Protein Structure Characterization From NMR Chemical Shifts

by

Noor E Hafsa

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science
University of Alberta

Abstract

In order to understand the complex biological functions of proteins, highly detailed, atomic resolution protein structures are needed. Experimental methods such as X-ray crystallography and NMR spectroscopy provide standard platforms for determining the atomic-resolution structures of proteins. However, a continuing bottleneck in conventional NOE-based NMR structure determination lies in the difficulty of measuring NOEs for medium-to-large proteins and the resulting time-costs and the corresponding reduction in structure accuracy and precision. This has led to an increased interest in using other easily identifiable NMR parameters, such as chemical shifts, to facilitate protein structure determination by NMR.

Chemical shifts, often considered as mileposts of NMR, have long been used to decipher the structures of small molecules. However, chemical shifts are much less frequently utilized for structural interpretation of larger macromolecules such as peptides and proteins. Most existing macromolecular methods use chemical shifts and various heuristic, rule-based algorithms to identify and determine a small number of structural parameters (such as secondary structure). Other methods, such as CS-Rosetta and CS23D, which attempt to determine 3D structures from chemical shifts alone, are only modestly successful (~50% success). So while good progress has been made, I believe that there is still substantial room for improvement and that the “Shift-to-Structure” problem has not yet been fully solved.

My PhD project involves investigating innovative computational and machine- learning approaches to develop chemical-shift based prediction models to determine protein structures with high efficiency and high accuracy (>90%). More specifically, my thesis consists of three major components: a) shift-based local protein structure prediction; b) prediction of protein local/non-local interactions from sequence and chemical shifts; and c) tertiary fold recognition from chemical shifts. Towards that goal, I have developed several chemical-shift based prediction models that exploit advanced computational and machine-learning algorithms. In particular, I developed a) CSI 2.0 - a multi-class prediction method for protein local structure prediction from chemical shift data; b) CSI 3.0 – a computational model that identifies detailed local structure and structural motifs in proteins using chemical shift data; c) ShiftASA – a boosted tree regression model for predicting accessible surface area from chemical shifts; and d) E-Thrifty - a protein fold recognition method that performs chemical shift threading to identify and generate the most probable fold or 3D structure that a query protein may have. Validation of these proposed methods was performed using several independent test sets and the results indicate substantial improvements over other state-of-the-art methods. Given their superior performance, I believe that these methods will be useful contributions to the field of NMR-based protein structure determination and will be fundamental to the development 3D structure determination protocols that use only chemical shift data.

Preface

The introductory discussion in Chapter 1 as well as the concluding analysis in Chapter 6 are my original work. Chapter 5, which is also my original work, is being prepared as a paper for submission to the Journal of Biomolecular NMR.

Chapter 2 of this thesis has been published as: Hafsa, N. E., & Wishart, D. S. (2015). "CSI 2.0: a significantly improved version of the Chemical Shift Index". *Journal of Biomolecular NMR*, *60*(2-3*),* 131-146. I was responsible for the algorithm creation, program development, experimental assessment and resulting analysis as well as the manuscript composition. Dr. Wishart was the supervisory author and was involved with the concept formation, testing and manuscript composition/editing.

Chapter 3 of this thesis has been published as: Hafsa, N. E., Arndt, D., & Wishart, D. S. (2015). "CSI 3.0: a web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts". *Nucleic Acids Research*, *43,* W370-377. I was responsible for the algorithm creation, program development, experimental assessment and resulting analysis as well as the manuscript composition. David Arndt assisted with the web server and program development. Dr. Wishart was the supervisory author and assisted with the concept formation, algorithm creation, testing and manuscript composition/editing.

Chapter 4 of this thesis has been published as: Hafsa, N. E., Arndt, D., & Wishart, D. S. (2015). "Accessible surface area from NMR chemical shifts". *Journal of Biomolecular NMR*, *62*(3*),* 387-401. I was responsible for the algorithm creation, program development, experimental assessment and resulting analysis as well as the manuscript composition. David Arndt assisted with the web server and program development. Dr. Wishart was the supervisory author and was involved with concept formation, testing and manuscript composition/editing.

Dedication

To my beloved parents, Helena Ahmed and Bayezid Ahmed

Acknowledgements

I would like to begin by acknowledging my extreme gratitude to Allah, the Lord, the Almighty, the Maintainer and the Sustainer of this universe. It was only through His unlimited mercy and bounties, that it was possible for me to complete this thesis. In every stressful and difficult situation during my PhD studies, I was able to turn towards Him and seek His help.

I would like to convey my special gratitude to my parents, especially to my beloved mother Helena Ahmed, whose inspiration and motivation were the driving forces for me to complete this enormous task. She is my role model. She taught me how to think positively in every situation and to work relentlessly towards solving every hard problem I encountered. My father, Bayezid Ahmed was my inspiration in patiently dealing with many difficult situations during my graduate studies. I would also like to thank my two sisters, Kashfia Bilkis and Jereen Perveen for their consistent encouragement throughout my PhD studies. My husband, Sayeed Rushd, also deserves my special thanks for his kind support and patience.

I would like to express my grateful acknowledgement to my supervisor, Dr. David Wishart. His superb guidance and experience greatly facilitated my PhD research. Initially he helped me improve my approach to solving research problems, and then with his expertise in protein structure, he helped me to choose the right projects. Later he taught me how to analyze and address my research problems in a more effective and efficient way. He critically analyzed my approaches, which helped me to improve my algorithms and, most importantly, my research outcomes. I am also thankful to him for teaching me how to write scientifically. Overall, thanks to his supervision, I was able to finish my PhD project and my PhD dissertation successfully. I also thank my two other supervisory committee members, Dr. Russ Greiner and Dr. Guo-hui Lin for providing their valuable suggestions and comments during the annual progress meetings and my candidacy examination.

I would also like to thank my colleagues, Dr. Mark Berjanskii, David Arndt and Jack Liang for their kind support and help. I would especially like to acknowledge Dr. Berjanskii, who helped to improve my research outcomes by generously sharing his valuable experience, ideas and suggestions. David Arndt and Jack Liang also helped me to solve a number of technical problems related to my research projects.

Last but not the least, I would like to gratefully acknowledge the financial support from various granting agencies which have helped to fund me during my PhD training. Most specifically, I deeply appreciate the financial support from the Natural Sciences and Engineering Research Council (NSERC), the Alberta Prion Research Institute (APRI), PrioNet and Genome Alberta. I also wish to thank the department of Computing Science for providing me with additional funding through various teaching assistantships during my PhD studies.

Table of Contents

Chapter 1 Introduction ……...……………………………………………. 1

1.1 **Proteins** ...……………………………………………………… 1

1.2 **Protein Structure** ……………………………………………... 3

1.2.1 Secondary Structure ……………………………………………. 5

1.2.2 Super-secondary Structure ……………………………………... 8

1.2.3 Accessible Surface Area ……………………………………...... 8

1.2.4 Protein Folding ……………………………………................... 9

1.3 **The Significance of Protein Structure** ……………................... 10

1.4 **Protein Structure Determination** ……………………………........ 11

1.5 **NMR Chemical Shifts** ………………………......................... 15

1.6 **Chemical Shifts to Protein Structure** ……………………......... 15

1.6.1 Secondary Structure and Structural Motifs from Chemical Shifts .. 18

1.6.2 Accessible Surface Area from Chemical Shifts ………………….. 19

1.6.3 Protein Structure Determination by Chemical Shifts …………….. 20

1.7 **Thesis Objectives** ...……………………..................................... 21

1.8 **Organization of Thesis** ……………………............................. 23

Chapter 2 CSI 2.0 – A Significantly Improved Version of the Chemical Shift Index ………………………………………………………………… 24

2.1 **Introduction** ……………………………………………………… 25

2.2 **Methods and Materials** ………….. 27

2.2.1 Data set Preparation ……………………………………………. 27

2.2.2 Feature Set ...……………………………………………………… 29

2.2.3 Feature Normalization …………... 34

2.2.4 Multi-class SVM Training …………………………….………. 35

2.2.5 A Multi-residue Markov Model for Post-assignment Filtering ……….. 36

2.2.6 Evaluation Metrics ……………………………………………………… 37

2.3 **Results and Discussion** ……………………………………………... 39

2.4 **Conclusion** ………………………………………………………………….. 55

Chapter 3 CSI 3.0: A Web Server for Identifying Secondary and Super-secondary Structure in Proteins using NMR Chemical Shifts ……. 57

3.1 **Introduction** ……………………………………………………………… 58

3.2 **Algorithm and Workflow** …………………………………………….. 61

3.3 **Results and Validation** …………………………………………………... 68

3.4 **Web Server Implementation** ………………………………………. 70

3.5 **Conclusion** ……………………………………….............................. 72

Chapter 4 Accessible Surface Area from NMR Chemical Shifts ………..……… 73

4.1 **Introduction** ……………………………………………………………… 74

4.2 **Materials and Methods** ……………………………………………… 77

4.2.1 Data set …………………………………………………………………… 77

4.2.2 Computation of Observed Fractional ASA …………………………. 77

4.2.3 Mapping fractional ASA Prediction as a Regression Task …………….. 78

4.2.4 Feature Set ………..………………………………………………… 79

4.2.5 Training the Prediction Model ……………………………………….. 82

4.2.6 Evaluation ………..…………………………………………………. 83

4.3 **Results and Discussion** ……………………………………………. 83

4.3.1 Training Performance and Feature Importance ……………………….. 83

4.3.2 Test Performance …………….…………………………………….. 85

4.3.3 Buried-Exposed and Buried-Intermediate-Exposed Classification …… 92

4.3.4 Discussion ……………………………………………..……………. 94

4.3.5 The ShiftASA Web Server …………………………………………….. 102

4.4 **Conclusion** ………………………………………………………………….. 103

Chapter 5 E-Thrifty - Chemical Shift Threading for Accurate Protein Fold Recognition ………………………………………………………. 105

5.1 **Introduction** ………………………………………………………… 106

5.2 **Materials and Methods** …………………………………………… 108

5.2.1 Measuring Local and Non-local Structure Similarity …..……………… 108

5.2.2 Gap Penalty Function in Sequence-Structure Alignment ……………… 114

5.2.3 Protein Local Alignment ..……………………………………………… 116

5.2.4 Chemical Shift Scoring and the Alignment Ranking …………………... 117

5.2.5 Optimization of E-Thrifty Parameters ..………………………………... 118

5.2.6 Statistical significance of E-Thrifty Alignments ...…………………….. 119

5.2.7 Generation of 3D Structures via MODELLER …..……………………. 120

5.2.8 Assessment of E-Thrifty Generated Structures ..……………………… 121

5.2.9 Clustering and Selection of E-Thrifty Templates ...……………………. 122

5.2.10 CS-GAMDy Refinement of E-Thrifty Models ..……………………….. 122

5.2.11 The E-Thrifty Web-server and Stand-alone Program ………………….. 123

5.3 **Results ……………………………………………………………….** 123

5.4 **Discussion …………………………………………………………….** 129

5.5 **Conclusion …………………………………………………………..** 138

Chapter 6 Contributions and Future Prospects ………………………………… 140

6.1 **Introduction** ..………………………………………………………….. 140

6.2 **Contributions of this Thesis** ..………………………………………… 141

6.3 **Future Prospects** ……………………………………………………… 146

Biblography …………………………………………..……………………… **149**

List of Tables

Table	Title	Page
1.1	The twenty amino acids and their abbreviations…………………………….	2
1.2	Major phases of the current thesis and corresponding sub-goals…………...	22
2.1	Weighting coefficients (|**w**|) of chemical-shift and sequence-derived features for CSI 2.0's SVM model ………………………………………………..	40
2.2	Performance of CSI 2.0 (shown in bold) and eight other chemical shift and sequence-based methods on an independent test set of 59 proteins (total 8,078 residues) when using "DSSP" (Kabsch et al. 1983) secondary structure assignments as the reference structure ……………………………	42
2.3	The p-values or probabilities of Student's two sample t-tests between CSI 2.0 and four other best performing methods are shown ……………………..	44
2.4	CSI 2.0 performance with selected chemical shift assignments and combinations of shift assignments …………………………………………...	45
2.5	Comparison of the performance of CSI 2.0 (shown in bold) versus other four methods (Psi-CSI, TALOSN, TALOS+, DANGLE) relative to the percentage of missing backbone ^{1}H, ^{13}C or ^{15}N chemical shift assignments ..	46
2.6	Percentage Q3-accuracies of the CSI 2.0 protocol (shown in bold) and eight other methods over an independent test set of 59 proteins using three different reference (DSSP (Kabsch et al. 1983), STRIDE (Frishman et al. 1995) and VADAR (Willard et al. 2003)) structures ……………………..	48
2.7	Confusion Matrix of secondary structure assignments generated by CSI 2.0 on the independent test set of 59 proteins …………………………………	50
3.1	Performance evaluation of CSI 3.0 on 13 selected proteins …………………	69
4.1	Normalized weighting coefficients (|**w**|) of chemical-shift and sequence-derived features are listed …………………………………………………	84

Table **Title** **Page**

4.2 The MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), Mean Spearman’s Correlation, and the standard deviation of Spearman’s correlation for all six fASA prediction methods (including ShiftASA, with and without SABLE) evaluated over the TEST1 set (65 test proteins) [64 proteins for side-chain RCI] ……………………… 87

4.3 The MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), Mean Spearman Correlation coefficient, and the standard deviation of Spearman correlation coefficient for ShiftASA (with and without SABLE), SABLE (Adamczak et al. 2004) and Side-chain RCI (Berjanskii et al. 2013) evaluated over the TEST2 set ……………………… 88

4.4 Spearman correlation coefficient between the actual fASA and the fASA values as predicted by ShiftASA (with and without SABLE), SABLE (Adamczak et al. 2004), RVPNet (Ahmad et al. 2003), SARpred (Garg et al. 2005) and side-chain RCI (Berjanskii et al. 2013) using TEST1 proteins …. 92

4.5 Two-state and three-state mean classification accuracy and precision of fASA for the TEST1 set reported by ShiftASA, SABLE and RVPNet for different threshold systems. The number of buried (#B), intermediate (#I) and exposed (#E) residues for each threshold system are shown in the first column ………………………………………………………………. 93

5.1 Template recognition performances of four threading programs on TEST1 proteins using sequence identity cutoff as ≤ 30% …………………………… 126

5.2 Template recognition performances of four threading and two protein structure prediction programs on TEST1 proteins using sequence identity cutoff as ≤ 95% ……………………………………………………………… 127

5.3 Template recognition performances of E-thrifty, POMONA, HHPred, and DALI protocols on TEST2 proteins are shown …………………………….. 128

Table **Title** **Page**

5.4 E-Thrifty performances for nine proteins randomly chosen from TEST1 and TEST2 sets using sequence only, sequence, shift-derived secondary structures and shift-derived torsion angles (POMONA features) and sequence and shift-derived secondary structures (HHpred features) with a sequence identity threshold of ≤30% are shown 135

List of Figures

Figure	Caption	Page
1.1	A generalized amino acid with its constituent parts i.e. amide nitrogen (*N*), carbonyl carbon (*CO)* and side-chain (*R*) is shown in the upper image and a protein backbone with peptide bond formed between two adjacent amino acids is depicted below …………………………………………………..	3
1.2	The secondary, tertiary and quaternary structure levels in the protein structure hierarchy are shown. Hydrogen bond formation in secondary structure elements such as α-helix and β-sheet is shown with yellow dashes	4
1.3	An α-helical structure is shown in a cartoon or "ribbon" representation (left) and the hydrogen bonds between backbone carbonyl carbon and nitrogen atoms (blue dashes) are shown in a ball and stick representation (right) …..	6
1.4	A mixed β-sheet structure consisting of three β-strands is shown on the left and the hydrogen bonds between adjacent β-strands are shown on the right	7
1.5	A β-turn comprising of three residues is shown in the left image. The right image shows the hydrogen bonding between residue *i* and residue *i+3* in the same β-turn ………………………………………………………………	7
1.6	A β-hairpin is shown on the left and the hydrogen bonded strands and turn region in the β-hairpin shown on the right ……………………………….	8
1.7	Accessible surface area and Van der Waals surface explained ……………..	9
1.8	The proton chemical shift spectrum for diacetone alcohol (4-hydroxy-4-methyl-2-pentanone) [Adapted from Becker 1999]. The chemical shift peaks, δ for functional groups are shown along the x-axis in ppm units…..	13
1.9	A general workflow of NMR structure determination process ……………...	14
2.1	A bar graph comparing CSI 2.0's Q3 accuracies with eight other chemical shift and sequence-based protocols over an independent test set of 59 proteins. The error bar (i.e. standard deviation in Q3- accuracy of each method) appears on top of each bar plot ……………………………………..	43

Figure **Caption** **Page**

2.2 Secondary structure prediction/assignment for BMRB 16116 (PDB ID: 2KDL) and BMRB 16117 (PDB ID: 2KDM) by CSI 2.0, TALOS-N and PSIPRED …………………………………………………………… 52

2.3 Secondary structure assignment/prediction for BMRB 4375 (Unfolded ubiquitin) as determined by the CSI 2.0, TALOS-N and PSIPRED programs 53

2.4 The CSI 2.0 (http://csi.wishartlab.com) web server image (a) homepage; and (b) Result page ………………………………………………………….. 55

3.1 Program flow chart for CSI 3.0 …………………………………………... 67

3.2 A montage of the CSI 3.0 webserver and typical output screen shots……… 71

4.1 Top ten relevant features in the SGBM method. The importance of these predictor variables or features is normalized to a scale of 1-100. The predictor variable names are shown on the vertical axis …………………… 85

4.2 Mean Spearman correlation coefficient and the standard deviation of correlations of all five fASA prediction models (including ShiftASA) are shown. The performance is measured over the TEST1 data set. The mean correlation associated with each method is shown at the top of each bar diagram ……………………………………………………………………… 86

4.3 Agreement between predicted and observed residue-specific fASA values by SABLE, RVPNet, Side-chain RCI and ShiftASA for the putative dinitrogenase iron-molybdenum cofactor from Thermotoga martima (PDB ID: 1O13, chain A) ..…………………………………………………………… 89

4.4 Agreement between the predicted and observed residue-specific fASA values for SABLE, RVPNet, Side-chain RCI and ShiftASA for ubiquitin bound to Hrs-UIM (PDB ID: 2D3G, chain B) 90

Figure	Caption	Page

4.5 Spearman correlation between observed and predicted ASA for TEST1 proteins using both sequence and chemical shift features [Hydrophobicity, RCI, secondary structure probability and conservation score and SABLE] (blue diamonds), sequence only [Hydrophobicity, conservation score and SABLE predicted ASA] (red rectangles) and chemical shift only [RCI (backbone and side-chain) and secondary structure probability] (green circles) ………………………………………………………………… 95

4.6 Percent complete chemical shift assignments, scaled in 0.0-1.0 (green circles) vs. Spearman's rank correlation coefficients (red rectangles and blue diamonds) for the TEST1 set ………………………………………………. 97

4.7 Standard deviation of ASA values in the test set for 20 different amino acids (blue circles) and prediction error (MAE) (red rectangles) are shown ……… 98

4.8 The relationship between fASA range vs. prediction error (MAE) (blue triangles) and ASA range vs. sample training size (red circles) is depicted … 100

4.9 Per-residue fASA values for unfolded ubiquitin (BMRB ID: 4357) ……….. 101

4.10 A montage of the ShiftASA webserver showing the home page (left) and the screenshot of an example output page (right) ………………………………. 103

5.1 The sequence-structure alignment concept used in E-Thrifty method ……… 109

5.2 Backbone Φ/Ψ torsion angles mapping into 9 overlapped regions in Ramachandran map with a letter assigned to each region …………………... 113

5.3 Definition of a gap, i, i', j and j' residue positions in sequence and structure block respectively. A single gap can include deletions or insertions on either in the sequence or structure and is defined by all four indices, i, j, i' and j' ... 115

5.4 Example sequence-structure alignments between query 2LCI and subject 2L82 (chain A) sequences are shown ……………………………………... 117

5.5 E-Thrifty template models after the CS-GAMDy refinement for eight query proteins are shown ……………………………………………………… 129

Figure **Caption** **Page**

5.6 TM-score distribution of full length models using input templates identified by DALI, E-Thrifty, HHpred and POMONA for NTPase [PDB ID: 2LCI] and Nkx protein [PDB ID: 2L9R] proteins …………………………………. 132

5.7 Alignment score vs. C_α – rmsd between equivalent residues of NTPase protein (PDB ID: 2LCI) and template proteins is shown ……………………. 133

5.8 E-Thrifty alignment score vs. TM-score is shown for top 500 templates identified by E-Thrifty for NTPase protein (PDB ID: 2LCI) ………………... 134

List of Abbreviations

3D	3 Dimensional
AGP	Affine Gap Penalty
AMBER	Assisted Model Building with Energy Refinement
ANN	Artificial Neural Networks
ASA	Accessible Surface Area
BLAST	Basic Local Alignment Search Tool
BLOSUM	BLOcks SUbstitution Matrix
BMRB	Biological Magnetic Resonance data Bank
CASD-NMR	Critical Assessment of Structure Determination by NMR
CD	Circular Dichroism
CGI	Common Gateway Interface
CS-GAMDy	Chemical Shift driven Genetic Algorithm for Molecular Dynamics
CSI	Chemical Shift Index
DALI	Distance matrix ALIgnment
DNA	Deoxyribonucleic Acid
DSS	4,4-Dimethyl-4-Silapentane-1-Sulfonic acid
DSSP	Dictionary of Secondary structure of Proteins
DYANA	DYnamics Algorithm for Nmr Applications
E-THRIFTY	Enhanced THRIFTY
EM	Electro-magnetic
fASA	Fractional Accessible Surface Area (ASA)
HMM	Hidden Markov Model
HSQC	Heteronuclear Single Quantum Coherence
IDP	Intrinsically Disordered Protein
MAE	Mean Absolute Error
MSE	Mean Square Error
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effects/Enhancement
NOESY	Nuclear Overhauser Effect Spectroscopy

nrPDB	Non-Redundant Protein Data Bank
PDB	Protein Data Bank
POMONA	Protein alignments Obtained by Matching Of NMR Assignments
PPM	Parts Per Million
PSI	Protein Structure Initiative
PSI-BLAST	Position Specific Iterated Basic Local Alignment Search Tool (BLAST)
PSSM	Position Specific Scoring Matrix
RBF	Radial Basis Function
RCI	Random Coil Index
RDC	Residual Dipolar Coupling
RMSD	Root Mean Square Deviation
RMSE	Root Mean Square Error
RNA	Ribonucleic Acid
RSA	Relative Solvent Accessibility
SAXS	Small-angle X-ray Scattering
SGBM	Stochastic Gradient Boosting Method
SOV	Segment OVerlap score
SRCC	Spearman's Rank Correlation Coefficient
STRIDE	STRuctural IDEntification
SVM	Support Vector Machine
THRIFTY	THReading with shIFTY
TM-score	Template Modeling score
TMS	Tetramethylsilane
TOCSY	Total Correlation Spectroscopy
VADAR	Volume, Area, Dihedral Angle Reporter
VGP	Variable Gap Penalty

List of Symbols

Å	RMSD unit
δ	Chemical Shift
$\Delta\delta$	Secondary Chemical Shift
ρ	Probability or Likelihood

Chapter 1

Introduction

1.1 Proteins

Every living cell contains three types of “information processing” macromolecules: DNA, RNA and proteins. Each DNA molecule carries genetic information that is transcribed into messenger RNA (mRNA) inside the nucleus. This mRNA is then translated (outside the nucleus) using specialized protein-RNA complexes called ribosomes that convert the genetic information contained in the mRNA into proteins. In all cases the gene sequence (in the DNA or its corresponding mRNA) determines the protein sequence. This information translation process – from DNA to protein -- is known as the “fundamental paradigm” of biology (Crick 1970).

Proteins serve as both the building blocks and the engines of life. Indeed they perform many vital tasks necessary for proper cell function. For instance, proteins, such as insulin, may regulate metabolism whereas other proteins, such as α-keratin, serve as building blocks for tissues. Proteins also transduce external stimuli from the environment (such as rhodopsin for light signal transduction in the eyes), catalyze important biochemical reactions (such as pepsin for digestion and DNA polymerase for DNA synthesis), or transport small molecules or ions (such as hemoglobin, which carries oxygen in the blood). Proteins also regulate gene activities and play key roles in transferring signals between or within the cell. In other words, proteins are involved in executing nearly every important cellular function. The enormous repertoire of functions performed by proteins arises from the enormous number of different three-dimensional structures they can adopt. These structures are determined by their amino acid sequences.

It is important to note that different proteins consist of different sequences derived from a standard set of 20 different, naturally occurring amino acids. As seen in Table 1.1, eleven amino acids are deemed as non-essential (i.e. our body can produce them on its own) and nine are considered to be essential (i.e. they must be obtained through the food we eat). Each protein is defined by its unique amino acid sequence, with the number of amino acids in any given protein ranging from as few as 40 to as many as 36,000 (Opitz et al. 2003). Table 1.1 lists the twenty naturally occurring amino acids, including their essentiality, their three letter codes and the single letter abbreviations used to represent them.

Chemical Name	Three-letter Code	One-letter Code
Alanine	Ala	A
Cysteine	Cys	C
Aspartic Acid	Asp	D
Glutamic Acid	Glu	E
Phenylalanine (essential)	Phe	F
Glycine	Gly	G
Histidine (essential)	His	H
Isoleucine (essential)	Ile	I
Lysine (essential)	Lys	K
Leucine (essential)	Leu	L
Methionine (essential)	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine (essential)	Thr	T
Valine (essential)	Val	V
Tryptophan (essential)	Trp	W
Tyrosine	Tyr	Y

Table 1.1: The twenty amino acids and their abbreviations

Using this list of amino acids it is possible to define the polymeric structures of proteins as a simple sequential list of letters and abbreviations. For an example, the amino acid sequence for the protein known as Ubiquitin can be described using the one-letter code.

MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSD YNIQKESTLHLVLRLRGG

When a protein has been assembled from its constituent amino acids (using the cell's translational engine called the ribosome), it spontaneously folds into a particular three-dimensional shape, called the tertiary structure. Typically each protein has a well-defined spatial arrangement into which it folds, and it will always fold into that arrangement (Anfinsen 1973, Lesk 2001). That three-dimensional spatial arrangement controls the protein's overall shape, surface, electrostatic and chemical features, which in turn determines the protein's function.

1.2 Protein Structure

Protein structure is commonly described using a four-level hierarchy. The first level is the **primary structure**, also referred to as the amino acid or protein sequence.

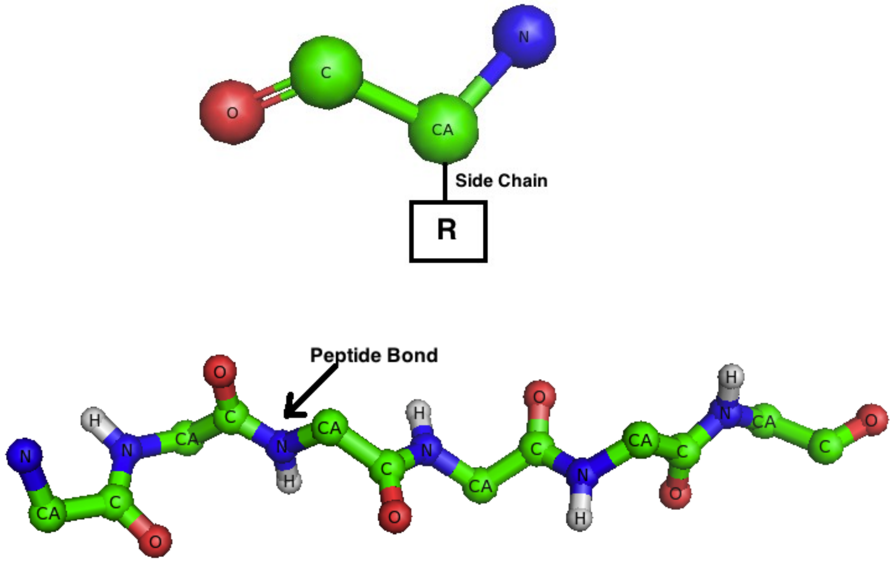


Figure 1.1: A generalized amino acid with its constituent parts i.e. amide nitrogen (*N*), carbonyl carbon (*CO)* and side-chain (*R*) is shown in the upper image and a protein backbone with peptide bond formed between several adjacent amino acids is depicted below.

Figure 1.1(a) depicts a generalized amino acid, with the colored spheres representing individual atoms and the letters within the circle identifying the type of atoms. The rectangle marked "R" symbolizes the "side chain" which gives an individual identity to each of twenty amino acids. Each amino acid consists of a nitrogen atom (or amino group NH_2) bonded to an "α-carbon" (the "CA" atom in the center of the Figure 1.1(a)), which in turn is bonded to another carbon atom, called the "carbonyl carbon" (or carboxylic acid (COOH) group). Proteins are formed by a complex enzymatic process that links these amino acid monomers together to form an amino acid polymer by creating what is called a “peptide bond” between the carbonyl carbon of one amino acid and the backbone nitrogen atom of the next amino acid (see Figure 1.1(b)). Thus, the N-CA-CO atoms of the individual amino acids are connected to form a continuous chain that constitutes the protein’s backbone. This linear assembly of amino acids is known as the protein’s **primary structure**.

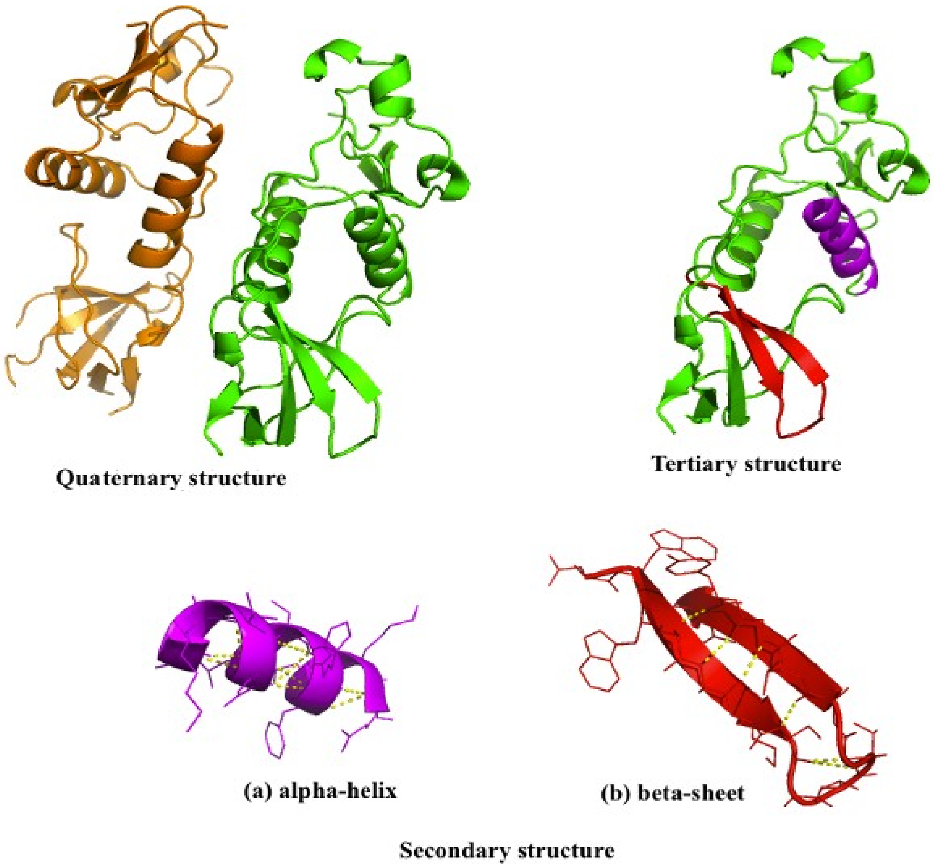


Figure 1.2: The secondary, tertiary and quaternary structure levels in the protein structure hierarchy are shown. Hydrogen bond formation in secondary structure elements such as α-helix and β-sheet is shown with yellow dashes.

The second level in the protein structure hierarchy is called **secondary structure**. Secondary structure is primarily stabilized by non-covalent interactions such as hydrogen bonds. Proteins can adopt four different types of secondary structures: α-helices, β-sheets, β-turns and coils. These secondary structure elements will be discussed in detail in next section. The third level in the standard protein structure hierarchy is called the **tertiary structure**. It is also called the three-dimensional (3D) structure and it refers to the global arrangement between secondary structure elements, which defines the overall protein fold. The tertiary structure largely determines the function of a protein. The final level in the standard protein structure hierarchy is called the **quaternary structure**. This refers to the organization of multiple folded proteins into a multi-subunit complex. Figure 1.2 illustrates the quaternary, tertiary and secondary levels in the protein structure hierarchy.

In order to understand how proteins can fold into a specific shape, it is important to understand some of the basics about secondary structure elements, super-secondary structure or structural motifs, accessible surface area and the concept of protein folding.

1.2.1 Secondary Structure

Beyond the primary structure or amino acid level, proteins can be viewed as segmented collections of "secondary structure". Unlike the strong covalent peptide bonds, which are responsible for maintaining the primary structure of proteins, the primary force that maintains the secondary and tertiary structure level integrity is the hydrogen bond, which is a weak non-covalent bond (Rose et al. 1993). A hydrogen bond is generally formed when an electronegative heavy atom (typically an oxygen atom) applies a strong pull on a nearby, non-covalently attached electropositive atom (typically a hydrogen atom). As a result the electropositive atom actually shares its electron with the electronegative partner – as long as the typical distance between two atoms is less than about 2.5 Å (Pimentel et al. 1971, Ippolito et al. 1990). This hydrogen bonding helps stabilize the proteins by creating regions of structural regularity, such as different secondary structures.

As stated earlier, protein secondary structure elements fall into four general classes: α-helices, β-sheets, β-turns and coils. The first three have a more regular or easily described shape due to their regular hydrogen bonding patterns while the fourth element (coil) lacks a well-defined structure due to the general absence of hydrogen bonds. A short description of these four elements is given below.

- α-helices, in which the protein backbone takes a "coiled spring" shape, are formed through hydrogen bond interactions between the backbone carbonyl carbon atom of residue *i* and the nitrogen atom of residue *i+4* (see Figure 1.3).

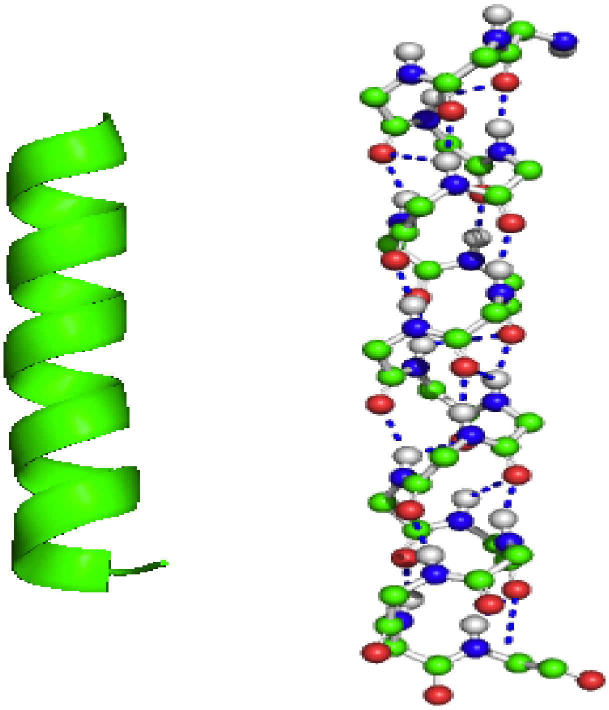


Figure 1.3: An α-helical structure is shown in a cartoon or "ribbon" representation (left) and the hydrogen bonds between backbone carbonyl carbon and nitrogen atoms (blue dashes) are shown in a ball and stick representation (right).

- β-sheets, in which the protein backbone assumes an extended or ribbon-like shape, look like aligned arrays of short polypeptides. These elements are made up of several β-strand that are stabilized by hydrogen bonds between the strands (see Figure 1.4). β-sheets can be subdivided into parallel, anti-parallel or mixed depending on whether the strands within the sheet run in one direction (from N- to C-terminus) or in opposite direction or a combination of both.

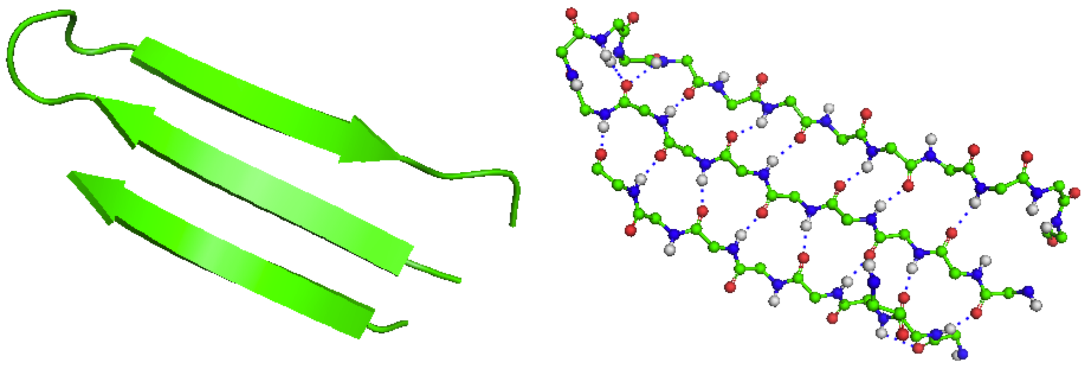


Figure 1.4: A mixed β-sheet structure consisting of three β-strands is shown on the left and the hydrogen bonds between adjacent β-strands are shown on the right.

