Published in final edited form as:

Biochim Biophys Acta. 2015 October ; 1854(10 0 0): 1545–1552. doi:10.1016/j.bbapap.2015.02.016.

Application of Data Mining Tools for Classification of Protein Structural Class from Residue Based Averaged NMR Chemical Shifts

Arun. V. Kumar¹, Rehana F.M. Ali¹, Yu Cao^{1,a}, and V.V. Krishnan^{2,3,*}

¹Department of Computer Science, California State University, Fresno CA 93740

²Department of Chemistry, California State University, Fresno CA 93740

³Department of Pathology and Laboratory Medicine, School of Medicine, University of California, Davis CA 95616

Abstract

The number of protein sequences deriving from genome sequencing projects is outpacing our knowledge about the function of these proteins. With the gap between experimentally characterized and uncharacterized proteins continuing to widen, it is necessary to develop new computational methods and tools for protein structural information that is directly related to function. Nuclear magnetic resonance (NMR) provides powerful means to determine three-dimensional structures of proteins in the solution state. However, translation of the NMR spectral parameters to even low-resolution structural information such as protein class requires multiple time consuming steps. In this paper, we present an unorthodox method to predict the protein structural class directly by using the residue's averaged chemical shifts (ACS) based on machine learning algorithms. Experimental chemical shift information from 1491 proteins obtained from Biological Magnetic Resonance Bank (BMRB) and their respective protein structural classes derived from structural classification of proteins (SCOP) were used to construct a data set with 119 attributes and 5 different classes. Twenty four different classification schemes were evaluated using several performance measures. Overall the residue based ACS values can predict the protein structural classes with 80 % accuracy measured by Matthew Correlation coefficient. Specifically protein classes defined by mixed $\alpha\beta$ or small proteins are classified with > 90% correlation. Our results indicate that this NMR-based method can be utilized as a low-resolution tool for protein structural class identification without any prior chemical shift assignments.

Keywords

Protein structural class; NMR; chemical shift; Data mining

*Corresponding author: Krish Krishnan (krish@csufresno.edu or vvk Krishnan@ucdavis.edu).

^aCurrent address: Department of Computer Science, The University of Massachusetts Lowell 198 Riverside Street, Lowell, Massachusetts 01854.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

I. Introduction

The secondary protein structure was postulated over 60 years ago by Pauling and Corey, who predicted the existence of two local periodic motifs: the α -helix and the β -sheet [1, 2]. By understanding the importance of the relationship between primary and secondary structures of proteins, it can aid in the types of ways that a protein folds. [3–12]. Specifically, the secondary structure is widely used in a number of structural biology applications, such as structure comparison [13], classification [14–16], and visualization [17]. The secondary structures can be used to determine the family, superfamily, and tertiary fold of the underlying protein [18, 19].

Knowledge of the three-dimensional structure of proteins is integral to understanding their functions. A wide range of computational methods are employed to estimate the properties of secondary, tertiary, and quaternary structures proteins. However, experimental methods to provide quantitative information at atomic resolution are limited to NMR spectroscopy and X-ray crystallography. Specifically, NMR spectroscopy, has a proven success at screening large number of proteins in the structural genomics pipeline [20]. However, both NMR and X-ray crystallography approaches are relatively more time and resource consuming procedures in comparison with computational methods. The demands of high throughput proteomics and structural genomics necessitate the development of new, faster experimental methods for providing structural information.

It was first observed in 1957 [21] that nuclear chemical shifts can be powerful indicators of biopolymer structural type. Over the years, chemical shifts have provided detailed information about the nature of hydrogen exchange dynamics, ionization and oxidation states, ring current influence of aromatic residues, and hydrogen bonding interactions [22]. Several review articles describe a wide variety of experimental and computational methods for correlating chemical shifts with protein three-dimensional structure [22–27].

We have demonstrated that the averaged chemical shift (ACS) of a protein backbone nucleus correlates well with both the secondary structure content (SSC) [28, 29] and structural class [30] of the protein. The correlation between the structures can enable the evaluation of the SSC of proteins can aid in the resonance assignment (unique identification of NMR spectral lines with particular nuclear spins within the protein). The ACS values used in these methods are average chemical shifts of all the amino acid residues in the protein. However, there is evidence in the literature that individual amino acids show intrinsic propensities towards certain secondary structure types [7, 31–34]. Furthermore, there is strong correlation between the number of ACS values to the total number of residue types in the given primary sequence to a maximum of 20 naturally occurring amino acids. Extensive research suggests five backbone ($^1\text{H}\alpha$, $^{13}\text{C}\alpha$, HN, ^{15}N and ^{13}C) and one side-chain ($^{13}\text{C}\beta$) nuclei are sensitive to changes in the protein structure [27]. Protein structure could be potentially sensitive to 119 ACS values (19 different amino acids and six nuclei and proline with 5 atoms). As the number of proteins with complete chemical shift assignments steadily increasing along with their respective three-dimensional structural information, correlating chemical shifts directly to protein structure is an excellent application for machine learning methods.

