Bruce Randall Donald
Professor
Department of Computer Science
Department of Chemistry
Department of Biological Sciences

DARTMOUTH COLLEGE
6211 Sudikoff Laboratory
Hanover, NH 03755-3510
Office: (603) 646-3173
Fax: (603) 646-1672
E-mail: brd@cs.dartmouth.edu
www.cs.dartmouth.edu/~brd/

3D Structural Homology Detection via *Unassigned* Residual Dipolar Couplings

Christopher James Langmead[*]   **Bruce Randall Donald** [*,†,‡,§,¶]

Abstract

Recognition of a protein's fold provides valuable information about its function. While many sequence-based homology prediction methods exist, an important challenge remains: two highly dissimilar sequences can have similar folds — how can we detect this rapidly, in the context of structural genomics? High-throughput NMR experiments, coupled with novel algorithms for data analysis, can address this challenge. We report an automated procedure for detecting 3D structural homologies from sparse, unassigned *protein NMR data.*

Our method identifies the 3D structural models in a protein structural database whose geometries best fit the unassigned experimental NMR data. It does not use sequence information and is thus not limited by sequence homology. The method can also be used to confirm or refute structural predictions made by other techniques such as protein threading or sequence homology. The algorithm runs in $O(pnk^3)$ time, where p is the number of proteins in the database, n is the number of residues in the target protein, and k is the resolution of a rotation search. The method requires only uniform ^{15}N-labelling of the protein and processes unassigned H^N-^{15}N residual dipolar couplings, which can be acquired in a couple of hours. Our experiments on NMR data from 5 different proteins demonstrate that the method identifies closely related protein folds, despite low-sequence homology between the target protein and the computed model.

[*]Dartmouth Computer Science Department, Hanover, NH 03755, USA.

[†]Dartmouth Chemistry Department, Hanover, NH 03755, USA.

[‡]Dartmouth Biological Sciences Department, Hanover, NH 03755, USA.

[§]Dartmouth Center for Structural Biology and Computational Chemistry, Hanover, NH 03755, USA.

[¶]Corresponding author: 6211 Sudikoff Laboratory, Dartmouth Computer Science Department, Hanover, NH 03755, USA. Phone: 603-646-3173. Fax: 603-646-1672. Email: brd@cs.dartmouth.edu

Abbreviations used: NMR, nuclear magnetic resonance; RDC, residual dipolar coupling; DOF, degrees of freedom; 3D, three-dimensional; HSQC, heteronuclear single-quantum coherence; H^N, amide proton; SAR, structure activity relation; $SO(3)$, special orthogonal (rotation) group in 3D.

1 Introduction

Current efforts in structural genomics are expected to determine experimentally many more protein structures, thereby populating the "space of protein structures" more densely [32]. However, the rate at which new fold families are discovered is decreasing. Thus, the structures of many proteins that have not yet been determined experimentally will likely fall into one of the existing families. Sequence homology can be used to predict a protein's fold, yielding important clues as to its function. However, it is possible for two dissimilar amino acid sequences to fold to the "same" tertiary structure. For example, the RMSD between the human ubiquitin structure (PDB Id 1D3Z) and the structure of the Ubx Domain from human Faf1 (PDB Id 1H8C) is quite small (1.9 Å), yet they have only 16% sequence identity. Detecting structural homology given low sequence identity poses a difficult challenge for sequence-based homology predictors. We ask: is there a set of very fast, cheap experiments that can be analyzed to rapidly compute 3D structural homology?

This paper presents a new method for homology detection, called GD, that takes advantage of high-throughput solution-state NMR. In particular, GD uses a class of NMR experiments that record backbone H^N-^{15}N Residual Dipolar Couplings (RDCs). H^N-^{15}N RDCs measure the global orientation of the backbone amide bond vector for each amino acid in the primary sequence (except prolines). RDCs can be recorded in a short amount of time, typically in under an hour. The GD method correlates the experimentally-measured backbone H^N-^{15}N bond orienta-

tions with the backbone H^N-^{15}N bonds in a putative homologous structure. In this way, GD can detect structural homologies from remote amino acid sequences.

Previous algorithms for identifying homologous structures using RDCs [4, 1] require resonance assignments beforehand. That is, they assume one has established the correspondence between each RDC D_i and the correct residue j in the primary sequence. Unfortunately, establishing this mapping is one of the key bottlenecks in NMR structural biology, requiring relatively expensive isotopic labelling, a variety of time-consuming triple-resonance experiments, and a combination of manual and only partially-automated computational analyses [50], typically entailing a non-trivial number of human-operator decisions and judgments. Our method, in contrast, is completely automated, and does not require resonance assignments. That is, it works on *unassigned* NMR data, thereby dramatically reducing the amount of experimental and computational time and effort required to identify homologies. The NMR spectra we use can be acquired in 1-2 hours, and we also require only ^{15}N-isotopic labelling, which is an order of magnitude cheaper than the $^{15}N/^{13}C$ double labelling usually required for assignments.

GD also has other applications. It may be used in conjunction with techniques such as protein threading [29, 49], and computational homology modelling [9, 17, 20, 25, 38], providing experimental validation of the computational predictions. Furthermore, GD can also be used to bootstrap the resonance assignment process by selecting models for structure-based resonance assignment methods [2, 24, 28]. These assignments, in turn, enable detailed studies of protein-protein interactions [18] (via chemical shift mapping [10]), protein-ligand binding (via SAR by NMR [43] or line-broadening analysis [16]), and dynamics (via, e.g., nuclear spin relaxation analysis [34]).