- β-turns are very small structures consisting of 4-5 consecutive amino acids that form a staple-like conformation, leading to a reversal in the polypeptide chain. β-turns are formed by creating hydrogen bond between the carbonyl carbon of residue *i* and the nitrogen atom of residue *i+3* (see Figure 1.5). Several types of β-turns such as type-I, type-I', type-II, type-II' and type-VIII can be formed, depending on the dihedral angles formed between residue *i+1* and residue *i+2* (Hutchinson et al. 1994). Turns play an important role by connecting β-strands together, β-strands to helices and helices to helices in protein structures.

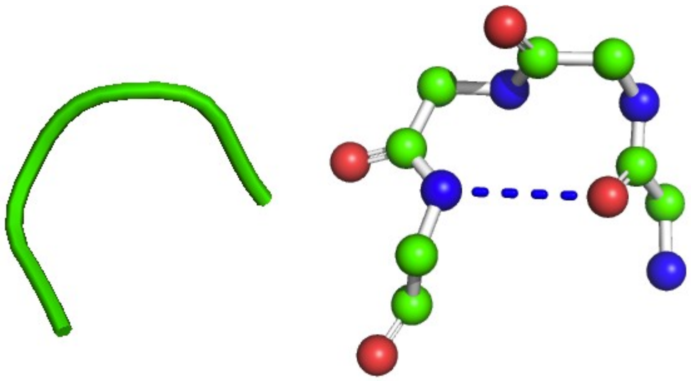


Figure 1.5: A β-turn comprising of three residues is shown in the left image. The right image shows the hydrogen bonding between residue *i* and residue *i+3* in the same β-turn.

- "coil", this is a catch-all phrase to refer to structures that lack any regular α-helical, β-turn or β-strand structure due to the absence of regular hydrogen bonding interactions.

1.2.2 Super-secondary structure

A super-secondary structure feature or a structural motif is formed when several adjacent secondary structures are packed together to form a well-defined or easily recognized compact shape. Examples of well-known structural motifs are β-hairpins, β-α-β units, Greek keys, β-meanders, β-barrels etc. These structural motifs are the basic building blocks for a protein's overall tertiary structure. The most abundant and simplest structural motif in globular proteins is the β-hairpin (shown in Figure 1.6). A β-hairpin, which is also called β-ribbon or β-β unit, involves two β-strands that look like hairpin used to hold one's hair in place. The structural motif consists of two β-strands that are adjacent in primary structure, having an anti-parallel direction (the N-terminus of one strand lies next to the C-terminus of the second strand) and connected by a small loop of two to five amino acids. β-hairpins can occur in isolation or as part of a sequence of hydrogen-bonded strands that together comprise a β-sheet. McCallister et al. (2000) has suggested that β-hairpin plays a critical role in the protein folding process. An NMR study by Blanco et al. (1994) demonstrated that short peptides can adopt stable β-hairpin conformations in aqueous solutions, suggesting that this structural motif can play an important role in the early stages of protein folding.

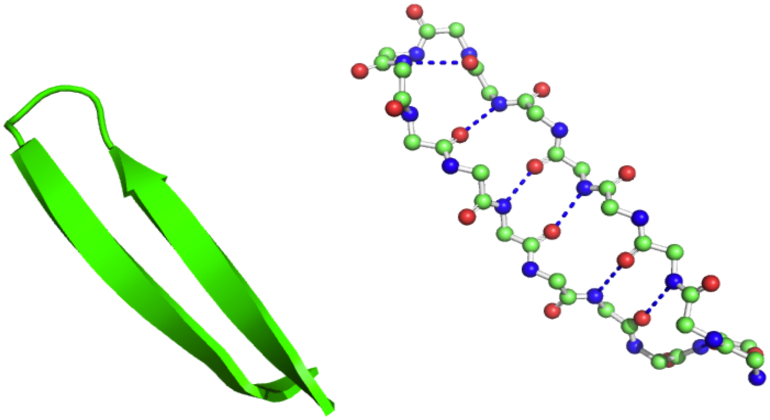


Figure 1.6: A β-hairpin is shown on the left and the hydrogen bonded strands and turn region in the β-hairpin shown on the right.

1.2.3 Accessible Surface Area

In the early 1970s, Dr. Fredric M. Richards and co-workers (Lee et al. 1970, Richards 1974 & 1977) first observed that certain parts of a folded protein seemed to be impermeable to water while other parts were highly accessible to water. This variance in exposure level seemed to be driven by the hydrophobicity or hydrophilicity of individual amino acid side chains as well as the 3D folded structure of the protein. Richards and colleagues also highlighted that the surface of a protein that was accessible to water was not equal to the van der Waals surface area but rather could be calculated by rolling a probe sphere of finite size (typically the size of an oxygen atom of 1.4 Å) over the entire van der Waals surface of a protein (Figure 1.7). The resulting swept-out "smoothed-surface" area is called the water accessible area or accessible surface area (ASA). ASA is a quantifiable property measured in square angstroms or $Å^2$. It can be measured for an entire protein molecule, for individual residues or even for specific atoms. ASA can also be represented as fractional ASA (fASA) that reports the percentage of ASA relative to a fully extended chain of amino acid residues. In addition to ASA, there is another measure called relative solvent accessibility or RSA, in which residues can be categorized into buried (B), partially buried (P) or exposed (E) based on their fASA values. Typically buried residues have a fASA of <0.25, partially buried have a fASA between 0.25 and 0.50 and exposed residues have a fASA of >0.50.

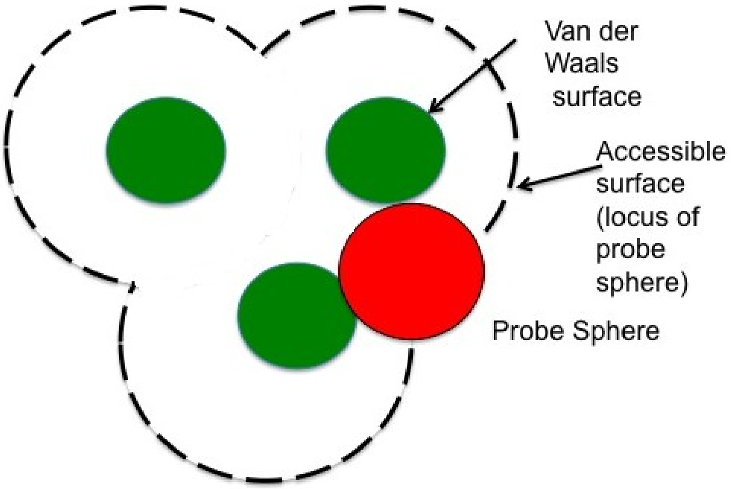


Figure 1.7: Accessible surface area and Van der Waals surface explained[1]

1.2.4 Protein Folding

Protein folding is a physical process by which a linear polypeptide is twisted, condensed or bent into a stable three-dimensional structure. It is through this folding process that a protein ultimately assumes its final, functional shape from an initial, non-functional random coil state (when it is just released from the ribosome). Understanding how proteins fold has been the

[1] Image Courtesy: http://www.ccp4.ac.uk/html/areaimol.html

subject of more than 50 years of intense research by tens of thousands of scientists. It is often referred to as the "protein folding problem" (Richards 1991, Chan et al. 1993). As yet, a complete answer to the protein folding problem has not been determined (i.e. we still cannot routinely predict the protein structure from a protein's amino acid sequence), but we are getting closer. In very general terms, during the folding process a protein goes through several different forms of molecular interactions, which include short range residue interactions and orientation dependent hydrogen bonds (found in secondary structures), residue-dependent long-range interactions and hydrophobic interactions, which stabilize tertiary structures (Hausrath 2003, Kadokura et al. 2006, Cieplak et al. 2009).

While short-range residue interactions and hydrogen bonding are largely responsible for forming secondary structures, long-range contacts due to hydrophobic interactions are generally responsible for forming and stabilizing a protein's tertiary structure. These long-range interactions assist in maintaining the integrity of the tertiary structure even in an aqueous environment. Water-soluble proteins largely obtain their unique native conformation due to the non-covalent hydrophobic effect (Cieplak et al. 2009). The hydrophobic effect is an entropic "force" that leads to hydrophobic molecules attracting one another, while at the same time excluding the surrounding water. The hydrophobic effect is commonly seen when one mixes oil and vinegar to make salad dressing. Mixing these two solvents leads to the formation of round oil droplets, which eventually coalesce back to a layer of oil that lies on top of the vinegar. In addition to non-covalent forces that drive the protein folding process, there are other types of covalent interactions that can stabilize tertiary structure. In particular, disulfide bonds are a type of covalent interaction that is known to help determine a protein's tertiary structure. Disulfide bonds form when non-adjacent cysteine residues are covalently linked through a sulfur-sulfur chemical interaction resulting from an oxidation process.

1.3 The Significance of Protein Structure

The preceding sections were intended to provide some basic information about protein structure and a brief backgrounder on how proteins fold. However, these sections did little to convey the motivation for studying protein structure. The reasons for studying protein structures are manifold. The first reason is that structure defines function. In other words, the shape and chemical characteristics of a protein give rise to its functions. The importance of protein function is profound considering that many human diseases, such as Alzheimer's disease, cystic fibrosis, Parkinson’s disease and sickle cell anemia result from disorders in both protein structure and protein function (Dobson 1999, Chaudhuri et al. 2006). The second reason is that studying protein structure gives us insight into the process of evolution. Proteins evolve over time and often protein structure is more conserved than sequence. By studying protein structures from diverse organisms, it is possible to understand more about their evolution and their distant relationships than by studying protein sequences (Yang et al. 2009). The third reason is that studying protein structure will ultimately help solve the protein folding problem. In principle, by solving enough protein structures it should be possible to “learn” the rules governing protein folding. If the rules could be learned, then it would be possible to predict the 3D structure of proteins from their amino acid sequence alone. This has been the motivation behind the multibillion-dollar Protein Structure Initiative (Smith et al. 2007). All three of these reasons provide a strong motivation for studying protein structure at an atomic level. However, determining the structure of proteins is not a trivial task.

1.4 Protein Structure Determination

To date, most protein structures have been determined by either X-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy. Of these two, X-ray crystallography is the older and more precise method of protein structure determination (Lesk 2001, Heinemann et al. 2001). NMR and X-ray crystallography, together, have been used to determine the structure of more than 100,000 proteins and peptides, with X-ray crystallography being responsible for about 90% of these structures and NMR being responsible for the rest. In X-ray crystallography, an X-ray beam of a particular wavelength is passed through a pure protein crystal containing large numbers of the molecules in a regular, crystallized lattice. The X-rays interact with the electrons in the crystallized molecular lattice producing a complex diffraction pattern that may be detected by a photographic film or an electronic area detector. This diffraction pattern, which is a collection of thousands of spots, can then be analyzed using mathematical methods, such as Fourier analysis, to produce a three-dimensional contour map that corresponds to the molecule’s electron density. The electron density map can then be re-interpreted into the positions and coordinates for each of the atoms in the molecule through careful (computer-aided) visual analysis. A more detailed explanation of the principles of X-ray crystallography can be found in Drenth (1994).

More recently, NMR spectroscopy has emerged as an alternative and complementary approach to macromolecular (especially protein) structure determination. NMR allows one to determine the three-dimensional structure of molecules dissolved in solutions (i.e. the liquid state) as opposed to molecules in the solid or crystalline state (as is required for X-ray crystallography). Because the primary focus of my research is on using NMR spectroscopic data to determine protein structures, most of the remaining discussion in this section will be concerned with NMR spectroscopy.

In NMR spectroscopy, a protein sample is typically dissolved in water and kept in a liquid state (as opposed to a crystalline state) and placed in a powerful magnetic field, supplied by a superconducting magnet. Radio-frequency electromagnetic (EM) radiation is pulsed into that sample and the absorption bands (resonances) corresponding to individual atoms in the protein are recorded, producing an NMR spectrum. An NMR spectrum of a molecule is typically characterized by a set of peaks, with the peak positions (or frequencies) corresponding to the chemical shifts (see section 1.5 for more details) of the individual atoms in that molecule. NMR spectra are also characterized by peak clusters (doublets, triplets, etc.) that correspond to scalar couplings between adjacent (geminal) or nearby (vicinal) atoms. The intensity of the peaks in an NMR spectrum correspond to the numbers of atoms found at a given resonance frequency or a given chemical shift. The intensity and shape of an NMR peak is also affected by the interactions between nearby atoms (via Nuclear Overhauser effects or NOEs) and the liquid lattice (relaxation effects). An example of a NMR spectrum for a simple molecule (diacetone alcohol) is shown in Figure 1.8, where the protons (or hydrogen atoms) from each of the functional groups of the molecule, $(CH_3)_2$, CH_3, CH_2, and OH are responsible for the characteristic peaks in the spectrum.

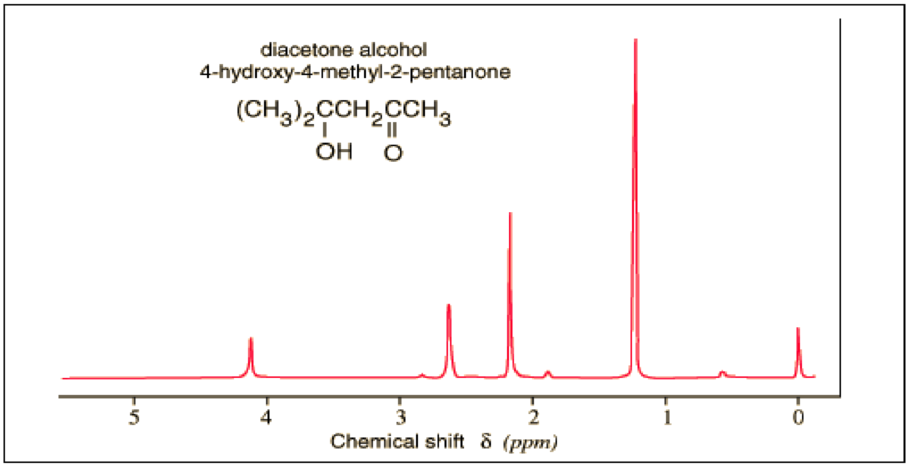


Figure 1.8: The proton chemical shift spectrum for diacetone alcohol (4-hydroxy-4-methyl-2-pentanone) [Adapted from Becker 1999]. The chemical shift peaks, δ for functional groups are shown along the x-axis in ppm unit.

Using specially designed combinations of radio frequency pulses with a defined length, strength, and orientation, it is possible to simultaneously record the NMR spectra of multiple types of atoms (hydrogen, carbon, nitrogen), coupled atoms or different NMR parameters such as NOEs, coupling constants, dipolar couplings, chemical shift anisotropy etc. Using two or three-dimensional NMR experiments with whimsical names such as TOCSY (Bax 1989), NOESY (Kumar et al. 1980), HSQC (Bodenhausen et al. 1980) and HNCAB (Wittekind et al. 1993), it is possible to determine the chemical shifts, coupling constants, and spatial proximity of most NMR active atoms in a protein. Once the sequence specific resonances or chemical shifts have been assigned and the spatial proximity of most hydrogen atoms has been determined through NOE measurements, then it is possible to use computational techniques such as distance geometry (Wuthrich 1986) or simulated annealing (Brunger 1993) to determine the three-dimensional structure of the protein. A general workflow of NMR structure determination procedure is shown in Figure 1.9. A more complete description of NMR spectroscopy can be found in Kemmink et al. 1996, Castellani et al. 2002 and Wüthrich 1990.

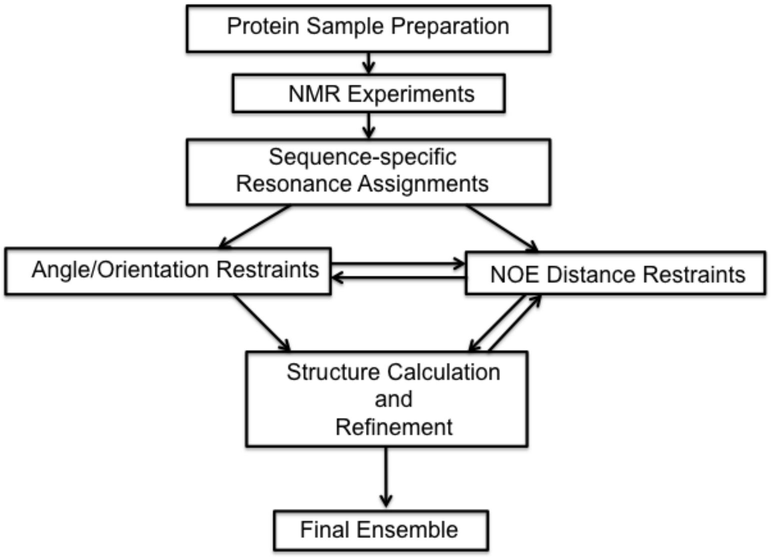

Protein Sample Preparation
NMR Experiments
Sequence-specific Resonance Assignments
Angle/Orientation Restraints
NOE Distance Restraints
Structure Calculation and Refinement
Final Ensemble

Figure 1.9: A general workflow of NMR structure determination process

As described above, one of the key steps in NMR structure determination involves a process known as chemical shift or resonance assignment. This involves assigning chemical shift values to each atom in the compound of interest (a small molecule or a protein). This process involves measuring NMR peak positions and using information about atom connectivity and, in the case of proteins, sequence information to determine which peak in the NMR spectrum corresponds to which atom in the molecule of interest. Chemical shift assignment is considered as an essential pre-requisite for Nuclear Overhauser Enhancement (NOE) peak assignment, which is a critical step in determining through-space connectivity between atoms in a given molecule or macromolecule. While NOE data can provide important structural constraints for structure generation, the difficulty and time required to collect and interpret NOE spectra have placed a serious limitation on the size and precision of protein structure determination via NMR. *It is my hypothesis* (and the motivation behind this thesis) that if chemical shifts could be more fully exploited in protein NMR, then many of these bottlenecks relating to protein size, analysis time and structural precision could be reduced and even eradicated. This is the primary motivation for developing chemical shift based structural parameter calculators described in Chapter 2, 3, 4 and 5 of this thesis.

In the following sections, I will provide some basic background about NMR chemical shifts and their use with regard to protein structure determination.

1.5 NMR Chemical Shifts

As stated earlier, when a chemical or protein sample is immersed into a powerful magnetic field and exposed to a pulse train of radio frequency electromagnetic (EM) signal, the atoms in the molecule will absorb certain frequencies from the impinging EM radiation. Each NMR-detectable nucleus (hydrogen or ^{1}H, carbon or ^{13}C and nitrogen or ^{15}N) absorbs at a characteristics EM frequency, which is dependent on the type of the nucleus and the strength of the magnetic field in which the nucleus is immersed. The fields "experienced" by atoms in molecules are modified by each atom's local electromagnetic field which, in turn, is affected by the atomic bonds and atomic neighbors in the molecule. The motions of the electrons orbiting the nucleus produce a secondary, smaller magnetic, that acts to oppose the applied external field. This change in the effective magnetic field on the nuclear spin causes the NMR signal to shift. This shift is called the "Chemical Shift", and it primarily depends on the type of nucleus being measured and the electron density in the nearby atoms and molecules. Chemical shifts are usually measured as the difference between the resonant frequency of a nucleus and that of a defined reference, relative to the reference's resonant frequency. This quantity is expressed by δ (chemical shift) and is reported in parts per million (ppm).

$$\delta = \frac{f - f_{ref}}{f_{ref}} \times 10^{6} \tag{1.1}$$

where f is the resonant frequency of the nucleus of interest and f_{ref} is the resonant frequency of the reference material. In NMR spectroscopy, the reference material is often TMS (TetraMethylSilane – for organic solvents) or DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid – for water), both of which are organo-silicon compounds that have a chemical shift defined as 0.00 ppm. Additional information about chemical shifts for a simple molecule is illustrated in Figure 1.8.

1.6 Chemical Shifts to Protein Structure

Chemical shifts are often called the "mileposts of NMR". They serve as reference points to help NMR spectroscopists map out atomic positions, identify chemical groups and determine molecular structures. Over the last few decades, chemical shifts have been used by chemists to successfully determine or confirm the covalent structure of tens of thousands of organic molecules. Not only are the chemical shifts sensitive to the type and character of nearby atoms but chemical shifts are also remarkably consistent or "predictive" for different chemical groups or chemical environments. This sensitivity and behavioral consistency has allowed chemists to produce well-defined chemical shift "principles" that allow them to deduce the identity of key chemical groups and thereby determine the precise structures of many small molecules (Brügel 1979, Biemann 1989, Steinbeck 2004). However, chemical shift assignments of large molecules, such as peptides or proteins, do not provide such precise or structurally obvious information. This is because the structure of large macromolecules, unlike small molecules, is primarily defined by non-covalent interactions, for which (until recently) very few chemical shift principles had been discovered or derived (Wishart 2011).

Nevertheless, the striking success of chemical shift-derived principles in characterizing small compounds actually led a number of biochemists and NMR spectroscopists (especially in the early 1960s) to try to develop similar principles or algorithms to interpret the relationship between chemical shifts and structure for larger molecules such as peptides and proteins. In the late 1960s and early 1970s, a number of theoretical models were developed that explained the influence of hydrogen bonds, dipole-dipole interactions and aromatic ring currents on protein proton shifts (Sternlicht et al. 1967, Tigelaar et al. 1972, Perkins et al. 1977). These models were refined and validated on the first published partial chemical shift assignments of peptides and proteins (Proctor et al. 1950, Arnold et al. 1951). For a number of years the use of chemical shifts was considered by many to be the best route to solving protein structures. Then in the early1980s NOE-based structure determination methods emerged (Jardetzky et al. 1981) and these new and powerful techniques largely supplanted the need for chemical shift interpretation. To date, NOE-based methods have allowed the determination of 11,093 protein structures (Source: PDB) and the assignment of more than 10,000 (Source: BMRB) protein and peptide resonances.

However, NOE-based methods are not completely without their problems nor are they necessarily routine or simple. NOE measurements, which are critical to the calculation of protein structures via conventional NMR methods, are responsible for a significant amount of overhead in the structure determination process. This is because measuring proton NOEs is both a tedious and a time consuming task. Another constraint with NOE-based methods lies in the determination of structures of large proteins (>200 residues). As the size of the protein grows, NOEs become gradually less useful and more difficult to measure and compute. Thus, the determination of large protein structures using NOE-based methods is still a difficult task (Wishart 2011). Consequently, NMR spectroscopists and computational biologists are still looking for potential improvements or possible transformative innovations to the NMR structure determination pipeline.

Interestingly, the explosion of NOE-based protein structures and assignments over the past 30 years has actually proven to be a real boon for protein chemical shift analysis. In particular, it has helped revive the idea that chemical shifts could be used to determine protein structures – independent of size. Indeed, the large quantity of available chemical shift assignments has allowed NMR spectroscopists to deduce a number of rules or algorithms concerning chemical shifts and their interpretation with regard to protein secondary structure, torsion angles, aromatic ring placement, oxidation states, solvent exposure and flexibility. In particular, chemical shifts have been successfully used to quantify protein secondary structure content (Mielke et al. 2009), to identify and assign secondary structure location (Wishart et al. 1992 & 1994b, Eghbalnia et al. 2005, Labudde et al. 2003) and structure classes (Mielke 2003), to identify structural motifs (Gronenborn et al. 1994, Shen et al. 2012), to determine redox states of cysteine residues (Wang et al. 2006), to predict disulfide bond information (Sharma et al. 2000), to measure surface accessibility (Vranken et al. 2009, Avbeli et al. 2004), to determine the orientation of aromatic rings (Perkins et al. 1979), to estimate the backbone torsion angles (Shen et al. 2009b, Berjanskii et al. 2006), to determine side-chain torsion angles and rotamer populations (London et al. 2008, Hansen et al. 2009, Shen et al. 2013), and to assess protein flexibility parameters (Berjanskii et al. 2005 and 2013). These studies have helped guide the development of powerful predictive algorithms that quantitatively infer the relationship of chemical shifts to the non-covalent structure and dynamics in proteins (Wishart 2011). Indeed, these advances led to the development of three programs in 2007 and 2008 including CSRosetta (Shen et al. 2008), CHESHIRE (Cavalli et al. 2007) and CS23D (Wishart et al. 2008)) that are able to predict or model the 3D structure of proteins using only chemical shift data as input.

While considerable excitement within the NMR community greeted these programs, they have not entirely lived up to expectations. Indeed, the performance or success rate of these techniques generally hovers around 50% (compared to 95%+ for NOE-based methods). This relatively poor performance indicates that there is still room for improvement and that the "Shift-to-Structure" problem has not yet been fully solved. Consequently, in the following sections we will discuss some of the problems relating to protein structure prediction in which chemical shifts are exploited.

1.6.1 Secondary Structure and Structural Motifs from Chemical Shifts

As mentioned earlier, proteins can be described as segmented collections of "secondary structures". Beyond the primary structure level, secondary structures are considered as essential to both describing and interpreting protein tertiary structures. As a result, protein chemists, and especially NMR spectroscopists, have shown a strong interest in identifying and characterizing secondary structure elements (Wuthrich 1986 & 1990). In the field of NMR, NOE based techniques are extensively used to identify and assign secondary structures in proteins (Wuthrich 1986). However it was not generally appreciated until the early 1990's that chemical shifts could also be used to identify secondary structures. Furthermore chemical shifts are very accurate and far easier to use than NOEs (Wishart et al. 1992 & 1994b). The "Chemical Shift Index" or the CSI was the first method that exploited the idea of using chemical shifts to identify protein secondary structures (Wishart et al. 1992). The CSI method was based on some simple observations regarding backbone chemical shift patterns in secondary structure locations in proteins. It uses a "numerical filter" to classify and cluster backbone ^{1}H and ^{13}C chemical shifts and thereby identify the type and location of protein secondary structure elements (α-helices, β-strands, and coils). Even though the CSI method is simple and easy to implement, it is surprisingly accurate. In particular, the reported agreement between the experimental and CSI-defined secondary structures is about 75-85% (Wishart et al. 1994b & 2002, Mielke et al. 2004 & 2009). As a result of its simplicity and accuracy, the CSI method has become one of the most popular methods in NMR to identify secondary structures from chemical shifts.

The success of CSI (along with some of its shortcomings) has inspired a number of CSI-like approaches to be developed over the past decade, including 1) a joint probabilistic model to secondary structure identification (PSSI) (Wang et al. 2002a); 2) an approach that combined CSI and a sequence-based secondary structure routine called PSIPRED (Jones 1999), which is called PsiCSI (Hung et al. 2003); 3) a method that uses pre-specified chemical shift patterns, PLATON (Labudde et al. 2003); 4) a prediction method that employs statistically derived chemical shift/structure potentials, PECAN (Eghbalnia et al. 2005); and 5) a two-dimensional cluster analysis method to analyze paired scattering diagrams of all six backbone chemical shifts, 2DCSi

(Wang et al. 2007). The above-mentioned methods typically incorporated advanced chemical shift model or additional statistical information as an extension to the original CSI algorithm.

More recently, chemical shift based secondary structure assignment approaches have emerged that exploit sophisticated machine learning techniques, torsion angle estimates, sequence homology information and extensive sequence-structure databases. These include: 1) a neural network method that predicts backbone and side-chain torsion angles, as well as secondary structure locations by matching chemical shift patterns over a five-residue window against a large database of previously assigned proteins with high-resolution structures (TALOS+ and TALOS-N) (Shen et al. 2009b & 2013); 2) a Bayesian-inference method that employs a similar concept to TALOS+ (DANGLE) (Cheung et al. 2010); and 3) a secondary structure identification method that identifies secondary structure population in both disordered and native-state proteins by analyzing the probability distribution of a large database of backbone chemical shifts (Delta2D) (Camilloni et al. 2012). The accuracies of these newer approaches range between 83% and 86%. In addition to these programs that identify regular secondary structures via chemical shifts, another recent program called, MICS identifies β turns and helix-capping motifs (Shen et al. 2012). MICS uses backbone chemical shifts, together with the PDB extracted amino acid preference to identify the locations of five types of β-turns and N-terminal and C-terminal helix capping motifs in a protein sequence via machine-learning methods such as artificial neural network.

Over the past 25 years it is clear that the field of chemical shifts and secondary structure identification has matured. However, it is also evident that there is still some room for improvement as no method has yet been described that achieves a level of accuracy >90% and few, if any, are able to identify super-secondary structures or motifs via chemical shifts. In this thesis I will describe two methods that achieve this goal (Chapters 2 and 3).

1.6.2 Accessible Surface Area from Chemical Shifts

Over the last two decades it has been observed that NMR chemical shifts correlate reasonably well with accessible surface area. The first evidence of such a phenomenon was first reported in 1994 (Wishart et al. 1994a). Almost a decade later Avbelj et al. (2004) studied the effect of secondary structure elements namely α-helices and β-sheet and solvent exposure on backbone chemical shifts. They showed that proton secondary shifts of smaller peptides have a characteristic chemical shift distribution that correlated with solvent exposure. In a later study by

Vranken et al. (2009), the effect of secondary structure and solvent exposure on chemical shift assignments was re-examined on a large database of proteins for which both reported atomic coordinates and chemical shift values were available. Two major findings from this study were: 1) Non-polar atoms have significantly larger chemical shift dispersion and a somewhat different chemical shift distribution compared to polar atoms; and 2) those atoms with greater atomic ASA, exhibited chemical shift values that tended towards random coil values. The relationship between chemical shifts and ASA was in fact used as one of the features in developing a significantly improved structure-based chemical shift prediction algorithm, called ShiftX2 in 2011 (Han et al. 2011). Most recently, Berjanskii et al. (2013) proposed a simple formula to calculate per-residue fractional accessible surface area from backbone and side-chain chemical shifts and observed a correlation of about 74% with the observed fASA values over a subset of 15 proteins.

Based on these examples it is evident that the field of chemical shifts and accessible surface area prediction is still maturing. As a result, there is still considerable room for improvement as no method has yet been described that achieves a level of accuracy >80%. In this thesis I will describe a method that achieves this goal (Chapter 4).

1.6.3 Protein Structure Determination by Chemical Shifts

Fast, automated protein structure determination is one of the ultimate goals for computational biologists. This is because high-resolution structures, obtained by X-ray crystallography or NMR spectroscopy, require considerable resources and are available only for a small fraction of all known protein sequences (<0.05%). It is universally agreed that computational methods, in combination with rapid experimental data acquisition, will be required to help generate or model the structures for the remaining sequences. These computational methods for structure generation can be broadly categorized into two types: 1) Comparative modeling; 2) *De novo* approaches. Comparative modeling algorithms usually perform structure determination in two steps. It starts by identifying related templates from known structures with modest (>30%) sequence identity through an optimal alignment between the query and template sequence(s). In the second stage, a complete three-dimensional model of the query protein is generated using the information from the aligned templates. On the other hand, *de novo* methods start with only amino acid sequence and no structural template. They use a combination of an effective conformation-searching algorithm and an energy function to generate structural models from the scratch. The key problem with *de novo* approaches is their massive computational cost. *De novo* methods are mostly restricted to solving smaller proteins and peptides because of the computational bottlenecks associated with the conformational sampling of proteins with large number of residues.

Protein NMR chemical shifts contain important structural information and if integrated effectively into the structure modeling process, can greatly improve the speed and accuracy of both *de novo* and comparative modeling approaches to structure determination. This idea motivated the development of three powerful *de novo* protein structure prediction programs: CHESHIRE (Cavalli et al. 2007), CS-ROSETTA (Shen et al. 2008), CS23D (Wishart et al. 2008), all of which exploit chemical shifts as the only input and generate high-quality all atom models for small to medium proteins with a diversity of folds.

Most recently Shen et al. (2015) described a system called POMONA (Protein alignments Obtained by Matching of Nmr Assignments) that identifies suitable homologs for query proteins from the sequence-structure database (PDB) using chemical shifts and NOE distance restraints when available, which is followed by a modified comparative modeling procedure to generate all-atom structures for protein. POMONA searches the PDB for suitable homologs that are well matched with backbone chemical shift predicted residue specific Φ/Ψ probability maps and secondary structures. The resulting structural templates are then clustered into groups (typically ten) using the normalized Cα- root mean square deviation as a metric. Representative homologs from these clusters are then used to build a structural pool for a comparative modeling using a modified RosettaCM procedure. Assessments of this method on a set of 16 proteins indicate that in majority cases the best alignments reported by POMONA have decent structural similarities with the native structures.

While good progress is being made in this area, it is evident that the field of chemical shifts and protein structure generation/prediction still has a long way to go. In this thesis I will describe a method, called E-Thrifty, that exhibits a significant improvement over existing methods (Chapter 5).

1.7 Thesis Objectives

I believe that further improvements are possible for the “Shift to Structure” problem and that these can be addressed using innovative computational and machine learning approaches. In particular, *my central hypothesis* is that novel algorithms can be developed that will allow chemical shifts to be used to determine and refine protein structures with high efficiency and high accuracy (>90%). More specifically, I believe that: 1) significant improvements to shift-based secondary structure prediction and identification can be made; 2) it is possible to develop algorithms that identify internal/external β-strands; 3) significant improvements to predicting accessible surface areas via shifts can be made; and 4) significant improvements in structure recognition via chemical shifts can be made. Together, these improvements and innovations will all lead to the creation of a robust framework to solve protein structures via NMR-chemical shifts.

My "chemical shift to structure" thesis project can be divided into three major components: a) the development of shift-based protein local structure prediction techniques; b) the development of prediction methods for protein local/non-local interactions from sequence data and chemical shifts; c) the development of a tertiary fold recognition technique from chemical shifts and sequence data. Many of these components or tools will be integrated into an updated 3D structure determination package, called "CS23D 3.0". These three major components or phases can be further sub-divided into several sub-goals. The sub-goals under each of the major phases are described in Table 1.2.

Major Phase	Sub-goals
1. Local protein structure identification	a) Chemical shift-based protein secondary structure determination
2. Identification of local/non-local protein interactions from protein sequence and chemical shift data	a) Internal/external β-strand identification from chemical shifts b) Fractional accessible surface area prediction using sequence and chemical shifts
3. Tertiary fold recognition from chemical shift and protein sequence data	a) Accurate template recognition by chemical shift and sequence-structure threading

Table 1.2: Major phases of the current thesis and corresponding sub-goals

1.8: Organization of Thesis

This document is organized as follows: Chapter 2 presents a novel method for chemical shift-based secondary structure identification, which includes the description and evaluation of the method using an independent test set. In particular, Chapter 2 describes a multi-class SVM model specially optimized to classify protein secondary structures into three major classes namely α-helices, β-sheets, and coils. This chapter also provides a detailed assessment of this program relative to other programs. Chapter 3 describes a web server that is designed to identify four major secondary structure elements; α-helices, β-sheets, β-turns and coils along with detailed classification of β-turns and β-sheets and other structural motif such as β-hairpins, a total of 11 types of secondary structure and super-secondary structural motifs, using chemical shift data. In this algorithm, we exploited the previously described multi-class SVM classifier to identify three major secondary structure classes. In addition we developed a rule-based algorithm to identify five types of β-turns and a SVM classifier to classify β-strands into edge and internal strands. This chapter provides a detailed assessment of this program relative to other programs. Chapter 4 describes a novel regression method for accurately estimating per-residue fractional accessible surface area using chemical shift and sequence data alone. Specifically, we developed a Stochastic Gradient Boosted Regression Tree method that has been optimized to estimate real-value accessible surface area. As with previous chapters, this chapter provides a detailed assessment of this program relative to other programs. Chapter 5 outlines a protein fold recognition method called "E-Thrifty", in which a parameterized sequence-structure threading method and a weighted chemical shift scoring function have been implemented. This chapter also provides a detailed assessment of the E-thrifty program relative to other programs. The document ends with Chapter 6 that summarizes the key results from the previous chapters and outlines some future research goals that can be commenced based on these results.

Chapter 2

CSI 2.0 – A Significantly Improved Version of the Chemical Shift Index[1]

Abstract

Protein chemical shifts have long been used by NMR spectroscopists to assist with secondary structure assignment and to provide useful distance and torsion angle constraint data for structure determination. One of the most widely used methods for secondary structure identification is called the Chemical Shift Index (CSI). The CSI method uses a simple digital chemical shift filter to locate secondary structures along the protein chain using backbone ^{13}C and ^{1}H chemical shifts. While the CSI method is simple to use and easy to implement, it is only about 75-80% accurate. Here we describe a significantly improved version of the Chemical Shift Index (CSI 2.0) that uses machine-learning techniques to combine all six backbone chemical shifts ($^{13}C_{\alpha}$, $^{13}C_{\beta}$, ^{13}C, ^{15}N, ^{1}HN, $^{1}H_{\alpha}$) with sequence-derived features to perform far more accurate secondary structure identification. Our tests indicate that CSI 2.0 achieved an average identification accuracy (Q3) of 90.56% for a training set of 181 proteins in a repeated 10-fold cross-validation and 89.35% for a test set of 59 proteins. This represents a significant improvement over other state-of-the-art chemical shift-based methods. In particular, the level of performance of CSI 2.0 is equal to that of standard methods, such as DSSP and STRIDE, used to identify secondary structures via 3D coordinate data. This suggests that CSI 2.0 could be used both in providing accurate NMR constraint data in the early stages of protein structure determination as well as in defining secondary structure locations in the final protein model(s). A CSI 2.0 web server (http://csi.wishartlab.com) is available for submitting the input queries for secondary structure identification.