In this manuscript we evaluate the application of data mining tools to predict the protein structural class. We have used experimental chemical shift information from BMRB and theoretically estimated values derived from three-dimensional structures. These data mining tools are used to predict protein structural class which validates the general conclusions derived by total ACS. Additionally these tools can provide information regarding the contributions from specific amino acid residues as well as predicting the protein structural class.

II. Materials and Methods

Protein structural information

Structure files were obtained from the Research Collaboratory for Structural Biology (RCSB) (PDB format, <http://www.rcsb.org/pdb/>) [35]. Since most NMR-STAR files identify several corresponding PDB (protein data bank) **structures**, it was necessary to examine each entry and choose by inspection the most appropriate PDB ID number. When possible, the PDB ID corresponding to the “best” NMR structure was chosen, though in some cases it was necessary to choose the best X-ray structure (resolution < 2.5 Å). A total of 1491 proteins were found to be suitable, and downloaded from the Protein Data Bank. The secondary structure content (SSC), the total percentage of sheet or helix (α and 3_{10}), was determined using the program PROMOTIF (<http://www.biochem.ucl.ac.uk/~gail/promotif/promotif.html>) [36].

Protein Chemical Shifts

Protein chemical shift information (NMR-STAR files) obtained from two databases, BioMagResBank (BMRB) (www.bmrwisc.edu) [37] and RefDB (www.redpoll.pharmacy.ualberta.ca/RefDB) [38]. BMRB is the first public database to collect chemical shift information from a large number of proteins and RefDB [38] fixes the errors on the files submitted at BMRB (e.g., reference issues and unassigned or missing resonances). Only proteins with 50 or more amino acid residues, and with at least 70% of their residues assigned chemical shifts, were considered.

In addition to the experimental chemical shifts, the 3D structural information from the respective PDB files were used to estimate the chemical shifts of the proteins using the program Sparta+ [39].

Both the experimental and calculated chemical shifts were then reduced to per-residue averaged chemical shift. The averaged chemical shift (ACS) of a nucleus species “i” was calculated using:

$$AA_{ACS}^k(i) = \frac{1}{M_k} \sum_{m=1}^{M_k} CS(i, m) \quad [1]$$

Here $i = {}^{13}\text{CO}$, ${}^{13}\text{C}\alpha$, ${}^{13}\text{C}\beta$, ${}^1\text{HN}$, ${}^1\text{H}\alpha$ or ${}^{15}\text{N}$; M_k denotes the total number of residues of type ‘k’ (20 AA’s) with CS values assigned for nucleus species i . $CS(i, m)$ denotes the CS

value of the i^{th} nucleus at the m^{th} residue of type 'k'. If a protein contains all the 20 amino acids then there will be 119 AAACS values.

Protein structural Class

Each protein can be cataloged into one of the six structural classes using SCOP (<http://scop.mrc-lmb.cam.ac.uk/scop/>): (1) all alpha proteins (α), (2) all beta proteins (β), (3) mixed $\alpha\beta$ proteins ($\alpha\beta$), (4) coiled coil, (5) small proteins (s) and membrane/cell surface proteins and multi-domain proteins (m) [40–42]. The distribution of the dataset is shown in Table I.

Data mining

Figure 1 shows the overview of the data mining approach. Six different protein classes defined by the protein structural class represent the classification labels and each protein is represented by 119-dimensional attribute vector $((19 \times 6) + (1 \times 5) = 119)$ made of chemical shifts from six different nuclei from BMRB. The number of proteins in the coiled-coil and multi-domain are low compared to other classes (Table I), and were not considered for rest of the analysis. Weka version 3.5.7 (<http://www.cs.waikato.ac.nz/ml/weka/>), developed by the University of Waikato in New Zealand, is a software collecting a variety of state-of-art machine learning algorithms was employed [43, 44]. 24 different algorithms were performed and the results indicated with a prediction rate higher than 80% in three of the four classes. The list of algorithms are: Bayes Net, Logit Boost, Ridor, NBTree, Multi Class Classifier, Ordinal Class Classifier, SMO, Simple Logistic, END, Random Forest, JRip, Data Near Balanced ND, ND, Decorate, J48, J48 graft, Class Balanced ND, PART, Decision Table, Filtered Classifier, IB1, IBk, Random Sub Space and Kstar. In ten-fold cross-validation, the dataset is split into 10 equal size partitions at random. Each partition is used for testing in turn and the rest is used for training, i.e., each time one-tenth of the dataset is used for testing and the rest for training, and the procedure is repeated 10 times so that each data is used for training and testing exactly once. In this research, 10-fold cross-validation was used to evaluate the classifiers for the basic dataset.

Performance Measures

True positive (TP) provides the measure of number of positive events positive for a virus infection and true negative (TN) provides the number of negative occurrences predicted correctly under a given classification scheme. False positive (FP) gives an estimate of negative events that are incorrectly predicted to be positive, while the false negative (FN) estimated the number of mice that were predicted negative but were positive [45].

