide bond vector orientations for a limited number of residues, e.g., template positions 8, 9, and 11 (Table 1). The extent to which these structural differences affect the fits to the dipolar couplings depends in part on the orientation of the PAS relative to the structures. Therefore, it is possible that data generated using different PAS orientations relative to 2PUE:A might not give equally strong hits at this location. We show below that this is in fact the case. In contrast, the angular similarity parameter for residues 193−210 of 2PUE:A (for which the fit is not statistically significant—see Figures 2 and 3) relative to 1RPE:R(17−34) is much smaller (Table 1), and we see no strong signal at this position for any PAS orientation.

The size of the template can have a major impact on the ability to recognize a structural motif. To investigate this, we compared the results obtained using the same data (synthetic data for 2PUE:A, $D_a = 5$, $R = 0.2$) with a full 18-residue HTH template 1RPE:R(17−34) versus a truncated version of the template consisting of the central nine residues 1RPE:R(22−30). The resulting relative probabilities of a match differ significantly: there is no longer a strong signal at the true HTH location for the truncated template, and the relative probabilities

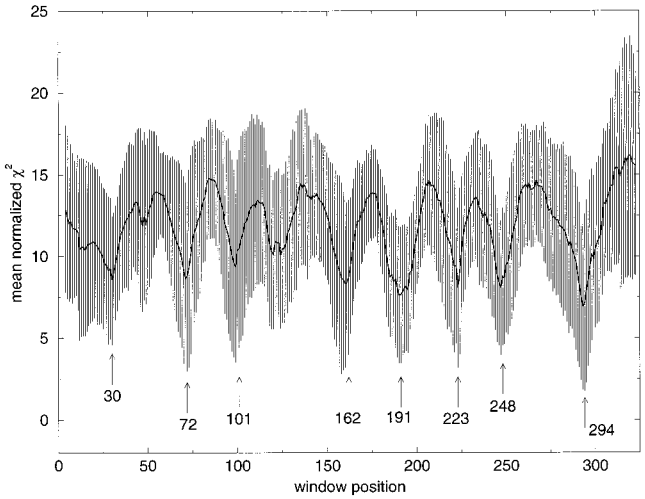

Figure 5. Plot of the mean (solid line) and standard deviation (error bars) of the distribution of mean normalized χ^2 values for each 18-residue window of data generated from 2PUE:A using 100 different PAS orientations fit by more than 1000 fragments chosen at random from the SCOP40 database. The locations indicated by the arrows represent the window positions corresponding to the start of 10-residue or longer helices in the 2PUE crystal structure.

of a match over all window positions are generally lower. This is because it is easier to fit a shorter data window with a randomly chosen structure, and one must have a correspondingly smaller template χ^2 to obtain a highly significant match. For moderately longer templates, the distributions of χ^2 for random structures shift away from zero more than the template χ^2 for the true HTH window, increasing its statistical significance. Further increases in the template length will make it harder to fit the template with a low χ^2, causing the statistical significance to decrease again. Thus, we expect that for the motif recognition procedure we propose there is an optimal template length for a given protein structural motif. This will be the subject of a future communication.

Structural Information Content and PAS Orientation. Figures 2 and 3 demonstrate that different windows in a given set of residual dipolar couplings contain varying degrees of structural information and that this information content is reflected in the distribution of the χ^2 statistic from fits of data to many randomly chosen protein fragments. We expect, however, that this structural information content will vary not only due to differences in the structure but also with changing orientations with respect to the PAS. One way to visualize these relative contributions is to calculate the mean of the χ^2 statistic for a given data window (normalized by the number of data in the window) for fits to a large number (> 1000) of protein fragments chosen at random for many data sets synthesized using different PAS orientations (100 in our case). The means and standard deviations of the resulting mean normalized χ^2 values then give a measure of the variability in structural information content intrinsic to the structure giving rise to the data and the variability for any given structure due to PAS orientation effects, respectively. In Figure 5, we show the results of such a calculation for data generated from 2PUE:A using 18-residue windows for 100 different PAS orientations.

Although there is a good deal of variability in the mean normalized χ^2 as a function of PAS orientation (as indicated by the error bars), it is clear that some regions tend to have larger mean normalized χ^2 values than other regions. For example, the mean normalized χ^2 for the window beginning at residue 193 shows a variability of ~ 4 Hz^2 centered around a mean of 8 Hz^2, while the window beginning at residue 175 has a comparable variability (3 Hz^2), but centered around a mean of 14 Hz^2. Therefore, some regions of 2PUE:A are inherently more difficult to fit and therefore likely to generate more structurally informative data (from the perspective of NMR dipolar couplings) than others even after accounting for the variability arising from different PAS orientations. These differences are related to the protein structure in those regions, as is demonstrated by comparing the structure of 2PUE:A in the residue windows 4−21 and 39−56 (for which the mean normalized χ^2 tends to be large, Figure 1A and C) with the residue window 193−210 (for which the mean normalized χ^2 tends to be small, Figure 1D). The difference is quite striking: windows 4−21 and 39−56 have a much greater diversity of amide vector orientations than window 193−210, where, due to the presence of an α-helix from residues 191 to 206, all of the amide vectors are nearly parallel to each other. Residue window 193−210 is not unique in this respect: there is in fact a very strong correlation between the starting residues of longer α-helices and the minima in the mean normalized χ^2 curve, as can be seen in Figure 5.