[1] Portions of this chapter were published as: Hafsa NE, Wishart DS (2014) CSI 2.0: a significantly improved version of the Chemical Shift Index. J Biomol NMR 60:131-146.

2.1 Introduction

Secondary structures are considered fundamental to both the description and the understanding of protein tertiary structures. Indeed, secondary structure maps and secondary structure ribbon diagrams are standardly used in almost all structural biology books, journals and databases (Wuthrich 1986, Berman et al. 2000). It is also notable that secondary structure assignments or predictions are still widely used as the basis to many protein fold recognition algorithms (Soding et al. 2005), protein threading methods (Jones et al. 1999), 3-D protein structure prediction algorithms (Wishart 2011, Soding et al. 2011) and intrinsically disordered protein (IDP) identification methods (He et al. 2009). Secondary structure is also key to many heuristic energy functions that are designed to assess, fold and/or refine protein structures (Wishart et al. 2008, Berjanskii et al. 2009, Adams et al. 2013). Furthermore, secondary structure provides not only approximate torsion angle and qualitative backbone flexibility data, it also provides hydrogen bonding information (for α-helices and β-strands), implied contact information (for β-strands) and important topological information (through β-turns). While increasing interest is turning to extracting or predicting more quantitative measures of protein structure (i.e. torsion angles, backbone order parameters, accessible surface area) it is important to note that the accuracy of these methods is not yet sufficient to permit their widespread use in 3D protein structure prediction or 3D structure calculation algorithms (Wishart 2011). Consequently the identification and delineation of secondary structure elements continues to be of interest to protein chemists, bioinformaticians, X-ray crystallographers and, of course, NMR spectroscopists (Wuthrich 1986 & 1990, Wishart 2011).

In the field of protein NMR, NOE-based methods are widely used to identify and assign secondary structures. Indeed, they continue to be the predominant method for identifying or delineating secondary structures in peptides and proteins (Wuthrich 1986). Less well known is the fact that NMR chemical shifts can also be used to identify secondary structures and that they are remarkably accurate and far easier to use than NOEs (Wishart et al. 1992, 1994a & 1994b). The idea of using chemical shifts to identify secondary structures was first exploited with the development of the Chemical Shift Index or CSI (Wishart et al. 1992). The CSI method applies a "digital filter" to backbone ^{1}H and ^{13}C chemical shifts to precisely identify the type and location of protein secondary structure elements (helices, β-strands, coils) along a protein chain (Wishart et al. 1994b). The CSI method is particularly popular because it is easy to implement and surprisingly accurate with the reported agreement between X-ray-defined secondary structures and CSI-identified secondary structure being about 75-85% (Wishart et al. 1994b & 2002, Mielke et al. 2004 & 2009).

However, the CSI method is not without some shortcomings. For instance, it requires near complete backbone assignments, it is sensitive to the choice of random coil shifts used to calculate the secondary shifts, and it identifies α-helices (>90% accuracy) more accurately than β-strands (<75%). Because of these limitations, a number of alternative CSI-like approaches have been developed over the past decade, including PSSI (Wang et al. 2002a), PsiCSI (Hung et al. 2003), PLATON (Labudde et al. 2003), PECAN (Eghbalnia et al. 2005), and 2DCSi (Wang et al. 2007). These methods typically extend the CSI concept by incorporating more advanced chemical shift models or additional statistical information. For instance, PSSI replaced CSI's simplistic digital filter with a more sophisticated joint probability model to improve its secondary structure identification accuracy. On the other hand, PsiCSI combined the basic CSI concept with a sequence-based secondary structure routine called PSIPRED (Jones 1999) to boost its performance. PLATON used a database consisting of reference chemical shift patterns from previously assigned proteins to improve its secondary structure calls, while PECAN employed a pseudo-energy model that combined sequence data with chemical shift data to more accurately identify secondary structure elements. Finally, 2DCSi used two-dimensional cluster analysis to analyze paired scattering diagrams of all six backbone chemical shifts to obtain improved secondary structure identification. All of these methods appear to achieve three-state secondary structure (Q3) accuracies better than 80%.

More recently, sophisticated chemical shift-based secondary structure assignment approaches that exploit machine-learning techniques, torsion angle estimates, sequence homology and far more extensive chemical shift-structure databases have appeared. These include TALOS+ (Shen et al. 2009), TALOS-N (Shen et al. 2013), DANGLE (Cheung et al. 2010) and Delta2D (Camilloni et al. 2012). Both TALOS+ and TALOS-N predict backbone torsion angles, as well as secondary structure locations by using neural networks to match chemical shift patterns over a five-residue window against a large database of previously assigned proteins with high-resolution structures. DANGLE exploits some of the same ideas as TALOS+ but employs Bayesian-inference techniques instead of neural nets to perform its analyses. Delta2D identifies secondary structure elements and secondary structure populations in both disordered and native-state proteins by analyzing the probability distribution of a very large database of backbone chemical shifts. In general, these newer approaches have average Q3 prediction accuracies between 83% and 86%.

With ongoing advances in machine learning and with continued improvements of our understanding of protein chemical shifts (Wishart 2011, Shen et al. 2012, Fesinmeyer et al. 2005), we believe that further improvements in shift-based secondary structure identification accuracy are possible. In particular, by making use of chemical shift information, sequence information and predicted backbone flexibility and then integrating this information using a multi-class Support Vector Machine (SVM) model (Weston et al. 1998), we found that it was possible to make statistically significant improvements (3-8%) in the accuracy of shift-based secondary structure assignments. Since this concept builds from our previous work on the Chemical Shift Index (CSI), we decided to call the new method CSI 2.0. The level of accuracy achieved by CSI 2.0 suggests that it could be used to assist with the initial stages of conventional NMR structure generation (i.e. fold identification via threading or providing useful torsion angle and distance restraints) as well as a robust alternative to standard coordinate-based methods for secondary structure identification.

2.2 Methods and Materials

2.2.1 Data set preparation

Training and testing data set

To construct the database needed to train and test our CSI 2.0 method, we chose a local, manually curated data set that we previously used to train and test the SHIFTX2 program (Han et al. 2011). An initial data set of ~300 X-ray protein structures with good quality NMR assignments was filtered based on following criteria: i) a resolution <2.1 Å, ii) largely monomeric, iii) free of bound DNA, RNA or large cofactors, iv) an average pairwise sequence identity < 33% to any other protein in the data set, v) nearly-complete (>90%) sequential assignment of ^{1}H, ^{13}C and/or ^{15}N backbone chemical shifts, and vi) must be a BMRB (Ulrich et al. 2008) entry. Several measures were taken to eliminate chemical shift re-referencing problems, check chemical shift quality and detect chemical shift outliers. A more detailed accounting of the data preparation protocol is provided in the SHIFTX2 paper (Han et al. 2011). The above selection and filtering process reduced the data set to 240 proteins. This data set was then divided into a training set and an independent test set. The training dataset consisted of 181 proteins (25,205 residues) whereas the test dataset contained 59 entries (8,078 residues). Among the training proteins, 146 proteins belonged to the α+β folding class, 15 proteins to the all-α, 18 proteins to the all-β and two proteins to the all-coil folding class. For the test proteins, 52 proteins had an α+β architecture, three were all-α and four were all-β. Note that there were no disordered proteins in the test set. The free parameters for the secondary structure assignment model were optimized on the training data set while the test set was used to perform an independent validation of the program's performance.

DSSP (Kabsch et al. 1983), STRIDE (Frishman et al. 1995) and VADAR (Willard et al. 2003) served as the three programs used to assign reference secondary structures ("α-helix", "β-strand", "coil") in both the training and test set proteins. These methods assign secondary structures based on the coordinates of the 3D structures as well as inferred H-bonds and torsion angles derived from those coordinates. The normal eight-state DSSP assignments were transformed into a three-state (helix, sheet, coil) assignment using the EVA convention (Eyrich et al. 2001). The same procedure was applied to the STRIDE output. No such transformation was required for the VADAR output. According to DSSP, there were a total of 2,335 β-strand residues (29%), 2,186 residues in α-helices (27%) and 3,557 coil assignments (44%) in the test set. STRIDE determined 2,499 residues as β-strands (31%), 2,677 as α-helices (33%) and 2,902 residues as coil structures (36%) in the test set. Finally, VADAR found 2,489 β-strands (31%), 2,720 α-helices (34%), and 2,869 coil structures (35%). According to DSSP, the training set had a total of 6,837 β-strand residues (28%), 7,588 residues in α-helices (29%) and 10,780 coil assignments (43%). STRIDE identified 7,368 residues in β-strands (29%), 8,857 residues in α-helices (35%), and 8,980 coil residues (36%). VADAR identified 7,196 β-strand residues (28%), 8,910 α-helical residues (36%), and 9,099 coil residues (36%).

Missing chemical shifts handling and neighbor residue correction

The completeness of a given protein's chemical shift assignments plays a crucial role in determining the performance of any chemical shift-based secondary structure assignment method (Shen et al. 2009). The current model is no exception. We assessed the performance of our CSI 2.0 program using both complete and incomplete shift assignments. Incomplete shift assignments were found to negatively affect the accuracy of the secondary structure assignments by up to 3%.

As mentioned in the previous section, because a small, but significant number (<10%) of chemical shift assignments were missing in some entries in our protein data set, we needed to take appropriate measures to handle the assignment gaps. This was done by searching through a sequence-chemical shift triplet database to fill in any missing assignments in a manner similar to that described by Shen et al. (2009). More specifically, each entry in our database was converted to an amino acid triplet and each had six backbone ($^{13}C_{\alpha}$, $^{13}C_{\beta}$, ^{13}C, ^{15}N, ^{1}HN, $^{1}H_{\alpha}$) experimental chemical shifts associated with it (except for Gly and Pro). To fill in the missing data, the query sequence triplet was compared with each triplet entry in the database and scored in terms of sequence and chemical shift similarity. The ten best scoring triplets were selected and the average of the ten central residue shifts was used as a proxy for the missing assignment. This process was repeated for all missing assignments (except $^{13}C_{\beta}$ for Glycine, ^{15}N and ^{1}HN for Proline)

Several studies have reported on the significant influence of the nearest neighbor residues on random coil chemical shifts (Wishart et al. 1998, Wang et al. 2002b, Wang et al. 2007). In particular, it has long been noted that the preceding amino acid type significantly affects the ^{15}N and amide proton chemical shift, while the ^{13}C and ^{1}H proton chemical shifts are largely affected by the identity of the following amino acid. Proper accounting for these nearest-neighbor effects is critical to accurately determining protein secondary and tertiary structures from chemical shift data (Wishart 2011). Hence, the random coil chemical shifts for all 20 amino acids were corrected by neighboring residue correction factors provided in Schwarzinger et al. (2001). Finally the secondary chemical shifts for all six-backbone atoms were calculated by subtracting the sequence-corrected random coil shift from the observed shift.

2.2.2 Feature Set

In developing any kind of machine-learning algorithm, it is necessary to extract a set of input features from the training data that will be used to infer or calculate the desired output (in this case, the secondary structure). Features can either be the raw data (e.g. sequence, chemical shifts, etc.) or derived data (e.g. estimated accessible surface area) that is calculated from the raw data. In developing CSI 2.0 we derived a set of eleven different features from our chemical shift and sequence data. These features included: 1) shift-derived β-strand propensity; 2) shift-derived helix propensity; 3) shift-derived coil propensity; 4) sequence-derived β-strand propensity; 5) sequence-derived helix propensity; 6) sequence-derived coil propensity; 7) random coil index (Berjanskii et al. 2005); 8) real-valued fractional accessible surface area; 9) two-state relative accessibility classification; 10) multi-sequence alignment-derived residue conservation score and 11) PSIPRED (Jones 1999) predicted secondary structure. Furthermore, for each data point in the protein sequence, a 5-residue window was evaluated, with the central residue being the residue of interest. This translates to a total of 55 features for each data point within the five-residue window, as each residue had 11 features. Note that all of the input features were derived from only the sequence and the backbone chemical shifts.

Secondary chemical shift-based probability of three-state secondary structure

The shift-based secondary structure probability of a residue is derived from the secondary chemical shift value of its constituent atoms. The secondary chemical shift ($\Delta\delta$) is defined as the difference between the absolute chemical shift (δ_{abs}) and the corresponding (neighbor-adjusted) random coil (δ_{rc}) shift (Wishart et al. 2011).

$$\Delta\delta = \delta_{abs} - \delta_{rc} \tag{2.1}$$

The probability of a residue being in one of the three secondary structure classes "α-helix", "β-strand" or "coil", is derived from its six backbone atom secondary chemical shifts, as described in Wang et al. (2002a). For each backbone atom, a Gaussian probability distribution is assumed, where the two parameters for the distribution (μ and σ) correspond to the average (μ) chemical shift value (for each of the three different secondary structure states) and the standard deviation (σ) of the chemical shift distribution respectively. These statistical parameters were derived from the "RefDB" database (Zhang et al. 2002). Therefore, given [$\Delta\delta_n$] {n = $^{13}C_{\alpha}$, $^{13}C_{\beta}$, ^{13}C, ^{15}N, ^{1}HN, $^{1}H_{\alpha}$}, the six experimental backbone secondary chemical shifts for a given residue *i*, the joint probability of being in one of three secondary structure states can be calculated from the Gaussian distributions of secondary chemical shifts of the six backbone atom types of non-Gly/Pro residues (five in case of Gly and four in case of Pro). The joint probability equation is formulated as:

$$P_i^s\ (\Delta\delta_n) = \rho \prod_n G_i^s\ (\Delta\delta_n) \tag{2.2}$$

where ρ represents the probability or likelihood for an amino acid of type i being in the secondary structure type s (s = ["α-helix", "β-strand", "coil"]), given its secondary chemical shifts. Note that this probability or likelihood ρ can also be described by amino acid conformational preference and is calculated using the same method described in the next paragraph. G_i^s represents the Gaussian distribution (see Eq. 2.3) of secondary backbone chemical shifts of a particular atom given amino acid type i and secondary structure type *s.* Note that for a given amino acid and secondary structure type, the six backbone secondary chemical shift distributions are independent and thus can be taken product over these distributions (Eq. 2.2).

$$G_i^s = \frac{1}{\sqrt{2\pi(\sigma_{i,n}^s)^2}} \exp(-\frac{(\Delta\delta_n - \overline{(\Delta\delta_n)}_{i,n}^s)^2}{2(\sigma_{i,n}^s)^2}) \tag{2.3}$$

where $\overline{(\Delta\delta_n)}_{i,n}^s$ corresponds to the average secondary chemical shift value (or μ) and $\sigma_{i,n}^s$ represents the standard deviation (or σ) of secondary chemical shift distribution of a particular atom *n,* given amino acid type *i* and secondary structure type *s.* The joint probability P^s for each residue is normalized so that its sum of three secondary structure types is equal to 1.0.

Sequence based probability of three-state secondary structure

The conformational preference for an amino acid is taken into account using this feature. Each amino acid has a predisposition to assume a specific secondary structure type, which is referred to as its conformational preference. We derived the secondary structure conformational preferences for all 20 amino acids using an in-house high-resolution sequence-structure database (the sequences and secondary structures in FASTA format are available on the CSI 2.0 website). This database contains 2100 X-ray structures that share no more than 33% sequence identity with each other, have an R-value ≤ 0.2 and a resolution ≤ 1.5 Å. These proteins were extracted using the PISCES server (Wang et al. 2003) via the PDB (Berman et al. 2000) and the secondary structures were assigned via DSSP (Kabsch et al. 1983). The conformational preference statistic was calculated as follows: given a residue *i*, and the available secondary structure conformation *s* (*s* = [“α-helix”, “β-strand”, “coil”]) that it can adopt, then the equation to calculate the conformational preference is given as (Levitt 1978):

$$C_i^s = \frac{T_i^s/T_i}{T^s/T} \tag{2.4}$$

where T_i^s denotes the total number of residues i adopting conformation *s,* while T_i is total number of residues of type i, T^s is the total number of times the conformation *s* observed in the database and T represents the total number of residues in the database. The conformational preference of each residue for three secondary structure types is then normalized so that its sum is equal to 1.0.

Random Coil Index (RCI) for backbone atoms

The Random Coil Index (RCI) for protein backbone atoms is an easily calculated measure that corresponds to the flexibility of an amino acid on a residue-level as derived from backbone chemical shifts (Berjanskii et al. 2005). The backbone RCI quantitatively traces the relative amount to which a protein backbone's chemical shifts match with the random coil values. Those that are closer to random coil values are the most flexible, while those that are most different from random coil values are least flexible. This feature was calculated using the RCI equation provided in the original RCI paper.

Relative accessible surface area

The solvent accessibility of a residue is a measure of an amino acid's (especially its side chain) solvent exposure. Generally unstructured coils or other highly hydrophilic regions are more accessible to water than hydrophobic helices or β-strands. This trend can be exploited to obtain useful information for identifying protein secondary structures. Recent publications suggest that including solvent accessibility along with sequence information can improve secondary structure prediction accuracy (Adamczak et al. 2005, Roknabadi et al. 2008). In an effort to include solvent accessibility in CSI 2.0, we developed a machine learning regression model that estimates real numerical value of each residue's fractional accessible surface area (fASA). The fASA is equal to the accessible surface area measured for a given residue (X) in a protein divided by the ASA for that residue in a *Gly-X-Gly* tripeptide. The fASA varies between 0.0 (fully buried) to 1.0 (fully exposed). The regression model we developed uses two sequence derived features (hydrophobicity and sequence conservation score) and two chemical shift-derived features (3-state structural probability using six backbone chemical shifts and the random coil index) to calculate the fASA. The model was trained on a dataset of 28 proteins with known 3D coordinates and near-complete ^{1}H, ^{13}C and ^{15}N chemical shift assignments and validated on a test set of 66 proteins (with known 3D coordinates and near-complete chemical shift assignments). The fASA for all training and test proteins was calculated using VADAR (Willard et al. 2003). The correlation between the observed fASA and the predicted fASA was 0.76. This fASA value was then incorporated into the CSI 2.0 feature set in the same manner as all other features.

Additional details regarding this shift-based fASA prediction method, its performance and its potential applications will be described in a forthcoming manuscript.

Two-state buried-exposed class

The two-state buried-exposed classification assignment is simply a transformation of the fractional ASA (fASA) into two discrete classes obtained by applying a 25% fASA cutoff. In other words, if the fASA is greater than 0.25, the residue is assigned to an "exposed" state, otherwise the residue is said to be "buried". This information was derived from the chemical shift-based fASA calculation described above.

Residue conservation score

Sequence conservation is a measure of how frequently a given residue is seen at an equivalent position, in an equivalent protein, across different species. Generally highly conserved residues are buried within the protein's core, and less conserved residues are more exposed (albeit with some exceptions). The conservation score for each residue position can be calculated as described by Valder (2002). First, a three-iteration PSI-BLAST (Altschul et al. 1997) search is performed on the UniRef90 clustered database (UniProt Consortium 2010). From the identified hits a multiple sequence alignment is then performed using ClustalOmega (Sievers et al. 2011). The conservation score for each non-gap column in the alignment (i.e. each residue in the target sequence) is then calculated using Shannon's entropy formula as described below,

$$s(x) = \lambda \sum_{a}^{K} p_a \log p_a \qquad (2.5)$$

where p_a is the probability of observing the a-th amino acid and λ is the scaling factor, which is defined as,

$$\lambda = [\log(\min(N, K))]^{-1} \qquad (2.6)$$

where N = number of residues in the alignment, K = 20 (length of the amino acid alphabet). The probability of observing the a-th amino acid is the summed weight of sequences having the symbol a in the position x in the sequence which is defined as,

$$p_a = \sum w_n \tag{2.7}$$

where w_n is the weight of the *n*-th sequence with w_n being defined as,

$$w_n = \frac{1}{L} \sum_x^L \frac{1}{k_x n_x} \tag{2.8}$$

where L = length of the alignment, k_x = the number of amino-acid types present at the *x*-th position, n_x = the number of times the a-th amino acid occurring in the *n*-th sequence at the *x*-th position.

PSIPRED predicted secondary structure

In an effort to boost the performance of CSI 2.0 we supplemented our method with another powerful secondary structure identification tool called PSIPRED (Jones 1999). PSIPRED is a pure sequence-based secondary structure prediction method developed in the 1990s. It has been refined and improved upon over the last decade and is generally considered one of the most accurate sequence-based prediction methods available, with a typical performance of >80% (Hung et al. 2002). Previous authors have observed a slight boost to the performance of their shift-based secondary structure assignment routines by including this information in their algorithm (Hung et al. 2002). Therefore, we also added a PSIPRED (sequence-based) prediction as one of the features to CSI 2.0. Therefore, PSIPRED (version 3.3) predicted secondary structure state for each residue is included in the CSI 2.0 feature vector for the training and test data points.

2.2.3 Feature normalization

A z-score normalization step was done to normalize the features in the training and the test data set. Assuming there are *m* (= R_1, ..., R_i, ..., R_m) rows in the training set, with each row corresponding to a particular data point and containing *n* different features (columns), then the normalized value of r_i^j for row R_i in the *j*-th column is calculated as:

$$Normalized(r_i^j) = \frac{r_i^j - \overline{R_i}}{std(R_i)} \tag{2.9}$$

where,

$$\bar{R}_\iota = \frac{1}{n}\sum_{j=1}^{n} r_i^j \quad \text{and}$$

$$std(R_i) = \sqrt{\frac{1}{(n-1)}\sum_{j=1}^{n}(r_i^j - \bar{R}_\iota)^2}$$

All test data points were normalized using the mean and standard deviation derived from the training feature distribution.

2.2.4 Multi-class SVM training

With a five-residue window, there were total 25,205 data points in our training set. All data points were normalized prior to the training. Two different normalization methods, a “Statistical Z-score” and a “Max/Min” score were assessed, with the “Statistical Z-score”, as described in section 2.2.3, ultimately being selected. A multi-class “kernelized” SVM classifier was used to train the model. For the classification of each data point, the multi-class SVM classifier fit three binary (two-class) sub-classifiers namely helix vs. strand, helix vs. random-coil and strand vs. random-coil classifiers and found the predicted class by a majority voting mechanism. For each binary classifier, a soft-margin classification approach was used. A soft-margin SVM classifier generally produces a wide decision- margin to separate the two classes by allowing some margin violations (i.e. permitting some noisy samples to be inside or the wrong side of the margin) in order to achieve a better generalization performance. In our “kernelized” SVM model, a Radial Basis Function (RBF) kernel (also known as Gaussian kernel) was used to compute the similarities (by computing the dot product) between two feature vectors in a higher dimensional space without explicitly computing the vectors in that space. With a RBF kernelized soft-margin classification framework, the performance of the SVM depends on two parameters: 1) the regularization parameter “C” (also known as the “cost” factor) and 2) the Gaussian kernel width “σ”. The “C” parameter allows one to adjust the trade-off between maximizing the decision-margin width vs. minimizing the number of misclassified samples in the training set. More specifically, large “C” values will cause a lower number of outliers, producing a narrower decision-margin hyperplane, whereas a small “C” will allow a large number of violations, resulting in a wider decision hyperplane. The “σ” parameter controls the width of Gaussian kernel and can be adjusted to achieve a smoother fit of the model. Both “C” and “σ” parameters were optimized using a repeated 10-fold cross validation (CV) on the training data. The goal of the parameter optimization was to find the optimal values that maximize the accuracy or Q3 score of the three-class secondary structure classification. The Multi-class SVM implementation in the R package "*kernlab*" was used to train the classifier. The optimization of "C" and "σ" through the "*repeatedcv*" method was performed using the *train()* function in the "caret" package in R.

2.2.5 A multi-residue Markov Model for post-assignment filtering

While the SVM classifier (described above) generally performs very well, it is still prone to making confusing, meaningless or "scrambled" secondary structure assignments such as: *CCBHH* or *BBHCC or HCHCH.* This is also a common problem for many other secondary structure prediction/assignment methods such as PSIPRED, TALOS-N or DANGLE. Most programs use heuristic "character smoothing" that employ "if-then-else" ladders or character averaging to correct or eliminate these problem assignments. However, these heuristic methods are not very robust nor are they very accurate. A more robust method to perform character smoothing or character correction is to use a Markov model (Durrett 2010). Markov or hidden Markov models are widely used methods for text filtering, pattern extraction and natural language processing. This also makes them ideally suited to treating the "scrambled" text problem. After assessing character window widths of three, five, seven and nine residues, a seven-residue Markov model was found to be optimal to handle scrambled or discontinuous segments of secondary structure. This Markov filtering involved sliding a trained, seven-residue Markov filter along the protein chain that identified scrambled secondary structure assignments and then corrected them as necessary. According to this multi-residue Markov model, if there are *n* residues i.e. $[t_1, t_2, \dots, t_n]$ in a single pattern along the protein chain, then the probability of observing *i*-th residue in that pattern depends on the observed probabilities of the preceding *(i-1)* residues. This can be expressed by the following equation,

$$P(t_1, t_2, \dots, t_n) = \prod_{i=1}^{n} P(t_i \mid t_1, \; t_2, \dots, t_{i-1}) \qquad (2.10)$$

The conditional probability of observing a residue in *i*-th location given the history of the preceding *(i-1)* residues is calculated from $[t_1, \; t_2, \dots, t_{i-1}]$ and $[t_1, \; t_2, \dots, t_i]$ pattern frequency counts.

$$P(t_i \mid t_1, \; t_2, \dots, t_{i-1}) = \frac{count\;(t_1, \dots, t_{i-1}, \; t_i)}{count(t_1, \; \dots, t_{i-1})} \qquad (2.11)$$

To calculate the probability of a seven-residue pattern, the frequencies of smaller patterns consisting of one, two, three, etc. up to six residues are extracted from the training database of reference structures. An example formula to calculate the probability of a five-residue pattern $HHBCC$ is as follows:

$$P(HHBCC) = \ P(C|HHBC) * \ P(C|HHB) * P(B|HH) * \ P(H|H) * P(H) \qquad (2.12)$$

and a probability value like $P(C|HHBC)$ can be calculated by following equation,

$$P(C|HHBC) = \frac{count(HHBCC)}{count(HHBC)} \qquad (2.13)$$

If the denominator in Eq. 2.13 (i.e. the count for a specific pattern in the database), is found to be 0, then it is set to minimum value of 1 to avoid the divide-by-zero error in calculating the probability on the left hand side. The probability cutoff to validate a pattern is chosen as 0 (i.e. if the probability of a multi-residue pattern, along with its two preceding and following patterns is found to be equal to the cut-off value, then the central pattern is considered to be "scrambled"). For a scrambled secondary structure pattern to be identified, the outlier must be either in the middle, or any of the two adjacent positions to the middle. The outlier is then corrected by looking at the secondary structure assignments of the four neighbor residues.

2.2.6 Evaluation metrics

Q3-accuracy

Q3-accuracy is the most widely used metric to evaluate three-state secondary structure predictions or assignments. It is the ratio of correctly predicted or identified states divided by the total number of amino acids or residues in the dataset. Q3-accuracy is simply defined as:

$$Q3 = \frac{N_p}{N} \qquad (2.14)$$

where N_p is the total number of residues for which secondary structure state is predicted correctly by the model and N is the total number of residues in the example set.

Segment-Overlap (SOV) score

The Segment-OVerlap score (SOV) is based on the average overlap between the observed and predicted segments. It is designed to evaluate the correctness of segment prediction with respect to a reference assignment (Rost et al. 1994, Zemla et al. 1999). The SOV score measures how much the predicted segments deviate from observed (experimental) segment length distributions. The definitions of the SOV score for single and multi-class secondary structure assignments are adapted from Zemla et al. 1999. Assuming, (s_1, s_2) represents a pair of overlapping segments of secondary structure in conformational state *c*, where $c \epsilon (H, B, C)$ and where s_1 and s_2 are the observed and predicted secondary structure segments. S_c is the set of all overlapping pairs of segments (s_1, s_2) in state *c* and is defined as:

$$S_c = \{(s_1, s_2): s_1 \cap s_2 \neq \emptyset, s_1 \text{ and } s_2 \text{ are both in conformational state } c\}$$

The complement of S_c or S'_c is the set of all segments s_1 for which there is no overlapping segment s_2 in state *c* and can be formulated as,

$$S'_c = \{s_1: \forall s_2, s_1 \cap s_2 = \emptyset, s_1 \text{ and } s_2 \text{ are both in conformational state } c\}$$

$$SOV_c = \frac{1}{N_c} \sum_{S_c} \left[\frac{minOV(s_1, s_2) + \delta(s_1, s_2)}{maxOV(s_1, s_2)} \times len(s_1) \right] \tag{2.15}$$

with the normalization term N_c defined as,

$$N_c = \sum_{S_c} len(s_1) + \sum_{S'_c} len(s_1) \tag{2.16}$$

The sum in Eq. 2.15 is taken over all the segment pairs in state *c*, which overlap by at least one residue. In the same equation, $len(s_1)$ is the number of residues in the segment s_1, $minOV(s_1, s_2)$ is the length of actual overlap of s_1 and s_2, $maxOV(s_1, s_2)$ is the length of the total extent for which either s_1 or s_2 segment has a residue in *c*-th state and $\delta(s_1, s_2)$ is defined as,

$$\delta(s_1, s_2) = \min \left\{ \begin{matrix} maxOV(s_1, s_2) - minOV(s_1, s_2) \\ minOV(s_1, s_2) \\ int(0.5 \times len(s_1)) \\ int(0.5 \times len(s_2)) \end{matrix} \right\} \tag{2.17}$$

The SOV measure defined in Eq. 2.15 can be extended for multi-class secondary structure assignments (α-helix (*H*), β-strand (*B*) and Coil (*C*)). It is denoted as SOV_{all} and can be defined as,

$$SOV_{all} = \frac{1}{N}\left(\sum_{i\in H,B,C}\sum_{S_i}\frac{minOV(\,s_1,s_2\,) + \,\delta\,(s_1,s_2)}{maxOV(s_1,s_2)} \times\, len(s_1)\right) \times\, 100 \tag{2.18}$$

where the normalization term *N* is a sum of N_c over all three conformational states and can be calculated as,

$$N = \sum_{c\epsilon(H,B,C)} N_c \tag{2.19}$$

2.3 Results and Discussion

As described earlier in the section 2.2.4, the "C" parameter (the "cost" value) in the SVM classifier and the kernel parameter, "*σ*" in the Gaussian RBF kernel, were optimized using 10-fold Cross Validation (CV). After achieving an optimal value of 0.0157, "*σ*" was held constant while "C" was iteratively changed to optimize its value. To achieve an unbiased training result, the *n*-fold cross validation (*n*=10) process was repeated five times. For each repetition, the accuracy of the three-state assignment of the training classes was measured. The optimal "cost" and "*σ*" values that were found to maximize the Q3 accuracy using this repeated training were 2.0 and 0.0157 respectively. The training accuracy was averaged over five repetitions of the 10-fold CV process. A training accuracy of Q3=90.56% on 181 training proteins was observed with the aforementioned optimized parameter values. A test accuracy of Q3=89.35% was achieved on an independent test set of 59 proteins.

The final set of weighting coefficients for the sequence and chemical shift-based features in our multi-class SVM model are listed in Table 2.1. The sum of all the weights (over the five residue positions) for chemical shift-derived features was 683 while the sum of all the weights for the sequence-derived features was 202 (a difference of 3.4X). Among individual features, the PSIPRED predicted secondary structure for the central residue (residue *i*) was found to have the largest single weight in the SVM formulation (|w|=58.0). The second largest weighted feature (|w|=54.0) was the β-strand propensity calculated from backbone chemical shifts at the central residue location. Chemical shift derived α-helix, β-strand and coil probability scores in the central residue or immediate neighbor locations were found to be moderately relevant in terms of their weighting. Both protein flexibility (RCI) and solvent accessibility (fASA) at the (*i-1*) location had larger weights than the same feature values at other residue positions.

	Weight Coeff.	**Feature**	**Weight Coeff.**
ProbBCS(i-2)	19.88598	ProbCAA(i+1)	14.47068
ProbBCS(i-1)	47.83482	ProbCAA(i+2)	12.65604
ProbBCS(i)	53.96416	RCI(i-2)	11.71170
ProbBCS(i+1)	21.52516	RCI(i-1)	29.60009
ProbBCS(i+2)	9.230448	RCI(i)	19.74642
ProbHCS(i-2)	8.925556	RCI(i+1)	14.32193
ProbHCS(i-1)	29.11158	RCI(i+2)	7.806794
ProbHCS(i)	25.97708	RSA(i-2)	4.939002
ProbHCS(i+1)	5.456130	RSA(i-1)	30.64792
ProbHCS(i+2)	14.17828	RSA(i)	22.54183
ProbCCS(i-2)	12.56523	RSA(i+1)	14.22241
ProbCCS(i-1)	19.28457	RSA(i+2)	10.04285
ProbCCS(i)	31.72981	BuriedExposed(i-2)	0.931595
ProbCCS(i+1)	19.65260	BuriedExposed(i-1)	14.87281
ProbCCS(i+2)	8.465084	BuriedExposed(i)	1.417325
ProbBAA(i-2)	2.680083	BuriedExposed(i+1)	15.17127
ProbBAA(i-1)	21.07358	BuriedExposed(i+2)	1.813477
ProbBAA(i)	9.595489	Scon(i-2)	10.84571
ProbBAA(i+1)	23.39969	Scon(i-1)	10.29277
ProbBAA(i+2)	7.901136	Scon(i)	18.86649
ProbHAA(i-2)	3.968213	Scon(i+1)	8.360907
ProbHAA(i-1)	6.390549	Scon(i+2)	22.17609
ProbHAA(i)	3.372710	PSIPRED(i-2)	17.87049
ProbHAA(i+1)	8.024158	PSIPRED(i-1)	28.21943
ProbHAA(i+2)	8.566371	PSIPRED(i)	57.99818
ProbCAA(i-2)	5.022616	PSIPRED(i+1)	20.72087
ProbCAA(i-1)	22.44050	PSIPRED(i+2)	6.860726
ProbCAA(i)	5.881907		

Table 2.1: Weighting coefficients (|**w**|) of chemical-shift and sequence-derived features for CSI 2.0's SVM model. The position of the feature over a five-residue window is given using standard indices in parentheses. The feature name abbreviations are as follows: ProbBCS = β-strand probability using chemical shift, ProbHCS = α-helix probability using chemical shift, ProbCCS = coil probability using chemical shift, ProbBAA = β-strand probability using amino acids, ProbHBAA = α-helix probability using amino acids, ProbCAA = coil probability using amino acids, RCI = Random Coil Index (protein flexibility), RSA = fractional or real-valued solvent accessibility, BuriedExposed= 2-state (Buried/ Exposed) solvent accessibility, Scon = residue conservation score, and PSIPRED = PSIPRED predicted secondary structure.

Interestingly the RCI and fASA weightings also proved to be more important than the sequence conservation scores. Given the 3.4X greater weight attached to shift-derived features in CSI 2.0's final SVM model,we believe it is fair to claim that CSI 2.0 is essentially a chemical shift based method that incorporates a small amount of sequence information. This assertion is also borne out by the fact that the Q3-accuracy of CSI 2.0 (without the sequence-based prediction) was only 2% worse than the version with sequence-based prediction (data not shown).

CSI 2.0 comparative performance

In Table 2.2, we compare the performance of our CSI 2.0 method with seven hybrid (chemical shift and sequence-based) and one pure sequence-based secondary structure identification/prediction programs. The eight programs are: TALOS+ (Shen et al. 2009), TALOS-N (Shen et al. 2013), DANGLE (Cheung et al. 2010), CSI (Wishart et al. 1992), PSSI (Wang et al. 2002a), Delta2D (Camilloni et al. 2012), Psi-CSI (Hung et al. 2002) and PSIPRED (Jones 1999). The performance of all eight programs was evaluated on the basis of: (i) Q3-accuracy of predicted three different structure states; (ii) individual structural state ("α-helix", "β-strand", "coil") prediction accuracy; (iii) Segment-Overlap or SOV score; and (iv) coverage (proportion of residues in the test set that were predicted). For Table 2.2, the first column indicates the name of the prediction model, while the second, third, and fourth columns indicate the accuracy or precision for each category of secondary structure. The fifth column presents the overall Q3-accuracy, while the last four columns indicate the individual and overall SOV-scores. The last column shows the percent coverage (proportion of residues of test data that were identified) by each method. As seen in this table, CSI 2.0 achieves the best overall Q3 and SOV scores while Psi-CSI and TALOS-N are essentially tied for second in their overall performance. With regard to the performance for individual secondary structure state (helix, sheet, coil) identification, CSI 2.0 also shows superior accuracy for all three-structure states. In particular, for DSSP-referenced structures, CSI 2.0's performance was an average of 10.87% better in case of β-sheet identification, and 8.59% better for coil identification, than the eight other chemical shift and sequence-based methods (see Table 2.2).