For multi-class classification schemes and the sum over rows (i) or columns (j) of the confusion matrix (M) should be considered. For a confusion matrix of dimension $k \times k$, the TP, TN, FP and FN for the measure (class) 'n' could be defined as follows:

$$TP = M_{ii}|_{i=n}; \quad TN = \sum_{i=1}^k M_{ii}|_{i \neq n}; \quad FP = \sum_{i=1}^k M_{ij}|_{i \neq n}; \quad FN = \sum_{j=1}^k M_{ij}|_{j \neq n} \quad [2]$$

These terms were combined to determine the performance of our testing *via* quantifiable categories such as sensitivity (SN), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), test efficiency/accuracy (TE) and Matthew correlation coefficient (MCC). These quantifiers are defined as follows:

Sensitivity (SN) gives an estimate of the percentage of actual positives identified, while specificity (SP) gives an estimate of the percentage of negatives identified.

$$SN(\%) = \frac{TP \times 100}{TP + FN} \quad [3]$$

$$SP(\%) = \frac{TN \times 100}{FP + TN} \quad [4]$$

The effectiveness of a test is evaluated based on two measures namely, positive predictive value (PPV) and negative predictive value (NPV). PPV gives an estimate of the percentage of positive samples that were correctly predicted and NPV gives the percentage of negative samples that were correctly predicted [46, 47].

$$PPV(\%) = \frac{TP \times 100}{TP + FP} \quad [5]$$

$$NPV(\%) = \frac{TN \times 100}{FN + TN} \quad [6]$$

The prediction power of a model can be evaluated either by test efficiency (TE) or Mathew correlation coefficient (MCC) [48]. Test efficiency is also referred as *test accuracy*. The MCC is in essence a correlation coefficient between the observed and predicted classifications; it returns a value between -100% and +100%. A coefficient of +100% represents a perfect prediction, 0% no better than random prediction and -100% indicates total disagreement between prediction and observation [49]. TE and MCC are defined as follows.

$$TE(\%) = \frac{(TP + TN) \times 100}{TP + TN + FP + FN} \quad [7]$$

$$MCC(\%) = \frac{(TP \times TN - FP \times FN) \times 100}{\sqrt{(TP + FP)(TP + FN)(TN + FN)(TN + FP)}} \quad [8]$$

The data analyses were performed using scripts written in *awk* and *perl* on the Linux workstation. These scripts, as well as a complete list of proteins studied, their BMRB accession numbers, PDB codes, and secondary structure contents, are available from the authors.

III. Results

Figure 2 shows the hierarchical clustering of the residue based averaged chemical shifts between α -helices and β -strands for ^{13}Ca (Fig. 2a) and ^1Ha (Fig. 2b). Hierarchical cluster analysis can be used to investigate how the individual chemical shifts of the different nuclei within a particular amino acids contribute to specific structural class of the protein. In this analysis all the proteins were classified using the corresponding amino acid chemical shift values into clusters. Protein secondary structural classes are indicated by dendrograms at the left, and their constituting amino acid chemical shifts (for each nuclei) clusters are indicated by color-coded dendrograms on the left side of the heat map. The extent of variation in the chemical shift values for each protein, is depicted by the intensity scale. The signal level in arbitrary units ranged from “-3 to 3,” with green as the minimum and red as the maximum signal intensities. The relative scaling of the ^{13}Ca chemical shifts among the 20 residues indicate the ACS values can be divided into two major groups. Nine amino acid residues Cys, Trp, Arg, Gln, Pro, Phe, His, Met and Tyr fall in one group while the rest in the second group based on the ^{13}Ca ACS values, as shown by the dendrograms on the top (Fig. 2). The residue based on ^1Ha ACS values were grouped with the remaining ten amino acids (Met, Cys, Trp, Phe, Tyr, His, Asn, Asp, Pro and Ser) in the first group. This combination of chemical shifts from particular groupings of the amino acids suggest specific combinations towards forming the structural classes.

The residue's specific average chemical shifts between the α (black symbols) and β (red symbols) class of the proteins, the ^1Ha and ^{13}Ca pairs for all the 20 amino acids are plotted in Figure 3. Figure 3 indicates both the experimental and calculated chemical shifts with the overall distributions similar between them. Long side-chain residues (Ile, Leu and Val) differentiate between the helical and strand shifts much more clearly as suggested by the hierarchical clustering (Fig. 2). The distribution of ACS values is narrow for most of the residues except for Cys in both the experimental and calculated values.

Figure 4 shows the experimental (Fig. 4a) and calculated (Fig. 4b) dispersions of ^1HN and ^{15}N chemical shift dispersions. Residue based ACS values of ^1HN and ^{15}N show a broader distribution than the corresponding ^{13}Ca - ^1Ha pairs. Chemical shift dispersions of other nuclei and in particular ^{13}C (carbonyl) spins, show a similar distinction between the helical and strand conformational classes (data not shown). Most of the major features, observed in the correlation between ACS values of all residues to protein structural class were observed [30]. Notably increasing the β -sheet characteristics shifts both the ^{13}Ca resonance upfield while that of the ^1Ha resonances downfield with an opposite effect between the ^{15}N and ^1HN pair (downfield shift of ^{15}N and upfield shift of ^1HN) (Figures 3–4).