This result has implications for the optimal choice of template in our motif-based approach. The less structural complexity a given structural fragment has (i.e., the closer its distribution of χ^2 values for random fragments is to zero), the harder it will be to find a statistically significant match, since such a match would require a correspondingly smaller template χ^2. Therefore, in choosing templates for structural motif recognition using dipolar couplings, one would like to find those that are maximally conserved among members of a given class (e.g., SCOP superfamily) and that at the same time have the maximal structural complexity (as measured by the mean of the mean normalized χ^2 over many PAS orientations). The mean normalized χ^2 calculation is straightforward to do for any set of proteins of interest, and the results can be taken into consideration when designing templates for structural motif recognition.

False Positive/Negative Rates and PAS Orientation. We can estimate the effectiveness of the protein structural motif recognition using NMR dipolar coupling data by calculating the rate of false positives and false negatives at various significance thresholds over many different data sets. Ideally, one would like to perform such an analysis for many different templates, target proteins, and order tensor parameters. We report here the results of a more limited study of false positive/false negative rates for five structurally similar templates and data calculated using one target protein for many different PAS orientations. Specifically, we generated synthetic data for 2PUE:A as above for 100 different orientations of the PAS relative to the protein coordinates. For a given template and any given threshold for the relative probability of a match T, we estimate the false negative rate to be the fraction of rotations for which the true HTH (window position 4) gives a relative probability of a match less than T and the false positive rate to be the fraction of window positions (other than at position 4) over all rotations that give a relative probability of a match greater than T. We can then plot the false negative rate versus the false positive rate parametrically as a function of T from $T = \infty$ (false positive rate $= 0$, false negative rate $= 1$) to $T = 0$ (false positive rate $= 1$, false negative rate $= 0$). Such a curve calculated using residual dipolar coupling data generated from 2PUE:A and fit using the 1RPE:R(17−34) template is shown as the solid curve in Figure 6. The result is quite encouraging: using a threshold of 1000, we can achieve a false positive rate

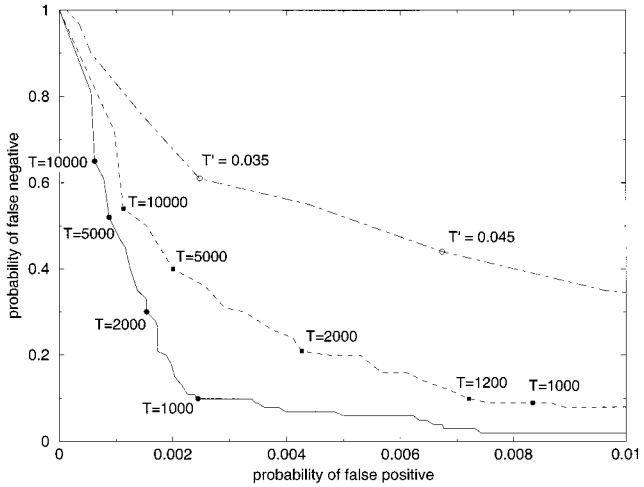

Figure 6. Plots of false positive vs false negative rates as a function of threshold T for $(1 - p)/p$ or T' for the normalized χ^2 estimated from 100 synthetic data sets generated from 2PUE:A using different PAS orientations and $D_a = 1$ and $R = 0$ as described in the text. The solid curve corresponds to the 1RPE:R template, the dashed line corresponds to a "consensus" based on the five HTH templates 1B0N:A, 1LCC:A, 1LMB:3, 1NEQ, and 1RPE:R, and the dot−dashed line corresponds to the "consensus" based on the χ^2 statistic alone.

Table 2. Distribution of False Positives (Non-HTH Positions with Relative Probabilities of a Match Greater than 1000) for 2PUE:A Synthetic Data and 1RPE:R(17−34) Template

window position(s)	no. of false positives out of 100 PAS orientations	fraction (%) of total false positives observed (out of 66 total)
39	16	24
208	12	18
19	7	11
276	6	9
67	4	6
54	3	5
36, 107, 268	2 each (6)	9
38, 65, 83, 153, 169, 170, 211, 223, 254, 265, 280, 303	1 each (12)	18
total	66	100

of less than 0.3% while still identifying more than 90% of the true positives.

