Probability Density Profile Analysis: A Method for Identifying Novel Protein Structures

Arjang Fahim[1], Stephanie Irausquin[1], Matthew Fawcett[1], Mikhail Simin[1], and Homayoun Valafar[1]
[1]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA

Abstract - *Although a number of scientific advances have been made in the area of structural biology, a few obstacles continue to impede our ability to quickly and efficiently characterize protein structure-function relationships. Probability Density Profile Analysis (PDPA) is a method which rapidly quantifies the structural novelty of a protein, based on the statistical analyses of a minimal amount of empirical data. Here we present findings related to the sensitivity and range of applicability of PDPA. Our results support the conclusion that two dimensional PDPA (2D-PDPA) can reliably be utilized for identification of a protein structure to within 3Å of the known structure, using a library of existing structures. Furthermore, the sensitivity of 2D-PDPA has been tested using proteins containing different secondary structural characteristics (α, β, and α/β) and our preliminary investigations support the conclusion that 2D-PDPA is equally applicable to all general classes of proteins.*

Keywords: Residual Dipolar Couplings, Parzen Density Estimation, Probability Density Profile Analysis, Structural Homology Detections

1 Introduction

Proteins are often referred to as the working molecules of a cell, performing many important structural, functional and regulatory processes [1]. Yet, revealing the function of proteins is a particularly challenging problem. Sequence-based approaches are an option, but identifying functionally characterized homologs is only feasible for less than half of the proteins predicted from genome sequencing projects [2] and is often compounded by the fact that proteins tend to be multi-functional [3]. Since a protein's structure often dictates its function, an alternative approach is to determine the structure of the protein of interest in order to identify functionally important sites [3]. This is believed to provide a solution for many of the remaining proteins, since structure is more evolutionarily conserved than sequence [2, 3].

Although the characterization of any protein adds to repositories of structural data, most structural biologists would concur that novel structures are particularly important for a number of reasons: they generate models of similar proteins for comparison; identify evolutionary relationships; further contribute to our understanding of protein function and mechanism; and allow for the fold of other family members to be inferred [4-6]. Considering the evolutionary mechanisms responsible for the generation of new structures in proteins, it has been speculated that there may be a limited number of unique protein folds - as few as ten thousand families [7-9]. Currently the Protein Data Bank (PDB; [10]) consists of nearly 68,000 protein structures, but less than 1,400 families are represented and approximately no new fold families have been reported since 2008 [11, 12]. Ideally, solved protein structures for new protein families [6] would be used as templates for *in silico* structure prediction methods [4, 13] and the results of both solved and predicted structures would in turn be used to infer function [2, 14, 15]. However, such an approach requires new, efficient and cost-effective computational methods for target selection and structure determination.

Traditional methods of structure determination, such as X-ray crystallography and NMR spectroscopy, are expensive and time-consuming techniques. Previously we presented a method, referred to as Probability Density Profile Analysis (PDPA), which rapidly quantifies the structural novelty of a protein using only a minimal amount of empirical data. PDPA is a potentially important tool that provides investigators with fast, cost-effective, easy to interpret results while also further contributing to our understanding of structure-function relationships in proteins. The interpretation of PDPA scores, as well as the effective applicable range of PDPA, had not been known previously. In this report, we provide the means to interpret PDPA results and establish both the sensitivity and applicability of this method [19, 23].

2 Methods

2.1 Residual Dipolar Coupling (RDC)

Residual Dipolar Couplings are the result of dipolar interactions in a partially ordered system [16] and are defined in Equation (1):

$$D_{ij} = D_{max} \cdot \left\langle \frac{3\cos^2(\theta) - 1}{2} \right\rangle \quad (1)$$

In this equation, D_{ij} is the magnitude of calculated RDC in hertz that is between two ½ spin nuclei in the presence of a magnetic field; θ signifies the angle between the magnetic field vector and the inter-spin vector (nuclei i and j); brackets represent the time average for a specific coupling; and D_{max} denotes the maximum magnitude of a coupling that is further defined in Equation (2).

$$D_{max} = -\left(\frac{\mu_0}{4\pi}\right)\frac{\gamma_i \gamma_j h}{2\pi^2 r_{ij}^3} \quad (2)$$

In Equation (2), μ_0 signifies the magnetic permeability; γ_i and γ_j are the gyromagnetic ratios of two nuclei (*i* and *j)*; *r* is the intranuclear distance between two nuclei; and *h* is Planck's constant.