Upon including all the structural classes (classification labels) and their qualitative attributes (all 119), these factors contributed to the efficiency of the proteins classification scheme. To determine the best machine learning algorithm for the predictive task, 24 classifiers from WEKA that gave positive results were used using their default parameters. The performance of the algorithms was determined by sensitivity (SN), specificity (SP), positive and negative predictive values (PPV and NPV), test efficiency (TE) and Matthews's correlation

coefficients (MCC) in a 10-fold cross-validation analysis using the experimental or calculated residue based chemical shifts. Of the various classification algorithms, Table II (experimental chemical shifts) and Table III (calculated chemical shifts) present the top performers that classifies all the 4 classes of proteins with at least 80% efficiency. Four algorithms (Bayes Net, Logit Boost, SMO and Random Forest) performed the best classification for the experimental data set, while three algorithms (Bayes Net, Logit Boost and Attribute Selected Classifier) produced good overall results ($> 80\%$ MCC) for the calculated chemical shifts. The performance measures of the residue based averaged chemical shifts for the calculated set (Table III) is slightly better than the experimental shifts.

Protein classes defined by either all β , $\alpha\beta/\alpha+\beta$ or small proteins were well classified, when experimental or calculated chemical shifts were used. In the case of experimental chemical shifts, Logit Boost algorithm is able to classify all the four classes with more than 80% efficiency according to MCC (Table II), with 80.9% for all α , 80.1% for all β , 89.6% for $\alpha\beta/\alpha+\beta$, and 94.1% for small proteins. For the calculated chemical shifts, Bayes Net produced the best results with 92.8% for all α , 84.8% for all β , 94.4% for $\alpha\beta/\alpha+\beta$, and 94.6% for small proteins. Residue based ACS values provided a higher correlation (based on MCC) in comparison to linear correlations derived with respect to complete average of chemical shifts. For example, the coefficients of correlation between ACS and sheet content is 0.84 for H^α , and 0.71 for H^N [29].

IV. Discussion and Conclusion

In an effort to explore new methods for the efficient identification of protein structures using NMR, we have investigated the degree to which residue based ACS can be used as a low-resolution structural parameter. The criteria defined to generate the empirical correlations (number of residues > 50 , at least 70 % complete chemical shift assignments and 3D structural resolution $< 2.5 \text{ \AA}$), establish a consistent and improved relationship between four different protein structural classes and residue based ACS values drawn from two separate databanks. Residue based ACS values increased the number of dimensions (attributes) by 20 fold leading to a better discrimination of the classification categories. The results obtained using either experimental chemical shifts directly from the deposited data, or chemical shifts calculated from the 3D structures provides comparable results, with the calculated shifts performing slightly better than experimental values (Table II and III). Figure 5 shows the correlation plot between the calculated and experimental chemical shifts for the nuclei $^1H^\alpha$, $^{13}C^\alpha$ and ^{13}CO for the all- α and all- β classes. This suggests that the residues based on ACS values are inherently sensitive to the protein structural class.

Determination of protein structural class, particularly in the absence of chemical shift assignments and primary sequence information could be valuable in a structural proteomics pipeline. One such notable approach was CSSI-PRO presented by Swain and Atreya [34]. Combination of Shifts for secondary structure Identification in Proteins (CSSI-PRO) is based on the detection of specific linear combination of backbone $^1H^\alpha$ and $^{13}C'$ chemical shifts in a two-dimensional (2D) NMR experiment. Linear combinations of shifts facilitated editing of residues belonging to α -helical/ β -strand regions into distinct spectral regions

nearly independent of the amino acid type, thereby allowing the estimation of overall secondary structure content of the protein. In this method a comparison of the predicted vs. experimental secondary structural content for 237 proteins provided a correlation of more than 90% and an overall rmsd of 7.0%. The hierarchical clustering analyses (Fig 2) reflect a similar principle; the profile (combination) of chemical shifts from each residue type contributes differently for the different protein structural classes.

Protein secondary structure prediction algorithms either estimate the backbone dihedral angles (ϕ and ψ) or actual secondary structure definition. The methods such as TALOS (TALOS+) [50, 51], SHIFTOR [52], PREDITOR (+) [53], DANGLE [54] or PROMEGA [55] estimate the dihedral angles while CSI (CSI 2.0) [56], PSSI [57], PsiCSI [58], PLATON [59], PECAN [60] or 2DCSI [61] estimate the secondary structure definition. The program that is relevant for chemical shift prediction includes SHIFTX [62], SPARTA (+) [39], CAMSHIFT [63], SHIFTS [64] or PROSHIFT [65]. In an authoritative review article, Wishart has compared the various prediction approaches (Table 11 [27]). SPARTA+ [39] is one of the top performing program is used here to estimate the chemical shifts from the three dimensional structure of the protein. SPARTA+ Uses an artificial neural network approach and includes a more complete consideration of various structural/dynamic parameters in proteins and able to predict chemical shifts for backbone and $^{13}\text{C}^\beta$ atoms with modestly improved accuracy, compared with other similar chemical shift prediction approaches listed above. SPARTA+ predicted chemical shifts includes structural/dynamic factors, i.e., χ_2 torsion angle, H-bonding and electric fields, as well as an averaging procedure over the outputs from three separated neural networks.