GD is demonstrated on NMR data from 5 proteins against a database of over 2,400 representative folds determined either by x-ray crystallography or by NMR. The method correctly identifies both the native folds and homologous structures.

1.1 Organization of paper

We begin, in Section 2, with a review of the specific NMR experiments used in our method, highlighting their information content. Section 3 describes existing applications of residual dipolar couplings, including homology detection. In section 4, we detail our algorithm and analyze its computational complexity. Section 5 presents the results of the application of GD on real biological NMR data. Finally, section 6 discusses these results.

2 Background

H^N-^{15}N RDCs can be obtained experimentally by recording a H^N-^{15}N Heteronuclear Single-Quantum Coherence (HSQC) spectrum of the target protein in the dilute liquid crystalline phase with the ^{15}N decoupling turned off. For each RDC D, we have

$$D = D_{\max}\mathbf{v}^T\mathbf{S}\mathbf{v}, \qquad (1)$$

where $D_{\max}$ is a constant, $\mathbf{v}$ is the internuclear bond vector orientation relative to an arbitrary coordinate frame, and $\mathbf{S}$ is the 3×3 *Saupe order matrix* [39]. $\mathbf{S}$ is a symmetric, traceless, rank 2 tensor with 5 degrees of freedom. $\mathbf{S}$ describes the average substructure alignment of the molecule. The measurement of five or more assigned RDCs in substructures of known geometry allows one to solve for $\mathbf{S}$ using singular value decomposition (SVD) [30]. Once $\mathbf{S}$ is determined, RDCs for other residues may be simulated (back-calculated) given any other internuclear bond vector $\mathbf{v}_j$. In particular, suppose an $(H^N, ^{15}N)$ peak i in an H^N-^{15}N HSQC spectrum is assigned to residue j of a protein, whose crystal structure is known. Let D_i be the measured RDC value corresponding to this peak. Then the RDC D_i is assigned to amide bond vector $\mathbf{v}_j$, and we should expect that $D_i \approx D_{\max}\mathbf{v}_j^T\mathbf{S}\mathbf{v}_j$ (modulo noise, dynamics, crystal contacts in the structural model, etc).

In the proposed method, the RDCs are unassigned and the geometry of the protein is unknown. Thus, $\mathbf{S}$ cannot be determined explicitly using SVD. We will show, however, that for any given 3D structural model m, a unique Saupe matrix, $\mathbf{S_m}$, can be estimated. $\mathbf{S_m}$ can, in turn, be used to generate a set of back-calculated RDCs using Eq. (1). Without resonance assignments it is not possible to compare an individual bond's predicted RDC to its corresponding experimentally measured RDC. However, the *distribution* of experimentally determined RDC values may be compared to the distribution of back-computed RDCs from a given model. Our algorithm is based upon a key fact:

Observation: *A model that is homologous to the target-protein will generate a distribution of simulated RDCs that is similar to the distribution of experimentally determined RDCs.*

A primary technical challenge is:

Problem: *How can we estimate the alignment tensor without knowing assignments?*

We solve this problem by employing the information in the (putative) *a priori* structural model. Thereby, one can identify homologous structures by comparing distributions of RDCs.

3 Previous Work

Previous applications of assigned RDCs include, structure refinement [11] and structure determination [23, 3, 46, 19, 37, 31, 15]. *Assigned* RDCs have also been used for homology detection [1, 4]. Unassigned RDCs have been used to expedite resonance assignments [51, 44, 2, 24]. These methods (except for [44]) require ^{13}C-labelling and RDCs from several different bonds (for example, ^{13}C$'$-^{15}N, ^{13}C$'$-H^{N}, ^{13}C$^{\alpha}$-H$^{\alpha}$, etc.). Donald and co-workers [28] have recently introduced a resonance assignment method, called Nuclear Vector Replacement, that requires only amide bond vector RDCs, no triple-resonance experiments, and no ^{13}C-labelling. In this paper, we extend some of the key techniques developed in [28] for a new application —homology detection. From a computational standpoint, GD adopts a minimalist approach [5], demonstrating the large amount of information available in a few key spectra. By eliminating the need for triple resonance experiments, our method saves many days of spectrometer time. Consequently, homology comparison can be made without resorting to full NMR-based structure calculation. Xu and co-workers [49] have also addressed the issue of homology detection using sparse NMR data. Their method extends protein threading by incorporating a sparse set of Nuclear Overhauser Effect (NOE) data. NOEs report the distances between pairs of protons while RDCs report the orientation of internuclear bond vectors. NOEs are local measurements while RDCs are global. [49] requires assigned NOEs while our method works on unassigned RDCs.

The main focus of this paper is to elucidate the information content of unassigned RDCs, and to exploit them to formulate fold detection as a geometric *matching problem* of unassigned experimental NMR data to a structural database. In this way, structural homology can be predicted based solely on experimental measurements. When this can be done, an empirical upper bound is obtained, in that unassigned RDCs alone are shown to be sufficient to define structural homology. This, in turn, implies that GD is an improvement upon previous algorithms that detect structural homology using *assigned* RDCs [1, 4], since GD can perform fold identification without assignments. A natural step for future work would be to combine GD with existing structural bioinformatics protocols (e.g., [29]) to see how experimental data could increase their accuracy. See [49] for allied work in this direction.