The RDC equation can be manipulated into a matrix form (Equation (4)) as shown in Equation (3):

$$D_{ij} = v_{ij} . S . v_{ij}^T \quad (3)$$

$$S = \begin{bmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{xy} & S_{yy} & S_{yz} \\ S_{xz} & S_{yz} & S_{zz} \end{bmatrix} \quad (4)$$

A unit vector that joins two corresponding nuclei is represented by v_{ij} and *S* is the traceless and symmetric Saupe order tensor matrix (OTM) [17]. *S* can be further decomposed into $S = RS'R^T$ such that *R* is a Euler rotation matrix, whose columns are the eigenvectors of *S*; and *S′* (Equation 5) is a traceless diagonal matrix of the eigenvalues of *S,* whose diagonal elements $S'_{xx}, S'_{yy}, S'_{zz}$ are the principle order parameters (POP).

$$S = \begin{bmatrix} S'_{xx} & 0 & 0 \\ 0 & S'_{yy} & 0 \\ 0 & 0 & S'_{zz} \end{bmatrix} \quad (5)$$

The rotation matrix *R* can be decomposed into three different rotations related to *x, y* and *z* as shown in Equation (6):

$$R(\alpha, \beta, \gamma) = R_z(\alpha) R_y(\beta) R_z(\gamma) \quad (6)$$

Using the previous equations, the order tensor can be rewritten in five parameters: $S'_{xx}, S'_{yy}, \alpha, \beta, \gamma$. This particular parameterization is used in our experiment to generate RDC data sets.

2.2 1D-PDP Analysis

Our initial work with PDPA was conducted using One Dimensional Probability Density Profile Analysis (1D-PDPA) and was based on unassigned RDC data from one alignment medium [18]. This proof of concept established the feasibility of identifying homologous structures from unassigned RDC data, however it lacked the potential for large scale applications. In summary, 1D-PDPA established structural similarity on the basis of comparing the distribution of experimental and computed RDC data. 1D-PDPA requires a collection of experimental unassigned RDCs as well as a library of potential structures.

2.3 2D-PDPA

2D-PDPA extends the analysis of 1D-PDPA by utilizing RDC data from two alignment media. The additional set of experimental RDC data has obvious advantages over 1D-PDPA. 2D-PDPA limits the search space to seven parameters [19] and is capable of generating a more accurate and unique PDP for a given structure. A 2D-PDPA analysis session requires a collection of RDC data from two alignment media along with a library of homologous structures. A two dimensional Parzen density estimation (or kernel density estimation) is used to generate a two dimensional PDP (2D-PDP) by considering both alignment media [19]. Figure 1 illustrates a sample 2D-PDP for the protein Pf2048, a structure which has not yet been characterized. The two dimensional distribution of RDCs that is generated from the experimental data is denoted as the query PDP (qPDP) and is used, in addition to the estimated order tensors, as input to the 2D-PDPA. Incorporation of RDC data from the second alignment medium requires an extension of the search space by three more variables representing possible orientations of the second alignment medium with respect to the first one. Traditional inclusion of these three additional variables would have increased computation time by a factor of 2.5657e+09. This intractable increase in computation time has been eliminated based on new technology that has been recently introduced [19, 20].