Computational methods often play a primary role in initial predictions of protein structure; specifically in regards to the protein structural class. These methods are typically invoked even before a protein is expressed or extracted for any biophysical characterization. Our results show that residue based ACS values clearly distinguish the four different protein classes, α , β , mixed $\alpha\beta$ and small proteins are classified either by SCOP. Multiple algorithms classify the protein structural classes with fairly good efficiency as described by the MCC. The quality of the protein structural class is affected by several known factors that include the size of the protein database, quality of the chemical shift data and classification algorithms. Most of the algorithms presented here show indicate moderate to good performance measured in terms of Mathew correlation coefficient (MCC 80–95%).

Prior to collecting several days' worth of NMR spectra for structure determination, other biophysical methods are generally adopted to infer secondary structural information about the protein of interest. In particular, circular dichroism (CD) spectroscopy is extensively used to estimate the secondary structure content of medium-sized proteins. However, CD spectrum does not provide information on protein structural class. Furthermore, NMR spectral information has seldom been used to obtain relatively low-resolution structural information, such as protein structural class. In some cases, the results of CD are used to determine whether it is feasible to obtain complete, three-dimensional structural information for a particular protein, using NMR. This suggests the critical importance of evaluating whether data obtained from NMR itself can be used to estimate secondary structure content. Lee and Cao have addressed this question extensively in their comprehensive study [66],

and have shown that the correlation between NMR- and CD-based secondary structure estimation is poor. Further, while CD spectroscopy is more suitable for studying relatively small proteins and polypeptides, the characterization of larger molecules requires NMR. Thus making NMR based low resolution approaches complementary to the CD experiments.

It must be emphasized that ACS-based methods do not provide an alternative to conventional NMR-based experiments, and should only be considered initial predictors of protein class or secondary structure content. ACS methods might provide a novel technique for monitoring protein structural changes in real time, such as in protein folding experiments. Such methods might also be used to detect major structural changes that occur upon protein-protein, protein-DNA/RNA, and other complex formations, to provide some direct experimental structural information in situations in which other techniques are incapable of doing so (e.g., in studies of large and/or highly disordered proteins), and to facilitate initial protein fold identification in high throughput proteomics applications.

Acknowledgments

The authors acknowledge A. Mani for critical reading. This research was in part supported by NIH grants P20 MD 002732 and P20 CA 138025.

References

1. Pauling L, Corey RB. The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A*. 1951; 37:251–256. [PubMed: 14834147]
2. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A*. 1951; 37:205–211. [PubMed: 14816373]
3. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*. 1991; 253:164–170. [PubMed: 1853201]
4. Chakrabarti P, Pal D. The interrelationships of side-chain and main-chain conformations in proteins. *Prog Biophys Mol Biol*. 2001; 76:1–102. [PubMed: 11389934]
5. Chen CC, Singh JP, Altman RB. Using imperfect secondary structure predictions to improve molecular structure computations. *Bioinformatics*. 1999; 15:53–65. [PubMed: 10068692]
6. Chou PY, Fasman GD. Prediction of protein conformation. *Biochemistry*. 1974; 13:222–245. [PubMed: 4358940]
7. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*. 1974; 13:211–222. [PubMed: 4358939]
8. Eyrich VA, Standley DM, Friesner RA. Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. *J Mol Biol*. 1999; 288:725–742. [PubMed: 10329175]
9. Eyrich VA, Standley DM, Felts AK, Friesner RA. Protein tertiary structure prediction using a branch and bound algorithm. *Proteins*. 1999; 35:41–57. [PubMed: 10090285]
10. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. *Protein Sci*. 1996; 5:947–955. [PubMed: 8732766]
11. Fischer D, Rice D, Bowie JU, Eisenberg D. Assigning amino acid sequences to 3-dimensional protein folds. *FASEB J*. 1996; 10:126–136. [PubMed: 8566533]
12. Lomize AL, Pogozheva ID, Mosberg HI. Prediction of protein structure: the problem of fold multiplicity. *Proteins, Suppl*. 1999; 3:199–203.
13. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. *Curr Opin Struct Biol*. 1996; 6:377–385. [PubMed: 8804824]