Methods	**Helix**	**Beta**	**Coil**	**Mean Q3-score**	**Min/ Max Q3-score**	**SOV (Helix)**	**SOV (Beta)**	**SOV (coil)**	**SOV (all)**	**% Cove-rage**
TALOS+	93.39	77.93	80.34	83.89±6.1	58.0/96.0	80.09	80.73	83.58	84.83	97.80
TALOS-N	95.54	82.65	79.08	86.39±6.1	58.0/96.0	88.78	85.71	83.08	87.85	97.70
DANGLE	95.88	80.0	76.44	83.00±5.1	68.0/95.0	80.61	80.58	81.46	83.66	98.60
CSI	84.17	67.40	84.23	80.33±6.4	53.3/89.0	76.01	69.34	71.53	75.18	100
PSSI	62.85	70.77	62.58	67.33±16.7	0.0/87.1	59.49	73.39	72.62	71.37	96.80
δ2D	43.29	33.17	36.73	42.24±9.4	0.0/90.5	42.58	38.20	42.28	42.82	48.24
Psi-CSI	92.88	80.0	85.53	86.20±5.2	73.0/96.0	89.02	81.43	83.06	86.94	100
PSIPRED	85.95	79.17	88.11	85.36±6.0	70.0/98.0	72.08	63.74	63.00	79.00	100
CSI 2.0	**93.41**	**86.50**	**87.80**	**89.35±3.9**	**79.0/97.0**	**90.76**	**85.34**	**82.75**	**88.45**	**100**

Table 2.2: Performance of CSI 2.0 (shown in bold) and eight other chemical shift and sequence-based methods on an independent test set of 59 proteins (total 8,078 residues) when using "DSSP" (Kabsch et al. 1983) secondary structure assignments as the reference structure. The reported Q3-accuracies in the corresponding publications of the eight methods are following: TALOS+= 91.0% (Shen et al. 2009), TALOS-N= 88.6% (Shen et al. 2013), DANGLE= 85.2% (Cheung et al. 2010), CSI= 92.0% (Wishart et al. 1992), PSSI= 88.0% (Wang et al. 2002a), δ2D= 86.4% (Camilloni et al. 2012), Psi-CSI= 89.0% (Hung et al. 2002), PSIPRED= 78.3% (Jones 1999)

For helix identification, the CSI 2.0 shows a comparable performance with respect to other methods. In terms of the SOV measure, the same trend is observed. Although the Q3 accuracy of CSI 2.0's residue-specific helix assignments was not much better than existing programs, its higher average SOV-score indicates a better agreement of predicted helix segments. The same is true for overall SOV-score for all three-secondary structure types.

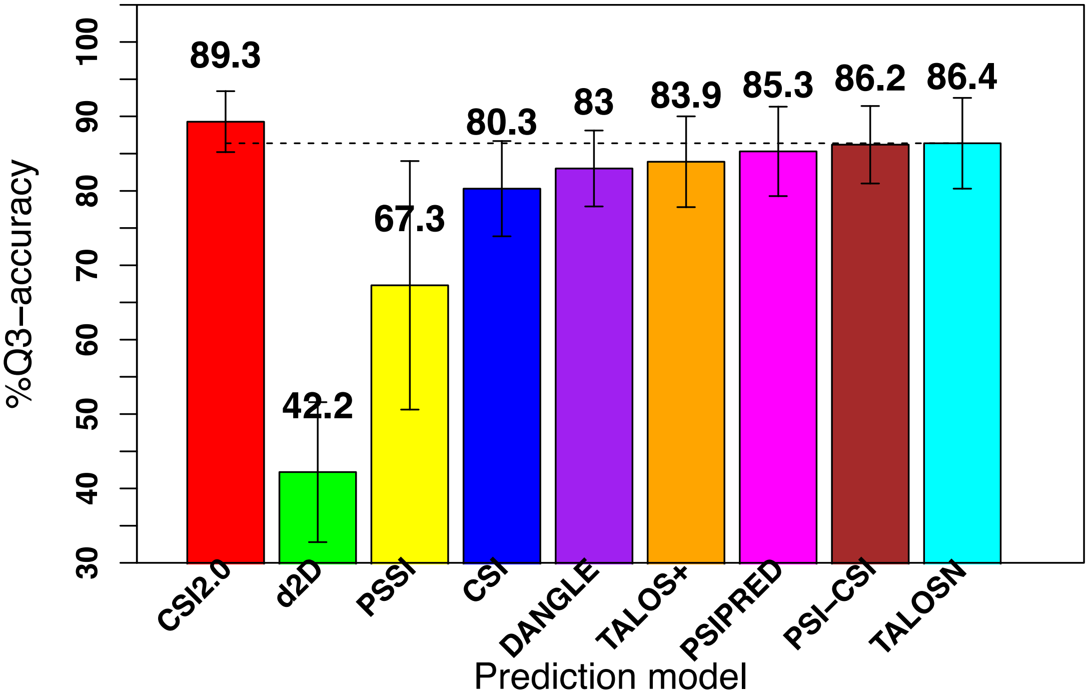


Figure 2.1: A bar graph comparing CSI 2.0's Q3 accuracies with eight other chemical shift and sequence-based protocols over an independent test set of 59 proteins. The error bar (i.e. standard deviation in Q3- accuracy distribution of each method) appears on top of each bar plot.

In the case of the SOV-score, the next best performance was seen for the most recent program, TALOS-N (Shen et al. 2013). CSI 2.0's assignments, unlike most of other programs, covers the full fraction (≈ 100%) of the test data points.

Statistical significance of CSI 2.0's improvement

As indicated in Table 2.2 and Figure 2.1, the best-performing methods all achieve Q3 accuracies above 80% and the difference between CSI 2.0 and the other top performing programs is only 3-4%. One may ask is this performance improvement statistically significant? To address this question we performed a Student's t-test to assess the p-value between CSI 2.0 and TALOS+, TALOS-N, DANGLE and Psi-CSI. The results are shown in Table 2.3.

Method1 vs. Method2	**p-value**
CSI 2.0 vs. TALOS+	6.575e-08
CSI 2.0 vs. TALOS-N	0.0016
CSI 2.0 vs. DANGLE	6.347e-08
CSI 2.0 vs. Psi-CSI	0.00087

Table 2.3: The p-values or probabilities of Student's two sample t-tests between CSI 2.0 and four other best performing methods are shown. Here the null hypothesis is that the difference between sample1 mean (mean accuracy of method1) and sample2 mean (mean accuracy of method2) is equal to zero. Alternative hypothesis indicates that the sample1 mean is greater than the sample2 mean.

These data confirm that the performance improvement seen in CSI 2.0 is, indeed, highly significant, with most p-values being << 0.001.

CSI 2.0 performance using selected and partial shift assignments

It is not particularly common for a protein to have all ^{1}H, ^{13}C and ^{15}N backbone shifts fully assigned. Indeed, many shorter peptides and proteins will only have their ^{1}H assignments completed, while larger proteins may only have their ^{1}H and ^{15}N shifts, ^{15}N and ^{13}C shifts or ^{1}H and ^{13}C shifts assigned. Given that only certain nuclei may be measured we decided to evaluate CSI 2.0's performance using only selected sets of chemical shifts or selected nuclei. The results are listed in Table 2.4 for five individual backbone nuclei ($^{13}C_\alpha$, ^{13}C, $^{13}C_\beta$, $^{1}H_\alpha$, ^{15}N) along with other common assignment combinations ($^{13}C_\alpha$, ^{13}C, $^{13}C_\beta$, $^{1}H_\alpha$ and $^{1}H_\alpha$, ^{15}N). As can be seen from this table, combinations of multiple nuclei give the best performance, but the performance for any single nucleus is surprisingly good (Q3 > 85%). This is because CSI 2.0 also uses sequence information (i.e. PSIPRED predicted secondary structures) to supplement its chemical shift-derived estimates. As was noted in the original CSI papers (Wishart et al. 1992 & 1994), certain nuclei carry more information about secondary structures than others. In particular, the ranking of nuclei for secondary structure information content, from most informative to least informative, is: $^{13}C_\alpha > {}^{13}C > {}^{13}C_\beta \approx {}^{1}H_\alpha > {}^{15}N$.

Shift Assignment	Helix	Beta	Coil	All
$^{13}C_{\alpha}$	84.38	93.36	87.90	88.72
^{13}C	83.41	90.22	87.95	87.37
$^{13}C_{\beta}$	84.37	87.03	86.59	86.50
$^{1}H_{\alpha}$	83.59	87.22	87.21	86.83
^{15}N	82.63	85.74	86.68	85.68
$^{13}C_{\alpha,}$ ^{13}C, $^{13}C_{\beta,}$ $^{1}H_{\alpha}$	81.46	92.89	98.98	90.92
$^{1}H_{\alpha,}$ ^{15}N	81.47	92.45	98.99	90.77

Table 2.4: CSI 2.0 performance with selected chemical shift assignments and combinations of shift assignments

Because it is often difficult to obtain complete chemical shift assignments for a protein (due to signal broadening from intermediate exchange events, signal overlap, solvent suppression, etc.) we were also interested to see how well CSI 2.0 performed with partial or incomplete chemical shift assignments. To do so we evaluated the performance of CSI 2.0 relative to the percentage of missing chemical shift assignments and compared its results to several other software packages. We analyzed a subset of 21 proteins (from our test set of 59 proteins) with a fraction of missing experimental shift assignments >15%. In particular, for this set, the percentage of incomplete or missing backbone ^{1}H, ^{13}C or ^{15}N assignments ranged from 16.7% to 37.0% (based on the total number of expected NMR signals from the protein's amino acid sequence). The secondary structures of these proteins were then determined using five different methods (including CSI 2.0) and evaluated against the observed secondary structures as determined by DSSP. The results are shown in Table 2.5. As can be seen from this table, CSI 2.0 does significantly better (~7-10%) in terms of Q3 accuracy than any of the other methods in terms of handling missing shift data.

PDB	BMRB	Percent Missing Shifts	CSI 2.0	Psi-CSI	TALOSN	TALOS+	DANGLE
1HQ2	4300	27.60	92.76	83.55	71.71	69.08	77.68
1T8L	5358	21.88	96.36	83.64	83.64	83.64	80.00
1JTG	6357	20.81	85.27	78.29	75.97	77.13	81.01
1UDR	4083	18.23	95.04	86.78	90.08	90.91	81.82
1ODV	6321	18.14	84.00	83.00	87.00	83.00	84.00
1W80	6034	23.24	88.74	76.62	75.76	73.16	74.03
1V9T	4037	16.68	86.50	82.82	81.60	79.75	76.07
2AOJ	5967	34.65	85.26	71.58	85.26	82.11	82.11
1CWC	2208	17.64	91.98	76.54	82.10	77.16	81.48
1YKY	4831	35.14	92.97	82.81	81.25	68.75	80.47
1KDB	6250	21.98	84.78	83.70	67.39	65.22	77.17
2A38	5760	37.01	91.62	85.86	74.35	79.06	79.06
256B	6560	17.68	93.07	92.08	94.06	95.05	93.07
1BT5	6024	19.32	88.03	81.85	84.17	81.08	78.38
1SGZ	6016	23.40	78.43	69.68	57.14	53.94	65.01
1SYD	15232	22.17	91.38	82.76	82.76	81.90	78.45
1JR2	7242	18.30	88.85	83.46	86.15	83.85	83.85
1U7B	15501	20.12	90.73	83.47	84.27	80.24	79.44
2A0N	15741	20.45	89.60	83.60	88.40	84.00	82.00
2DYI	10139	24.69	86.84	73.03	78.29	72.37	68.42
1B1H	10053	18.02	86.62	81.10	81.95	77.92	79.83
Average		**22.72±6.0**	**88.99±4.2**	**81.25±5.3**	**80.63±8.2**	**78.06±8.9**	**79.21±5.6**

Table 2.5: Comparison of the performance of CSI 2.0 (shown in bold) versus other four methods (Psi-CSI, TALOSN, TALOS+, DANGLE) relative to the percentage of missing backbone ^{1}H, ^{13}C or ^{15}N chemical shift assignments.

Furthermore, for all of the methods (except CSI 2.0) there is a general trend (r<0.5) showing a degradation in their performance with an increasing fraction of missing chemical shifts. Interestingly, CSI 2.0 seems to be largely immune to any detectable performance degradation with respect to missing chemical shifts (at least up to a level of ~35% missing shifts). This appears to be due to its robust handling of missing shift data (described earlier) as well as its use of sequence-based secondary structure prediction from PSIPRED.

Different definitions of secondary structure

Secondary structure is not an absolute quantity nor is it universally defined. In other words, there is no gold standard for secondary structure. Different definitions exist of helices, β-strands, β-turns and coils (Zhang et al. 2008). As a result, no two individuals and no two coordinate-based secondary structure assignment programs will agree on the exact start and end locations of many secondary structure elements (Tyagi et al. 2009, Shen et al. 2009). Likewise some programs (or some individuals) will invariably classify short helices and short β-strands as coil structures and vice versa. Given the variation in secondary structure "calling" from well-defined 3D structures and the fact that there are several different secondary structure identification algorithms that are widely used by structural biologists, we decided to investigate the performance of CSI 2.0 and the other eight programs against three of the most commonly used coordinate-based secondary structure assignment algorithms: DSSP (Kabsch et al. 1983), STRIDE (Frishman et al. 1995) and VADAR (Willard et al. 2003). Table 2.6 lists the Q3-accuracies of the eight secondary structure prediction/identification programs when compared against the calls made by locally installed versions of DSSP, STRIDE and VADAR. As can be seen in this table, CSI 2.0 agrees best with the DSSP secondary structure assignments while its performance drops slightly with the STRIDE or VADAR calls. The same trend is seen with the other eight programs as well. This is largely due to the fact that essentially all of these programs were trained using DSSP data, as opposed to STRIDE or VADAR data. It is worth noting that PSIPRED (which is used by both Psi-CSI and CSI 2.0) was also trained exclusively on DSSP data. Attempts to train CSI 2.0 with STRIDE or VADAR secondary structure calls yielded no overall improvement in the performance.

Assignment Method	**DSSP**	**STRIDE**	**VADAR**
TALOS+	83.89	82.07	81.47
TALOS-N	86.36	85.62	83.89
DANGLE	83.0	81.40	81.0
CSI	80.33	74.64	76.29
PSSI	67.33	65.15	64.0
δ2D	42.24	40.66	41.16
Psi-CSI	86.20	83.53	82.17
PSIPRED	85.36	79.81	78.0
CSI 2.0	**89.35**	**86.72**	**86.10**

Table 2.6: Percentage Q3-accuracies of the CSI 2.0 protocol (shown in bold) and eight other methods over an independent test set of 59 proteins using three different reference (DSSP (Kabsch et al. 1983), STRIDE (Frishman et al. 1995) and VADAR (Willard et al. 2003)) structures.

It is also interesting to note that the pairwise agreement between the three different secondary structure assignment methods (DSSP, STRIDE and VADAR) in our independent test dataset ranged from 85-90% with an average pairwise agreement of 87.63%. Furthermore, the overall agreement between all three methods was only 82%. This suggests that secondary structure identification is inherently imprecise and that the best possible performance that a secondary structure identifier (or predictor) could attain is probably no better than 90%. Given that all of the proteins we studied had both X-ray structures and NMR structures, we also investigated the level of agreement between the secondary structure assigned via more conventional NMR approaches (NOEs, J-couplings) or via author-assigned secondary structure assignments with those generated from the coordinate data (determined by DSSP, VADAR or STRIDE). Among the coordinate–based assignment methods, STRIDE showed the highest level of agreement (90.05%) with the author assignments, while DSSP and VADAR had slightly lower levels of agreement (88.54% and 84.54% respectively). Again, this level of agreement between secondary structure assignment methods (human vs. computer) suggests that CSI 2.0 is performing near the maximum level of accuracy achievable for secondary structure assignment.

Local interaction effects

Regular secondary structure is formed when the local environment induces nearby residues to interact and adopt a specific pattern such as an "α-helix" or a "β-strand". Hence, local interactions and nearest neighbor data (such as nearby shifts and amino acids) can provide important information about the secondary structure propensity of a certain region. To capture these local interaction effects, we assessed CSI 2.0's performance using several different residue window lengths (three, five and seven residues). Our data indicated that CSI 2.0 achieved its best performance, in terms of Q3-accuracy, when using a five-residue window (data not shown for other windows). No significant improvement was achieved by including more than four neighbors (two preceding and two following). This indicates that the features of immediately nearby residues provide the most useful secondary structure information.

Mis-assigned secondary structures

As accurate as CSI 2.0 appears to be, it still exhibits less-than-ideal performance with regard to distinguishing between β-strands and coil regions. In our test data set, there were a total of 2,335 residues in β-strands, of which CSI 2.0 correctly identified 2,019 of them (see Table 2.7). However, it also mis-identified 316 residues as "coils", or about 13.5% of the β-strand population. On the other hand, a somewhat smaller percentage of coil residues (7.9%) were also incorrectly identified as being in β-strands. The probable reason for this is the high degree of chemical shift and amino acid compositional similarity between these two structure types. Indeed, the chemical shifts in β-strands and coil regions tend to exhibit more similarity to each other than to helices. Furthermore, as we discovered on further inspection, many of the mis-identifications occurred at the borders or edges of β-strands and coil regions.

Secondary Structure	**H_{pred}**	**B_{pred}**	**C_{pred}**	**Total**
H_{obs}	2,043	0	143	2,186
B_{obs}	0	2,019	316	2,335
C_{obs}	281	153	3,123	3,557

Table 2.7: Confusion Matrix of secondary structure assignments generated by CSI 2.0 on the independent test set of 59 proteins.

While some ambiguity or mis-identification would be expected between the borders of secondary structure elements or short β-strands and extended coil regions, one would hope that there would be no ambiguity between β-strands and helical regions. Therefore it is worth noting that CSI 2.0 did not confuse any β-strands with α-helices and vice versa. In a few cases (6.5%), CSI 2.0 failed to recognize α-helical residues and identified them as "coil". Likewise, about 4% of coil residues were mis-identified as α-helices. Once again, many of the mis-identifications occurred at the borders or edges of α-helices and coil regions. In all likelihood these misidentified helices were somewhat flexible or only partially helical under the solution conditions that were originally used to collect the NMR data. The fact that protein structures do sometimes differ between crystal forms (solved by X-ray methods) and in solution (solved by NMR) has been noted for many years. Indeed, there are many examples showing these discrepancies (Andrec et al. 2007, Ratnaparkhi et al. 1998). It is also important to remember the agreement between the secondary structures determined by conventional NMR methods and those determined using X-ray data typically differ by 5-10%.

Identification of 3_{10} helices and β-bridges

3_{10}-helices are short helical structures with an average length of three residues and a distorted hydrogen-bonding network, whereas β-bridges are single-residue β-strands. Only the DSSP program identifies these structures and consolidates them into α-helices and β-strands. On the other hand, STRIDE and VADAR often characterize them simply as "coil". In looking more closely at our results, we found that CSI 2.0, regardless of its training set, would identify isolated 3_{10} helices and β-bridges as simple "coil" structures. This underscores one of the challenges with secondary structure identification, namely the fact that different programs (and different structural biologists) have different opinions or different definitions of what secondary structures are. Interestingly CSI 2.0 still performed best when it was working with DSSP assigned secondary structures (as opposed to VADAR or STRIDE assignments) – even with the presence of these "hard-to-identify" 3_{10} helices and β-bridges.

PSIPRED improves performance

CSI 2.0 was originally intended to be a chemical shift-only method. However, the exceptional performance of Psi-CSI (Hung et al. 2002) led us to reconsider the use of sequence information. Indeed, the inclusion of PSIPRED (Jones 1999) into the CSI 2.0 algorithm improved the Q3-accuracy from 87.3% (chemical shift only) to 89.35%. This improvement is statistically significant ($p<0.001$). More specifically, the inclusion of PSIPRED was found to improve the "β-strand" accuracy by 4% and the "coil" accuracy by 2.3%. On the other hand, the identification accuracy of α-helices was not improved in any substantial way. Given that chemical shift-based methods tend to confuse some β-strand residues with coil residues (and vice versa), it appears that PSIPRED helps to remove this chemical shift ambiguity.

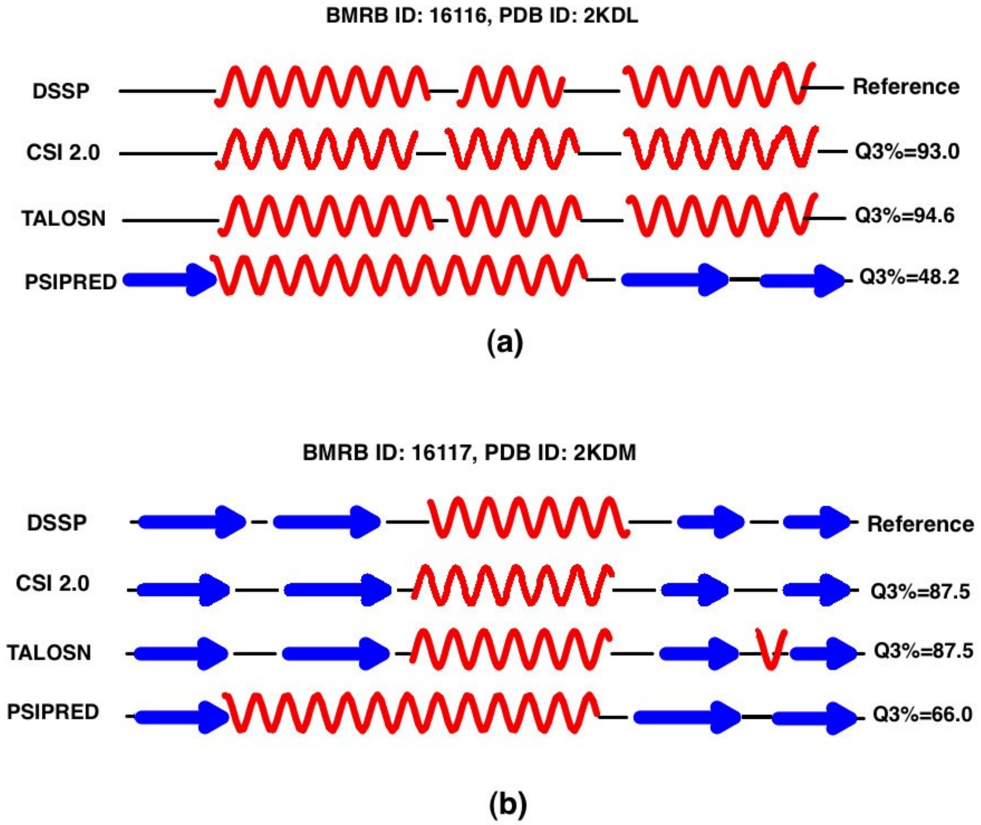


Figure 2.2: Secondary structure prediction/assignment for BMRB 16116 (PDB ID: 2KDL) and BMRB 16117 (PDB ID: 2KDM) by CSI 2.0, TALOS-N and PSIPRED.

CSI 2.0 accurately identifies secondary structure with "trick" proteins

Proteins with high sequence identity but very different folds pose special challenges for sequence-based structure prediction methods (Shen et al. 2010). One example of note is the protein G pair known as GA (95) and GB (95) (Alexander et al. 2009). Protein GA (95) is a specially designed, mostly helical protein, that shares a high degree of sequence identity (95%) with the native, β-rich protein GB. Here, we investigated how CSI 2.0 performed in distinguishing the local structures of these two proteins when compared to other methods (TALOS-N (Shen et al. 2013) and PSIPRED (Jones 1999)). As seen in Figure 2.2, and as expected, PSIPRED did quite well with its secondary structure prediction for GB but not so well with GA. On the other hand, CSI 2.0 and TALOS-N performed comparably well and were able to correctly identify the secondary structures in both proteins. The fact that CSI 2.0 uses PSIPRED in its determination of secondary structure, but its performance was not compromised in this "GA vs. GB test", illustrates how CSI 2.0 is able to appropriately balance experimental chemical shift information with sequence/PSIPRED information.

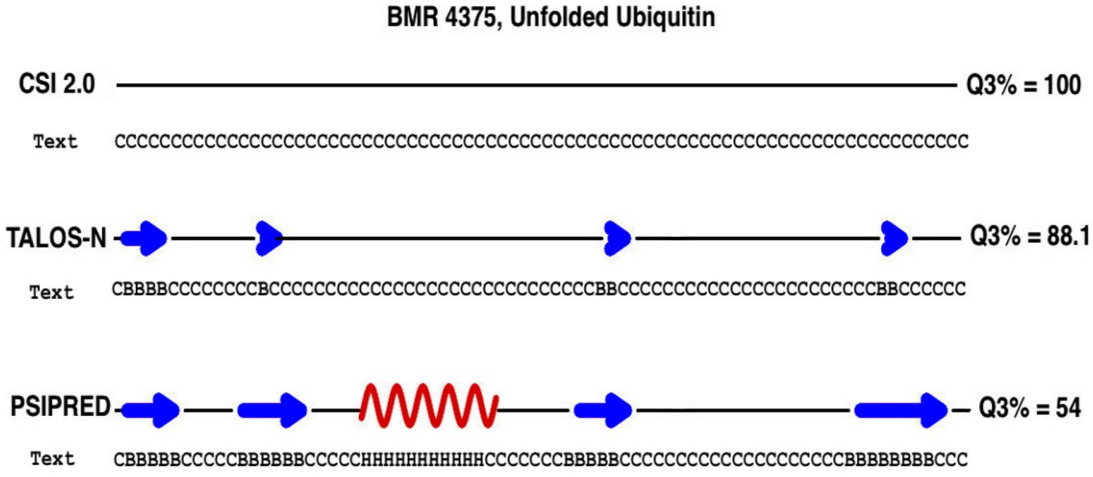

BMR 4375, Unfolded Ubiquitin
CSI 2.0
Q3% = 100
Text
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
TALOS-N
Q3% = 88.1
Text
CBBBBCCCCCCCBCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBBCCCCCCCCCCCCCCCCCCCCCCBBCCCCCCC
PSIPRED
Q3% = 54
Text
CBBBBBCCCCCBBBBBBCCCCCHHHHHHHHHHHCCCCCCCBBBBBCCCCCCCCCCCCCCCCCCCCBBBBBBBBCCC

Figure 2.3: Secondary structure assignment/prediction for BMRB 4375 (Unfolded ubiquitin) as determined by the CSI 2.0, TALOS-N and PSIPRED programs.

We also investigated the performance of CSI 2.0 for assigning the secondary structure for a completely unfolded protein (i.e. unfolded ubiquitin in 8 M urea - BMRB 4375). As seen in Figure 2.3, CSI 2.0 was able to accurately identify the disordered structure of this protein, whereas PSIPRED and TALOS-N proved to be somewhat less accurate than CSI 2.0. As PSIPRED predicts the secondary structure from sequence, it just reported the folded ubiquitin structure retrieved by a PSI-BLAST search. In case of TALOS-N, nine "coil" regions were incorrectly predicted as "β-strands". In the case of PSIPRED most of the secondary structure that was predicted contain a high proportion of helices and β-strands. However, because CSI 2.0 weighs both the chemical shift information with PSIPRED predictions, its performance was not compromised.

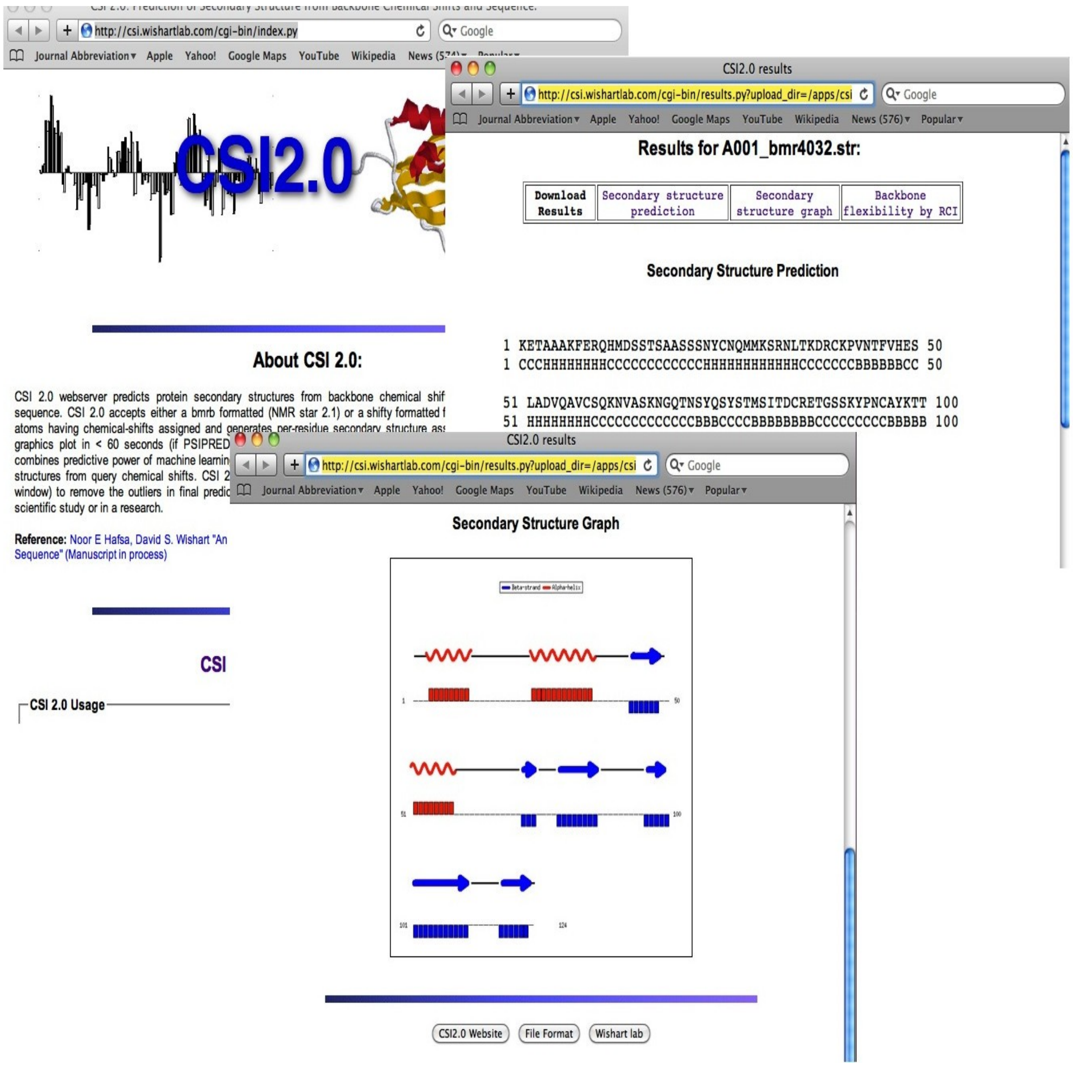
Potential improvements

J-coupling constants and NOE data can obviously aid in inferring the existence or delineation of secondary structure. This is why conventional NMR methods have traditionally depended so heavily on these NMR-derived parameters to identify secondary structures. Potentially some improvement in CSI 2.0's performance could be achieved if these parameters were also included in the model, particularly in cases when chemical shift data is missing or ambiguous. However, our focus was primarily on developing a simple approach that requires only sequence data and backbone chemical shift information to accurately identify protein secondary structures. The advantage of using chemical shifts is that these are the first pieces of experimental data that one obtains when studying proteins by NMR. Chemical shifts are also far easier to measure and far more accurately measured than NOEs and J-coupling data.

Instead of adding more experimental data, another approach that could potentially improve the performance of CSI 2.0 is to include sequence homology information from previously solved protein structures. With more than 100,000 protein structures in the PDB, this represents a significant and largely untapped information resource on secondary structure. The use of sequence homology from solved structures has been shown to substantially improve the performance of sequence-only secondary structure prediction methods (Montgomery et al. 2006, Cole et al. 2008). However, it is not clear whether the same level of improvement could be achieved when working with data that already has some experimental information concerning the secondary structure (i.e. chemical shifts).

The CSI 2.0 Webserver

A web server (http://csi.wishartlab.com) has been developed that accepts a BMRB (NMR–Star 2.1 or NMR-Star 3.1) or SHIFTY-formatted chemical shift file and generates secondary structure assignments along with a colorful CSI bar graph plot with secondary structure icons marked above the bar graph. The server supports a number of user-selectable options including the choice of running with or without PSIPRED. The web server is implemented using Python CGI-scripting and is hosted on a system with 4GHz 2-Core processor and a CentOS operating system. With the available computing infrastructure, the web server takes <60 seconds (if PSIPRED is off) or >140 seconds (if PSIPRED is on) to calculate secondary structures for a single query protein. A screen shot of the CSI 2.0 web server and its output is shown in Figure 2.4.


http://csi.wishartlab.com/cgi-bin/index.py
Google
Journal Abbreviation▾ Apple Yahoo! Google Maps YouTube Wikipedia
CSI2.0
CSI2.0 results
http://csi.wishartlab.com/cgi-bin/results.py?upload_dir=/apps/csi
Journal Abbreviation▾ Apple Yahoo! Google Maps YouTube Wikipedia News (576)▾ Popular▾
Download Results
Secondary structure prediction
Secondary structure graph
Backbone flexibility by RCI
Beta-strand
Alpha-helix
1
50
51
100
101
124
CSI2.0 Website
File Format
Wishart lab

About CSI 2.0:

CSI 2.0 webserver predicts protein secondary structures from backbone chemical shif sequence. CSI 2.0 accepts either a bmrb formatted (NMR star 2.1) or a shifty formatted f atoms having chemical-shifts assigned and generates per-residue secondary structure as graphics plot in < 60 seconds (if PSIPRED combines predictive power of machine learnin structures from query chemical shifts. CSI 2 window) to remove the outliers in final predic scientific study or in a research.

Reference: Noor E Hafsa, David S. Wishart "An Sequence" (Manuscript in process)

CSI

CSI 2.0 Usage

Results for A001_bmr4032.str:

Secondary Structure Prediction

```
1  KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHES 50
1  CCCHHHHHHHHCCCCCCCCCCCCHHHHHHHHHHHHCCCCCCCBBBBBBCC 50

51  LADVQAVCSQKNVASKNGQTNSYQSYSTMSITDCRETGSSKYPNCAYKTT 100
51  HHHHHHHHCCCCCCCCCCCCCBBBCCCCBBBBBBBBCCCCCCCCCBBBBB 100
```

Secondary Structure Graph

Figure 2.4: The CSI 2.0 (http://csi.wishartlab.com) web server image (a) homepage; (b) Result page.

2.4 Conclusion

CSI 2.0 represents a substantial improvement over the original CSI concept. In particular it uses an extended feature set derived from chemical shift and sequence data. It also replaces the simple digital filtering used in the original CSI algorithm with a more powerful "feature filter" that uses machine learning. Using the standard 3-state criteria (α-helix, β-strand and coil) and standard evaluation method such as Q3-accuracy, CSI 2.0 shows a significantly improved performance over the original CSI (89% vs. 79%) as well as significantly improved performance over other available state-of-the-art secondary structure identification methods (89% vs. ~86%). This performance improvement was statistically significant not only for the most common secondary structure assignment method (DSSP) but also for two other commonly used secondary structure assignment methods, VADAR and STRIDE. Based on data presented here concerning the level of agreement between different secondary structure identification methods (NMR vs. X-ray vs. different programs), we suspect that we are at or near the maximum performance that secondary structure assignment methods can achieve. In addition to the performance improvement seen with CSI 2.0, we also showed that CSI 2.0 successfully detected different secondary structures in structurally dissimilar proteins sharing high sequence identity – something that commonly fools other programs. We also showed that CSI 2.0 is able to identify the (lack of) secondary structure of unfolded proteins.

To make this method publicly accessible, a CSI 2.0 web-server (http://csi.wishartlab.com) has been developed. It accepts chemical shift assignments in a variety of formats and generates colorful graphical output describing the identity and location of all secondary structure elements. We believe that CSI 2.0, with its superior performance will be a useful contribution to the field of biomolecular NMR. It should be particularly useful in the initial stages of conventional NMR structure generation (i.e. providing useful torsion and distance restraints) as well as serving as a robust alternative to standard coordinate-based methods for secondary structure identification. CSI 2.0 is currently being used in the development of improved chemical shift-only 3D structure determination methods.

Chapter 3

CSI 3.0: A Web Server for Identifying Secondary and Super-secondary Structure in Proteins using NMR Chemical Shifts[1]

Abstract

CSI 3.0 (http://csi3.wishartlab.com) is a web server designed to accurately identify the location of secondary and super-secondary structures in protein chains using only Nuclear Magnetic Resonance (NMR) backbone chemical shifts and their corresponding protein sequence data. Unlike earlier versions of CSI (the Chemical Shift Index), which only identified three types of secondary structure (helix, β-strand and coil), CSI 3.0 now identifies total of 11 types of secondary and super-secondary structures, including helices, β-strands, coil regions, five common β-turns (type I, II, I', II' and VIII), β-hairpins as well as interior and edge β-strands. CSI 3.0 accepts experimental NMR chemical shift data in multiple formats (NMR Star 2.1, NMR Star 3.1 and SHIFTY) and generates colourful CSI plots (bar graphs) and secondary/super-secondary structure assignments. The output can be readily used as constraints for structure determination and refinement or the images may be used for presentations and publications. CSI 3.0 uses a pipeline of several well-tested, previously published programs to identify the secondary and super-secondary structures in protein chains. Comparisons with secondary and super-secondary structure assignments made via standard coordinate analysis programs such as DSSP, STRIDE and VADAR on high resolution protein structures solved by X-ray and NMR show >90% agreement between those made with CSI 3.0.