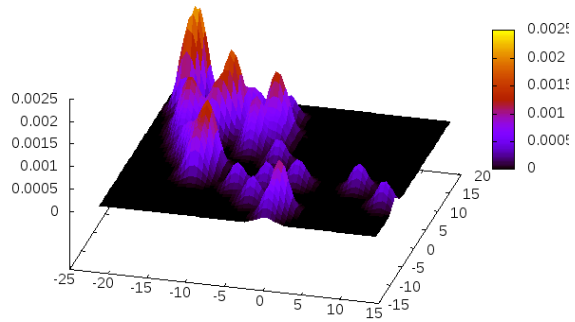


Figure 1. *An example 2D-PDP signature for a protein (Pf2048) of unknown structure.*

2D-PDPA calculates PDP for every rotation and a scoring method is used to find the best structure in terms of the similarity to the qPDP. To calculate fitness scores we consider three metric systems: Manhattan Block, Chi-Square, and Modified Chi-Square. The Manhattan Block method is defined in Equation (7):

$$S(qPDP, cPDP) = \sum_{i \in M} |(q_i - c_i)| \quad (7)$$

In Equation (7), q_i represents the i^{th} value of *qPDP* and c_i represents the i^{th} value of computed PDP (*cPDP*). *M* denotes the number of sampled points in both query and calculated

PDP sets. The Chi-Square method is defined in Equations (8) and (9):

$$\delta^2_i(q_i, c_i) = \frac{(q_i - c_i)^2}{q_i} \quad (8)$$

$$X^2(qPDP, cPDP) = \sum_{i \in M} \delta^2_i(q_i, c_i) \quad (9)$$

In Equation (8), q_i represents the i^{th} value of *qPDP* and c_i represents the i^{th} value of *cPDP*. Due to the asymmetric nature of the χ^2 metric $\left(X^2(A,B) \neq X^2(B,A)\right)$, a modified Chi-Square has been introduced and shown in Equation (10):

$${}_mX^2(qPDP, cPDP) = \frac{\left[X^2(qPDP, cPDP) + X^2(cPDP, qPDP)\right]}{2} \quad (10)$$

In equation (10), ${}_m\chi^2$ denotes the modified χ^2 metric and *qPDP* and *cPDP* represent the experimental and computed PDPs, respectively. The modified χ^2 metric is a symmetric measure of distance and it therefore constitutes a formal metric space. During our early investigations, no preference was given to any one of the scoring metrics described above. However, based on the investigation that is presented here, the Manhattan Block metric was able to demonstrate slightly better results (shown in Figure 3a-c) in terms of the distribution of scores over bb-rmsd and as well as greater R^2 values.

2.4 Data Preparation

In this experiment, three reference proteins of different sizes and structural types (Table 1) were utilized in order to assess the sensitivity and selectivity of 2D-PDPA. This step is necessary due to the influence of secondary structures on orientation of the backbone N-H vectors. Traditionally, RDC data from helical regions have been reported to carry less information relative to other secondary strcutres. The proteins listed in (Table 1), were obtained from PDB [10]; Figure 2 provides a cartoon representation of each of the structures listed in Table 1.

Table 1. *Protein structures obtained from the Protein Data Bank.*

Protein	Secondary Structure	Number of Residues	CATH Classification
1A1Z	α	91	1.10.533
1OUR	β	114	2.60.120.400
1GB1	α/β	56	3.10.20.10

For each protein structure listed in Table 1, a set of one thousand structural variations were created by randomly altering the backbone ϕ and ψ torsion angles. Each dataset represented structural variations in the range of 0-8 Å with respect to the corresponding reference structure and were generated in the PDB file format. To obtain the RDC data for the three reference proteins, we utilized REDCAT [21]. The assignment information was discarded prior to providing these data to 2D-PDPA. The PDB files were exported in REDCAT [21] to retrieve the RDC data in the 2D-PDPA program. Two sets of ^{15}N-^{1}H backbone RDC data, representing two typical alignment media, were calculated for each reference protein by using REDCAT and the initial order parameters shown in Table 2. The RDC sets were calculated separately under three conditions: with one set containing no error, the second set corrupted through the addition of uniform noise in the range of ±1Hz, and the third set consisted of randomly eliminating 15% of RDC data that is normally expected during pragmatic conditions. The first set serves to simulate the ideal conditions (no error) versus the real conditions (±1Hz and 15% of RDC gap for the second and third sets).