14. 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; 247:536–540. [PubMed: 7723011]
15. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. *Structure.* 1997; 5:1093–1108. [PubMed: 9309224]
16. Liao B, Peng T, Chen H, Lin Y. Incorporating secondary structural features into sequence information for predicting protein structural class. *Protein and peptide letters.* 2013
17. Ordog R. PyDeT, a PyMOL plug-in for visualizing geometric concepts around proteins. *Bioinformatics.* 2008; 2:346–347. [PubMed: 18685723]
18. Gong H, Rose GD. Does secondary structure determine tertiary structure in proteins? *Proteins.* 2005; 61:338–343. [PubMed: 16104021]
19. Fitzkee NC, Fleming PJ, Gong H, Panasik N Jr, Street TO, Rose GD. Are proteins made from a limited parts list? *Trends Biochem Sci.* 2005; 30:73–80. [PubMed: 15691652]
20. Page R, Peti W, Wilson IA, Stevens RC, Wuthrich K. NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline. *Proc Natl Acad Sci U S A.* 2005; 102:1901–1905. [PubMed: 15677718]
21. Gutowsky HS, Saika A, Takeda M, Woessner DE. Proton magnetic resonance studies on natural rubber. II. Line shape and T_1 measurements. *J Chem Phys.* 1957; 27:534–542.
22. Szilagyi L. Chemical Shifts in Proteins Come of Age. *Prog Nucl Magn Reson Spectros.* 1995; 27:325–443.
23. Case DA. Interpretation of chemical shifts and coupling constants in macromolecules. *Curr Opin Struct Biol.* 2000; 10:197–203. [PubMed: 10753812]
24. Ando I, Kuroki S, Kurosu H, Yamanobe T. NMR chemical shift calculations and structural characterizations of polymers. *Prog Nucl Magn Reson Spectros.* 2001; 39:79–133.
25. Wishart DS, Case DA. Use of chemical shifts in macromolecular structure determination. *Methods Enzymol.* 2001; 338:3–34. [PubMed: 11460554]
26. Mielke SP, Krishnan VV. Characterization of protein secondary structure from NMR chemical shifts. *Prog Nucl Magn Reson Spectrosc.* 2009; 54:141–165. [PubMed: 20160946]
27. Wishart DS. Interpreting protein chemical shift data. *Prog Nucl Magn Reson Spectrosc.* 2011; 58:62–87. [PubMed: 21241884]
28. Sibley AB, Cosman M, Krishnan VV. An empirical correlation between secondary structure content and averaged chemical shifts in proteins. *Biophys J.* 2003; 84:1223–1227. [PubMed: 12547802]
29. Mielke SP, Krishnan VV. Estimation of protein secondary structure content directly from NMR spectra using an improved empirical correlation with averaged chemical shift. *J Struct Funct Genomics.* 2005; 6:281–285. [PubMed: 16283427]
30. Mielke SP, Krishnan VV. Protein structural class identification directly from NMR spectra using averaged chemical shifts. *Bioinformatics.* 2003; 19:2054–2064. [PubMed: 14594710]
31. Levitt M. Conformational preferences of amino acids in globular proteins. *Biochemistry.* 1978; 17:4277–4285. [PubMed: 708713]
32. Minor DL Jr, Kim PS. Context is a major determinant of beta-sheet propensity. *Nature.* 1994; 371:264–267. [PubMed: 8078589]
33. Minor DL Jr, Kim PS. Measurement of the beta-sheet-forming propensities of amino acids. *Nature.* 1994; 367:660–663. [PubMed: 8107853]
34. Swain M, Atreya HS. CSSI-PRO: a method for secondary structure type editing, assignment and estimation in proteins using linear combination of backbone chemical shifts. *J Biomol NMR.* 2009; 44:185–194. [PubMed: 19529884]
35. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucl Acid Res.* 2000; 28:235–242.
36. Hutchinson EG, Thornton JM. Promotif - a Program to Identify and Analyze Structural Motifs in Proteins. *Protein Sci.* 1996; 5:212–220. [PubMed: 8745398]
37. Seavey BR, Farr EA, Westler WM, Markley JL. A relational database for sequence-specific protein NMR data. *J Biomol NMR.* 1991; 1:217–236. [PubMed: 1841696]