4 Algorithm

The experimental inputs to GD are backbone H^{N}-^{15}N Residual Dipolar Couplings (RDCs) [45] recorded in two different aligning media. Proteins align differently in different media, yielding two different alignment tensors. The use of multiple tensors for interpreting RDCs is a standard technique. The total data acquisition time is approximately 2 hours. We record two sets of RDCs (one in each of two aligning media) for each backbone amide bond vector in the protein (modulo missing data). The secondary structure for each target protein was predicted from its primary sequence using the program JPRED [13]. The native fold was not used to estimate secondary structure. The percentage of predicted α and β secondary structure (from JPRED) and the length of the target protein are also used as input to GD.

We have assembled a database of 2,456 structural models from the Protein Data Bank (PDB [8]) representing a variety of different fold-families. Protons were added to the x-ray models using the Protonate module from the program AMBER [35]. Next, the backbone amide bond vectors were extracted from each model. Finally, the length of the primary sequence and percentage of α and β secondary structure were extracted for each protein in the database.

An alignment tensor is a symmetric and traceless 3×3 matrix with five degrees of freedom. The five degrees of freedom correspond to three Euler angles (α, β and γ), describing the average partial alignment of the protein, and the axial (D_a) and rhombic (D_r) components of the tensor. When resonance assignments and the structure of the macromolecule are known, all five parameters can be computed by solving a system of linear equations [30]. If the resonance assignments are not known, as in our case, these parameters must be estimated. It has been shown [30] that D_a and D_r can be decoupled from the Euler angles by diagonalizing the alignment tensor:

$$\mathbf{S} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T \qquad (2)$$

Here, $\mathbf{V} \in SO(3)$ is a 3×3 rotation matrix* that defines a coordinate system called the *principal order frame.* $\mathbf{\Sigma}$ is a 3×3 diagonal and traceless matrix containing the eigenvalues of $\mathbf{S}$. The diagonal elements of $\mathbf{\Sigma}$ encode D_a and D_r: $D_a = \frac{S_{zz}}{2}, D_r = \frac{S_{xx} - S_{yy}}{3}$ where $|S_{yy}| < |S_{xx}| < |S_{zz}|$. S_{yy}, S_{xx} and S_{zz} are the diagonal elements of Σ and therefore the eigenvalues of $\mathbf{S}$. It has been shown that D_a and D_r can be estimated, using only unassigned experimentally recorded RDCs, by the powder pattern method [46]. The axial and rhombic components of the tensor can be computed in time $O(nk^2)$, where n is the number of observed RDCs and k is the resolution of the search-grid over D_a and D_r.

Once the axial and rhombic components have been estimated, matrix $\mathbf{\Sigma}$ in Eq. (2) can be constructed using the relationship [30, 46] between the D_a and D_r and the diagonal elements of $\mathbf{\Sigma}$. Next, the Euler angles α, β and γ of the principal order frame are estimated by consider-

---

*While any representation of rotations may be employed, we use Euler angles (α, β, γ).

ing rotations of the model. Given Σ (Eq. 2), for each rotation $V(\alpha, \beta, \gamma)$ of the model, a new Saupe matrix $\mathbf{S}$ is computed using Eq. (2). That matrix $\mathbf{S}$ is used to compute a set of back-computed RDCs using the amide bond vectors extracted from the model and Eq. (1). The relative entropy, also known as the Kullback-Leibler distance [26], is computed between the histogram of the observed RDCs and the histogram of the back-computed RDCs. The rotation of the model that minimizes the relative entropy is chosen as the estimate for the Euler angles. The comparison of distributions to evaluate Euler angles is conceptually related to the premise used by the powder pattern method [46] to estimate the axial and rhombic components of the tensor. In the powder pattern method, the observed RDCs are implicitly compared to a distribution of RDCs generated by a uniform distribution of bond vectors. When estimating the Euler angles, GD explicitly compares the distributions using a relative entropy measure. Intuitively, the correct rotation of the model will generate a simulated distribution of unassigned RDCs that is similar to the unassigned distribution of experimentally measured RDCs.

The rotation search takes $O(nk^3)$ time for n residues on a $k \times k \times k$ grid over $SO(3)$. Thus, we can estimate alignment tensors in $O(nk^3)$ time. In practice, it takes about a minute to estimate the alignment tensor for a given medium on a Pentium 4 class processor. Thus, for p protein models in the database, the total run time is $O(pnk^3)$. Each model can be processed independently and thus the computation can be run in parallel on a cluster of machines. Further performance enhancements can be obtained by restricting the search to models that have similar lengths, or α/β mixtures. Intuitively, a model that is significantly larger/smaller, or has radically different percentages of α/β secondary structure than the target protein is less likely to have a significant structural homology. If a homology prediction has been made using protein threading or homology modelling, one need not search the entire database. Rather, these predictions can be evaluated for how well they fit the experimental data using the same method.