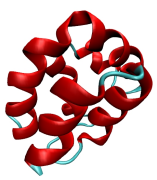
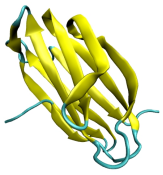
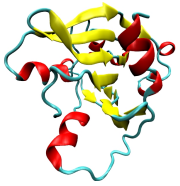
1A1Z	1OUR	1G1B
		

Figure 2*. Illustrates the structures listed in Table 1.*

Table 2. *List of initial order parameters used to calculate two RDC sets.*

	Sxx	Syy	Szz	Alpha	Beta	Gamma
Set1	3.00e-4	5.00e-4	-8.00e-4	0º	0º	0º
Set2	-4.00e-4	-6.00e-4	1.00e-3	40º	50º	-60º

The 2D Parzen Density Estimation [18] program was used to analyze RDC data and to create the 2D Probability Density Profile (2D-PDPA) [16] finger prints of each protein. Order tensors were calculated in two ways: First, the optimal 2D order tensors are obtained from REDCAT using structure and calculated RDC data; Second, order tensors are estimated using RDC data from two alignment media [22]. These two approaches represent a transition from ideal to more pragmatic conditions.

3 Results and Discussion

3.1 Experiment 1

The main objective of this experiment was to identify differences between the various metrics in order to establish the most appropriate metric for use. Experiment 1 used protein 1GB1 and its corresponding calculated RDC data, using no error or noise to demonstrate the ideal conditions. The experiment was repeated 3 times with different metrics each time. Order tensor matrices were obtained from REDCAT [21] for each RDC set (Table 3).

The relationship between 2D-PDPA structure scores and bb-rmsd for the one thousand variable structures generated is shown for each metric in Figure 3; the corresponding least squares regression logarithmic line and R^2 values are also shown for each metric. For bb-rmsd values up to 2.5Å, a linear correlation between PDPA scores and bb-rmsd exists

(Figure 3). For structures with bb-rmsd greater than 2.5Å, PDPA scores remain in the same range: [0.8-1] for Manhattan Block, [2-3] for Modified Chi-Square, and [10-15] for Chi-Square (Figure 3). For all metrics tested, the Manhattan Block obtained the highest R^2 value (0.65, Figure 3). Therefore, the Manhattan-Black metric was selected and utilized exclusively for all remaining experiments.

Table 3. *List of order parameters for each RDC set (alignment medium) obtained from REDCAT.*

Order Tensor	No Error (1G1B)	±1Hz Error (1G1B)	15 RDC Gap (1G1B)	±1Hz Error (1A1Z)	±1Hz Error (1OUR)
Sxx1	3e−4	2.966e−4	2.967e−4	3.091e−4	3.022e−4
Syy1	4e−4	5.08e−4	5.061e−4	4.985e−4	5.053e−4
Sxx2	7.99e−5	9.624e−5	9.726e−5	8.665e−5	−3.235e−5
Sxy2	3.89e−4	3.863e−4	3.795e−4	3.936e−4	4.05e−4
Sxz2	5.42e−4	5.412e−4	5.428e−4	5.44e−4	6.325e−4
Syy2	−1.70e−4	−1.784e−4	−1.82e−4	−1.856e−4	−5.998e−5
Syz2	5.414e−4	5.445e−4	5.396e−4	5.369e−4	4.348e−4