[1] Portions of this chapter were published as: Hafsa NE, Arndt D, Wishart DS (2015) CSI 3.0: a web server for identifying secondary and super-secondary structure in protein using NMR chemical shifts. Nucleic Acids Res. 42(W1):W370-377.

3.1 Introduction

Secondary structures such as α-helices, β-strands and coils are commonly used to describe, understand and visualize protein tertiary structures (Richardson 1981). Because of their importance, the identification and delineation of secondary structure elements has long been an integral part of the protein structure determination process. This is particularly true for NMR-based protein structure determination where secondary structure is used to help in structure generation and refinement (Wuthrich 1986 & 1990). In protein NMR, secondary structures are traditionally identified and assigned using NOE-based (Nuclear Overhauser Effect) methods. By manually analyzing the positions and patterns of weak, medium or strong NOEs it is possible identify helices, β-turns and β-strands with reasonably good accuracy. Even today NOE pattern measurements continue to be the most commonly used method for identifying secondary structures in peptides and proteins (Wuthrich 1986). However, in addition to NOEs, NMR chemical shifts can also be used to identify secondary structures. The use of chemical shifts to identify protein secondary structures was first demonstrated in the early 1990s with the development of a technique called the Chemical Shift Index or CSI (Wishart et al. 1992). The CSI method applies a three-part or ternary "digital filter" to backbone ^{1}H and ^{13}C chemical shifts as a way of simplifying the chemical shift information. By comparing the experimentally observed chemical shifts to a set of residue-specific "random coil" chemical shifts and converting significant downfield secondary shifts to "1's", significant upfield secondary shifts to "-1's" and small secondary chemical shifts to "0's", a simple bar graph can be generated. By observing how the 1's or -1's or 0's cluster together in the graph it is possible to accurately identify the type and location of protein secondary structure elements (helices, β-strands, coils) along the length of a protein chain (Wishart et al. 1992 & 1994). The CSI method is particularly popular because it is fast, easy to perform and surprisingly accurate -- exhibiting a ~80% agreement with secondary structures determined from PDB coordinate analysis.

However, the CSI method is far from perfect. For instance, it requires nearly complete backbone assignments to obtain good results. Furthermore, it is quite sensitive to the choice of random coil or reference chemical shifts used to calculate the secondary shifts and it tends to be less accurate for β-strand identification than helix. Because of these limitations, a number of alternative CSI-like methods have been proposed. These include PSSI (Wang et al. 2002), PsiCSI (Hung et al. 2003), PLATON (Labudde et al. 2003), PECAN (Eghbalnia et al. 2005), and 2DCSi (Wang et al. 2007). Most of these newer methods extend the basic CSI protocol by including more sophisticated chemical shift models or more elaborate statistical calculations. For instance, the developers of PSSI chose to discard CSI's simplistic digital filter and replace it with a sophisticated joint probability model to enhance PSSI's secondary structure identification accuracy. On the other hand, the developers of PsiCSI kept the basic CSI protocol but combined it with a well-known sequence-based secondary structure prediction program called PSIPRED (Jones 1999) to enhance its performance. In contrast to PSSI and PsiCSI, PLATON uses a database of well-defined reference chemical shift patterns to help identify secondary structures. This pattern database appears to boost its secondary structure identification performance. The program PECAN uses a chemical shift "energy function" that combines sequence information with chemical shift information to improve its secondary structure identification accuracy. Finally, 2DCSi uses cluster analysis to extract information for chemical shift scatter diagrams derived from all six backbone chemical shifts to improve its secondary structure identification performance. Most of these methods achieve a three-state secondary structure (Q3) accuracy better than 80%, with some reaching as high as 85%.

Over the past 5 years, many chemical shift-based secondary structure assignment methods have begun to exploit machine-learning techniques, torsion angle estimates, sequence similarity assessments, chemical-shift derived flexibility and larger chemical shift-structure databases to improve their performance. These newer methods include TALOS+ (Shen et al. 2009), TALOS-N (Shen et al. 2013), DANGLE (Cheung et al. 2010) and CSI 2.0 (Hafsa et al. 2014). The TALOS+ and TALOS-N packages use chemical shifts to calculate backbone torsion angles. This information is then used to identify secondary structure locations by exploiting the power of artificial neural networks (ANNs) to match chemical shift patterns against a large database of previously assigned proteins with high-resolution 3D structures. DANGLE employs some of the same concepts found in TALOS+ but instead of ANNs it uses Bayesian-inference techniques to help identify secondary structures. Like the TALOS and DANGLE programs, CSI 2.0 makes use of machine learning algorithms to integrate multiple pieces of information together but unlike TALOS it also combines more extensive sequence information with additional data regarding chemical-shift derived flexibility. The performance of these newer "shift-to-structure" programs is now quite impressive with most reporting Q3 accuracies above 85% and with CSI 2.0 achieving a Q3 score of 88-90%. This kind of performance generally exceeds the performance of NOE-only based methods for secondary structure assignment or identification (Hafsa et al. 2014). Furthermore, a Q3 score of 88-90% essentially matches the level of agreement that one achieves by comparing the results of different coordinate-based secondary structure assignment programs such as DSSP (Kabsch et al. 1983), STRIDE (Frishman et al. 1995) or VADAR (Willard et al. 2003) on the same PDB coordinate set (Hafsa et al. 2014).

While the performance of the most recent shift-based secondary structure assignment programs is very impressive, they are still missing a significant amount of information that can be easily derived from chemical shifts. This includes such useful information as flexibility, backbone torsion angles and accessible surface area (Shen et al. 2013, Berjanskii et al. 2005 & 2013). Furthermore, the traditional 3-state model of secondary structure assignments (helix, β-strand and coil) is often considered rather "dated" and somewhat inadequate with regard to modern expectations of detailed protein topology diagrams or information-rich protein structure descriptions. Three-state secondary structure assignments are also insufficiently precise for many 3D structure generation or 3D structure refinement programs such as XPLOR (Schwieters et al. 2003), CYANA (Guntert 2004), CHESHIRE (Cavalli et al. 2007), CS-Rosetta (Shen et al. 2008) and CS23D (Wishart et al. 2008). Ideally if NMR chemical shifts could be used to identify other kinds of secondary or super-secondary structure features such as β-turns, β-hairpins or more complex β-strand topologies, then they could be more fully exploited as additional constraints for NMR structure generation and refinement. This same information could also be used to create far more informative protein secondary structure and topology diagrams.

Given the need for this kind of information and given the availability of high performing tools to calculate these features from NMR chemical shift data, we decided to create a new kind of "shift-to-structure" tool. In particular we combined a high-end secondary structure calculation algorithm (CSI 2.0) with a high-performing torsion angle calculator (TALOS-N), an accurate measurement method for backbone flexibility (RCI) and a robust method for calculating fractional accessible surface areas (Side-chain RCI) – all of which use NMR chemical shifts as input. By linking these four tools together into a single structure determination pipeline and intelligently processing their respective structure assignments we found that it was possible to create a program that accurately identifies 11 types of secondary and super-secondary structures using only backbone NMR chemical shift data. These shift-derived structures include helices, β-strands, coil regions, five common β-turns (type I, II, I', II' and VIII), β-hairpins as well as interior and edge β-strands. Since this concept builds from our previous work on the Chemical Shift Index (CSI) and an earlier program called CSI 2.0, we decided to call the new method CSI

3.0. A detailed description of the CSI 3.0 web server along with a discussion of its capabilities and overall performance is given below.

3.2 Algorithm and Workflow

The CSI 3.0 system consists of four well-tested and previously published programs, namely CSI 2.0 (Hafsa et al. 2014), TALOS-N (Shen et al. 2013), RCI (Berjanskii et al. 2005) and Side-chain RCI (Berjanskii et al. 2013). CSI 2.0 uses chemical shift and sequence data to accurately identify three types of secondary structures: helices, β-strands and coil regions. Extensive tests have shown that it has a Q3 accuracy (agreement between identified by shifts and those determined by coordinate analysis) of 88-90% depending on the dataset and coordinate assignment algorithm that is chosen (Hafsa et al. 2014). TALOS-N uses chemical shift and sequence data to calculate backbone torsion angles. It can routinely determine backbone torsion angles for more than 90% of amino acid residues, with a root mean square difference between estimated and X-ray observed (ϕ, ψ) torsion angles of about 12º (Shen et al. 2013). The RCI or random coil index technique uses backbone chemical shifts to calculate the flexibility or order parameters of a protein sequence. The RCI method is frequently used to identify ordered and disordered segments in proteins. The agreement between RCI-calculated order parameters or RMSFs and observed order parameters or RMSFs ranges between 77-82% (Berjanskii et al. 2005). The side-chain RCI or the side-chain random coil index is a technique that can be used to calculate residue-specific fractional accessible surface area (fASA) using side-chain chemical shifts. The original paper reported a correlation coefficient between the shift-calculated fractional ASA and the coordinate measured fASA of approximately 0.76 (Berjanskii et al. 2013). Recent improvements to the algorithm (named as ShiftASA) now allow backbone (only) shifts to be used and the correlation between observed and shift-calculated fASAs is now 0.82 (Hafsa et al. 2015).

The central concept behind the CSI 3.0 algorithm is to intelligently combine each of the four shift-based calculators into a more comprehensive or more fully integrated structure assignment program that is “greater than the sum of its parts”. Specifically, by starting with the most accurate method first (secondary structure assignment with CSI 2.0) and then filtering out protein sequence segments that were already assigned a clear secondary structure (helix or β-strand) we found we could selectively apply the less accurate methods (torsion angle, flexibility and fASA calculations) to the remaining regions to identify other secondary or super-secondary structures. For instance, to identify a β-hairpin it is better to start with the precise location of the two sequentially adjacent β-strands and to determine if the “coil” residues between the β-strands have the appropriate torsion angles and sequence characteristics to form a β-hairpin. Similarly, the identification of edge or interior β-strands can only be determined once a β-strand is identified and only then should the fASA, flexibility or other characteristics of the entire β-strand be calculated. Similarly, the identification of β-turns and β-turn types should only be conducted in regions initially identified as “coil” regions (since β-turns are not found in helices or β-strands) and only in regions where the chain is well defined (i.e. a RCI-calculated order parameter >0.7).

A flow chart describing the CSI 3.0 algorithm is shown in Figure 3.1. As can be seen in this diagram the user first provides a file (NMRStar 2.1, NMRStar 3.1 or SHIFTY format) containing the protein sequence and the assigned chemical shifts. Complete and properly referenced (Wishart et al. 2001) ^{1}H, ^{13}C and/or ^{15}N chemical shifts are strongly preferred. However, ^{15}N chemical shift data is not required and the lack of ^{15}N shift data typically does not reduce the overall program performance. Once the chemical shift file is provided, CSI 2.0 is called to perform a per-residue, three-class secondary structure assignment. Extensive studies have shown than CSI 2.0 is the most accurate method for identifying secondary structures using only chemical shift data (Hafsa et al. 2014). Additional details regarding the algorithm and its performance with regard to missing assignments or chemical shift completeness are fully described in the original publication (Hafsa et al. 2014) and in Chapter 2 of this thesis. Once the helices, β-strands and coil regions are identified, the RCI (random coil index) program is run. The RCI program calculates backbone flexibility from backbone chemical shifts. Additional details regarding its algorithm, its applications and overall performance are also described in the original publication (Berjanskii et al. 2005). The purpose of the RCI program is to identify CSI 2.0 annotated coil regions that are too flexible to produce reliable torsion angles (for β-turn identification). Residues that have a RCI-calculated order parameter (S^2) ≤ 0.7 are excluded from further analysis. The choice of $S^2 \leq 0.7$ is based on observations from many NMR protein structures that have intrinsically disordered or poorly defined regions and these are usually characterized by S^2 values of less than 0.7. After the RCI filtering step is performed all remaining coil regions have their φ, ψ backbone torsion angles calculated by TALOS-N (Shen et al. 2013). TALOS-N is widely regarded as the most accurate, chemical shift-based backbone torsion angle calculator. Details of the algorithm and its performance with regard to missing assignments or chemical shift completeness are fully described in the original publication (Shen et al. 2013). Finally the last program (ShiftASA) is used to calculate the fASA (fractional accessible surface area) for all β-strand residues initially identified by CSI 2.0. More details about this program are described in Chapter 4 of this thesis and in a subsequent publication (Hafsa et al. 2015)

Once the initial per-residue assignment phase (helices, β-strands, coil, order parameters, φ, ψ angles, fASA) has been completed, the algorithm moves to the second phase, which involves identifying β-turn types (type I, II, I', II' and VIII), β-hairpins and edge/interior β-strands. This "contextual assignment" phase employs the per-residue assignment data from the first phase along with the contextual data from the neighbouring residue assignments, local sequence (hydrophobicity) data and additional chemical shift pattern information.

The first part of the contextual assignment phase involves the identification of β-turns. β-turns can be classified into five different types, i.e. type I, II, I', II' and VIII, based on the characteristic backbone torsion angles for the central two residue *(i+1)* and *(i+2)* locations (Hutchinson et al. 1994). A simple heuristic rule-based algorithm was designed to identify these five turn types along the protein chain. Coil regions were identified using CSI 2.0 (Hafsa et al. 2014). Those coil regions with 2 or more consecutive coil assignments having RCI-estimated order parameters > 0.7 were then analysed further. In particular, the φ/ψ torsion angles and amino acid types in these coil residues were analysed to identify the presence of a β-turn and to assign the appropriate turn type to the residue locations. According to Hutchinson et al. (1994), the five common types of β-turns have propensities not only to adopt very specific backbone φ/ψ torsion angles but also to have very specific positional amino acid preferences. In particular, the torsion angles of the two central residues in a β-turn segment must fall within 30° of their characteristic backbone φ/ψ values and that certain preferred amino acids must be observed over a four-residue window, centered about the turn region. Consequently we developed a simple equation that accommodates these criteria to identify and as well as to classify β-turns.

$$S_{bturn} = (1 - f_{helical}) * (f_{\phi\psi} + f_{aa}) \tag{3.1}$$

This equation was applied over a sliding four-residue window over the sequence segment of interest. For this equation, $f_{helical}$ is a simple binary function that outputs a zero or one value based on the secondary structure class of two central positions over a four-residue segment. The term $f_{helical}$ is one if two central residues ($i + 1$ and $i + 2$) either have helix or β-strand assignments and zero otherwise. The term f_{aa} is the preferred amino acid content in the four-residue segment. The range of f_{aa} is [0.0-1.0]. If there is no favourable amino acids in any of the four positions then f_{aa} = 0 whereas if all four positions have preferred amino acids, the result is f_{aa} = 1.0. The term $f_{\phi\psi}$ is a numerical value related to the five β-turn types (I, II, I', II' and VIII) and can adopt a discrete value, m = 2×n+1, where n ϵ [0,4] represents five turn types in order, if three of the four torsion angles fall within 30° of their characteristic φ or ψ angles, with one φ/ψ angle deviating by up to 45°. According to the above formulation, a valid β-turn has a minimum S_{bturn} score of 1 and the required range of scores for the five turn types are: type I ≈ [1-2], type II ≈ [3-4], type I' ≈ [5-6], type II' ≈ [7-8] and type VIII ≈ [9-10].

The second part of the contextual assignment phase involves identifying β-hairpins. A β-hairpin is formed when a β-turn connects and aligns two anti-parallel β-strands. CSI 3.0's β-hairpin algorithm simply searches for two sequential β-strands that are connected by 6 or fewer residues containing an appropriate reverse β-turn.

The third part of the contextual assignment phase involves identifying edge (exterior) and interior β-strands. Those β-strands located on the "outside" edges of β-sheets with inter-strand hydrogen bonds only on one side are called edge strands. Those β-strands that have inter-strand hydrogen bonds on both sides are called interior β-strands. Therefore β-sheets with just two β-strands would have two edge strands, β-sheets with three β-strands would have one interior and two edge strands, and so on. In general, edge β-strands and interior β-strands are distinguishable by their length (edge strands tend to be shorter), rigidity (interior strands have higher order parameters), repeating patterns of hydrophobic/hydrophilic residues, charged residue distribution, distinct hydrogen bonding patterns and their level of solvent exposure (Siepen et al. 2003) -- all of which can be identified via chemical shift data and sequence information. However, the complexity of the features and patterns led us to develop a machine learning algorithm to identify interior and edge β-strands more accurately.

To construct the database needed to train and test the β-strand classification model in CSI 3.0, we chose a local, manually curated data set that we previously used to train and test the SHIFTX2 (Han et al. 2011) and CSI 2.0 (Hafsa et al. 2014) programs. Similar filtering criteria were used to eliminate chemical shift re-referencing problems, check chemical shift quality and detect chemical shift outliers from an initial data set of ~300 X-ray protein structures with good quality NMR assignments. A more detailed accounting of the data preparation protocol is provided in the SHIFTX2 and CSI 2.0 papers. The above selection and filtering process reduced the data set to 171 proteins. This data set was then divided into a training set and an independent test set. The training dataset consisted of 150 proteins (263 β-strands) whereas the test dataset contained 21 entries (38 β-strands). PyMOL was used to visually inspect and assign the appropriate class (interior or edge) to each β-strand in the data set. β-strands that were located on the "outside" edges of β-sheets with inter-strand hydrogen bonds only on one side were assigned as edge strands and those β-strands that had inter-strand hydrogen bonds on both sides were classified as interior β-strands. In the training set, the class ratio was edge: interior = 131: 132 and in the test set the ratio was 19:19, which represents a balanced class distribution in both training and testing set. Among the training and testing proteins, 136 proteins belonged to the α + β folding class, 15 proteins to the all-α class, 18 proteins to the all-β class and two proteins to the all-coil folding class. The free parameters for the secondary structure assignment model were optimized on the training data set while the test set was used to perform an independent validation of the program's performance. DSSP (Kabsch et al. 1983) was used to assign reference secondary structures ("α-helix", "β-strand", "coil") in both the training and test set proteins. DSSP assigns secondary structures based on the coordinates of the 3D structures as well as inferred H-bonds and torsion angles derived from those coordinates. The normal eight-state DSSP assignments were transformed into a three-state (helix, sheet, coil) assignment using the EVA convention (Eyrich et al. 2001).

We then extracted a set of input features from the training and testing data that were used to classify each β-strand into one of two classes – interior and edge. The features were based on observations and data provided in the literature regarding certain distinguishing characteristics of β-strands. For instance, the differing pattern of hydrogen bonding between edge strands and interior strands generates distinct ^{1}Hα chemical shift patterns. In particular, the ^{1}Hα protons of residues engaged in inter-strand hydrogen bonds tend to be deshielded, leading to downfield secondary chemical shifts. On the other hand, the ^{1}Hα protons of residues that are only hydrogen bonded to water (i.e. edge) tend to be shielded, leading to slight upfield or far weaker downfield secondary chemical shifts. Therefore an alternating pattern of upfield/downfield secondary chemical shifts is often seen in edge strands. This pattern was also noted by others as early as 1994 (Ösapay et al. 1994). An interior β-strand, on the other hand, will not exhibit this pattern. In addition, to these distinct chemical shift patterns, residues in edge β-strands tend to have greater average accessible surface area (fASA) than interior β-strands, which are usually buried in the protein core. Because the fASA of a residue can be reasonably well determined by its chemical shifts, we can use the shift-derived fASA to calculate the average exposure of each β-strand. Because interior strands tend to be more rigid, they often have comparatively higher S^2 order parameters than edge strands. Likewise edge β-strands are often characterized by a pattern of alternating hydrophilic and hydrophobic residues. Interestingly, edge strands are also characterized by a higher proportion and a more central positioning of charged residues along the strand (Siepen et al. 2003). For example, charged residues are often found at the middle of an edge strand, whereas they are almost never found in the middle of an interior strand. As observed by many others, the length of edge strands also tends to be much shorter than interior strands. Based on these data we derived a set of nine (9) different features that could be derived from backbone chemical shifts and/or sequence data. These features included: 1) proportion of exposed residues; 2) proportion of residues with high S2 order (> 0.7); 3) periodicity in hydrophilicity; 4) periodicity in polarity; 5) proportion of hydrophilic residues; 6) charge score; 7) ^{1}Hα chemical shift periodicity; 8) ^{1}HN chemical shift periodicity; and 9) β-strand strand length.

Once the features were determined for all 263 training β-strands, a binary kernelized SVM classifier was used to train the edge/interior β-strand model. All data points were normalized using a "Statistical Z-score" method prior to training. For our binary-class SVM model, a soft-margin classification approach was used. Unlike a hard-margin SVM, a soft-margin SVM classifier generally produces a wide decision-margin to separate the two classes by allowing some noisy examples inside or on the wrong side of the margin in order to achieve a better test performance. A Radial Basis Function (RBF) kernel (also known as Gaussian kernel) was used to map the feature vectors in a higher dimensional space. With a Gaussian kernelized soft-margin classification framework, the performance of the SVM depends on the following two parameters: 1) the regularization parameter "C" (also known as the "cost" factor) and 2) the Gaussian kernel width "σ". The "C" parameter allows one to adjust the trade-off between maximizing the decision-margin width vs. minimizing the number of misclassified samples in the training set. Selecting a large "C" value will cause a smaller number of misclassified samples, leading to a smaller decision boundary, whereas choosing a very small "C" will allow a large number of training errors, resulting in a wider decision margin. The "σ" parameter controls the width of the Gaussian kernel and can be adjusted to achieve a smoother fit for the model. Both "C" and "σ" parameters were optimized using a repeated 10-fold cross validation (CV) on the training data. The goal of this parameter optimization was to find the optimal values that maximize the accuracy or Q2 score of the two-class β-strand classification. The binary-class SVM implementation in the R package "*kernlab*" was used to train the classifier. The optimization of "C" and "σ" through the "*repeatedcv*" method was performed using the *train()* function in the R "caret" package.

The "C" parameter (the "cost" value) in the SVM classifier and the kernel parameter, "σ" in the Gaussian RBF kernel were optimized using a repeated 10-fold Cross Validation (CV). After achieving an optimal value of 0.076, "σ" was held constant while "C" was iteratively changed to optimize its value. To achieve an unbiased training result, the whole process was repeated five times. For each repetition, the accuracy of the two-state assignment of the training classes was measured. The optimal "cost" and "σ" values that were found to maximize the Q2 accuracy using this repeated training were 0.25 and 0.076 respectively. The training accuracy was averaged over five repetitions of the 10-fold CV process. An average training accuracy of Q2=77% with 263 data points (or β-strands) on 150 training proteins was observed with the aforementioned optimized parameter values. A test accuracy of Q2=79% was achieved on an independent test set of 21 proteins with 38 data points (or β-strands).

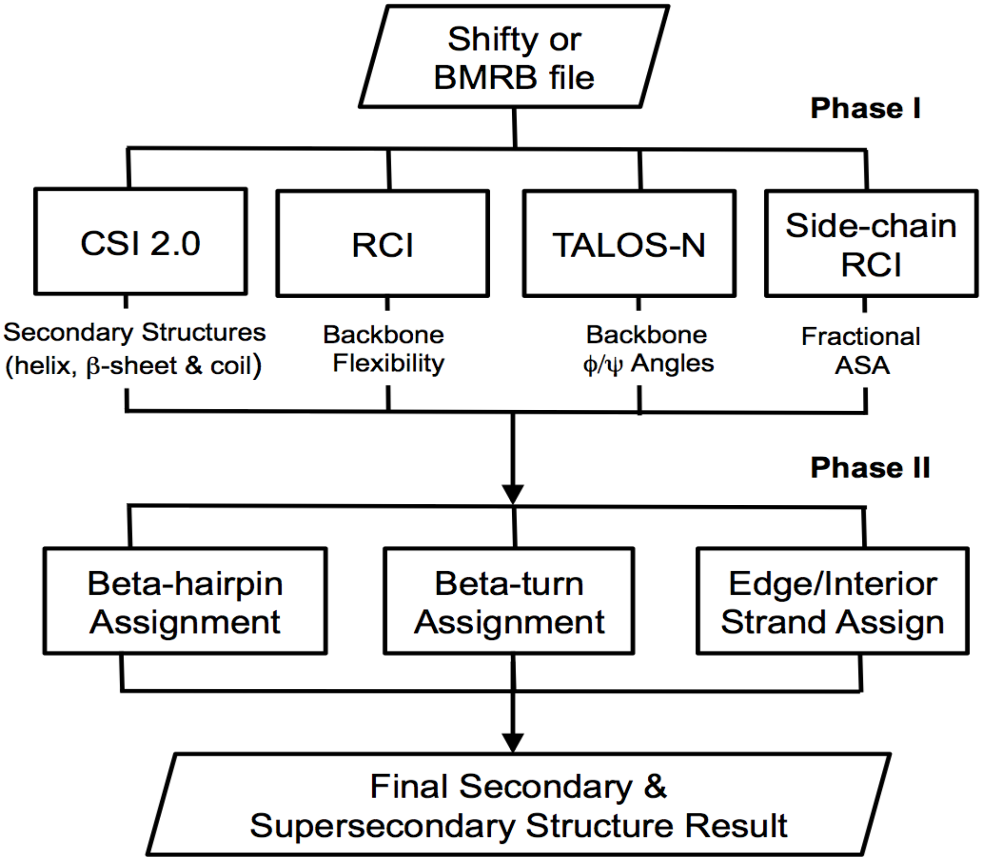


Figure 3.1: Program flow chart for CSI 3.0

3.3 Results and Validation

To demonstrate the utility of CSI 3.0 we evaluated its performance for both secondary and super-secondary structure identification using a set of 13 proteins with known 3D structures. The proteins were chosen to span a broad range of sizes (50 – 200 residues), secondary structure content, turn types, super-secondary structure features and 3D folds. These proteins also had an average level of backbone chemical shift completeness of 95% (which is relatively high). For each of the corresponding 3D structures, the identification of the consensus secondary structures (type and location) was performed by carefully combining the DSSP, STRIDE, VADAR, and author assignments together (Kabsch et al. 1983, Frishman et al. 1995, Willard et al. 2003). For this particular evaluation, β-strands with 2 or fewer residues (more formally known as β-bridges) were classified as coil regions. The identification of β-hairpins as well as the identification of edge and interior β-strands was done through visual inspection of the 3D structures using PyMOL. The complete set of proteins along with their consensus 3D structural assignments (as well as with their CSI 3.0 identified structural elements) is available on the CSI 3.0 website. For the entire set of 13 proteins there were 444 residues in helices, 349 residues in β-strands and 635 residues in coil regions. Within the coil regions there were 160 residues in type I turns, 12 residues in type I' turns, 36 residues in type II turns, 8 residues in type II' turns, and 4 residues in type VIII turns. Additionally there were a total of 14 β-hairpins, 31 edge β-strands and 28 interior β-strands. Note that only regions that are well defined (as identified by the RCI-derived order parameter score >0.7) and which had non-overlapping β-turns were used in the evaluation of the β-turn performance. The evaluation metric for the secondary structures (i.e. helices, β-strands, coils) was the standard Q3-score evaluated over all residues. The evaluation metric for β-turns (type I, I', II, II', VIII β-turns and non-turns) was a simple Q6 score evaluated over all residues. The evaluation metric for the β-hairpins (hairpins and non-hairpins) was a Q2 score while the evaluation metric for the edge, interior and non-edge/non-interior strands was a Q3 score. The "Qn" score is essentially a percent correct score similar to a multiple-choice exam where n is the number of possible answers for each question. The results are shown in Table 3.1.

This table shows that, as expected, the agreement between the per-residue secondary structure assignments derived by chemical shifts matches very well to those determined by analysing the coordinate data. The average Q3 score for the three main secondary structure types was 97%, which actually exceeds the performance of other state-of-the-art chemical shift-based methods.

Protein ID	Number of residues	Q3 score (H,B,C)	Q6 score (I,I',II,II',VIII turns, non-turns)	Q3 score (edge/interior β-strand, non-strand)	Q2 score (β-hairpins, non-hairpins)
Ubiquitin (Human) PDB: 1UBQ; BMRB: 5387	76	99	100	98	99
GB1 domain (Streptococcus) PDB: 1GB1; BMRB: 7280	58	96	100	91	96
Parvalbumin (Human) PDB: 1RK9; BMRB: 6049	110	97	100	100	95
Dinitrogenase (T. Mortima) PDB: 1O13; BMRB: 6198	124	97	100	85	97
Cyclic nucleotide protein (M. loti) PDB:1VP6; BMRB:15249	142	98	100	90	94
Glutaredoxin (Poxvirus) PDB: 2HZE; BMRB: 4113	108	100	100	96	100
SH3 domain Myo3 (Yeast) PDB: 1RUW; BMRB:6197	70	100	97	80	100
Acyltransferase (A. thaliana) PDB: 1XMT; BMRB: 6338	103	96	100	93	98
Cytosine Deaminase (Yeast) PDB: 1YSB; BMRB: 6223	158	96	100	95	100
Sortase A (Staphylococcus) PDB: 1T2W; BMRB:4879	148	91	97	85	95
Peptidyl-tRNA hydrolase (M. tuberculosis) PDB: 2Z2I; BMRB: 7055	191	96	97	97	100
Photoactive Yellow Protein (H. halophila) PDB: 1ODV; BMRB: 6321	100	100	96	80	100
Calmodulin (Bovine) PDB: 1A29; BMRB: 547	148	96	100	98	100

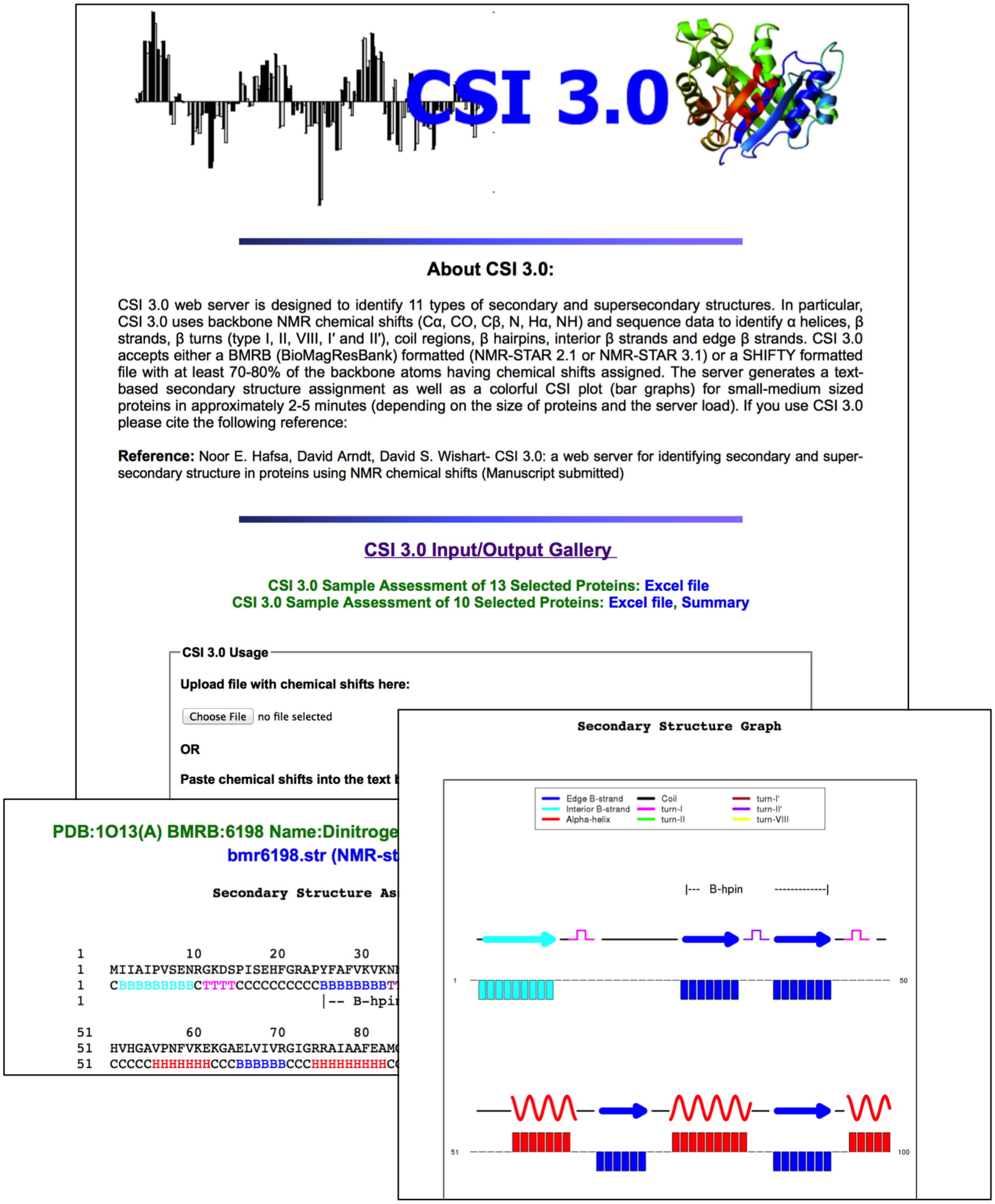
Table 3.1: Performance evaluation of CSI 3.0 on 13 selected proteins

The average agreement between the observed structure and the CSI 3.0 identified structure was 98% for helices, 96% for β-strands and 96% for coil regions (prior to β-turn ID). Likewise the average Q6 score for β-turns/non-turns was 99%, with a range spanning between 98% (type I) to 100% (type II, I', II', VIII). In terms of the super- secondary structure identification (edge, interior and β-hairpins), the average Q2 score for CSI 3.0 for hairpins/non-hairpins was 98%. For edge/interior β-strands/non-strands, the average Q3 score was 91%. CSI 3.0 achieved an edge β-strand accuracy of 73%, an interior β-strand accuracy of 88% and a non-strand accuracy of 97%. Closer inspection of the results shows that the disagreements between 3D structure-generated assignments and those derived by CSI 3.0 were often ambiguous or "close calls". This was particularly true with regard to the identification of edge strands. In many cases, edge and interior β-strands have a "dual" nature with some regions of any given β-strand being exposed and others being hydrogen bonded. In certain cases, it appears that CSI 3.0 struggled with identifying these hybrid β-strands. However, it is important to remember that this level of topological information is rarely obtained from preliminary NOE data or NOE pattern matching methods and is often not revealed until the final 3D structure is generated and thoroughly refined. Overall, we believe CSI 3.0's level of performance greatly exceeds what is achievable from NOE pattern matching methods and it is certainly sufficient to provide a useful topologically rich picture of protein structures (for illustrative or publication purposes) and to provide useful constraint data that could be used to generate and refine 3D protein structures using additional NOE or chemical shift (only) data from any number of packages.

3.4 Web Server Implementation

In developing the CSI 3.0 server we endeavoured to create a simple graphical interface that allows users to submit experimental NMR chemical shift data (from a single contiguous polypeptide) by either uploading the files or pasting them into a text box. Multiple chemical shift assignment formats (NMRStar 2.1, NMRStar 3.1 or SHIFTY) are accepted and examples of these formats are provided on the website. After submitting the shift file, the server generates colourful CSI plots or bar graphs (generated by the R package, version 3.0.2) with annotated helices, strands and colour-coded indications of β-turns, β-hairpins or edge and interior β-strands. The images are available in a PNG (Portable Network Garphics) format. A text file with the secondary and super-secondary structure assignments is also generated. The CSI 3.0 web server has been implemented as a Python CGI script (v. 1.1). The component programs were written in

Python (CSI 2.0, Side-chain RCI, RCI) or in C++ (TALOS-N). The web application is platform



About CSI 3.0:

CSI 3.0 web server is designed to identify 11 types of secondary and supersecondary structures. In particular, CSI 3.0 uses backbone NMR chemical shifts (Cα, CO, Cβ, N, Hα, NH) and sequence data to identify α helices, β strands, β turns (type I, II, VIII, I′ and II′), coil regions, β hairpins, interior β strands and edge β strands. CSI 3.0 accepts either a BMRB (BioMagResBank) formatted (NMR-STAR 2.1 or NMR-STAR 3.1) or a SHIFTY formatted file with at least 70-80% of the backbone atoms having chemical shifts assigned. The server generates a text-based secondary structure assignment as well as a colorful CSI plot (bar graphs) for small-medium sized proteins in approximately 2-5 minutes (depending on the size of proteins and the server load). If you use CSI 3.0 please cite the following reference:

Reference: Noor E. Hafsa, David Arndt, David S. Wishart- CSI 3.0: a web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts (Manuscript submitted)

CSI 3.0 Input/Output Gallery

CSI 3.0 Sample Assessment of 13 Selected Proteins: Excel file
CSI 3.0 Sample Assessment of 10 Selected Proteins: Excel file, Summary

CSI 3.0 Usage

Upload file with chemical shifts here:

no file selected

OR

Paste chemical shifts into the text b

PDB:1O13(A) BMRB:6198 Name:Dinitroge
bmr6198.str (NMR-st

Secondary Structure As

```
1              10        20        30
1    MIIAIPVSENRGKDSPISEHFGRAPYFAFVKVKN
1    CBBBBBBBBBCTTTTCCCCCCCCCCBBBBBBBBT
1                             |-- B-hpi

51             60        70        80
51   HVHGAVPNFVKEKGAELVIVRGIGRRAIAAFEAM
51   CCCCCHHHHHHHCCCBBBBBBCCCHHHHHHHHHC
```

Secondary Structure Graph

Figure 3.2: A montage of the CSI 3.0 webserver and typical output screen shots.

independent and has been tested successfully under Linux, Windows and Mac operating systems. CSI 3.0 has also been tested and found to be compatible with most modern web browsers including: Google Chrome (v. 31 and above), Internet Explorer (v. 9 and above), Safari (v. 7 and above), and Firefox (v. 23 and above). The CSI 3.0 web server is hosted on a system with 4GHz 2-Core processor and a CentOS operating system. With this computing infrastructure, the web server takes approximately 2-5 minutes to complete its calculations depending on the server load and the length of the protein query sequence. The server is freely available at http://csi3.wishartlab.com. A montage view of the CSI 3.0 web server along with screenshot examples of its output are shown in Figure 3.2.