Finally, each model is assigned a score. Let $\Delta_\alpha = |\alpha_t - \alpha_m|$ and $\Delta_\beta = |\beta_t - \beta_m|$, where α_t and β_t are the predicted percentages of α and β structure for the target protein, t, and α_m and β_m are the actual percentages of α and β structure taken from the model, m. Let Δ_l be the difference in length between t and m. Finally, let KL_1 and KL_2 be the Kullback-Leibler distances of the two tensor estimates. A model's score is computed as follows:[†]

$$I_m = \Delta_\alpha + \Delta_\beta + \Delta_l + KL_1 + KL_2. \tag{3}$$

Each model is then ranked according to its score.

[†]Δ_α and Δ_β are multiplied by 100 so that they have the same order of magnitude as Δ_l, KL_1, and KL_2.

4.1 Improved Algorithm

We now show how the rotation minimizing the Kullback-Leibler distance can be computed in polynomial time (without a grid search) using the first-order theory of real-closed fields [21, 22, 7, 6]. Hence the $O(nk^3)$ discrete-grid rotation search can be replaced by a combinatorially precise algorithm, eliminating all dependence of the rotation search upon the resolution k.

Suppose two variables of the same type are characterized by their probability distributions f and f'. The relative entropy formula is given by $KL(f, f') = \sum_{i=1}^s f_i \ln(f_i/f'_i)$, where s is the number of levels of the variables. We will use a polynomial approximation to $\ln(\cdot)$. Let us represent rotations by unit quaternions, and use the substitution $u = \tan(\theta/2)$ to 'rationalize' the equations using rotations, thereby yielding purely algebraic (polynomial) equations. Let V be such a rotation (quaternion), D be the unassigned experimentally-measured RDCs, E be the set of model NH vectors and $B(V)$ be the set of unassigned, back-computed RDCs (parameterized by V). Hence, from Eqs. (1,2), $B(V) = E^T \mathbf{S} E = (E^T (V^T \Sigma V) E) = \{\mathbf{w}^T (V^T \Sigma V) \mathbf{w} \mid \mathbf{w} \in E\}$. (We have ignored $D_{\max}$ here for the simplicity of exposition). We wish to compute

$$\operatorname*{argmin}_{V \in S^3} KL(D, B(V)) \tag{4}$$

(We use the unit 3-sphere S^3 instead of $SO(3)$, since the quaternions are a double-covering of rotation space). Eq. (4) can be transformed into a sentence in the language of semi-algebraic sets (the first order theory of real closed fields):

$$\exists V_0 \in S^3, \ \forall V \in S^3 : \ KL(D, B(V_0)) \le KL(D, B(V)). \tag{5}$$

S^3 and $SO(3)$ are semi-algebraic sets, and Eq. (5) is a polynomial inequality with bounded quantifier alternation ($a = 1$). The number of DOF (the number of variables) is constant ($r = 3$ DOF for rotations), and the size of the equations is $O(n)$. Hence Eq. (5) can be decided exactly, in polynomial time, using the theory of real-closed fields. We will use Grigor'ev's algorithm [21, 22] for deciding a Tarski sentence, which is singly-exponential in the number of variables, and doubly-exponential only in the number of quantifer alternations. The time complexity of Grigor'ev's algorithm is $n^{O(r)^{4a-2}}$, which in our case ($a = 1$, $r = 3$) reduces to $n^{O(1)}$ which is polynomial time.

5 Results and Discussion

RDCs in two media were obtained for five different proteins; the 76-residue human ubiquitin (PDB Id 1D3Z

[12]), the 56-residue streptococcal protein G (SPG) (PDB Id 3GB1 [27]), the 129-residue hen lysozyme (PDB Id 1E8L [40]), the 81-residue DNA-Damage-Inducible Protein I (Dini) (PDB Id 1GHH [36]), and the 152-residue Galpha Interacting Protein (Gaip) (PDB Id 1CMZ [14]). Using the program CE [42], 5 structural homologs were identified for each protein. These homologous structures have low sequence identity to the target protein (Table 1). The five test proteins and their structural homologs were added to the database prior to the experiment.