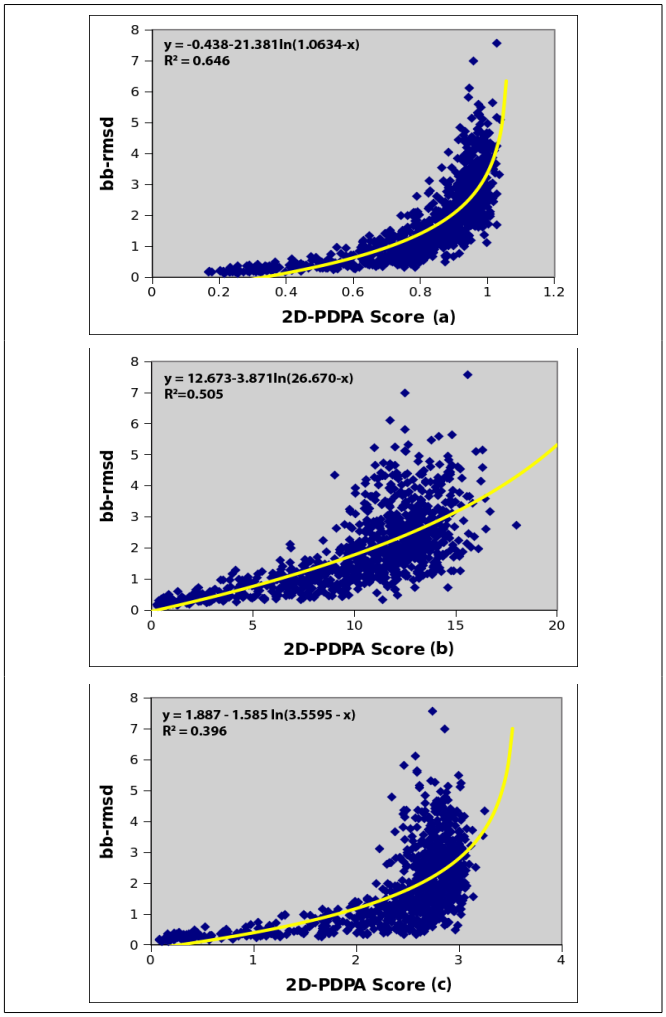


Figure 3. *Calculated 2D-PDPA scores vs bb-rmsd using different scoring methods for 1GB1 protein: (a) Block scoring method, (b) Chi-square scoring method, and (c) Modified chi-square scoring method.*

3.2 Experiment 2

The objective of this experiment was to study the behavior of 2D-PDPA as a function of experimental noise. Experiment 2 used 1GB1 and calculated the RDC data using ±1Hz error to demonstrate noisy conditions. Order tensor matrices were obtained from REDCAT [21] for each RDC set (Table 3).

Figure 4 shows the relationship between the 2D-PDPA's score (Manhattan-Block distance) for one thousand structures and their corresponding bb-rmsd with respect to the original structure; the least squares regression line and R^2 for the data are also shown in Figure 4.

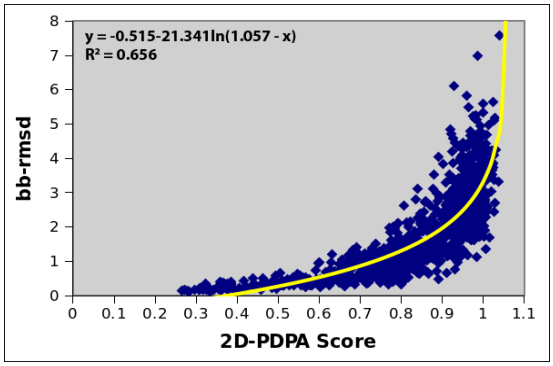


Figure 4. *Calculated 2D-PDPA scores vs bb-rmsd using Manhattan Block metric for 1GB1 with ±1hz error added.*

This experiment was repeated by randomly removing 15 (28%) RDC values from both synthetic RDC data sets. Order tensor matrices were obtained from REDCAT for each RDC set (Table 3). The plot of the 2D-PDPA scores using block metric against the bb-rmsd is seen in Figure 5 along with the least squares regression line and R^2 value. The PDPA scores increase as a result of the random removal of RDC data, however a correlation still exists between PDPA score and bb-rmsd (R^2= 0.573, Figure 5).