38. Zhang HY, Neal S, Wishart DS. RefDB: A database of uniformly referenced protein chemical shifts. *J Biomol NMR*. 2003; 25:173–195. [PubMed: 12652131]
39. Shen Y, Bax A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J Biomol NMR*. 2010; 48:13–22. [PubMed: 20628786]
40. Gewehr JE, Hintermair V, Zimmer R. AutoSCOP: automated prediction of SCOP classifications using unique pattern-class mappings. *Bioinformatics*. 2007; 23:1203–1210. [PubMed: 17379694]
41. Dubchak I, Muchnik I, Mayor C, Dralyuk I, Kim SH. Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. *Proteins*. 1999; 35:401–407. [PubMed: 10382667]
42. Dubchak I, Muchnik I, Kim SH. Protein folding class predictor for SCOP: approach based on global descriptors. *Proc Int Conf Intell Syst Mol Biol*. 1997; 5:104–107. [PubMed: 9322023]
43. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. *Bioinformatics*. 2004; 20:2479–2481. [PubMed: 15073010]
44. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*. 2009; 11:10–18.
45. Carugo O. Detailed estimation of bioinformatics prediction reliability through the Fragmented Prediction Performance Plots. *BMC bioinformatics*. 2007; 8:380. [PubMed: 17931407]
46. Gunnarsson RK, Lanke J. The predictive value of microbiologic diagnostic tests if asymptomatic carriers are present. *Statistics in medicine*. 2002; 21:1773–1785. [PubMed: 12111911]
47. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*. 1994; 308:1552. [PubMed: 8019315]
48. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*. 1975; 405:442–451. [PubMed: 1180967]
49. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res*. 2008; 36:D419–425. [PubMed: 18000004]
50. Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR*. 1999; 13:289–302. [PubMed: 10212987]
51. Shen Y, Delaglio F, Cornilescu G, Bax A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol NMR*. 2009; 44:213–223. [PubMed: 19548092]
52. Neal S, Berjanskii M, Zhang H, Wishart DS. Accurate prediction of protein torsion angles using chemical shifts and sequence homology. *Magn Reson Chem*. 2006; 44(Spec No):S158–167. [PubMed: 16823900]
53. Berjanskii MV, Neal S, Wishart DS. PREDITOR: a web server for predicting protein torsion angle restraints. *Nucleic Acids Res*. 2006; 34:W63–69. [PubMed: 16845087]
54. Cheung MS, Maguire ML, Stevens TJ, Broadhurst RW. DANGLE: A Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure. *J Magn Reson*. 2010; 202:223–233. [PubMed: 20015671]
55. Shen Y, Bax A. Prediction of Xaa-Pro peptide bond conformation from sequence and chemical shifts. *J Biomol NMR*. 2010; 46:199–204. [PubMed: 20041279]
56. Hafsa NE, Wishart DS. CSI 2.0: a significantly improved version of the Chemical Shift Index. *J Biomol NMR*. 2014; 60:131–146. [PubMed: 25273503]
57. Wang YJ, Jardetzky O. Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Sci*. 2002; 11:852–861. [PubMed: 11910028]
58. Hung LH, Samudrala R. Accurate and automated classification of protein secondary structure with PsiCSI. *Protein Sci*. 2003; 12:288–295. [PubMed: 12538892]
59. Labudde D, Leitner D, Kruger M, Oschkinat H. Prediction algorithm for amino acid types with their secondary structure in proteins (PLATON) using chemical shifts. *J Biomol NMR*. 2003; 25:41–53. [PubMed: 12566998]
60. Eghbalian HR, Wang L, Bahrami A, Assadi A, Markley JL. Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements. *J Biomol NMR*. 2005; 32:71–81. [PubMed: 16041485]

61. Wang CC, Chen JH, Lai WC, Chuang WJ. 2DCSi: identification of protein secondary structure and redox state using 2D cluster analysis of NMR chemical shifts. *J Biomol NMR*. 2007; 38:57–63. [PubMed: 17333485]
62. Neal S, Nip AM, Zhang H, Wishart DS. Rapid and accurate calculation of protein ¹H, ¹³C and ¹⁵N chemical shifts. *J Biomol NMR*. 2003; 26:215–240. [PubMed: 12766419]
63. Kohlhoff KJ, Robustelli P, Cavalli A, Salvatella X, Vendruscolo M. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J Am Chem Soc*. 2009; 131:13894–13895. [PubMed: 19739624]
64. Xu XP, Case DA. Automated prediction of ¹⁵N, ¹³C_{alpha}, ¹³C_{beta} and ¹³C' chemical shifts in proteins using a density functional database. *J Biomol NMR*. 2001; 21:321–333. [PubMed: 11824752]
65. Meiler J. PROSHIFT: protein chemical shift prediction using artificial neural networks. *J Biomol NMR*. 2003; 26:25–37. [PubMed: 12766400]
66. Lee MS, Cao B. Nuclear magnetic resonance chemical shift: comparison of estimated secondary structures in peptides by nuclear magnetic resonance and circular dichroism. *Protein Eng*. 1996; 9:15–25. [PubMed: 9053898]

Highlights

- Data mining methods to protein chemical shifts to structural class identification
- NMR without chemical shift assignments for protein structural class prediction.
- Averaged chemical shift (ACS) based approach provides 80%–90% accuracy via MCC.

Figure 1.

Overview of the data mining approach to predict protein structural class from the residue based averaged chemical shift values (attributes). DM stands of data mining.

Figure 2.

Hierarchical clustering of $^{13}\text{C}\alpha$ and $^1\text{H}\alpha$ residue based ACS values. Natural grouping of the amino acids residues (top) with respect to the experimentally determined residue based ACS values (along the side). Euclidian distance metric was used for clustering. The clustering processes generated natural groupings proteins (secondary structure content) in the form of dendrograms (left color coded) and the respective contributions of the amino acids that constitute the proteins (top dendrograms color coded) with the total changes presented as heat maps. Intensity scales are shown in arbitrary units shown by the scale at the bottom of the panel ranging from green to red in a relative scale. Names of amino acid residues above each heat map represent various combinations and those on the right (color coded) represent either predominantly α helical or β -sheet. The amino acids are grouped into two major clusters based on the profiles. All α and β proteins groups separately as marked for both the $^{13}\text{C}\alpha$ and $^1\text{H}\alpha$ nuclei.