We note that while there are over 18,000 protein structures deposited in the PDB to date, only a small handful of these proteins have RDC data published in the BioMagRes-Bank (BMRB) [41]. This is due, in part, to the fact that the recording of RDCs in solution has only recently been perfected. In contrast, NOE data is available for thousands of proteins. Simulating RDC data is difficult for two reasons. First, one needs to predict the alignment tensor for a given medium. This devolves to simulating the tumbling dynamics for the interaction of the protein with the aligning medium in solution. This is, in general, difficult to do. Furthermore, it is difficult to create an accurate noise model because the noise in real experimental RDC data is governed in part by such factors as the internal dynamics of the protein. We felt that we could not reasonably simulate realistic RDC data. Thus, the number of proteins we tested was limited by the contents of the BMRB. However, experimental data for 5 independent proteins is considered to be a more than adequate test suite by the NMR community [48], and many new computational protocols are tested on only one protein.
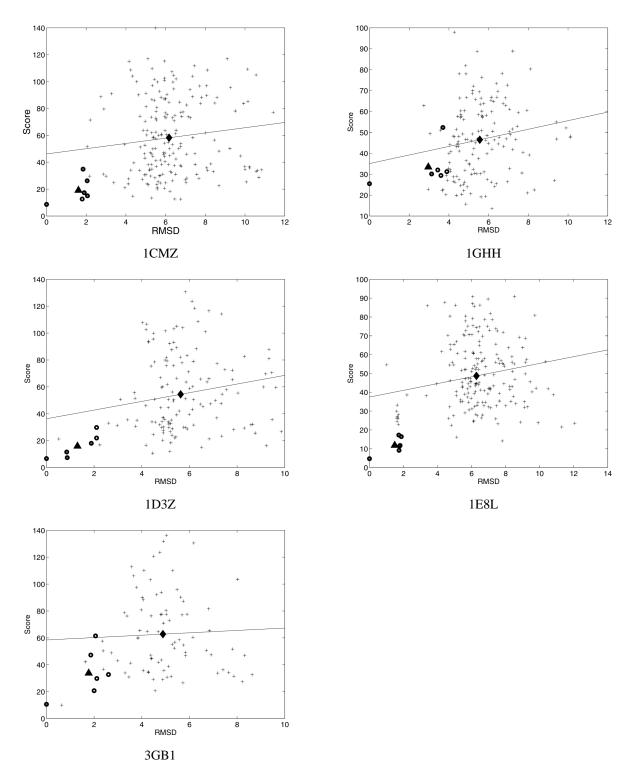
As shown in Table 1, GD identifies both the native structure and its structural homologs. The native structure and its 5 structural homologs are highly ranked among the 2,456 proteins in the structural database. In all but one case, the native fold is the top ranked model. The one exception, 1GHH, was due to the fact that the secondary structure prediction for that protein was inaccurate. We repeated the experiment on 1GHH using the correct percentages of secondary structure. The top ranking model in that experiment is the native fold. This highlights a certain sensitivity to the quality of the secondary structure prediction. One could imagine supplementing the prediction with circular dichroism (CD) data to address this issue. While it is not unexpected that the native fold is often the top ranked model, it is noteworthy that the homology detection is done without any comparison of primary sequence.

The overall rankings of the 5 selected homologs are also good. The Ubx Domain from human Faf1 (PDB Id 1H8C) (discussed in Sec. 1) was identified as a structural homolog of ubiquitin. Lysozyme (1E8L) does the best, with the native structure and 5 homologous structures occupying the top 6 places. Once again, the homologs for 1GHH do com-

paratively worse than those of the other proteins. This is due to both the inaccuracy of the secondary structure prediction and the relatively low structural similarity between 1GHH and its 5 homologs (1DHM, 1DT4, 1DV5, 1KDX and 1QR5). Note that the average RMSD between 1GHH and its homologs is 3.4 Å, while the average RMSDs between 1CMZ, 1D3Z, 1E8L and 3GB1 and their respective homologs are 1.9, 1.5, 1.8, and 2.2 Å, respectively. A subsequent analysis of 1GHH with the program CE revealed that there are no significant homologs for 1GHH in the PDB for proteins of its size. In particular, the CE significance scores computed between the 5 homologs and 1GHH are marginal. Thus, while these 5 proteins have the lowest RMSD to 1GHH of any others in the PDB, they are not necessarily related to 1GHH. This suggests that GD works best when there is a close homolog in the database.

Figure 1 contains scatter plots of the results. For all 5 proteins, the scores associated with the native fold and the 5 homologs are statistically significantly lower than the scores of unrelated proteins (p-values of 2.6×10^{-5}, 2.6×10^{-5}, 4.2×10^{-5}, 2.3×10^{-5}, and 2.9×10^{-5} for 1CMZ, 1D3Z, 1GHH, 1E8L, and 3GB1, respectively). Note the clustering of the homologs and the native structure in the lower left-hand corner. The relationship between the score computed by GD and RMSD is most highly correlated in the lower left-hand corner of the scatter plots, in the vicinity of 0-3 Å RMSD. Above about 5 Å RMSD, the correlation between the score computed by GD and RMSD is much lower. Indeed, there is no reason to expect any correlation because these proteins are unrelated to the target. Let U be the set of proteins that are unrelated to the target. Let $L \subset U$ be the proteins that have a similar length to the target, $A \subset U$ be the proteins that have a similar percentage of α structure, and $B \subset U$ be the proteins that have a similar percentage of β structure. A protein chosen at random from U will randomly fall into one or more of L, A, or B. Similarly, the bond vector orientations of unrelated proteins are only randomly correlated to the target protein. Consequently, the histograms of their back-computed RDCs are only randomly correlated to the histograms of the experimentally measured RDCs. Thus, at high RMSD, the terms Δ_α, Δ_β, Δ_l, KL_1 and KL_2 in Eq. (3) become, in effect, random variables.

6 Conclusion

We have described a fast, automated procedure for homology detection from unassigned NMR data. The relationship between structure and function is strong, thus GD can be used to help characterize the function of new proteins. GD identifies the 3D structural models in a protein structural database whose geometries best fit the unassigned experimental NMR data. It does not use sequence informa-

Figure 1. RMSD vs. GD score. Scatter plots of the RMSD vs. the score computed by GD. Only those proteins whose length is within 10% of the target protein are shown. The open circles are the data points for the native structure and five homologous structures. The + signs are the data points associated with non-homologous proteins. The diamond is the 2D mean of the +'s while the triangle is the 2D mean of the open circles. The trend line shows the correlation between the score computed by GD and RMSD for all the data points. The scores associated with the native fold and the 5 homologs are statistically significantly lower than the scores of unrelated proteins (p-values of 2.6×10^{-5}, 2.6×10^{-5}, 4.2×10^{-5}, 2.3×10^{-5}, and 2.9×10^{-5} for 1CMZ, 1D3Z, 1GHH, 1E8L, and 3GB1, respectively).