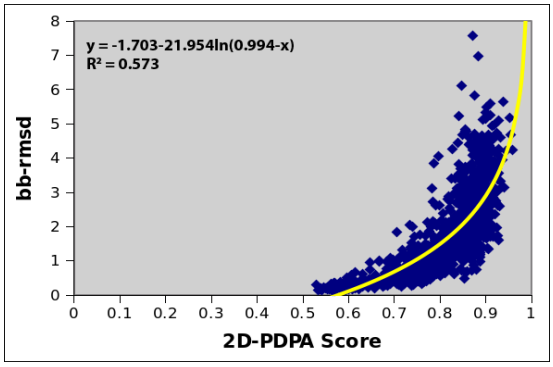


Figure 5. *Calculated 2D-PDPA scores vs bb-rmsd using Manhattan Block metric for 1GB1 with 15 (28%) of RDC data removed from RDC sets.*

3.3 Experiment 3

Experiment 3 used protein 1A1Z, which is an α–helical structure, and calculated the RDC data with ±1Hz of uniformly added error. Order tensors were obtained from REDCAT for each RDC set (Table 3). Figure 6 shows the correlation between the bb-rmsd of the structures and the 2D-PDPA scores; leasts squares linear regression line and R^2 values are included.

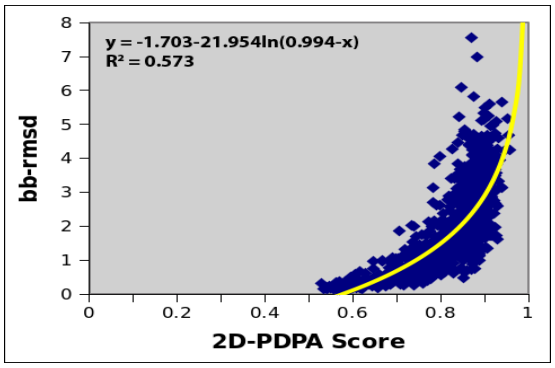


Figure 6. *Calculated 2D-PDPA scores vs bb-rmsd using Manhattan Block metric for 1A1Z with ±1hz error added.*

3.4 Experiment 4

Experiment 4 used protein 1OUR and calculated the RDC data using ±1Hz error to demonstrate noisy conditions. Order tensor matrices were obtained from REDCAT [21]for each RDC set (Table 3). Figure 7 shows the relationship between one thousand 2D-PDP structure scores and bb-rmsd with the Manhattan Block metric. The least squares regression line and the R^2 value are also shown in Figure 7.

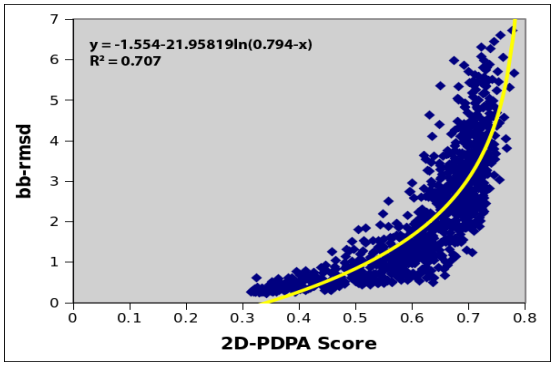


Figure 7. *Calculated 2D-PDPA scores vs bb-rmsd using Manhattan Block metric for 1OUR with ±1hz error added.*

4 Conclusion

2D-PDPA is a powerful method which can be utilized to identify homologous structures using only a minimal set of experimental data prior to a full structure determination protocol. Therefore, 2D-PDPA is a viable method for ascertaining a protein's structural novelty to within 3Å, relative to the existing library of structures. The main contribution of our method demonstrates the correlation between scored PDP and bb-rmsd of the corresponding structure. This also confirms the reliability of the 2D-PDPA identification and scoring, up to a threshold of 3Å. To conduct our experiments we chose 3 structures representing three distinct CATH families. The experiment repeated for RDCs with no error and RDCs with error and missing data has confirmed 2D-PDPA's capability for pragmatic conditions. In all cases, the correlation between bb-rmsd and calculated PDP scores are clear. In the case of noisy RDCs data, our experiments show a slight shift of 2D-PDPA's score, yet a correlation is maintained. A-priori determination of score thresholds allows for interpretation and reliability of the 2D-PDPA's performance. The observed threshold of 3Å also extends the use of the presented method to confirmation of computationally modeled structures. A hybrid approach of 2D-PDPA based selection of best computed structures can be envisioned, which allows for combined strengths of computational and experimental methods of structure determination while maintaining low cost.