The distribution of the false positives over window positions at a significance threshold of 1000 is shown in Table 2. These results have implications with respect to how much information might be gained by collecting additional data for the same target protein in different ordering media (i.e., with different values of R and PAS orientation). It has been recognized in the literature that collection of two or more data sets with different PAS orientations can reduce the structural ambiguities inherent in residual dipolar couplings.[25,26] Table 2 suggests that the additional information that could be obtained from a small number of data sets with different PAS orientations is quite significant for the motif recognition approach as well. One would expect the most "added value" from multiple-PAS data sets to occur when the false positives are evenly distributed throughout the protein, and the least for cases where the false positives are completely dominated by one window position. For 2PUE:A vs 1RPE:R(17−34), the false positives are not dominated by one window and are on the whole relatively well-dispersed throughout the protein. If this is characteristic for other templates and targets, then we can expect a relatively high "added value" from multiple-PAS data sets.

In practice, we might wish to construct a "basis set" of templates that encompass the known structural variability of a given motif of interest, since we would not know a priori which is the optimal template. To investigate this, we repeated the search for the HTH motif using four other templates (1B0N:A(17−34), 1LCC:A(6−23), 1LMB:3(33−50), 1NEQ-(25−42), all members of the λ-repressor SCOP superfamily) in addition to 1RPE:R(17−34), and constructed a "consensus" profile of the relative probabilities of a match as follows: for each window we use the result of the template that gives the smallest p value (or the lowest χ^2, since its distribution under the null hypothesis is constant for any given window). This is a reasonable procedure if we assume that each template is a priori an equally good model for an HTH. One can then use this consensus profile to calculate a false positive/false negative curve as a function of T as above. The result is shown as the dashed line in Figure 6. Although the false positive rate has gone up slightly as expected (based on the increase in "degrees of freedom"), it is still possible to obtain a false positive rate of less than 1% and still identify the true HTH more than 90% of the time, which strongly suggests that p value-based motif recognition can be practical for structural proteomics.

We can also construct a false positive/false negative curve based on the χ^2 statistic alone by defining a "hit" to be any window with a normalized χ^2 statistic smaller than a threshold T'. This allows us to directly compare the statistical power of our p value approach with one based on the χ^2 statistic. The resulting false positive/false negative curves for the consensus fit is shown as the dot−dashed line in Figure 6 and shows that a χ^2-based motif recognition strategy is significantly less powerful than one based on the p value. It should be noted that previous workers[10] have made use of the "Q factor" statistic, which is equal to χ^2 divided by the sum of the squares of the data (originally proposed by Cornilescu et al.[27] in analogy to the X-ray crystallographic R factor). We have found that this scale factor is approximately proportional to the mean of the χ^2 statistic for a given set of data over many randomly chosen protein structures and that this can be regarded as a first-order correction to the simple χ^2 statistic.

Conclusions

We have shown that amide $N-H_N$ residual dipolar couplings can be used to reliably detect and locate protein structural motifs consisting of gapless templates ~20 residues in length and that such a strategy represents a viable approach to the structural interpretation of residual dipolar couplings that differs from those currently in use. The structural information content and discriminatory ability of the data vary quite strongly; we account for this by formulating a suitable measure of statistical significance for the fit of the template motif to the target dipolar couplings. The results are quite encouraging: we can obtain 90% identification of the true positive with a less than 1% chance per window position of a false positive. Nonetheless, this error rate may not be sufficiently robust for application of this method on a genomic scale. For example, given a novel protein with 100 data window positions, even at a 0.3% false positive rate there is still a 26% chance that there will be at least one false positive and a 50% chance at a 0.7% false positive rate. It should be noted that these probabilities increase strongly

(25) Ramirez, B. E.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 9106−9107.
(26) Al-Hashimi, H. M.; Valafar, H.; Terrell, M.; Zartler, E. R.; Eidsness, M. K.; Prestegard, J. H. *J. Magn. Reson.* **2000**, *143*, 402−406.
(27) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836−6837.

with false positive rate: a 3% false positive rate (such as we obtain using the "consensus" method using the χ^2 statistic alone) would result in a 95% chance of at least one false positive, a situation that would be extremely difficult to tolerate in practice.

The false positive rates that we observe can be decreased further by the use of additional information. In our examples, we have only made use of the information inherent in the amide $N-H_N$ dipolar couplings. However, for each window position we obtain not only a χ^2 and p value but also values of the best-fit order tensor magnitudes D_a and R. One can test for the consistency of these D_a and R values with rough estimates derived from the distribution of dipolar couplings over the entire protein.[28] Furthermore, if the target protein is already ^{13}C-labeled, collection of additional residual dipolar couplings associated with other internuclear vectors does not represent a large investment of resources and may also substantially reduce the false positive rate. Backbone chemical shifts and the amino acid sequence itself provide considerable information about the secondary structure, which could be used to reduce error rates still further. We are currently investigating these possibilities, as well as performing more comprehensive analyses of the statistical power and radius of convergence of the method and exploring a variety of ways in which this methodology could be applied to practical problems in structural biology and proteomics.

Acknowledgment. This research was supported by the National Institutes of Health (NRSA Fellowship GM19856-02 to M.A. and Grant GM-30580 to R.M.L.).

JA003979X

(28) Clore, G. M.; Gronenborn, A. M.; Bax, A. *J. Magn. Reson.* **1998**, *133*, 216−221.