PDB ID	Homolog	Sequence Identity	RMSD	Rank
1CMZ		100%	0Å	1
	1FQI	37.8%	1.9Å	5
	1FQJ	38.2%	1.8Å	6
	1DK8	28.7%	1.9Å	8
	1EZT	44.9%	2.0Å	16
	1FQK	38.6%	1.8Å	49
1D3Z		100%	0Å	1
	1NDD	55.6%	0.6Å	2
	1BT0	61.0%	0.7Å	3
	1H8C	15.7%	1.9Å	11
	1GUA	11.6%	2.1Å	19
	1C1Y	11.6%	2.1Å	38

PDB ID	Homolog	Sequence Identity	RMSD	Rank
1GHH		100%	0Å	19
	1DHM	11.9%	3.6Å	35
	1DV5	8.5%	3.0Å	36
	1QR5	9.7%	3.5Å	37
	1DT4	12.1%	3.4Å	42
	1KDX	7.3%	3.3Å	97
1E8L		100%	0Å	1
	2EQL	49.2%	1.8Å	2
	1ALC	35.8%	1.8Å	3
	1HFZ	38.3%	1.8Å	4
	1A4V	38.2%	1.8Å	5
	1F6S	38.7%	1.7Å	6
3GB1		100%	0Å	1
	1HZ5	14.5%	2.2Å	2
	1JML	12.8%	1.8Å	5
	1HEZ	12.7%	2.0Å	12
	2GCC	10.0%	2.6Å	24
	1HZ6	14.5%	2.2Å	55

Table 1. Test Proteins and Results. The sequence identity and RMSD of the five test proteins and their respective five homologs. The final column is the rank of that model, based on the score computed by GD.

tion and is thus not limited by sequence homology. The algorithm runs in $O(pnk^3)$ time, where p is the number of proteins in the database, n is the number of residues in the target protein, and k is the resolution of a rotation search. GD requires only uniform ^{15}N-labelling of the protein and processes unassigned H^N-^{15}N residual dipolar couplings, which can be acquired in a couple of hours.

GD has been tested on NMR data from 5 test proteins against a protein structure database containing over 2,400 models. In all cases, the scores computed by GD for the native structure and its five homologs were statistically significantly lower than the scores for the unrelated proteins. In most cases, the highest ranking model is the native structure, while close structural homologs were also highly ranked. GD performs well even though it uses a very sparse set of experimental data and does not incorporate sequence homology. The GD algorithm also handles missing and noisy data well, and all our results are reported using the raw published data sets, which contain RDCs for most, but not all residues in the 5 test proteins.

We have shown that GD works well on proteins in the 56-152 residue range. It is to be expected that some modifications may be needed when scaling GD to larger proteins. The accuracy of the powder pattern method is known to increase as the number of RDCs increases. Thus, our ability to estimate the axial and rhombic components of the alignment tensors should increase with protein size. Estimating the eigenvectors of the tensors, however, will become harder as the distribution of amide bond vectors becomes more uniform. We are now exploring ^{1}H and ^{15}N chemical shift prediction [33, 47] for GD, which might be incorpo-

rated into GD as a probabilistic constraint on assignment and alignment. These assignments provide an alternative means [30] for estimating the tensors that is independent of the protein's size. We are also combining our technique with existing bioinformatics techniques in order to increase accuracy. In particular, we are incorporating protein threading (e.g., [29]) as a complementary means to identify structural homologies. The results from protein threading will be used to prune the set of candidates from the structural database. GD will, in turn, be used to evaluate how well each candidate explains the experimental data.

7 Software

The GD software is available by contacting the authors.

References

[1] MEILER, J. AND BLOMBERG, N. AND NILGES, M. AND GRIESINGER, C. A new approach for applying residual dipolar couplings as restraints in structure elucidation. *Journal of Biomolecular NMR 16* (2000), 245–252.

[2] AL-HASHIMI, H.M. AND GORIN, A. AND MAJUMDAR, A. AND GOSSER, Y. AND PATEL, D.J. Towards Structural Genomics of RNA: Rapid NMR Resonance Assignment and Simultaneous RNA Tertiary Structure Determination Using Residual Dipolar Couplings. *J. Mol. Biol. 318* (2002), 637–649.

[3] ANDREC, M. AND DU, P. AND LEVY, R.M. Protein backbone structure determination using only resid-

ual dipolar couplings from one ordering medium. *J Biomol NMR 21*, 4 (2001), 335–347.

[4] ANNILA, A. AND AITIO, H. AND THULIN, E. AND DRAKENBERG. T. Recognition of protein folds via dipolar couplings. *J. Biom. NMR 14* (1999), 223–230.

[5] BAILEY-KELLOGG, C. AND WIDGE, A. AND KELLEY III, J.J. AND BERARDI, M.J. AND BUSHWELLER, J.H. AND DONALD, B.R. The NOESY Jigsaw: automated protein secondary structure and mainchain assignment from sparse, unassigned NMR data. *J Comput Biol 7*, 3-4 (2000), 537–58.

[6] BASU, S. An Improved Algorithm for Quantifier Elimination Over Real Closed Fields. *IEEE FOCS* (1997), 56–65.

[7] BASU, S. AND ROY, M.F. On the combinatorial and algebraic complexity of quantifier elimination. *Journal of the ACM (JACM) 43*, 6 (1996), 1002–1045.