3.5 Conclusion

CSI 3.0 is an accurate, automated, easy-to-use web service for calculating structural information from chemical shift data. In particular, CSI 3.0 accurately determines 8 types of local secondary structures (helices, β-strands, coils and 5 types of β-turns) as well as 3 types of super-secondary structures or topological features (β-hairpins, edge strands and interior strands). This represents nearly a fourfold increase in the number of secondary structure types identified by any other shift-analysis tool that we are aware of – including its predecessor, CSI 2.0. We believe that the additional secondary structure data, along with the useful topological information and colourful graphical output generated by CSI 3.0 will not only improve the quality of preliminary protein structure descriptions (often obtained shortly after chemical shift assignments are completed) but also facilitate the protein structure determination by NMR. In particular, with the recent trends towards protein structure determination and refinement using chemical shifts (only), chemical shift threading or minimal numbers of NOEs, this added information could prove to be particularly useful to a growing number of NMR spectroscopists.

Chapter 4

Accessible Surface Area from NMR Chemical Shifts[1]

Abstract

Accessible Surface Area (ASA) is the surface area of an atom, amino acid or biomolecule that is exposed to solvent. The calculation of a molecule's ASA requires three-dimensional coordinate data and the use of a "rolling ball" algorithm to both define and calculate the accessible surface area. For polymers such as proteins, the ASA for individual amino acids is closely related to the hydrophobicity of the amino acid as well as its local secondary and tertiary structure. For proteins, ASA is a structural descriptor that can often be as informative as secondary structure. Consequently there has been considerable effort over the past two decades to try to predict ASA from protein sequence data and to use ASA information (derived from chemical modification studies) as a structure constraint. Recently it has become evident that protein chemical shifts are also sensitive to ASA. Given the potential utility of ASA estimates as structural constraints for NMR, we decided to explore this relationship further. Using machine learning techniques (specifically a boosted tree regression model) we developed an algorithm called "ShiftASA" that combines chemical-shift and sequence derived features to accurately estimate per-residue fractional ASA (fASA) values of water-soluble proteins. This method showed a correlation coefficient between predicted and experimental values of 0.79 when evaluated on a set of 65 independent test proteins, which was an 8.2% improvement over the next best performing (sequence-only) method. On a separate test set of 92 proteins, ShiftASA reported a mean correlation coefficient of 0.82, which was 12.3% better than the next best performing method. ShiftASA is available as a web server (http://shiftasa.wishartlab.com) for submitting input queries for fractional ASA calculation.

[1] Portions of this chapter were published as: Hafsa NE, Arndt D, Wishart DS (2015) Accessible surface area from NMR chemical shifts. J Biomol NMR 62:387-401.

4.1 Introduction

Accessible surface area is a concept first introduced and popularized by Dr. Frederic M. Richards and co-workers in the early 1970s (Lee et al. 1971, Richards 1974 & 1977). It grew from the observation that certain parts of a folded protein seemed to be impenetrable to water while other parts were highly exposed. This differential exposure seemed to be driven by the hydrophobicity or hydrophilicity of individual amino acid side chains, the 3D structure of the protein and the influence that the hydrophobic effect had on the overall protein folding process. Richards and colleagues also pointed out that water molecules are not infinitely small point particles and that the surface of a protein that was water accessible was not equal to the van der Waals surface area but rather could be calculated by rolling a ball of finite size (roughly the size of an oxygen atom of 1.4 Å) over the entire van der Waals surface of a protein. The resulting, "smoothed-surface" defined the water accessible area or the accessible surface area (ASA). ASA is a quantifiable property measured in square Angstroms (Å^2). It can be determined for entire proteins or for individual residues or even atoms. ASA can also be re-cast as a fractional accessible surface area (fASA) that reports the percentage of ASA relative to a fully exposed protein (or residue). This concept can be carried further to a relative accessibility, or RSA, which is a more qualitative measure of surface accessibility. With the RSA concept, residues are considered buried (B), partially buried (P) or exposed (E) based on their fASA. Typically buried residues have a fASA of <0.25, partially buried have a fASA between 0.25 and 0.5 and exposed residues have a fASA of >0.50.

Since its first description, the concept of ASA has proven to be extremely useful for assessing the quality of protein folds and for scoring protein structure predictions (Benkert et al. 2008), for assessing conformational changes upon protein or ligand binding, for calculating protein folding energies, for determining protein-ligand binding constants and for calculating protein enthalpy and entropy changes (Lavigne et al. 2000). More recently, indirect measurements of residue-specific ASAs through targeted chemical modification or partial proteolysis have been used to provide constraints for low-resolution protein structure determination efforts by mass spectrometry (Serpa et al. 2014). Indeed since its first description some 40 years ago, the concept of ASA has probably been proven to be among the most useful concepts for understanding, comparing and evaluating protein folds and protein functions.

Quantitative ASA measurements can only be determined from protein coordinate data (i.e. solved structures). However, given the utility of ASA measurements as structural constraints or for evaluating structural/thermodynamic properties of proteins, there has been a growing interest in finding ways of predicting ASA, fASA or RSA from sequence data alone. As a result there have been a number of published studies that describe methods for predicting accessible surface area and relative surface accessibility from sequence (Ahmad et al. 2002 & 2003, Wagner et al. 2005, Petersen et al. 2009, Nguyen et al. 2005, Li et al. 2001, Pollastri et al. 2002, Chen et al. 2005, Naderi-Manesh et al. 2001, Thompson et al. 1996, Rost et al. 1994, Garg et al. 2005, Yuan et al. 2004, Holbrook et al. 1990, Adamczak et al. 2004). The majority of these prediction systems rely on using multiple sequence alignments, pairwise residue assessments and the predictive power of machine-learning algorithms. The best performance reported by these sequence-only methods using a two-state (Buried, Exposed) and a three-state RSA measure (Buried, Partially Buried, Exposed) yielded Q_2 and Q_3 scores of 88% and 63% respectively (Ahmad et al. 2002). For real-value ASA predictions, the best performance so far reported used PSSM matrices from PSI-BLAST (Altschul et al. 1997) profiles in a two-stage support-vector regressor to achieve a correlation coefficient between observed and calculated fASA of 0.68 (Nguyen et al. 2005).

While these sequence-only results are promising, Rost et al. (1994) pointed out that surface accessibility is less conserved in structural homologs than secondary structure and therefore ASA would be predicted less accurately from homology modeling. The Rost et al. study also showed that the correlation coefficient of relative solvent accessibility between 3D homologues (by structural alignment) is only 0.77, whereas prediction of accessibility by homology modeling (sequence alignment) resulted in a correlation coefficient of about 0.68. This suggests that the upper limit of ASA prediction that could be achieved by sequence-only methods would yield a correlation of 0.70-0.75.

Over the last two decades, it has been observed that a number of experimentally measurable properties in proteins correlate reasonably well with accessible surface areas. For instance, folding and unfolding free energies as measured through calorimetry appear to correlate quite well with ASA or fASA (Myers et al. 1995). Protease cleavage sites or protease susceptibility along with chemical modification susceptibility also appears to map with solvent accessibility (RSA or ASA) (Croy et al. 2004). Hydrogen exchange, as measured by MS (Mass-Spectrometry) or NMR also allows the identification of buried and exposed residues in proteins (Huyghues-Despointes et al. 1999). NMR chemical shifts also appear to be influenced by ASA effects. The first evidence of such a phenomenon was reported in 1994 (Wishart et al. 1994a).

Nearly a decade later Avbeli et al. (2004) studied the effect of secondary structure and solvent exposure on backbone chemical shifts. They demonstrated that proton secondary shifts have a different chemical shift distribution for solvent exposed residues, particularly in smaller peptides. In a later study by Vranken et al. (2009), the effect of secondary structure and solvent exposure on chemical shift assignments was re-examined on a large database of proteins for which both reported atomic coordinates and chemical shift values were available. There were two major findings from this study. First, they found that non-polar atoms have significantly larger chemical shift dispersion and a somewhat different chemical shift distribution compared to polar atoms. Secondly those atoms with greater atomic ASA, exhibited chemical shift values that tended towards random coil values. The relationship between chemical shifts and ASA was actually used to develop a significantly improved structure-based chemical shift prediction algorithm, called ShiftX2 in 2011 (Han et al. 2011). Most recently, Berjanskii et al. (2013) proposed a simple formula to calculate per-residue fractional accessible surface area from side-chain chemical shifts and observed a correlation of more than 70% with the observed fASA values over a subset of 15 proteins.

These studies demonstrate that both sequence and chemical shift information can be used individually to estimate the ASA values with reasonable accuracy. Now the question is: Can one develop more accurate fASA estimation by more intelligently combining sequence AND chemical shift information? Here we report the development of a machine-learning based method that can be used to accurately estimate per-residue fractional ASA of water-soluble proteins using sequence and chemical shifts. After training on a set of 30 fully assigned proteins, the performance of the resulting model, called ShiftASA, was compared with other sequence-based and chemical-shift based methods over a test set of 65 proteins. For this test set ShiftASA achieved a mean correlation coefficient of 0.79 compared to correlation coefficients of 0.73 and 0.59 found for sequence-only and chemical shift-only methods respectively. On a separate test set of 92 proteins, ShiftASA attained a correlation coefficient of 0.82. A number of other statistical measures were also used to prove that this method shows a consistently better performance than any existing method.

4.2 Materials and Methods

4.2.1 Dataset

A set of 30 proteins with complete experimental NMR chemical shift assignments and available high-resolution X-ray structures was chosen for training purposes. The list of proteins along with their PDB and BMRB identifiers is provided on the ShiftASA website. Note that the number of training proteins was varied to examine any enhancement in training and test performance. However no (or very little) improvement was observed with an increased number of proteins. Two separate sets of 65 and 92 proteins with available experimental chemical shifts and high-resolution X-ray structures were used as independent test sets. Henceforth we shall refer to the training data set and two test data sets as TRAIN, TEST1 and TEST2, respectively. The list of the TEST1 and TEST2 proteins along with their PDB and BMRB identifiers is provided on the ShiftASA website. No two proteins shared more than 40% sequence identity in the TRAIN set. Similarly, no two proteins shared more than 40% sequence identity in the TEST1 and TEST2 sets. The TRAIN proteins had ~92% and ~83% of their backbone and side-chain chemical shifts assigned, respectively. The TEST1 proteins had on average ~90% (max = 100%, min = 49%) and ~60% (max = 91%, min = 0%) of their backbone and side-chain chemical shifts assigned while those in TEST2 had an average of ~97.50% (max = 100%, min = 85%) and ~83.5% (max = 89%, min = 53%) of their backbone and side-chain chemical shifts assigned. Note that no attempt was made to handle missing assignments in either the training or the test data sets. The TRAIN proteins had ~49% of their residues in regular secondary structure while the TEST1 and TEST2 proteins had ~63% and ~44% (respectively) of their residues in regular secondary structure as assessed by STRIDE (Frishman et al. 1995).

4.2.2 Computation of Observed Fractional ASA

Most predictive studies associated with ASA prediction have focused on generating RSA or binary/ternary class predictions. However, in the majority of cases, real-valued or fractional ASA is more informative than the binary/ternary classification of residues into buried or exposed states. This is because the threshold for classifying residues in a protein into two or three exposure classes is subjective and often depends on the mean ASA over all the residues in a particular protein (Ahmad et al. 2003). In the absence of a universal threshold for categorical prediction of buried and exposed states, fractional ASA (fASA) is considered to be more reliable or useful estimation of residue-specific solvation status. Therefore for this study we focused on developing a predictor for fASA. The fractional ASA of a residue is defined as the ratio between absolute ASA (aASA) calculated within a three-dimensional structure and that is observed for a central residue location in an extended tri-peptide (*Ala-X-Ala*) conformation, denoted as mASA:

$$fASA_i = \frac{aASA_i}{mASA_i} \tag{4.1}$$

Hence, fASA values range between 0.0 and 1.0, with 0.0 corresponding to a fully buried and 1.0 to a fully exposed residue, respectively. Absolute ASA values were calculated using the Dictionary of Secondary Structure Prediction (DSSP) (Kabsch et al. 1983) program. The values of the extended state ASAs for all 20 residues were extracted from Eisenhaber et al. (1993).

4.2.3 Mapping Fractional ASA Prediction as a Regression Task

Given a protein with a length of *n* amino acids, the task is to estimate the fractional accessible surface area at each residue. We initially mapped the estimation problem as a regression task and then employed a Stochastic Gradient Boosting Tree model to solve the regression problem as outlined by Ridgeway (2007) and Trevor et al. (2001). To map the problem as a regression task, we defined an error function as the square of the difference between the observed per-residue fASA values and the predicted per-residue fASA values over the length of the training set sequences. The predicted per-residue fASA was calculated from a set of features (see below) and expressed as function $f^*(x)$, of amino acid position or sequence length. In stochastic gradient boosting, the method approximates the function $f^*(x)$, in an iterative fashion through fitting the solution tree in each step that maximally reduces the expectation of the error function. The gradient step in each iteration *m* (*m*= *1…T,* where T= total number of iterations), updates the model according to a learning rate or a shrinkage parameter that controls the rate at which the boosting algorithm descends upon the error surface. For each iteration, only a fraction *p* of the *N* training observations is randomly sampled (without replacement) and the next solution tree is grown with that subsample. The solution tree that is generated for each boosting iteration is a *K*-terminal node regression tree.

After mapping the fractional ASA prediction problem into a Stochastic Gradient Boosted Tree Model (SGBM), the model was optimized on the protein data in the training set. The

"GBM" package (Ridgeway 2007), written in R (Team R.D.C. 2008) was used for optimizing the training model.

4.2.4 Feature Set

To use or develop machine-learning algorithms it is necessary to extract a set of input features from the training data that will be used to infer or calculate the desired output (i.e. the fractional ASA). Features can either be the raw data (i.e. sequence, NMR chemical shifts, etc.) or derived data (i.e. estimated hydrophobicity) that is calculated from the raw data. We derived a set of five different feature types from our chemical shift and sequence data. The features included: (1) residue specific hydrophobicity, (2) chemical shift-derived three-state secondary structure probability, (3) random coil index values relating to flexibility using backbone and side-chain chemical shifts (Berjanskii et al. 2005, Berjanskii et al. 2013), (4) multiple sequence alignment derived residue conservation score (Valdar 2002, Mayrose et al. 2004), and (5) SABLE predicted ASA (Adamczak et al. 2004). These features are explained in more detail below.

Residue specific hydrophobicity

Hydrophobicity is a widely used physico-chemical characteristic of amino acids that is used to measure their relative water aversion. Hydrophobicity scales are numeric scales that define the relative hydrophobicity of amino acid residues. In general terms, the more positive the number, the more hydrophobic the amino acid, and consequently the more buried it is likely to be. Over the past few decades, a number of different hydrophobicity scales have been developed. We investigated six different hydrophobicity scales to see which was the most useful on a validation set when combined with other features. The scales we examined included Janin's scale (Janin 1979), Kyte and Doolittles's scale (Kyte et al. 1982), Eisenberg's scale (Eisenberg et al. 1984), Engelman's scale (Engelman et al. 1986), Hopp and Woods scale (Hopp et al. 1981) and Manavalan's scale (Manavalan et al. 1978). The best correlation was achieved using Janin's hydrophobicity values (data not shown). Interestingly Janin's scale was developed by analyzing the relative surface accessibility of all 20 amino acid residues from solved protein structures. In this regard Janin's scale is more a solvent accessibility scale than a hydrophobicity scale.

Two different approaches were examined regarding how to use hydrophobicity as a feature: i) single-residue hydrophobicity and ii) a running average of hydrophobicity over a 3-residue window. The first approach exhibited a comparatively better correlation than the second one (data not shown) and so this was incorporated in our feature set.

Chemical-shift derived secondary structure probability

The secondary structure probability of a residue is derived from the secondary chemical shift value of its constituent atoms. The secondary chemical shift ($\Delta\delta$) is defined as the difference between the absolute chemical shift (δ_{abs}) and the corresponding random coil (δ_{rc}) shift (Wishart et al. 2011).

$$\Delta\delta = \delta_{abs} - \delta_{rc} \tag{4.2}$$

The probability of a residue being in one of the three states "α-helix", "β-strand" or "coil" is derived from its six backbone atom secondary chemical shifts, as described in Wang et al. (2002a). For each backbone atom, a Gaussian probability distribution was assumed, where the two parameters of the distribution corresponded to 1) the average secondary chemical shift value for each of three different secondary structure states and 2) the standard deviation of the distribution. These statistical parameters were derived from the "RefDB" database (Zhang et al. 2002). A more detailed description of the secondary structure probability method is given by Wang et al. (2002).

Random Coil Index

The Random Coil Index (RCI) is a technique that can be used to determine the flexibility of an amino acid residue in a polypeptide chain from its backbone and side-chain chemical shifts (Berjanskii et al. 2005 and 2013). Both the backbone and side-chain RCI quantitatively trace the relative amount to which a protein backbone and side-chain's chemical shifts match with the random coil values. These features were calculated using the RCI equations provided in the original RCI papers.

Residue conservation score

Residue conservation is a measure of how often a given residue is seen at an equivalent position, in an equivalent protein, across different species. Generally highly conserved residues are buried within the protein core, while less conserved residues are generally exposed or found in loops (albeit with some exceptions). The conservation score for each residue position is calculated as described by Valdar (2002). First, a PSI-BLAST (Altschul et al. 1997) search with three iterations for query sequence is done on UniREf90 clustered database (UniProt Consortium. 2010). Then a multiple sequence alignment is performed using ClustalOmega (Sievers et al. 2011). The conservation score for each non-gap column in the alignment (i.e. each residue in the target sequence) is then calculated using Shannon's entropy formula as described below,

$$s(x) = \lambda \sum_{a}^{K} p_a \log p_a \tag{4.3}$$

where p_a is the probability of observing the a-th amino acid and λ is the scaling factor and defined as,

$$\lambda = [\log(\min(N, K))]^{-1} \tag{4.4}$$

where N = number of residues in the alignment, K = 20 (length of amino acid alphabet). The probability of observing a-th amino acid is the summed weight of sequences having the symbol *a* in the position *x* in the sequence which is defined as,

$$p_a = \sum w_i \tag{4.5}$$

where, w_i is the weight of the *i*-th sequence. w_i is defined as,

$$w_i = \frac{1}{L} \sum_{x}^{L} \frac{1}{k_x n_x} \tag{4.6}$$

where, L = length of the alignment, k_x = the number of amino-acid types present at the *x*-th position, n_x = the number of times the a-th amino acid occurring in the *i*-th sequence at the *x*-th position.

SABLE-predicted ASA

To further improve the performance of ShiftASA we supplemented our method with another sequence-only ASA prediction tool called SABLE (Adamczak et al. 2004). SABLE is a pure sequence-based method for predicting real-valued relative solvent accessibilities of amino acid residues in proteins. It was initially developed using neural network based regression models and later refined using other linear regression models (Wagner et al. 2005). It has a reported correlation coefficient between predicted and experimental values of 0.64-0.67 on various test sets. Because SABLE's correlation coefficient was comparable to the reported correlation of shift-based ASA estimations, it was expected that including sequence estimated ASA would enhance the performance of ShifASA. Therefore the SABLE predicted real valued ASA for each residue was included in the ShiftASA feature vector for the training and test data points.

Local residue interactions

To take into account the local-residue interaction in the protein structure, a 3-residue window feature set was used throughout this study. Accounting for nearby residue-interactions provides important information about local geometry and the local environment that is accessible/non-accessible to solvent.

4.2.5 Training the Prediction Model

The prediction model parameters were optimized so as to obtain an estimator that minimized the (absolute) difference between actual output and predicted ASA values. The model was also optimized to achieve a better correlation between the observed and response (i.e. predicted) variables. With those two objectives in mind, a repeated 10-fold cross-validation (CV) was performed to estimate the optimal number of iterations (*T*) and interaction depth of each regression tree (*K*) for our SGBM. This was done after the model had been initially fit on the set of 30 sample observations.

Optimization using 10-fold repeated cross-validation (CV) suggested that the optimal number of iterations should be 180. That is, the final regression model best approximates the response value after 180 gradient steps. The second parameter estimated by the cross-validated optimization process was the optimal depth of interaction among the predictor variables in each regression tree. The optimal depth of interaction was found to be eight (8). Specifically, the loss function was minimized when eight predictor variables were split in each regression tree during the optimization steps.

Analysis of feature influence

During the optimization of ShiftASA, an analysis of the feature influence was performed as a part of the boosting process. The top ten features are shown in Figure 1. The influence of the predictor variable (~X) indicates the relative importance or contribution of that variable in predicting the response (~Y) and can be estimated by the weighting coefficient associated with that variable in the method formulation. This analysis helped to identify those variables that had the most significant influence on the response. The weighting coefficients of all features are described in Table 4.1.

4.2.6 Evaluation

The performance of ShiftASA was evaluated using several different metrics. This was done to more completely ascertain its performance against other methods as well as to better assess the effects brought on by using different weighting protocols. Specifically the following metrics were used:

1) Root Mean Square Error (RMSE) - RMSE is a statistical measure that calculates the difference between the values predicted by an estimator model and the actual observed values. RMSE is the square root of the average squared deviation between predicted and actual values, and thus gives larger deviations more weight. A smaller value indicates a better model performance*,*

2) R^2 or the coefficient of determination - R^2 is a statistical measure that indicates how well a set of data points fit to a regression line or curve;

3) Spearman's rank correlation coefficient (SRCC) - SRCC is a non-parametric measure of the monotonic relationship between two variables, irrespective of whether their relationship is linear;

4) Mean absolute error (MAE) - MAE is the average of the absolute errors in a prediction i.e. the absolute difference between predicted and true values in a set of outcomes. Unlike other measures, larger deviations are not given additional weight;

5) Mean squared error (MSE)- MSE measures the average of the square of the "error" or deviation of the estimator from the quantity being estimated. MSE tends to heavily weight outliers.

4.3 Results and Discussion

4.3.1 Training Performance and Feature Importance

During the optimization process, a 10-fold repeated cross validation protocol yielded the lowest RMSE (0.18) and the best R-squared values (0.65) for the training data. The weighting coefficients of all features are described in Table 4.1. These data indicate that the SABLE (Adamczak et al. 2004) estimated ASA at the central *(i)-th* residue is the most informative ASA predictor. The side-chain random coil index, backbone random coil index and hydrophobicity, were found to be next three most influential variables in our fASA estimation. The next most important feature was the random coil index value of the *(i-1)-th* residue followed by helix propensity of the *(i)-th* and β-strand propensity of the *(i)-th* residue. The helix and β-strand propensities of the central residue have comparatively higher importance, they often indicate that this residue is buried as buried residues have a higher propensity to form □-helices and β-sheets in proteins and have a tendency to interact with the residues in the core region. Our analysis shows that central residue features carry the most information content (occupying six of the top seven positions), with exceptions of the flexibility information of neighboring residues (the RCI value of the *(i-1)-th* residue).

Feature	**Weighting coefficient, |w|**
ProbBCS(i-1)	1.114
ProbBCS(i)	2.465
ProbBCS(i+1)	0.849
ProbHCS(i-1)	0.860
ProbHCS(i)	2.541
ProbHCS(i+1)	1.692
ProbCCS(i-1)	1.089
ProbCCS(i)	1.577
ProbCCS(i+1)	2.294
hydro(i-1)	0.651
hydro(i)	3.557
hydro(i+1)	0.579
scon(i-1)	1.063
scon(i)	1.355
scon(i+1)	1.316
BackBoneRCI(i-1)	3.164
BackBoneRCI(i)	5.381
BackBoneRCI(i+1)	1.947
SideChainRCI(i-1)	1.124
SideChainRCI(i)	27.482
SideChainRCI(i+1)	0.839
sable(i-1)	0.560
sable(i)	35.991
sable(i+1)	0.505

Table 4.1: Normalized weighting coefficients (|**w**|) of chemical-shift and sequence-derived features are listed. (i-1) and (i+1) in parentheses represents the neighbor residue locations, whereas (i) indicates the central residue. The feature name abbreviations are as follows: ProbBCS = β-strand probability using chemical shift, ProbHCS = α-helix probability using chemical shift, ProbCCS = coil probability using chemical shift, hydro = residue-specific hydrophobicity, scon = residue conservation score, BackBoneRCI = Random Coil Index (protein flexibility) from backbone chemical shifts, SideChainRCI = Random Coil Index from side-chain chemical shifts, and sable = SABLE predicted real-value solvent accessibility.

Although the SABLE estimation is found to be the most relevant feature, the chemical shift features also provide significant contribution, roughly equal to the SABLE feature at central residue position. Residue hydrophobicity also carries useful information to estimate the fASA.

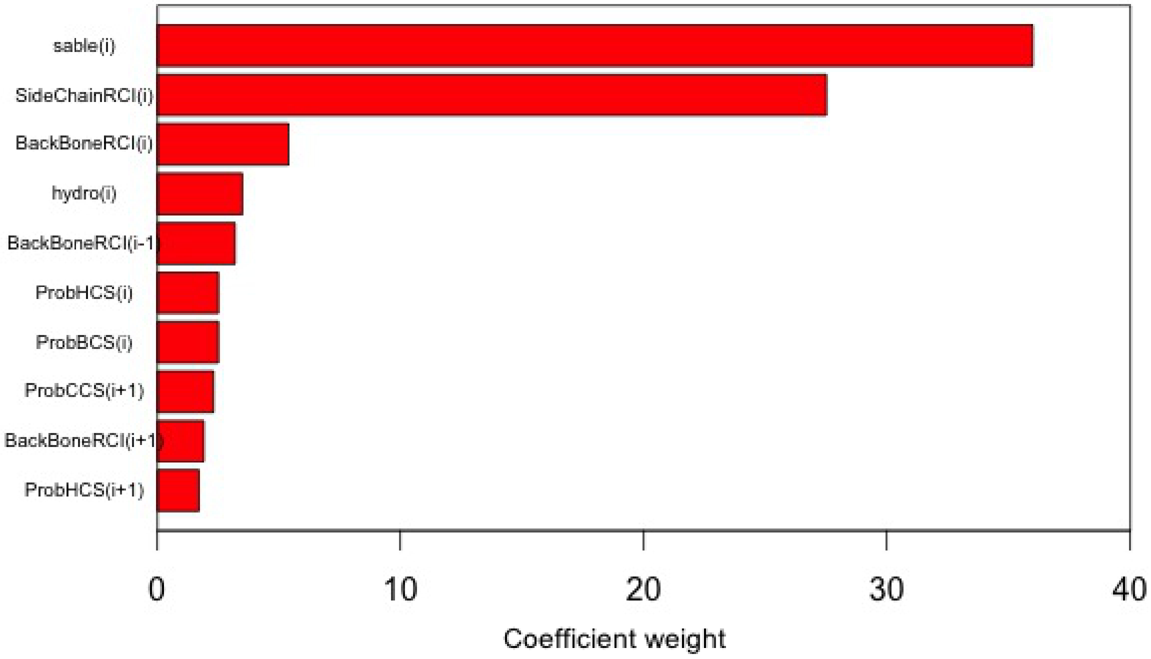


Figure 4.1: Top ten relevant features in the SGBM method. The importance of these predictor variables or features is normalized to a scale of 1-100. The predictor variable names are shown on the vertical axis.

Other than the SABLE ASA estimation and hydrophobicity, eight of the top ten features are chemical-shift features, and have collectively larger weights in the final formulation. It is notable that residue conservation scores are not present among ten most relevant features, which indicates their somewhat smaller contribution to the feature set.

4.3.2 Test Performance

The final parametric regression tree model generated by repeated cross-validated optimization of the TRAIN set was used to predict the fractional ASA values for proteins in the TEST1 and TEST2 data sets. The Spearman correlation coefficient was calculated between the actual fASA and the predicted fASA using both our ShiftASA method and five other models. The results are shown in Table 4.2 and Table 4.3 (Table 4.4 lists the individual performance of all TEST1 proteins). As seen in Table 4.2, the mean correlation coefficient for the predicted fASA values for ShiftASA of 0.79. This corresponds to an 8.2% improvement over the best sequence-only and a 22% improvement over chemical shift-only prediction methods. The prediction accuracy of the different methods was also evaluated using other statistical metrics and is shown in the same table.

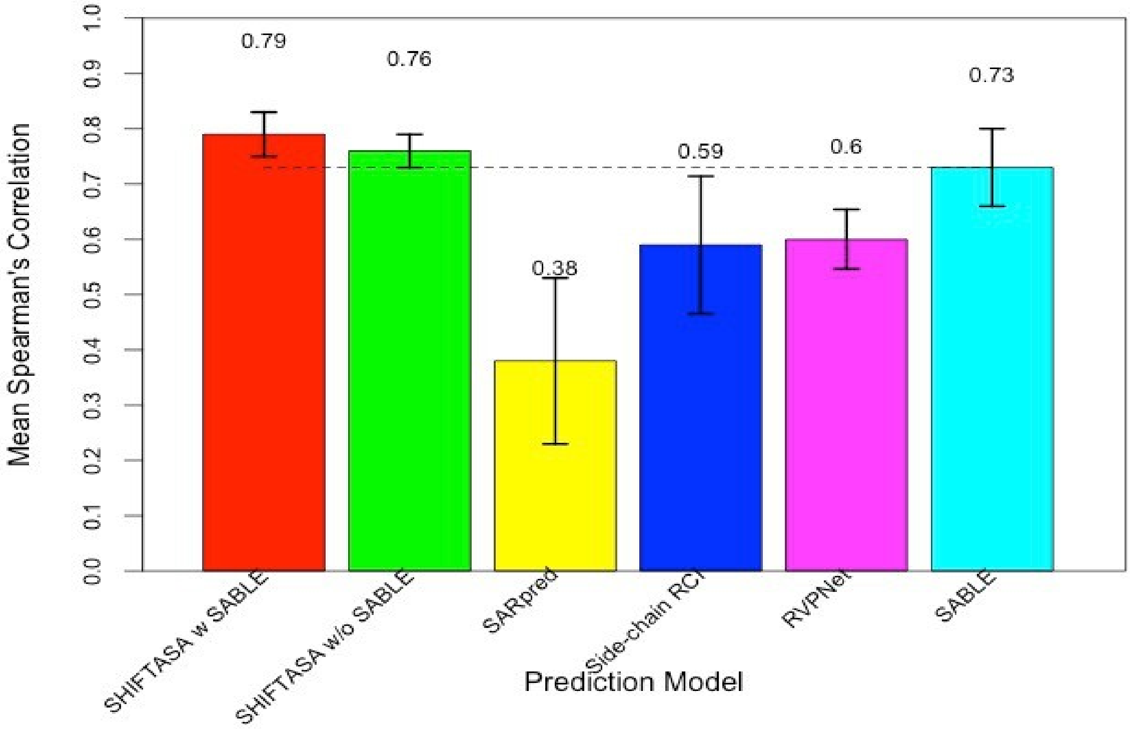


Figure 4.2: Mean Spearman correlation coefficient and the standard deviation of correlations of all five fASA prediction models (including ShiftASA) are shown. The performance is measured over the TEST1 data set. The mean correlation associated with each method is shown at the top of each bar diagram.

Table 4.2 also shows that ShiftASA reports the highest mean prediction accuracy among all five methods that were evaluated. The mean absolute error was decreased from 0.20 (the best MAE among other methods) to 0.14 Å^2 with ShiftASA, which is a 26% improvement over the best sequence-only method. The mean squared error also decreased to 0.03 from 0.07 Å^2 as measured over all TEST1 proteins. Moreover, ShiftASA shows the lowest deviation in Spearman's rank correlations. These data indicate that ShiftASA is not only the most accurate, but also the most consistent among the five methods. Bar plots exhibiting the mean Spearman's rank correlations and the corresponding standard deviations reported by the five methods are shown in Figure 4.2.

Evaluation Metric	**ShiftASA (w SABLE)**	**ShiftASA (w/o SABLE)**	**SABLE (Seq.)**	**RVPNet (Seq.)**	**SARpred (Seq.)**	**Side-chain RCI (Chem. Shift)**
MAE	0.14	0.16	0.19	0.20	0.24	0.20
MSE	0.03	0.04	0.07	0.07	0.09	0.07
RMSE	0.19	0.20	0.26	0.26	0.31	0.26
Minimum Spearman Correlation	0.72	0.70	0.54	0.47	0.21	0.22
Mean Spearman Correlation	0.79	0.76	0.73	0.60	0.38	0.59
Maximum Spearman Correlation	0.86	0.83	0.82	0.70	0.67	0.77
Standard deviation (Spearman Correlation)	0.04	0.03	0.07	0.05	0.15	0.12

Table 4.2: The MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), Mean Spearman's Correlation, and the standard deviation of Spearman's correlation for all six fASA prediction methods (including ShiftASA, with and without SABLE) evaluated over the TEST1 set (65 test proteins) [64 proteins for side-chain RCI]

Evaluation Metric	**ShiftASA (w SABLE)**	**ShiftASA (w/o SABLE)**	**SABLE (Seq.)**	**Side-chain RCI (Chem. Shift)**
MAE	0.14	0.15	0.31	0.23
MSE	0.03	0.04	0.16	0.09
RMSE	0.17	0.20	0.41	0.30
Minimum Spearman Correlation	0.67	0.76	0.20	0.27
Mean Spearman Correlation	0.82	0.79	0.67	0.73
Maximum Spearman Correlation	0.89	0.88	0.85	0.84
Standard deviation (Spearman Correlation)	0.04	0.03	0.12	0.07

Table 4.3: The MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), Mean Spearman Correlation coefficient, and the standard deviation of Spearman correlation coefficient for ShiftASA (with and without SABLE), SABLE (Adamczak et al. 2004) and Side-chain RCI (Berjanskii et al. 2013) evaluated over the TEST2 set.

The Spearman correlation coefficients for the TEST2 proteins as well as other statistical measures from predictions derived by ShiftASA, SABLE (Adamczak et al. 2004) and Side-chain RCI (Berjanskii et al. 2013) are shown in Table 4.3. As seen in this table, ShiftASA estimation has a correlation of 0.82, whereas SABLE's and Side-chain RCI's correlations are 0.67 and 0.73 respectively. The mean absolute error is also significantly decreased (0.14 compared to 0.31 and 0.14 compared to 0.23).

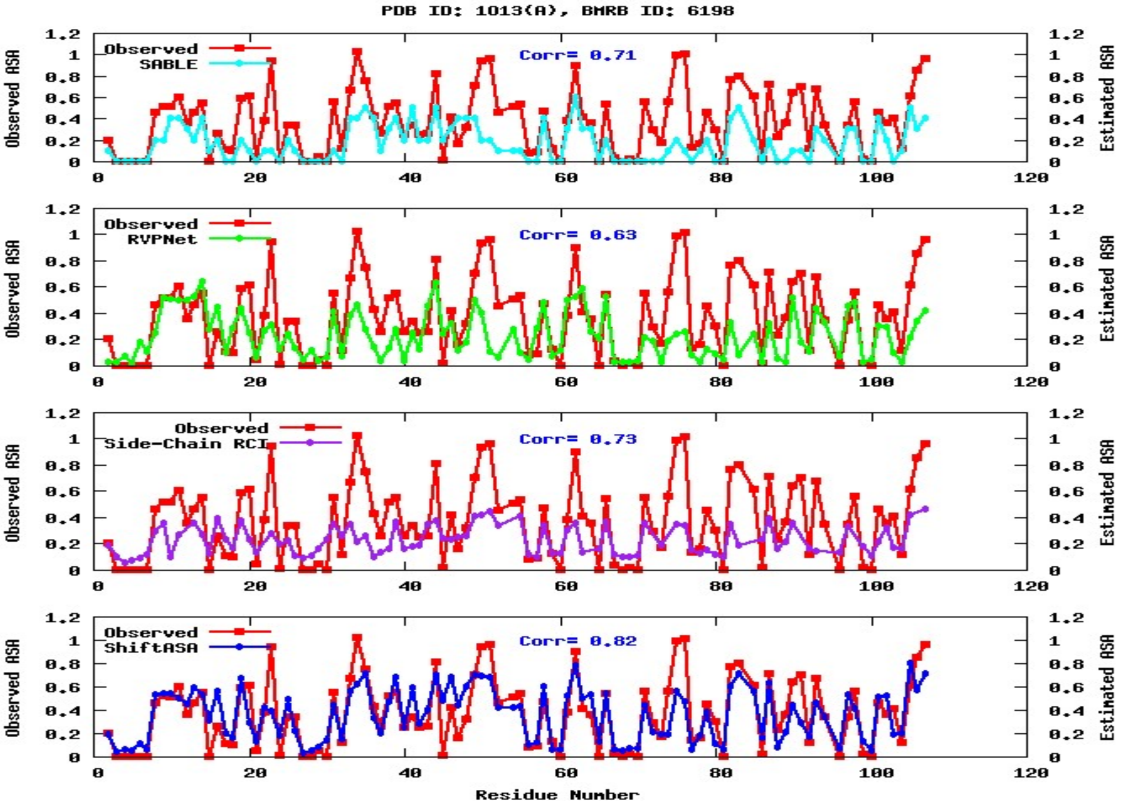


Figure 4.3: Agreement between predicted and observed residue-specific fASA values by SABLE, RVPNet, Side-chain RCI and ShiftASA for the putative dinitrogenase iron-molybdenum cofactor from Thermotoga martima (PDB ID: 1O13, chain A). The corresponding BMRB ID is 6198. The Spearman correlation coefficient is shown in the centre of each graph.