5 Acknowledgements

This work was supported by NSF-Career grant MCB-0644195 from NSF to Dr. Homayoun Valafar. The high-performance computational environment used for this work was funded by NSF grant CNS-0708391.

6 References

[1] Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D & Darnell J. Molecular Cell Biology, 4th edition. . W.H. Freeman, 2000.

[2] Hvidsten TR, Laegreid A, Kryshtafovych A, Andersson G, Fidelis K & Komorowski J. A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. *PLoS One* (2009) **4**: p. p. e6266.

[3] Skolnick J & Fetrow JS. From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends in Biotechnology* (2000) **18**: pp. 34-39.

[4] Baker D & Sali A. Protein structure prediction and structural genomics. *Science* (2001) **294**: p. pp. 93-96.

[5] Brenner SE, Chothia C, Hubbard TJ & Murzin AG. Understanding protein structure: using scop for fold interpretation. *Methods Enzymol* (1996) **266**: p. pp. 635-643.

[6] Chandonia J & Brenner SE. The impact of structural genomics: expectations and outcomes. *Science* (2006) **311**: p. pp. 347-351.

[7] Orengo CA, Todd AE & Thornton JM. From protein structure to function. *Current Opinion in Structural Biology* (1999) **9**: pp. 374-382.

[8] Sali A & Kuriyan J. Challenges at the frontiers of structural biology. *Trends in Biochemical Sciences* (1999) **24**: p. M20-M24.

[9] Service RF. A dearth of new folds. (2005) **307**: pp. 1555-1555.

[10] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN & Bourne PE. The Protein Data Bank. *Nucleic Acids Res* (2000) **28**: pp. 235-242.

[11] Pearl FMG, Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP, Thornton JM & Orengo CA. Assigning genomic sequences to CATH. *Nucleic Acids Research* (2000) **28**: pp. 277-282.

[12] Murzin AG, Brenner SE, Hubbard T & Chothia C. SCOP - A Structural Classification Of Proteins Database For The Investigation Of Sequences And Structures. *J Mol Biol* (1995) **247**: pp. 536-540.

[13] Kopp J, Bordoli L, Battey JND, Kiefer F & Schwede T. Assessment of CASP7 predictions for template-based modeling targets. *Proteins* (2007) **69 Suppl 8**: p. pp. 38-56.

[14] Murzin AG & Patthyt L. Sequences and topology: From sequence to structure to function. *Curr Opin Struct Biol* (1999) **9**: p. p. 359–362.

JH & Valafar H. Rapid classification of protein structure models using unassigned backbone RDCs and probability density profile analysis (PDPA). *J Magn Reson* (2008) **192**: pp. 60-68.

[17] Saupe, A & Englert, G. Phys. Rev. Lett;High-Resolution Nuclear Magnetic Resonance Spectra of Orientated Molecules. *Phys. Rev. Lett* (1963) **11**: pp. 462-464.

[18] Valafar H & Prestegard JH. Rapid classification of a protein fold family using a statistical analysis of dipolar couplings. *Bioinformatics* (2003) **19**: pp. 1549-1555.

[19] Yandle R MR&VH. Using Residual Dipolar Coupling from two Alignment Media to Detect Structural Homology. *BIOCOMP* (2009) : p. pp. 90-95.

[20] Mukhopadhyay R, Miao X, Shealy P & Valafar H. Efficient and accurate estimation of relative order tensors from lambda-maps. *J Magn Reson* (2009) **198**: pp. 236-247.

[21] Valafar H & Prestegard J. REDCAT: a residual dipolar coupling analysis tool. *J Magn Reson* (2004) **167**: pp. 228-241.

[22] Miao X, Mukhopadhyay R & Valafar H. Estimation of relative order tensors, and reconstruction of vectors in space using unassigned RDC data and its application. *J Magn Reson* (2008) **194**: pp. 202-211.

[23] Bansal S,Miao X,Adam M,Prestegard J & Valafar H Rapid classification of protein structure models using unassigned backbone RDCs and probability density profile analysis (PDPA). *J Magan Reson* (2008) **192:** pp. 60-68.