[8] BERMAN, H.M AND WESTBROOK, J. AND FENG, Z. AND GILLILAND, G. AND BHAT, T.N. AND WEISSIG, H. AND SHINDYALOV, I.N. AND BOURNE, P.E. The Protein Data Bank. *Nucl. Acids Res. 28* (2000), 235–242.

[9] BLUNDELL, T.L. AND SIBANDA, B.L. AND STERNBERG, M.J. AND THORNTON, J.M. Knowledge-Based Prediction of Protein Structures and the Design of Novel Molecules. *Nature 326* (1987), 347–352.

[10] CHEN, Y. AND REIZER, J. AND SAIER JR., M. H. AND FAIRBROTHER, W. J. AND WRIGHT, P. E. Mapping of the binding interfaces of the proteins of the bacterial phosphotransferase system, HPr and IIAglc. *Biochemistry 32*, 1 (1993), 32–37.

[11] CHOU, J.J AND LI, S. AND BAX, A. Study of conformational rearrangement and refinement of structural homology models by the use of heteronuclear dipolar couplings. *J. Biom. NMR 18* (2000), 217–227.

[12] CORNILESCU, G. AND MARQUARDT, J. L. AND OTTIGER, M. AND BAX, A. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *J.Am.Chem.Soc. 120* (1998), 6836–6837.

[13] CUFF, J.A. AND CLAMP, M.E. AND SIDDIQUI, A.S. AND FINLAY, M. AND BARTON, G.J. Jpred: A Consensus Secondary Structure Prediction Server. *Bioinformatics 14* (1998), 892–893.

[14] DE ALBA, E. AND DE VRIES, L. AND FARQUHAR, M. G. AND TJANDRA, N. Solution Structure of Gaip (Galpha Interacting Protein): A Regulator of G Protein Signaling. *J.Mol.Biol. 291* (1999), 927.

[15] DELAGLIO, F. AND KONTAXIS, G. AND BAX, A. Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *J. Am. Chem. Soc 122* (2000), 2142–2143.

[16] FEJZO, J. AND LEPRE, C.A. AND PENG, J.W. AND BEMIS, G.W. AND AJAY AND MURCKO, M.A. AND MOORE, J.M. The SHAPES strategy: An NMR-based approach for lead generation in drug discovery. *Chem. and Biol. 6* (1999), 755–769.

[17] FETROW, J.S. AND BRYANT, S.H. New Programs for Protein Tertiary Structure Prediction. *Bio/Technology 11* (1993), 479–484.

[18] FIAUX, J. AND BERTELSEN, E. B. AND HORWICH, A. L. AND WÜTHRICH, K. NMR analysis of a 900K GroELGroES complex. *Nature 418* (2002), 207 – 211.

[19] FOWLER, C.A. AND TIAN, F. AND AL-HASHIMI, H. M. AND PRESTEGARD, J. H. Rapid Determination of Protein Folds Using Residual Dipolar Couplings. *J. Mol. Bio 304*, 3 (2000), 447–460.

[20] GREER, J. Comparative Modeling of Homologous Proteins. . *Meth. Enzymol. 202* (1991), 239–252.

[21] GRIGOR'EV, D.Y. Complexity of deciding Tarski algebra. *Journal of Symbolic Computation 5*, 1-2 (February/April 1988), 65–108.

[22] GRIGOR'EV, D.Y. AND VOROBJOV,N.N. Solving systems of polynomial inequalities in subexponential time. *Journal of Symbolic Computation 5*, 1-2 (February/April 1988), 37–64.

[23] HUS, J.C. AND MARION, D. AND BLACKLEDGE, M. *De novo* Determination of Protein Structure by NMR using Orientational and Long-range Order Restraints. *J. Mol. Bio 298*, 5 (2000), 927–936.

[24] HUS, J.C. AND PROPMERS, J. AND BRÜSCHWEILER, R. Assignment strategy for proteins of known structure. *J. Mag. Res 157* (2002), 119–125.

[25] JOHNSON, M.S. AND SRINIVASAN, N. AND SOWDHAMINI, R. AND BLUNDELL, T.L. Knowledge-Based Protein Modeling. *Mol. Biochem. 29* (1994), 1–68.

[26] KULLBACK, S. AND LEIBLER, R. A. On Information and Sufficiency. *Annals of Math. Stats. 22* (1951), 79–86.

[27] KUSZEWSKI, J. AND GRONENBORN, A. M. AND CLORE, G. M. Improving the Packing and Accuracy of NMR Structures with a Pseudopotential for the Radius of Gyration. *J. Am. Chem. Soc. 121* (1999), 2337–2338.

[28] LANGMEAD, C. J., YAN, A. K., WANG, L., LILIEN, R. H., AND DONALD, B. R. A Polynomial-Time Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments. *Proc. of the 7th Ann. Intl. Conf. on Research in Comput. Biol. (RECOMB) Berlin, Germany, April 10-13* (2003), 176–187.

[29] LATHROP, R.H. AND SMITH, T.F. Global Optimum Protein Threading with Gapped Alignment and Empirical Pair Score Functions. *J. Mol. Biol. 255* (1996), 641–665.

IEEE
COMPUTER
SOCIETY

[30] LOSONCZI, J.A. AND ANDREC, M. AND FISCHER, W.F. AND PRESTEGARD J.H. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Magn Reson 138*, 2 (1999), 334–42.