Examples of the per-residue correlation for the predicted fASA values of two protein chains, the double-sided ubiquitin binding of Hrs-UIM (PDB ID: 2D3G(B)) and a putative dinitrogenase iron-molybdenum cofactor from *Thermotoga martima* (PDB ID: 1O13(A)) are displayed in Figure 4.3 and Figure 4.4 respectively. The first example shows a stronger correlation (0.82) between SHIFTASA and the observed fASA, compared to SABLE, RVP-Net and Side-chain RCI reported values (0.71, 0.63 and 0.73 respectively).

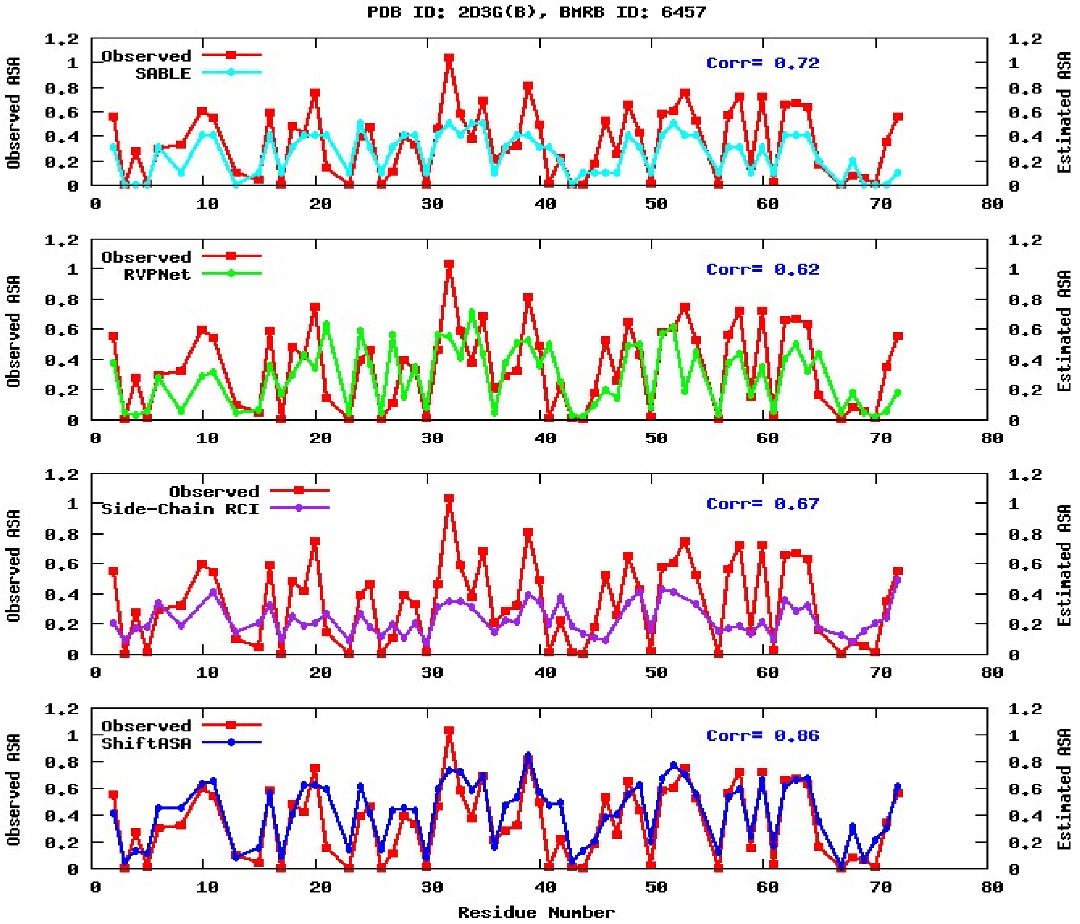


Figure 4.4: Agreement between the predicted and observed residue-specific fASA values for SABLE, RVPNet, Side-chain RCI and ShiftASA for ubiquitin bound to Hrs-UIM (PDB ID: 2D3G, chain B). The corresponding BMRB ID is 6457. The Spearman correlation coefficient is shown on the right of each graph.

For the second example, a stronger correlation (0.86) is also evident, compared to the correlations (0.72, 0.62 and 0.67) reported by three other methods, namely SABLE (Adamczak et al. 2004), RVPNet (Ahmad et al. 2003), and Side-chain RCI (Berjanskii et al. 2013). As seen in Figures 4.3 and 4.4, ShiftASA yields better agreement in matching the observed ASA amplitude, which certainly contributes to its higher correlation coefficients.

BMRB ID	ShiftASA w SABLE (Seq. + Chem. Shift)	ShiftASA w/o SABLE (Seq. + Chem. Shift)	SABLE (Seq.)	RVPNet (Seq.)	SARpred (Seq.)	Side-chain RCI (Chem. Shift)
6338	0.855	0.755	0.823	0.575	0.540	0.555
6457	0.865	0.823	0.724	0.621	0.597	0.675
6198	0.823	0.751	0.706	0.634	0.341	0.730
7322	0.818	0.728	0.789	0.599	0.350	0.660
15517	0.841	0.729	0.787	0.613	0.346	0.585
15249	0.803	0.725	0.798	0.666	0.318	0.593
16007	0.809	0.728	0.805	0.627	0.214	0.679
4113	0.795	0.777	0.761	0.669	0.359	0.571
6197	0.785	0.792	0.768	0.563	0.641	0.617
6032	0.820	0.779	0.738	0.551	0.393	0.704
4083	0.782	0.744	0.754	0.574	0.421	0.587
7242	0.791	0.734	0.768	0.610	0.398	0.643
15501	0.807	0.78	0.803	0.662	0.387	0.404
15741	0.745	0.721	0.743	0.650	0.249	0.354
4091	0.813	0.785	0.666	0.606	0.440	0.722
4077	0.783	0.747	0.550	0.595	0.334	0.650
4562	0.760	0.744	0.755	0.570	0.358	0.534
4371	0.835	0.73	0.726	0.472	0.320	0.770
4031	0.800	0.762	0.786	0.631	0.514	0.580
4082	0.786	0.748	0.681	0.546	0.253	0.685
4091	0.813	0.742	0.660	0.624	0.430	0.710
4421	0.817	0.803	0.696	0.632	0.396	0.645
5571	0.825	0.772	0.743	0.699	0.274	0.628
5623	0.833	0.809	0.823	0.649	0.229	0.702
5756	0.813	0.77	0.723	0.572	0.433	0.631
5799	0.744	0.742	0.757	0.670	0.243	0.426
5921	0.793	0.738	0.780	0.614	0.494	NA
6075	0.764	0.757	0.703	0.619	0.272	0.647
6122	0.774	0.715	0.656	0.556	0.451	0.701
6375	0.721	0.756	0.694	0.466	0.293	0.475
6494	0.815	0.722	0.786	0.695	0.377	0.620
6503	0.797	0.771	0.747	0.636	0.533	0.675
4031	0.806	0.761	0.787	0.618	0.500	0.560
4566	0.832	0.763	0.793	0.613	0.441	0.754
4296	0.813	0.708	0.718	0.600	0.591	0.702
4094	0.740	0.715	0.653	0.508	0.329	0.594
4019	0.778	0.746	0.653	0.639	0.315	0.551
5211	0.811	0.776	0.775	0.548	0.489	0.514
1062	0.805	0.746	0.757	0.581	0.524	0.684
6776	0.815	0.773	0.755	0.560	0.471	0.589
6575	0.851	0.789	0.781	0.544	0.430	0.657
7086	0.833	0.758	0.802	0.644	-0.464	0.663

BMRB ID	ShiftASA w SABLE (Seq. + Chem. Shift)	ShiftASA w/o SABLE (Seq. + Chem. Shift)	SABLE (Seq.)	RVPNet (Seq.)	SARpred (Seq.)	Side-chain RCI (Chem. Shift)
4077	0.779	0.743	0.704	0.602	0.317	0.652
10096	0.742	0.754	0.763	0.574	0.214	0.246
4717	0.745	0.69	0.682	0.584	0.223	0.602
7086	0.832	0.746	0.821	0.648	0.323	0.649
6122	0.772	0.75	0.634	0.505	0.450	0.694
4717	0.786	0.735	0.753	0.568	0.284	0.643
5540	0.802	0.803	0.707	0.676	0.263	0.654
5571	0.797	0.786	0.707	0.676	0.263	0.613
5529	0.703	0.747	0.575	0.561	0.362	0.522
4039	0.772	0.747	0.708	0.523	0.350	0.605
4041	0.781	0.76	0.708	0.523	0.350	0.595
4554	0.781	0.769	0.822	0.654	0.371	0.318
5740	0.808	0.777	0.832	0.657	0.372	0.516
15084	0.822	0.774	0.667	0.643	0.352	0.607
15854	0.789	0.76	0.667	0.598	0.352	0.570
5387	0.734	0.734	0.548	0.507	0.642	0.523
6779	0.746	0.746	0.558	0.630	0.380	0.555
5226	0.764	0.777	0.678	0.660	0.561	0.633
6019	0.836	0.826	0.788	0.663	0.533	0.701
5286	0.741	0.784	0.737	0.657	0.519	0.248
6541	0.75	0.752	0.737	0.657	0.519	0.580
15650	0.731	0.743	0.756	0.624	0.464	0.223
15852	0.722	0.754	0.740	0.615	0.455	0.287
Average	**0.79±0.04**	**0.76±0.03**	**0.73±0.07**	**0.60±0.05**	**0.38±0.15**	**0.59±0.12**

Table 4.4: Spearman correlation coefficient between the actual fASA and the fASA values as predicted by ShiftASA (with and without SABLE), SABLE (Adamczak et al. 2004), RVPNet (Ahmad et al. 2003), SARpred (Garg et al. 2005) and side-chain RCI (Berjanskii et al. 2013) using TEST1 proteins. The "NA" value in any column indicates that no result is available for that particular protein using that specific method.

4.3.3 Buried-Exposed and Buried-Intermediate-Exposed Classification

Categorical ASA measures are still commonly used in the field of ASA prediction and evaluation. However, no universal threshold for categorical prediction of buried and exposed states exists and so fractional ASA (fASA) is generally considered to be a more reliable estimation of residue-specific solvation status. Nevertheless, we performed a detailed evaluation of ShiftASA's performance for categorical ASA prediction. Two-state and three- state classification of residue fractional ASA values for different threshold systems were calculated based on the real-value fASA predictions by ShiftASA, SABLE (Adamczak et al. 2004) and RVPNet (Ahmad et al. 2003). The number of residues in each (Exposed, Intermediate or Buried) class using different threshold cutoffs including the accuracy and precision of classification results are described in Table 4.5.

Threshold System	ShiftASA		SABLE		RVPNet	
	Accuracy	Precision	Accuracy	Precision	Accuracy	Precision
0% (2-state) #B=823, #E=6711	0.90	0.89	0.80	0.80	0.89	0.90
5% (2-state) #B=1672, #E=5862	0.81	0.80	0.84	0.82	0.79	0.80
10% (2-state) #B=2202,#E=5332	0.82	0.78	0.76	0.75	0.77	0.77
15% (2-state) #B=2603,#E=4931	0.82	0.80	0.78	0.78	0.76	0.76
25% (2-state) #B=3420, #E=4114	0.81	0.78	0.76	0.73	0.74	0.74
50% (2-state) #B=5246, #E=2288	0.81	0.80	0.70	0.66	0.72	0.73
10%-20% (3-state) #B=2202,#I=808,#E=4524	0.71	0.68	0.72	0.62	0.67	0.67
15%-25% (3-state) #B=2603,#I=817,#E=4114	0.72	0.69	0.66	0.67	0.65	0.65
25%-50% (3-state) #B=3420,#I=1826,#E=2288	0.66	0.62	0.58	0.50	0.54	0.55

Table 4.5: Two-state and three-state mean classification accuracy and precision of fASA for the TEST1 set reported by ShiftASA, SABLE and RVPNet for different threshold systems. The number of buried (#B), intermediate (#I) and exposed (#E) residues for each threshold system are shown in the first column

The performance of ShiftASA for two-state and three-state classification of real-value solvent accessibility using different threshold values was found to be comparable or higher (in most cases) than that of RVPNet (Table 4.5). ShiftASA also showed consistently high accuracy (≥80%) for all threshold values in the two-state classification. Three-state classifications (buried-intermediate-exposed) were challenging for the current method (although ShiftASA reports better accuracies than RVPNet). The probable reason might be the lower estimation accuracies associated with more exposed residues (fASA range ≈ 0.6-1.0 -- see Discussion for more details).

4.3.4 Discussion

The performance of ShiftASA is clearly superior to other methods for fASA prediction. Obviously the inclusion of experimental information (i.e. NMR chemical shifts) means that additional information, beyond sequence data, is being exploited in the prediction process. However, unlike most other sequence-based fASA predictors, ShiftASA also makes use of residue-specific hydrophobicity to help with its prediction. This is based on the fact that there is a strong relationship between residue-specific hydrophobicity and solvent exposure (Manavalan et al. 1978). Indeed, several hydrophobicity scales have been derived by calculating the solvent accessible surface area for residues in solved proteins or by employing empirical solvation parameters derived from calculated surface areas (Chothia 1976 and Biswas et al. 2003). Because ShiftASA employs both chemical shifts and residue-specific hydrophobicity, it would be of interest to analyze the predictive ability of sequence alone and chemical shift alone to estimate fractional ASA values. In addition it would also be useful to explore how ShiftASA's prediction accuracy varies as the fraction of complete shift assignment changes. In the following subsections, we investigated these two issues along with other issues based on the performance of the TEST1 proteins.

Sequence and chemical-shift based prediction -- combined vs. alone

To address the issue of predictive accuracy for sequence-only vs. shift-only vs. combined, another two stochastic gradient boosting regression tree models were developed and trained using sequence-only and chemical shift-only training features for each residue in a 3-residue window. Parameter optimization indicated the optimal "number of trees" as 150 and 225 and the optimal "interaction depth" as six (6) and eight (8) respectively for the final regression trees of these sequence-only and chemical shift-only models. The optimized models were then evaluated on TEST1 proteins. Figure 4.5 shows the correlation between the actual and predicted fASA values using the sequence-only and chemical shift-only prediction models. For comparative purposes, the correlations of ShiftASA's predictions are also shown. The graph clearly shows a significant performance difference. Note that, the mean correlation for our sequence-only prediction is 0.60 (for the 65 protein test set). The green line in the graph shows the correlation between chemical shift derived parameters and fASA values, which is 0.46. These results show that the performance improvement seen for ShiftASA was not just achieved through the use of sequence-derived parameters, but also by the sensible use of chemical shift data.

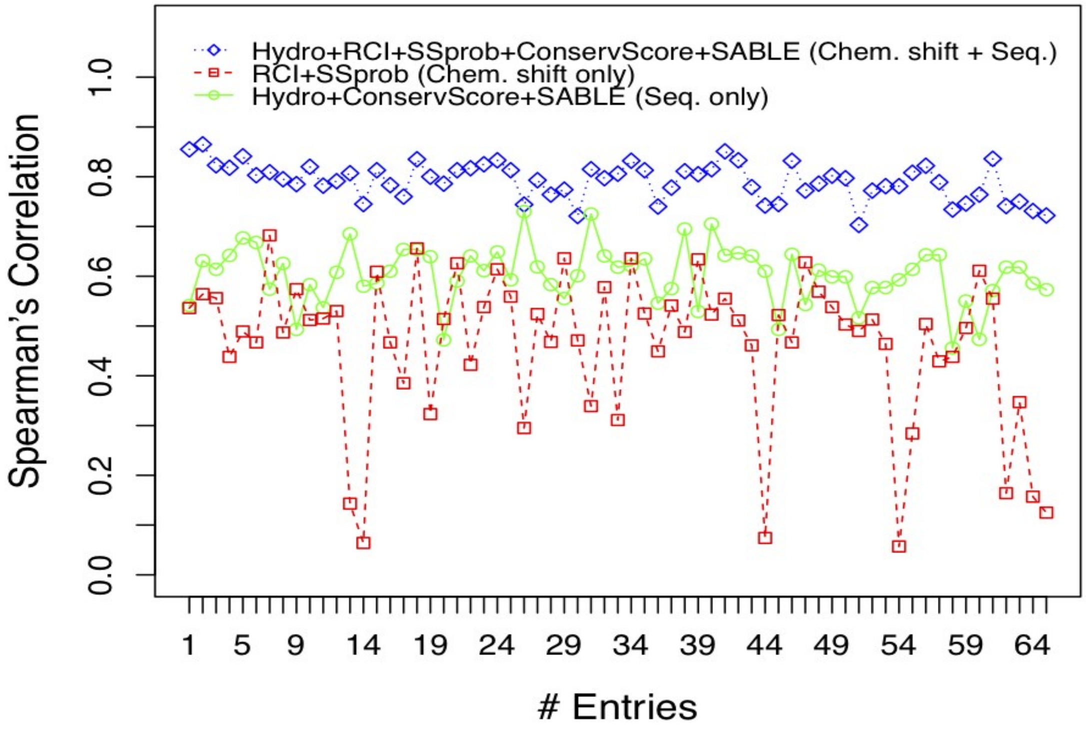


Figure 4.5: Spearman correlation between observed and predicted ASA for TEST1 proteins using both sequence and chemical shift features [Hydrophobicity, RCI, secondary structure probability and conservation score and SABLE] (blue diamonds), sequence only [Hydrophobicity, conservation score and SABLE predicted ASA] (red rectangles) and chemical shift only [RCI (backbone and side-chain) and secondary structure probability] (green circles). The average correlation between observed and predicted ASA using combined features is 0.79 over all TEST1 proteins, whereas the average correlation using sequence-only features is 0.60 and 0.46 when using chemical-shift-only features. As seen in Figure 4.6, the low correlation when using only chemical shifts is attributable to incomplete chemical shift assignments.

It is also interesting to compare this sequence-only method with the three sequence-only fASA prediction systems we evaluated in this study, namely, SABLE (Adamczak et al. 2004), RVPNet (Ahmad et al. 2003) and SARpred (Garg et al. 2005). SABLE, RVPNet and SARpred all are neural network-based prediction systems. SABLE uses feed-forward neural networks that estimate real value RSA's based on information derived from the PSI-BLAST (Altschul et al. 1997) position specific scoring matrix (PSSM), hydrophobicity, volume, entropy and secondary structure propensity of amino acids in a running window of 11 residues. The second method (RVPNet) uses only sequence data with adjacent neighbor information encoded in a binary (0/1) sequence, and the last method (SARpred) method feeds multiple sequence alignments into a two stage neural network to predict fASA. For the TEST1 proteins, SABLE achieved a mean correlation of 0.73 and RVPNet achieved a correlation of 0.60, whereas the correlation of SARpred was found to be 0.38. With the exception of SABLE, the other two methods (RVPNet and SARpred) appear to be either comparable or significantly worse than our sequence-only approach.

Prediction error vs. complete shift assignments

Recent studies (Marsh 2013, Berjanskii et al. 2013) have demonstrated a correlation between fractional accessible surface area and local flexibility as well as global flexibility. Marsh (2013) found a mean correlation of 0.61 between RCI-predicted local flexibility and residue-specific fractional ASA over a set of monomeric proteins. Likewise, Berjanskii et al. (2013) found a correlation of 0.74 between the side chain RCI (a chemical shift-derived parameter) and residue specific fASA over a set of 15 proteins. However, one of the limitations of these RCI-based methods is that a complete or near-complete chemical shift assignment is required to achieve relatively moderate prediction accuracy. In the present study, we found the mean Spearman correlation of the side-chain RCI method over proteins in TEST1 to be 0.59, which was somewhat less than what was originally reported (albeit using a different set of proteins). It was also found that the side-chain RCI method was particularly sensitive to missing or incomplete assignments. This is reflected in the spread of up to 12% in the Spearman correlation coefficient distributed over the TEST1 set. Fortunately, one of the strengths of ShiftASA is the fact that it is not solely dependent on side-chain chemical shifts but also on the relatively more easily measured backbone chemical shifts. Furthermore, when all of the described sequence and chemical-shift derived features (see section 4.2.4) are combined, ShiftASA's accuracy does not vary significantly in the absence of complete shift assignments.

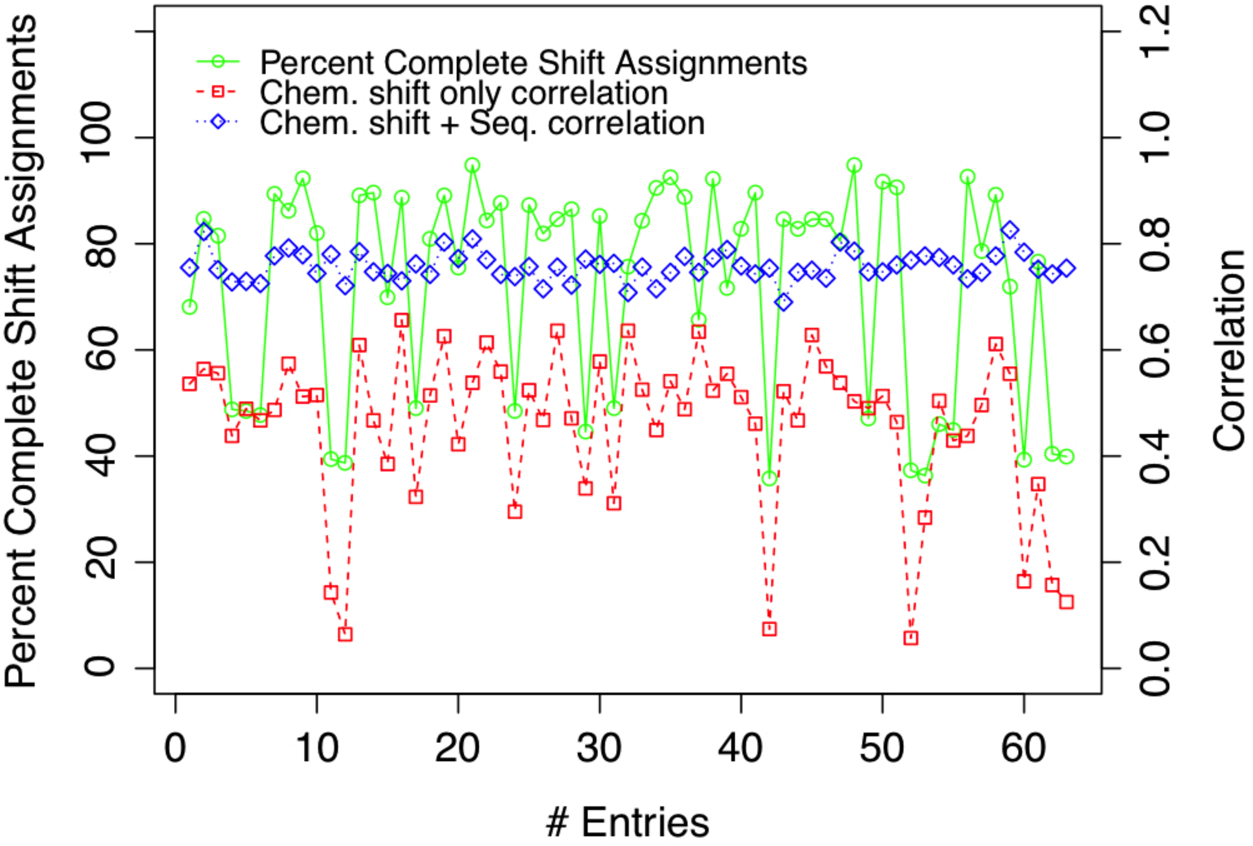


Figure 4.6: Percent complete chemical shift assignments, scaled in 0-100 (green circles) vs. Spearman's rank correlation coefficients (red rectangles and blue diamonds) for the TEST1 set. This graph indicates that the chemical shift and sequence-based estimation error is relatively insensitive to complete shift assignments, whereas only chemical shift-based correlation shows a relatively high sensitivity (Pearson correlation coefficient of 0.75) to the shift assignment completeness.

This invariance is shown using the line connected by blue diamonds in Figure 4.6. We believe the robustness that ShiftASA exhibits to missing chemical shifts is due to the redundancy in information that is available from both sequence and neighboring residue chemical shift data. On the other hand, chemical shift-only estimation performance varies with the amount of complete shift assignments and shows a Pearson correlation coefficient of 0.75 (depicted by a line connected by red rectangles in Figure 4.6)

Prediction error vs. residue specific variance in test set ASA distribution

Figure 4.7 illustrates the relationship between the standard deviation of the fASA value for each of the 20 different amino acids in the test set and the corresponding fASA prediction error. The variance in the fASA values shows a relatively good agreement with the prediction error (MAE), yielding a Spearman correlation coefficient of 0.92. In general, the prediction error in fASA values for exposed residues is higher than for buried residues. Frequently buried and partially buried residues such as CYS, ILE, VAL, PHE, TYR, TRP and LEU have comparatively lower variability in the observed fASA values, leading to the lower associated prediction errors.

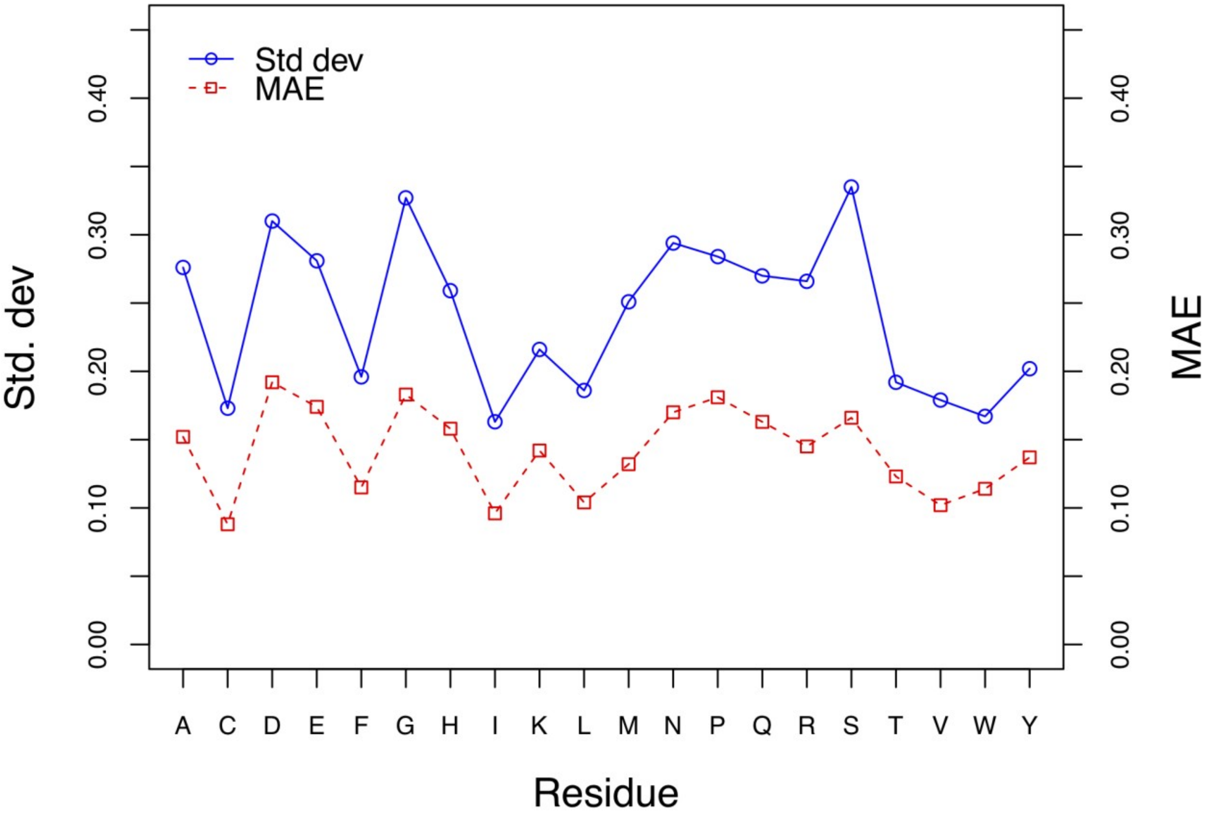


Figure 4.7: Standard deviation of ASA values in the test set for 20 different amino acids (blue circles) and prediction error (MAE) (red rectangles) are shown. A strong correlation (Pearson correlation coefficient =0.92) is observed between the variance of observed ASA in the test data and the associated ASA prediction error for different amino acids.

Among these residues, CYS, ILE and VAL have less than a 10% mean prediction error, while others are within a 10-13% error range. This might be because buried residues generally have a more conserved fASA distribution, as can be seen in Figure 4.7. In contrast, exposed and partially exposed residues such as ASP, GLU, PRO and GLY have a much higher ($\geq$17%) mean estimation error. ASN, GLN, SER, HIS and ALA fall into the medium range of prediction errors (15-16%). These increased prediction errors might be a consequence of the high fASA variability seen in exposed residues (see Figure 4.7). The most difficult residue to predict is ASP, which produces the highest mean prediction error of 19.3%. All aromatic residues (PHE ~ 11%, TRP ~ 12%, TYR ~ 13%) are within a 13% error limit, which again confirms their relatively buried nature or their affinity to associate with residues in buried regions. Overall our data show that buried and partially buried residues are predicted with relatively higher accuracy than exposed, partially exposed or charged residues. More exposed residues tend to have fewer assignments due to their higher mobility, higher overlap, and lower importance to researchers.

ASA range vs. prediction error vs. training point fractions

The error distribution with the fASA value range and the corresponding sample training size revealed some interesting and unexpected trends. These are shown in Figure 4.8. The training fraction curve reveals that there is a relative abundance of chemical shift (and ASA) data for buried and partially buried regions of proteins, which facilitates higher prediction accuracies in those regions. As training fractions slowly decrease for higher fASA ranges (partially exposed and fully exposed residues), so does the prediction accuracy for those residues.

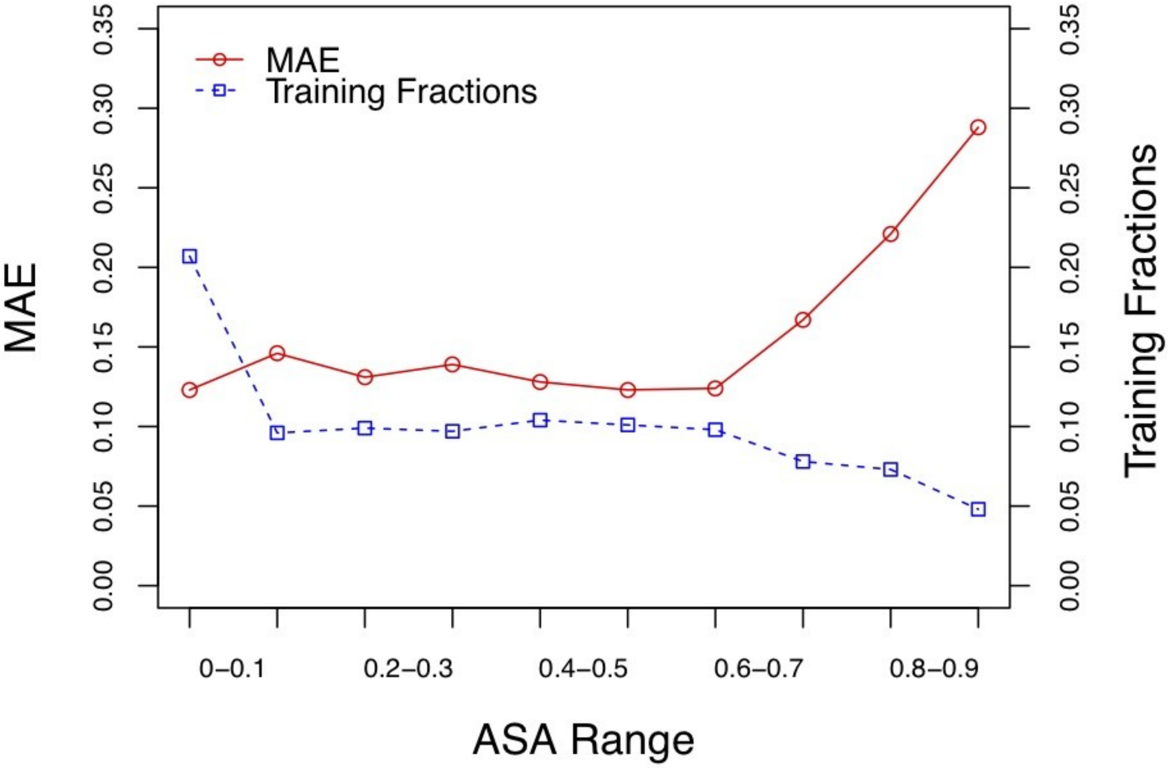


Figure 4.8: The relationship between fASA range vs. prediction error (MAE) (blue triangles) and ASA range vs. sample training size (red circles) is depicted. The prediction error increases (red line), as the number of observations (blue line) decreases in the training data.

This trend partially explains why ShiftASA performs somewhat differently in estimating the accessible surface area of buried, partially buried, partially exposed and fully exposed residues in proteins.

SABLE improves prediction performance

In ShiftASA, we tried to incorporate as much information as available both from sequence and chemical shifts in order to achieve optimal performance. Because of the excellent performance of the sequence-only method SABLE (Adamczak et al. 2004) we decided to include its sequence-based prediction into the ShiftASA algorithm. Indeed, this addition led to an increase of mean correlation coefficients between predicted and experimental values from 0.76 to 0.79 (TEST1) and 0.79 to 0.82 (TEST2). This improvement is statistically significant ($p<0.001$). Evidently

SABLE's sequence-driven structural homology and evolutionary profile based prediction provides additional information that helps to accurately estimate the buried/exposed states of residues.

ShiftASA accurately estimates fractional ASA of "unfolded" proteins

We also investigated the performance of ShiftASA for estimating fractional ASA values for a completely unfolded protein (i.e. unfolded ubiquitin in 8 M urea - BMRB 4375).