Figure 3.

Residue based ACS distribution of the heteronuclear pair $^{13}\text{C}\alpha$ and $^1\text{H}\alpha$ for each residue type. Experimental (left) and Calculated ACS (right) values for each residue noted by three letter amino acid code on each frame. Chemical shift scaling for all the frames is same except for the Gly residues along the $^{13}\text{C}\alpha$ axis.

Figure 4.

Residue based ACS distribution of the heteronuclear pair ^{15}N and $^1\text{H}\text{N}$ for each residue type. Experimental (left) and Calculated ACS (right) values for each residue noted by three letter amino acid code on each frame. The α and β class proteins are differentiated by black and red symbols, respectively.

Figure 5.

Comparison of the experimental (along the Y-axis) and calculated (along X-axis) residue based chemical shifts for helical (top row) and strand conformations (bottom row). Left, middle and the right row represent correspond to $^1\text{H}\alpha$, $^{13}\text{C}\alpha$ and ^{13}C (carbonyl) nuclei. Each residue is identified by different symbol as noted on the right side of the plot.

Table I

Number of proteins in each structural class.

Protein Structural Class	Abbreviation	Number
All alpha proteins	α	267
All beta proteins	β	317
Alpha and beta proteins (a/b)	$\alpha\beta$	527
Small proteins	s	289
Coiled coil	c	31
Multi-domain proteins	o	60
Total		1491

Table II

Performance of protein structural class prediction using experimental chemical shifts¹

Algorithm	All agr;							All β						
	SN	SP	PVP	NPV	TE	MCC		SN	SP	PVP	NPV	TE	MCC	
Bayes Net	58.8%	69.7%	69.7%	61.7%	63.0%	86.8%		93.5%	89.6%	88.2%	81.3%	76.5%	78.2%	
Logit Boost	76.7%	80.1%	68.5%	25.0%	78.5%	80.9%		94.2%	91.6%	98.2%	92.1%	71.8%	80.1%	
SMO	76.1%	78.7%	71.3%	1.7%	48.4%	68.4%		93.9%	91.5%	99.8%	90.0%	60.0%	80.8%	
Random Forest	86.0%	74.5%	60.9%	0.0%	52.9%	64.9%		95.5%	94.7%	100.0%	91.0%	60.6%	84.0%	
Algorithm	Mixed α/β or $\alpha+\beta$							Small Proteins						
	SN	SP	PVP	NPV	TE	MCC		SN	SP	PVP	NPV	TE	MCC	
Bayes Net	72.7%	23.6%	51.1%	74.3%	90.3%	88.2%		97.5%	87.6%	75.0%	87.6%	83.9%	86.7%	
Logit Boost	73.3%	44.1%	75.7%	84.5%	94.2%	89.6%		95.9%	93.2%	79.2%	91.0%	85.7%	94.4%	
SMO	76.4%	33.3%	60.1%	82.1%	93.1%	89.2%		94.3%	84.9%	71.3%	90.1%	85.9%	94.1%	
Random Forest	81.1%	NA	64.6%	88.0%	92.2%	86.7%		94.3%	86.1%	73.0%	90.4%	85.5%	94.3%	

¹ SN (sensitivity), SP (Specificity), PVP (Positive predictive value), NPV (Negative predictive value), TE (Test efficiency) and MCC (Mathew Correlation Coefficient).

Table III

Performance of protein structural class prediction using calculated chemical shifts

Algorithm	All α						All β					
	SN	SP	PVP	PVN	TE	MCC	SN	SP	PVP	PVN	TE	MCC
Bayes Net	74.9%	89.3%	83.2%	50.9%	78.1%	92.8%	95.5%	91.7%	96.8%	91.3%	87.0%	84.8%
Logit Boost	81.0%	83.6%	70.2%	36.8%	69.1%	82.9%	95.1%	89.3%	99.2%	92.9%	75.6%	83.3%
Attribute Selected Classifier	77.2%	76.0%	71.2%	24.6%	58.7%	81.6%	93.1%	89.7%	97.2%	87.5%	74.6%	77.1%
Algorithm	Mixed α/β or $\alpha+\beta$						Small Proteins					
	SN	SP	PVP	PVN	TE	MCC	SN	SP	PVP	PVN	TE	MCC
Bayes Net	76.5%	44.6%	72.3%	85.1%	97.0%	94.4%	97.5%	93.5%	85.8%	94.2%	89.6%	94.6%
Logit Boost	67.8%	70.0%	74.8%	87.0%	95.2%	90.3%	96.7%	90.8%	82.2%	92.5%	84.6%	96.0%
Attribute Selected Classifier	71.0%	33.3%	58.7%	83.6%	92.7%	89.8%	95.8%	87.5%	79.8%	89.1%	84.8%	93.4%

¹ SN (sensitivity), SP (Specificity), PVP (Positive predictive value), NPV (Negative predictive value), TE (Test efficiency) and MCC (Mathew Correlation Coefficient).