[31] MUELLER, G.A. AND CHOY, W.Y. AND YANG, D. AND FORMAN-KAY, J.D. AND VENTERS, R.A. AND KAY, L.E. Global Folds of Proteins with Low Densities of NOEs Using Residual Dipolar Couplings: Application to the 370-Residue Maltodextrin-binding Protein. *J. Mol. Biol. 300* (2000), 197–212.

[32] NATIONAL INSTITUTE OF GENERAL MEDICAL SCIENCES. The Protein Structure Initiative. The National Institute of General Medical Sciences, 2002. URL: http://www.nigms.nih.gov/funding/psi.html.

[33] OSAPAY, K. AND CASE,D.A. A new analysis of proton chemical shifts in proteins. *J. Am. Chem. Soc. 113* (1991), 9436–9444.

[34] PALMER III, A. G. Probing Molecular Motion By NMR. *Current Opinion in Structural Biology 7* (1997), 732–737.

[35] PEARLMAN, D.A. AND CASE, D.A. AND CALDWELL, J.W. AND ROSS, W.S. AND CHEATHAM, T.E. AND DEBOLT, S. AND FERGUSON, D. AND SEIBEL, G. AND KOLLMAN, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structures and energies of molecules. *Comp. Phy. Comm. 91* (1995), 1–41.

[36] RAMIREZ, B. E. AND VOLOSHIN, O. N. AND CAMERINI-OTERO, R. D. AND BAX, A. Solution structure of DinI provides insight into its mode of RecA inactivation. *Protein Sci. 9* (2000), 2161.

[37] ROHL, C.A AND BAKER, D. De Novo Determination of Protein Backbone Structure from Residual Dipolar Couplings Using Rosetta. *J. Am. Chem. Soc. 124*, 11 (2002), 2723–2729.

[38] SALI, A. AND OVERINGTON, J.P. AND JOHNSON, M.S. AND BLUNDELL, T.L. From Comparisons of Protein Sequences and Structures to Protein Modelling and Design. *Trends Biochem. Sci. 15* (1990), 235–240.

[39] SAUPE, A. Recent Results in the field of liquid crystals. *Angew. Chem. 7* (1968), 97–112.

[40] SCHWALBE, H. AND GRIMSHAW, S. B. AND SPENCER, A. AND BUCK, M. AND BOYD, J. AND DOBSON, C. M. AND REDFIELD, C. AND SMITH, L. J. A Refined Solution Structure of Hen Lysozyme Determined Using Residual Dipolar Coupling Data. *Protein Sci. 10* (2001), 677–688.

[41] SEAVEY, B.R. AND FARR, E.A. AND WESTLER, W.M. AND MARKLEY, J.L. A Relational Database for Sequence-Specific Protein NMR Data. *J. Biom.*

NMR 1 (1991), 217–236.

[42] SHINDYALOV, I., AND BOURNE, P. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Engineering 11*, 9 (1998), 739–747.

[43] SHUKER, S. B. AND HAJDUK, P. J. AND MEADOWS, R. P. AND FESIK, S. W. Discovering high affinity ligands for proteins: SAR by NMR. *Science 274* (1996), 1531–1534.

[44] TIAN, F. AND VALAFAR, H. AND PRESTEGARD, J. H. A Dipolar Coupling Based Strategy for Simultaneous Resonance Assignment and Structure Determination of Protein Backbones. *J. Am. Chem. Soc. 123* (2001), 11791–11796.

[45] TJANDRA, N. AND BAX, A. Direct Measurement of Distances and Angles in Biomolecules by NMR in a Dilute Liquid Crystalline Medium. *Science 278* (1997), 1111–1114.

[46] WEDEMEYER, W. J. AND ROHL, C. A. AND SCHERAGA, H. A. Exact solutions for chemical bond orientations from residual dipolar couplings. *J. Biom. NMR 22* (2002), 137–151.

[47] WISHART, D.S. AND WATSON, M.S. AND BOYKO, R.F. AND SYKES, B.D. Automated 1H and 13C Chemical Shift Prediction Using the BioMagResBank. *J.Biomol. NMR 10* (1997), 329–336.

[48] WÜTHRICH, K., Ed. *The Journal of Biomolecular NMR*. Kluwer Academic Publishers, Van Godewijckstraat 30, P.O. Box 17, 3300 AA Dordrecht, The Netherlands, 1997-2003.

[49] XU, Y. AND XU, D. AND CRAWFORD, O. H. AND EINSTEIN, J. R. AND SERPERSU, E. Protein Structure Determination Using Protein Threading and Sparse NMR Data. In *Proc. RECOMB* (2000), pp. 299–307.

[50] ZIMMERMAN, D.E. AND KULIKOWSKI, C.A. AND FENG, W. AND TASHIRO, M. AND CHIEN, C-Y. AND Z ROS, C.B. AND MOY, F.J. AND POWERS, R. AND MONTELIONE G.T. Artificial intelligence methods for automated analysis of protein resonance assignments. *J. Mol. Biol. 269* (1997), 592 – 610.

[51] ZWECKSTETTER, M. AND BAX, A. Single-step determination of protein substructures using dipolar couplings: aid to structural genomics. *J Am Chem Soc 123*, 38 (2001), 9490–1.