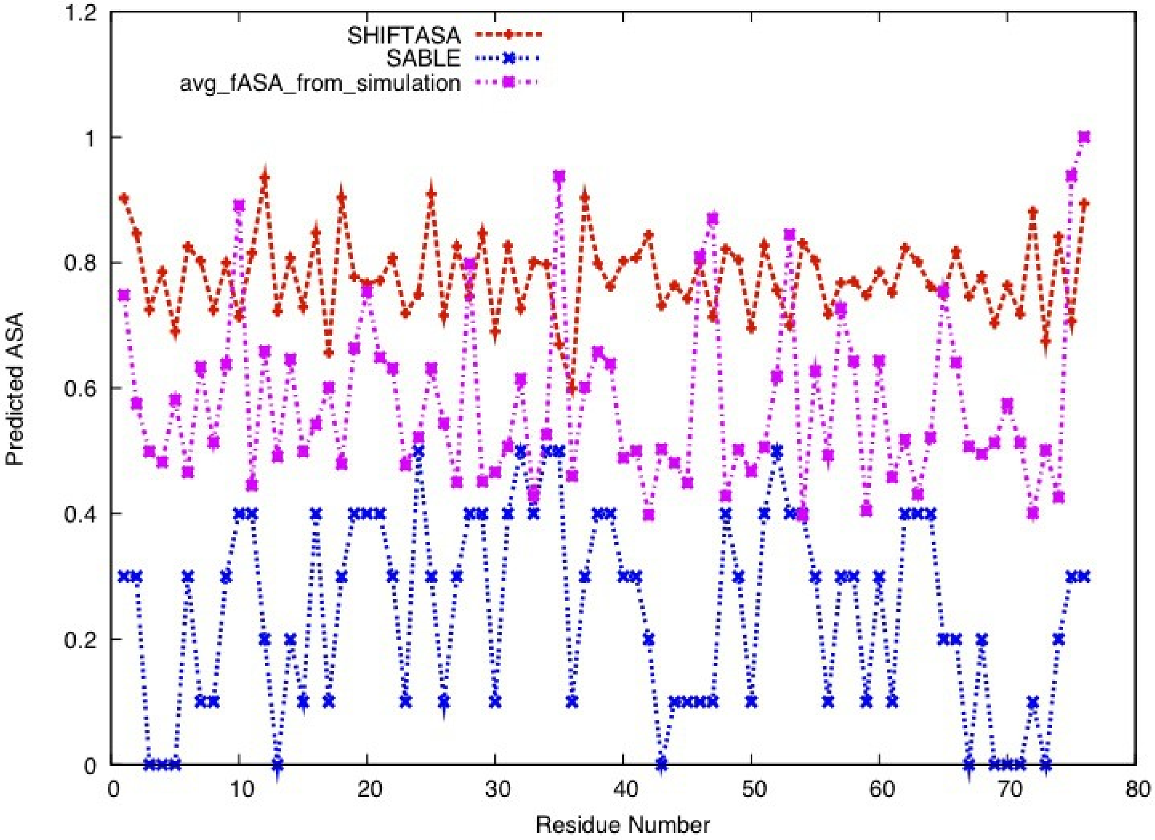


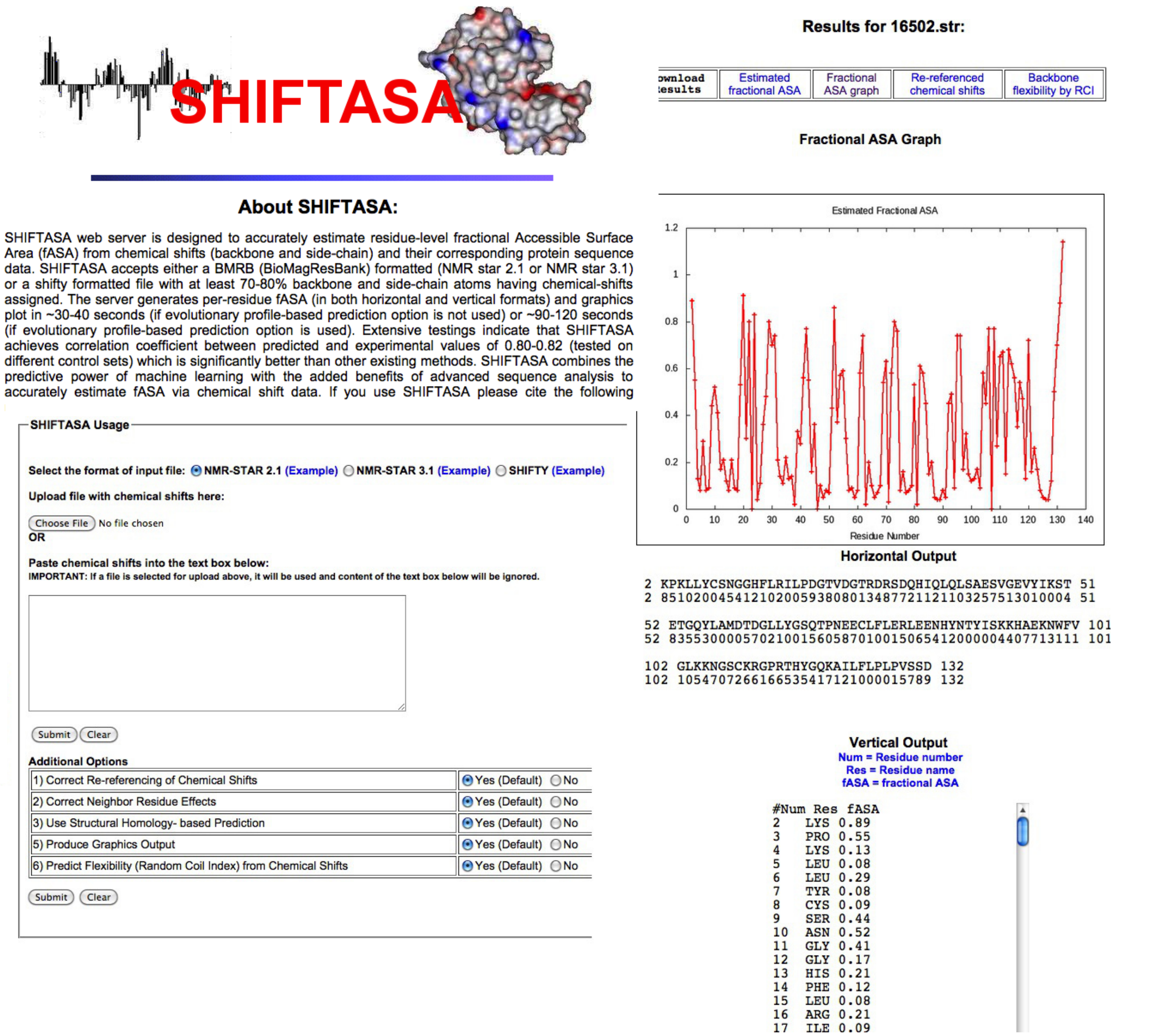
Figure 4.9: Per-residue fASA values for unfolded ubiquitin (BMRB ID: 4357). The red, magenta and blue lines indicate the estimated fASA by ShiftASA, the average fASA from 10,000 simulated unfolded structures of ubiquitin and the estimated fASA by SABLE respectively.

As a substitute for observed fASA values, an average per-residue fASA value is calculated from 10,000 unfolded structures of ubiquitin generated using the computer program Flexible Meccano (Ozenne et al. 2012). As seen in Figure 4.9, ShiftASA was able to estimate the exposed state of this protein with a moderate accuracy. In contrast to the sequence-only method, SABLE

(Adamczak et al. 2004) most of the protein was estimated to contain a high proportion of buried regions. Because SABLE predicts the fractional ASA from sequence, it simply reported the ASA states of the folded ubiquitin structure retrieved by a PSI-BLAST (Altschul et al. 1997) search. However, because ShiftASA weighs both the experimental chemical shift information with sequence-derived features, its performance was not compromised.

4.3.5 The ShiftASA Web Server

A web server (http://shiftasa.wishartlab.com) has been developed that accepts a BMRB (NMR–Star 2.1 or NMR-Star 3.1) or SHIFTY-formatted chemical shift file and generates per-residue fractional ASA (in both horizontal and vertical formats) along with a fractional ASA plot. The server supports a number of user-selectable options including the choice of using sequence homology for the SABLE (Adamczak et al. 2004) prediction. The web server has been implemented as a Python CGI-script and is hosted on a system with 4GHz 2-Core processor and a CentOS operating system. With the available computing infrastructure, the web server takes <60 seconds (if homology is off) or >140 seconds (if homology is on) to calculate the fASA for a single query protein. A screen shot of the ShiftASA web server and its output is shown in Figure 4.10.



About SHIFTASA:

SHIFTASA web server is designed to accurately estimate residue-level fractional Accessible Surface Area (fASA) from chemical shifts (backbone and side-chain) and their corresponding protein sequence data. SHIFTASA accepts either a BMRB (BioMagResBank) formatted (NMR star 2.1 or NMR star 3.1) or a shifty formatted file with at least 70-80% backbone and side-chain atoms having chemical-shifts assigned. The server generates per-residue fASA (in both horizontal and vertical formats) and graphics plot in ~30-40 seconds (if evolutionary profile-based prediction option is not used) or ~90-120 seconds (if evolutionary profile-based prediction option is used). Extensive testings indicate that SHIFTASA achieves correlation coefficient between predicted and experimental values of 0.80-0.82 (tested on different control sets) which is significantly better than other existing methods. SHIFTASA combines the predictive power of machine learning with the added benefits of advanced sequence analysis to accurately estimate fASA via chemical shift data. If you use SHIFTASA please cite the following

SHIFTASA Usage

Select the format of input file: ● NMR-STAR 2.1 (Example) ○ NMR-STAR 3.1 (Example) ○ SHIFTY (Example)

Upload file with chemical shifts here:

No file chosen

OR

Paste chemical shifts into the text box below:

IMPORTANT: If a file is selected for upload above, it will be used and content of the text box below will be ignored.

Additional Options

1) Correct Re-referencing of Chemical Shifts	● Yes (Default) ○ No
2) Correct Neighbor Residue Effects	● Yes (Default) ○ No
3) Use Structural Homology- based Prediction	● Yes (Default) ○ No
5) Produce Graphics Output	● Yes (Default) ○ No
6) Predict Flexibility (Random Coil Index) from Chemical Shifts	● Yes (Default) ○ No

Results for 16502.str:

Fractional ASA Graph

Horizontal Output

```
2 KPKLLYCSNGGHFLRILPDGTVDGTRDRSDQHIQLQLSAESVGEVYIKST 51
2 85102004541210200593808013487721121103257513010004 51

52 ETGQYLAMDTDGLLYGSQTPNEECLFLERLEENHYNTYISKKHAEKNWFV 101
52 83553000057021001560587010015065412000004407713111 101

102 GLKKNGSCKRGPRTHYGQKAILFLPLPVSSD 132
102 1054707266166535417121000015789 132
```

Vertical Output

Num = Residue number
Res = Residue name
fASA = fractional ASA

```
#Num Res fASA
2    LYS 0.89
3    PRO 0.55
4    LYS 0.13
5    LEU 0.08
6    LEU 0.29
7    TYR 0.08
8    CYS 0.09
9    SER 0.44
10   ASN 0.52
11   GLY 0.41
12   GLY 0.17
13   HIS 0.21
14   PHE 0.12
15   LEU 0.08
16   ARG 0.21
17   ILE 0.09
```

Figure 4.10: A montage of the ShiftASA webserver showing the home page (left) and the screenshot of an example output page (right).

4.4. Conclusion

We have developed a method that accurately predicts the per-residue fractional accessible surface area (fASA) of water-soluble proteins using a combination of both sequence and chemical shift data. Our prediction method, called ShiftASA, demonstrates superior performance relative to sequence-only or chemical shift-only methods in two independent test sets of 65 and 92 proteins (TEST1 and TEST2, respectively). In particular, with the TEST1 data set, ShiftASA showed a mean Spearman's rank correlation coefficient between predicted and experimental values of 0.79, which is a 8.2% improvement over the best performing method. The mean absolute error was found to drop from 0.19 to 0.14 Å^2 and the root mean squared error fell from 0.26 to 0.19 Å^2

compared to its sequence-only and chemical shift-only counterparts. On the TEST2 set, ShiftASA attained a mean correlation coefficient of 0.82, a clear improvement over correlation coefficients of 0.67 and 0.73 reported by the best performing sequence-only and chemical-shift-only methods, respectively. In addition, the real-value fASA prediction by ShiftASA allows flexible, categorical prediction of binary or ternary ASA states. Overall, we believe that ShiftASA, with its improved prediction of ASA parameters, will not only facilitate protein fold recognition and *de novo* protein structure prediction methods, but as we will show in upcoming papers, contribute to the generation and refinement of protein structures by NMR and the calculation of useful thermodynamic parameters from chemical shift data.

Chapter 5

E-Thrifty - Chemical Shift Threading for Accurate Protein Fold Recognition

Abstract

Protein structure determination using Nuclear Magnetic Resonance (NMR) spectroscopy can be both time-consuming and labor intensive. One approach that may reduce this time and cost burden is a relatively old bioinformatics technique called “threading”. Threading uses a combination of sequence information and predicted secondary structure to generate 3D protein structures from both closely related and remote structural homologs. While sequence information is the primary input for most threading methods, using other experimentally measurable parameters, such as circular dichroism (CD) data, small angle X-ray scattering (SAXS) data or NMR chemical shifts, could potentially improve threading performance. The key motivations behind using NMR chemical shifts lie in the fact that they are easy to measure, they are available prior to 3D structure determination and they contain vital structural information. Here we describe a novel, chemical shift-based threading method called “Enhanced-Thrifty” or “E-Thrifty” that not only uses sequence and chemical shift similarity but also chemical shift-derived secondary structure, super-secondary structure and accessible surface area to identify and determine the most likely 3D structure that a query protein may have. E-Thrifty was optimized on a training set of >1700 alignments and subsequently evaluated on 25 “difficult” test cases including a number of recent “Critical Assessment of Structure Determination by NMR” (CASD-NMR-2013) targets. E-Thrifty was found to significantly outperform other shift-based or threading-based structure determination methods with an average TM-score performance of 0.66. Tests indicate that E-Thrifty’s performance is actually comparable to using coordinate data (i.e. knowing the answer ahead of time) to identify structurally similar proteins via 3D superposition. Coupled with recent developments in chemical shift refinement, these results suggest that protein structure determination, using only NMR chemical shifts, is becoming increasingly practical and reliable. E-Thrifty is available both as a standalone program and as a web server at http://ethrifty.ca.

5.1 Introduction

One of the long-term goals in protein NMR is to be able to generate accurate, atomic-resolution protein structures using only chemical shift data. Protein chemical shifts can provide accurate information about secondary structure (Wishart et al. 1992; 1994a; 1994b, Shen et al. 2009; 2013, Hafsa et al. 2014; 2015a), torsion angles (Berjanskii et al. 2006, Shen et al. 2013), hydrogen bonds (Wishart et al. 1998; 2001), dynamics (Berjanskii et al. 2005; 2013), disulfide bonds (Sharma et al. 2000), charge states (Osapay et al. 1991), accessible surface area (Vranken et al. 2009, Berjanskii et al. 2013, Hafsa et al. 2015b), ligand interactions (Medek et al. 2000) and aromatic ring proximity (Osapay et al. 1994, Kuszewski et al. 1995). The fact that protein chemical shifts have been shown to provide such a rich diversity of structural information has inspired the development of several chemical shift based protein structure prediction methods such as CS-Rosetta (Shen et al. 2008), Cheshire (Cavalli et al. 2007) and CS23D (Wishart et al. 2008). The CS-Rosetta and Cheshire methods generally follow an *ab initio* approach and attempt to model protein structures by generating large numbers of possible structures from the observed chemical shift data and then ranking the structures based on a knowledge-based potential. CS23D differs from CS-Rosetta and Cheshire in that it also attempts to use comparative (i.e. homology) modeling along with chemical shift "threading" to identify potential known protein folds that may be similar to that of the query protein whose structure is being determined.

Sequence "threading" is relatively old comparative modeling technique that can be used to detect very remote structural homologs or to predict protein fold similarities (Rost et al. 1995; 1997, Karplus et al. 1998, Peng et al. 2010). However, threading by sequence similarity, alone, is often not sufficient to routinely identify remote structural homologues. As a result, other information must be used. In particular, secondary structure information (predicted or calculated) and fractional accessible surface area (predicted or calculated) can substantially improve threading performance (Bowie et al. 1991, Jones et al. 1992, Rost et al. 1997). Indeed, studies by Jones et al. (1992) and Rost et al. (1997) suggested that the environment of an individual residue described by its (sequence-predicted) secondary structure, (sequence-predicted) torsion angles and (sequence-predicted) solvent accessibility are particularly useful.

The accuracy for sequence-based prediction of secondary structure is now exceeding 80% (Montgomerie et al. 2008) while sequence-based fractional accessible surface area prediction and sequence-based torsion angle prediction is typically hovering at 60-75%

(Heffernan et al. 2015, Singh et al. 2014). As impressive as these results are, they are not yet sufficiently accurate to make threading as useful as hoped. However, if sequence information can be supplemented with other experimental observations (circular dichroism data, FTIR data, SAXS data, NMR data) to make these "predictions" more accurate, it is possible to make threading quite effective (Shen et al. 2015). In particular, protein chemical shifts have been shown to provide very accurate readouts of protein secondary structure (Wishart et al. 1992; 1994, Shen et al. 2009; 2013, Hafsa et al. 2014; 2015a), torsion angles (Berjanskii et al. 2006, Shen et al. 2009; 2013) and fractional accessible surface area (Vranken et al. 2009, Berjanskii et al. 2013, Hafsa et al. 2015b). Because protein chemical shifts are often determined far before NOE measurements can be completed, the use of chemical-shift threading could potentially be used to guide or even completely solve protein structures by NMR. This concept was the basis to CS23D. Originally described in 2008 (Wishart et al. 2008), CS23D used a chemical shift threading program called THRIFTY (THReading with shIFTY) to help generate 3D protein structures from chemical shfits. THRIFTY essentially uses torsion angles predicted via chemical shifts and chemical-shift predicted secondary structures to identify related distant homologous templates or potential structural homologs that already exist in the PDB. THRIFTY has been used extensively in several other shift-based structure programs including GeNMR (Berjanskii et al. 2009).

The concept of chemical shift threading is not new. The first description of this technique was made more than 15 years ago (Wishart et al. 2001). Five years later, another chemical shift threading program called SimShift appeared (Ginzinger et al. 2006), which was followed by CS23D (Wishart et al. 2008). Most recently Shen et al. (2015) described a threading-like system called POMONA (Protein alignments Obtained by Matching Of NMR Assignments) that identifies suitable PDB homologs for query proteins using chemical shift data (and NOE distance restraints when available), which is followed by a modified comparative modeling procedure to generate all-atom structures for proteins. In particular, POMONA searches the PDB for suitable homologs that are well matched with backbone chemical shift-predicted, residue-specific Φ/Ψ probability maps and chemical-shift derived secondary structures. The resulting structural templates are then clustered into groups (typically ten) using a normalized Cα–rmsd as a distance metric. Representative homologs from these clusters are then used to build a structural pool for comparative modeling using a modified RosettaCM procedure (Song et al. 2013). POMONA was evaluated on a set of 16 proteins and in most cases the best alignments found by POMONA have good (an average MaxSub score (Siew et al. 2000) of 0.49) structural similarity with the native structures even when there is no detectable sequence similarity (≤20% sequence identity).

Published results from SimShift (Ginzinger et al. 2006), THRIFTY/CS23D (Wishart et al. 2008) and POMONA (Shen et al. 2015) -- all strongly suggest that the structural information encoded by chemical shifts can help to identify structurally similar template(s) even in the absence of detectable sequence similarity. Inspired by these studies, we have developed a method called "E-Thrifty" (Enhanced-Thrifty) that employs a more advanced version of chemical shift threading to more accurately identify the most likely fold and to generate a 3D structure that a given query protein may have. In particular, E-Thrifty uses significantly enhanced secondary structure identification (Hafsa et al. 2014) as well as recently developed shift-based super-secondary and structural motif identification (Hafsa et al. 2015a) to improve its performance. It also uses a newly developed shift-based accessible surface area prediction method (Hafsa et al. 2015b) as well as shift-based torsion angle predictions (Shen et al. 2013) and very accurate secondary chemical shift calculations (Han et al. 2012). These are combined to perform a modified threading protocol using a specially constructed, non-redundant database of known protein structures (a modified version of the PDB). When compared to the state-of-the-art threading programs or chemical shift-based structure generation programs such as POMONA (Shen et al. 2015), HHpred (Söding 2005), CS-Rosetta (Shen et al. 2008), and CS23D (Wishart et al. 2008) on two different test data sets, E-Thrifty exhibits a 10-20% improvement in overall performance. Details describing the E-Thrifty algorithm, its performance and its implementation as both a web server and a standalone program are given in the following pages.

5.2 Materials and Methods

5.2.1 Measuring Local and Non-local Structure Similarity

While amino acid substitution scores are normally used to guide the local alignment between two protein sequences, sequence alignment alone does not necessarily guarantee optimal structural or topological alignment between two proteins. This is particularly true when the sequence identity between two proteins drops below 35%. To perform sequence alignments or sequence threading for distantly related proteins, additional information such as (predicted or calculated) backbone Φ/Ψ angles, secondary structure, structural motifs, secondary chemical shifts and accessible surface area (ASA) are often needed to guide the alignment process. This is because these structural states tend to be more conserved than sequence among remote structural homologues

(Rost et al. 1997). Therefore, to perform a threading calculation it is necessary to have three things: 1) a database of solved structures where all of the threading parameters (sequence, torsion angles, secondary structure, structure motifs, ASA and secondary shifts) are pre-calculated; 2) a series of programs where the same parameters (torsion angles, ASA, etc.) are predicted and/or calculated for the query protein and 3) an alignment algorithm that scores, aligns and matches the query protein by taking into account all of the calculated and/or predicted parameters. If the structural parameters used for threading can be converted to letters or character strings (similar to the sequence), the sequence-structure alignment process can employ the Smith-Waterman alignment algorithm (Rost et al. 1997). The sequence-structure alignment concept used in the E-Thrifty method is depicted in Figure 5.1. The E-Thrifty algorithm, the scoring scheme and the parameter mapping are explained in more detail below.

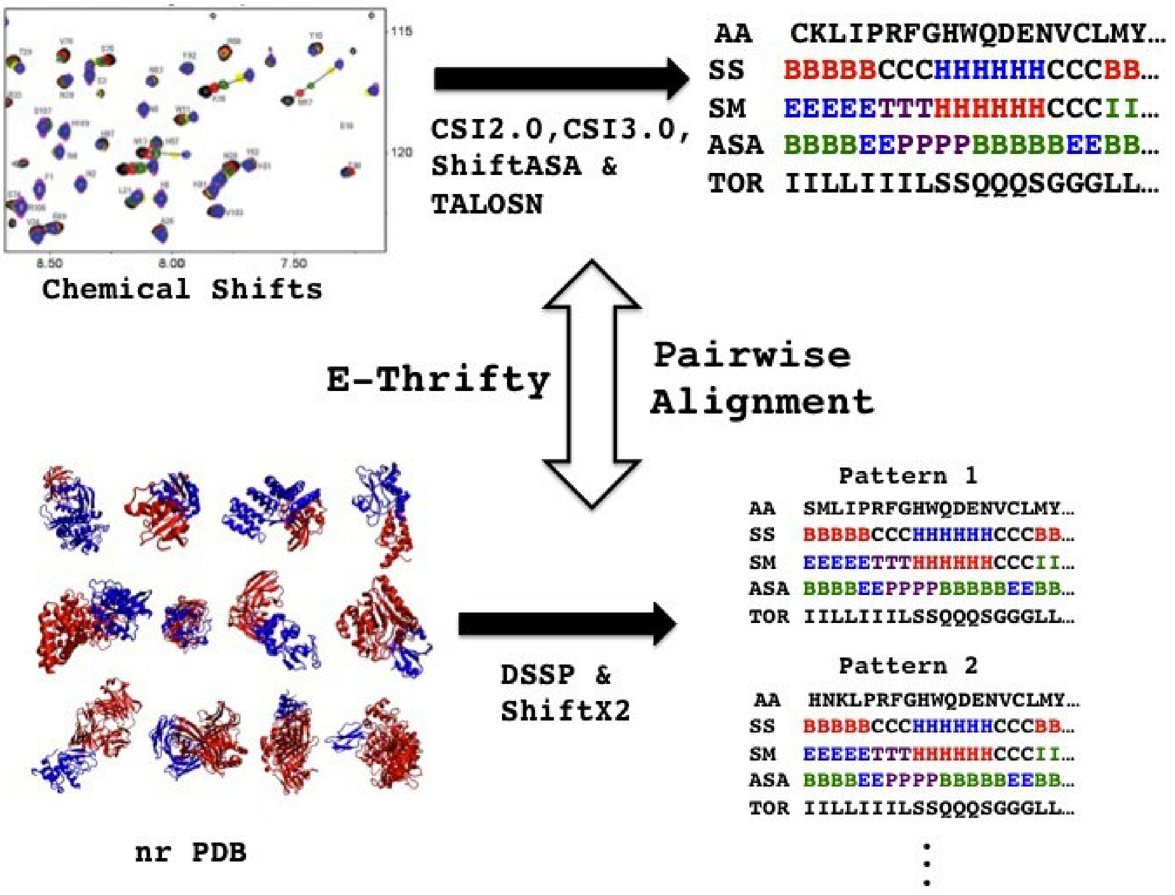

Chemical Shifts
8.50
8.00
7.50
115
120
CSI2.0,CSI3.0, ShiftASA & TALOSN
AA CKLIPRFGHWQDENVCLMY...
SS BBBBBCCCHHHHHHCCCBB...
SM EEEEETTTHHHHHHCCCII...
ASA BBBBEEPPPPBBBBBEEBB...
TOR IILLIIILSSQQQSGGGLL...
E-Thrifty
Pairwise Alignment
nr PDB
DSSP & ShiftX2
Pattern 1
AA SMLIPRFGHWQDENVCLMY...
SS BBBBBCCCHHHHHHCCCBB...
SM EEEEETTTHHHHHHCCCII...
ASA BBBBEEPPPPBBBBBEEBB...
TOR IILLIIILSSQQQSGGGLL...
Pattern 2
AA HNKLPRFGHWQDENVCLMY...
SS BBBBBCCCHHHHHHCCCBB...
SM EEEEETTTHHHHHHCCCII...
ASA BBBBEEPPPPBBBBBEEBB...
TOR IILLIIILSSQQQSGGGLL...

Figure 5.1: The sequence-structure alignment concept used in E-Thrifty method. Here AA, SS, SM, ASA and TOR represent the amino acid, secondary structure, structure motif, accessible surface area and backbone torsion angle sequences respectively. nrPDB stands for non-redundant protein data bank.

Substitution matrices for structural descriptors

Structural parameters such as secondary structure, structural motifs, torsion angles and fASA classes that describe the structural environments around each residue in the query and the database proteins are represented by simple one-letter codes. During the alignment of the query and the database sequences, these letters are compared in a way to maximally match the local and non-local structural similarity. The one-letter codes of the query and the database sequences are either matched or substituted by another letter. Matched structural states are given a high score whereas a unmatched states are given a low score (a small positive or a negative value). For example, a negative score is assigned to a helical secondary structure class (represented by "H") being substituted/replaced by a β-strand class (represented by "E" or "I"), whereas a β-turn replacement by a coil assignment is given a small positive value (i.e., a lower penalty). A substitution matrix can be used to compactly represent this scoring scheme. A 3×3 substitution matrix for the three-state secondary structure states describes the substitution/matching scores of the three secondary structure classes. Similarly, a 5×5 substitution matrix is used for the five structural motif states, a 3×3 substitution matrix is used for the three fASA categorical states, and a 9×9 substitution matrix is used for the nine torsion angle states. Substitution matrix values were initially chosen from the BLOSUM62 matrix (Henikoff et al. 1992) and then optimized through trial-and-error grid search methods on the training alignments.

Scoring local and non-local structural similarity

After defining substitution matrices for different structural descriptors, the structural similarity between the query residue i and the database residue j is calculated using the following equation.

$$S(i,j) = w_{AA} \times AA_{score} + w_{SS} \times SS_{score} + w_{SM} \times SM_{score} + w_{ASA} \times fASA_{score} + w_{Torsion} \times Torsion_{score} \tag{5.1}$$

where AA_{score} is the amino acid similarity score, $Torsion_{score}$ is the torsion letter similarity score, SS_{score} is the secondary structure similarity score, SM_{score} is the structural motif similarity score, $fASA_{score}$ is the ASA state similarity score and the w^{*} terms represent the corresponding weighting coefficients. The structural similarity scores of the central residue position takes into account the structural letter substitution scores in the preceding and the following neighbor locations. Each $S(i,j)$ entry in the calculated scoring matrix is rescaled to a range (-2.0, 3.0) so as to obtain a uniform distribution of alignment scores. The rescaling is performed using the following equation.

$$S_{i,j}^{scaled} = ((max_{new} - min_{new}) * \frac{S[i][j] - min_{old}}{max_{old} - min_{old}}) + min_{new} \qquad (5.2)$$

The scoring components described in Eq. 5.1 are briefly explained in the following paragraphs.

Amino acid similarity

The aligned amino acids are scored using the BLOSUM62 (Henikoff et al. 1992) substitution matrix. In this 20×20 matrix, every possible amino acid substitution is assigned a score based on its observed frequencies derived from careful alignment of evolutionarily related proteins (with no more than 62% sequence identity). A positive score is given to more probable substitutions while a negative score is given to less probable substitutions. The amino acid similarity score (AA_{score}) is given as:

$$AA_{score}(i,j) = BLOSUM62(aa_i, aa_j) \qquad (5.3)$$

where aa_i is the query residue in *i*-th position and aa_j is the database residue in *j*-th position.

Secondary structure similarity

Secondary structures for the query protein are calculated using the CSI 2.0 program (Hafsa et al. 2014). CSI 2.0 is a multi-class, machine-learning algorithm that determines the extent and location of α-helices, β-strands and coil regions based on $^{13}C_{\alpha}$, $^{13}C_{\beta}$, ^{13}C, $^{1}H_{N}$, $^{1}H_{\alpha}$, ^{15}N backbone chemical shifts and sequence. For the E-Thrifty threading algorithm, the secondary structure similarity between secondary structure of the query residue *i* and the database residue *j* is calculated over a 3-residue window using the following formula.

$$SS_{score}(i,j) = \sum_{n \epsilon \{-1,0,1\}} S_{i+n,j+n} \quad \{S_{i,j} = SS_{matrix}(ss_i, ss_j) \qquad (5.4)$$

where SS_{matrix} is a 3 × 3 substitution matrix for secondary structure states (H, B and C), which describes the substitution scores for the replacement of one secondary structure state with another.

Structural motif similarity

The classification of β-strands and β-turns in the query sequence is performed by CSI 3.0, a chemical shift based super-secondary structure identification program, described by Hafsa et al. (2015a). The CSI 3.0 output is mapped to 5 letters H, E, I, C and T which stands for helix, edge β-strand, interior β-strand, coil and β-turn respectively. For the E-Thrifty threading algorithm the structural motif similarity between the query residue *i* and the database residue *j* is calculated over a 3-residue window using the formula.

$$SM_{score}(i,j) = \sum_{n\epsilon\{-1,0,1\}} M_{i+n,j+n} \quad \{M_{i,j} = SM_{matrix}(sm_i\,, sm_j) \qquad (5.5)$$

where SM_{matrix} is a 5 × 5 substitution matrix for five structure motif states (H, E, I, C, T), which describes the substitution scores for the replacement of one of the five structural motif letters with another.

Fractional accessible surface area (fASA) similarity

The fractional ASA (fASA) is an ASA descriptor that describes the percentage of accessible surface area for a given residue relative to a fully exposed residue. Residue-specific fASAs for the query protein are calculated using the ShiftASA program (Hafsa et al. 2015b). Residues are assigned three letters such as B (Buried) (fASA<=0.25), P (Partially buried) (0.50=>fASA>0.25), and E (Exposed) (fASA>0.50) based on the predicted/calculated fASA range. For the E-Thrifty threading algorithm the similarity between fASA categorical states of the query residue *i* and the database residue *j* is calculated as below.

$$fASA_{score}(i,j) = \sum_{n\epsilon\{-1,0,1\}} F_{i+n,j+n} \qquad \{F_{i,j} = fASA_{matrix}(f_i\,, f_j) \qquad (5.6)$$

where $fASA_{matrix}$ is a 3 × 3 substitution matrix for three fASA categorical states {B, P, E} that describes the substitution scores for the replacement of one fASA-state letter with another.

Torsion letter similarity

Backbone Φ/Ψ torsion angles from experimental $^{13}C_\alpha$, $^{13}C_\beta$, ^{13}C, $^{1}H_N$, $^{1}H_\alpha$, ^{15}N chemical shifts are predicted by TALOS-N (Shen et al. 2013) and converted into a 9-letter torsion angle alphabet.

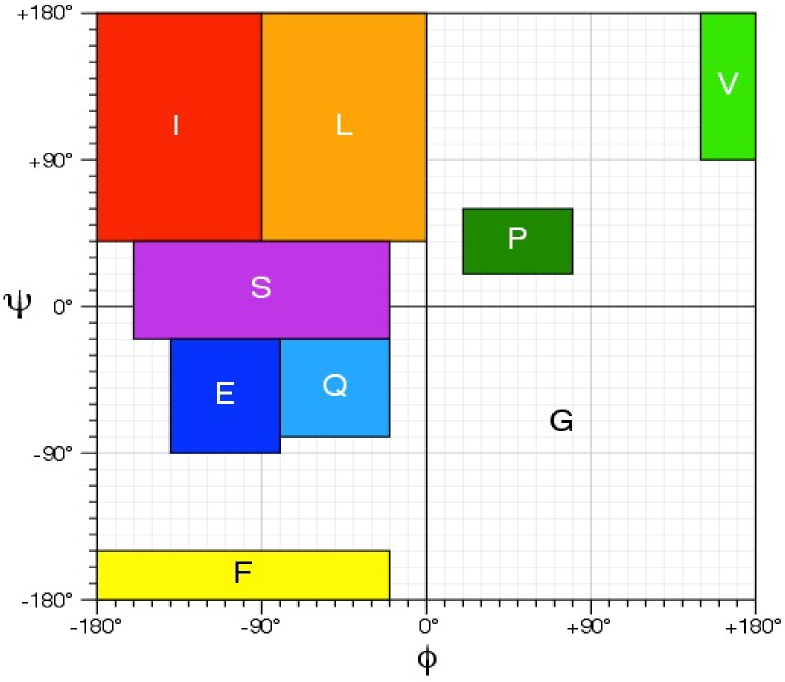


Figure 5.2: Backbone Φ/Ψ torsion angles mapping into 9 overlapped regions in Ramachandran map with a letter assigned to each region. Letters I, L, F and V represent the preferred conformational spaces for β-sheet and β-turns; S, E and Q indicate the favorable regions for right-hand α-helical conformation; P indicates the preferred left-hand helical conformation and lastly G represents all other flexible conformations that a protein can adopt.

This so-called torsion angle alphabet, which is very similar to the THRIFTY alphabet used in CS23D (Wishart et al. 2008), splits the Ramachandran map into 9 non-overlapped regions based on the Φ/Ψ propensity of common secondary structural classes, with a letter is assigned to each region (Figure 5.2). For the E-Thrifty threading algorithm the torsion letter similarity between the query residue *i* and the database residue *j* is calculated over a 3-residue window using the following formula.

$$Torsion_{score}(i,j) = \sum_{n\epsilon\{-1,0,1\}} T_{i+n,j+n} \quad \{T_{i,j} = Torsion_{matrix}(tor_i\,, tor_j) \qquad (5.7)$$

where $Torsion_{matrix}$ is a 9 × 9 matrix that describes the substitution scores for the replacement of one torsion angle letter with another.

Structural annotation of the database proteins

A non-redundant (nr) version of the PDB (Berman et al. 2000) was generated using the Pisces server (Wang et al. 2003). As of February 1st 2016, there were a total of 71,100 sequences and coordinate files in this nrPDB data set. This database of known (or previously solved) structures was then annotated using a series of programs so that every residue was assigned a secondary structure, specific secondary structure motifs, a set of backbone torsion angles and a fractional accessible surface area. The secondary structures, torsion angles and accessible surface areas were generated from the DSSP (Kabsch et al. 1983) program. Other secondary structure elements such as β-turns and edge/internal strand information are obtained using methods described in Hafsa et al. (2015a). Fractional accessible surface areas or fASA values for each residue were derived from the DSSP output using a method described in Hafsa et al. (2015b). After calculating these data, we generated four "pseudo-sequences" associated with each entry in our nrPDB data set. These pseudo-sequences correspond to: 1) a secondary structure sequence; 2) a structure motif sequence; 3) a torsion angle sequence and 4) a fASA sequence. These, along with the amino acid sequence of each protein describe its local and non-local structural states.

5.2.2 Gap Penalty Function in Sequence-structure Alignment

In order to perform a proper sequence or even a sequence-structure alignment of two protein sequences, it is important to develop a scoring function to properly handle the insertion or deletion of gaps in either sequence. Gaps are usually counted as a penalty in the total alignment score. Typically an affine gap penalty or AGP function of the form, $g = u + vl$ is used in most sequence-only alignment algorithms. This kind of function depends on the gap initiation (u) and gap extension (v) parameters, and the length of the gap in the alignment (l). However, previous studies suggest that including a conformation specific gap penalty in sequence-structure alignment increases the accuracy of the correctly aligned residues (Madhusudhan et al. 2006). Hence in our work, we adopted a conformation specific gap penalty function called a variable gap penalty or VGP (Madhusudhan et al. 2006, Shen et al. 2015), in which the gaps that are introduced in regular secondary structure regions (contiguous helices and β-strands) and between two spatially distant residues are penalized. Details of the algorithm are outlined below.

```
                      i'          i
Sequence    Q H M L F T N - - - N Q S L P
Structure   K M N G H L - - - - - H P K Y
                      j'          j
```

Figure 5.3: Definition of a gap, i, i', j and j' residue positions in sequence and structure block respectively. A single gap can include deletions or insertions in either the sequence or the structure and is defined by all four indices, i, j, i' and j'.

In VGP, to initiate a gap that extends from the i to i' positions in the query sequence (referred to as the sequence block) and from the j to j' positions in the database sequence (with known structure -- referred to as the structure block), as illustrated in Figure 5.3, a gap penalty function of the following form is used.

$$G(i,j,i',j') = \begin{cases} 0 & if\ l = 0\ and\ l' = 0 \\ R \times u + (l + l') \times v & if\ l > 0\ or\ l' > 0 \end{cases} \tag{5.8}$$

$$l = i' - i - 1$$

$$l' = j' - j - 1$$

$$R = 1 + w_{HBS} \times [HBS(j,j') + HBS(i,i')] + w_d \times D(j,j')$$

where l and l' are the gap lengths in the query sequence and the database structure blocks, u is the gap opening parameter and v is the gap extension parameter. R is a modulating function that controls the gap-opening penalty depending on the structural environment at the position of insertion. The value of R is at least 1 and can be larger to make the opening of gaps more difficult in certain circumstances. In particular, these penalties are larger within regular secondary structure regions such as α-helices and β-strands, and between two spatially distant residues. Also, w_{HBS} and w_d are the corresponding weights of different terms in R. The term $HBS(j,j')$ takes a binary value of 1 or 0 depending on whether the segment in the structure block spanning from residue j to j' adopts a helical or a β-strand conformation or not. $HBS(i,i')$ is the same function applied to the sequence block, however it is based on the CSI 2.0 (Hafsa et al. 2014) predicted secondary structure for query residues i to i'. The term $D(j,j')$ is a function that depends on the spatial proximity of the two database residues spanning the gap and is defined as below.

$$D(j,j') = \max(0, d(j,j') - d_0)^{\gamma} \tag{5.9}$$

where $d(j,\ j')$ is the distance between Cα− atoms at residue positions j and j' in the structure block, d_0 is an empirical distance cutoff and γ is an empirical constant. For two database residues having the distance below d_0, there is no increase in the gap-opening penalty.

5.2.3 Protein Local Alignment

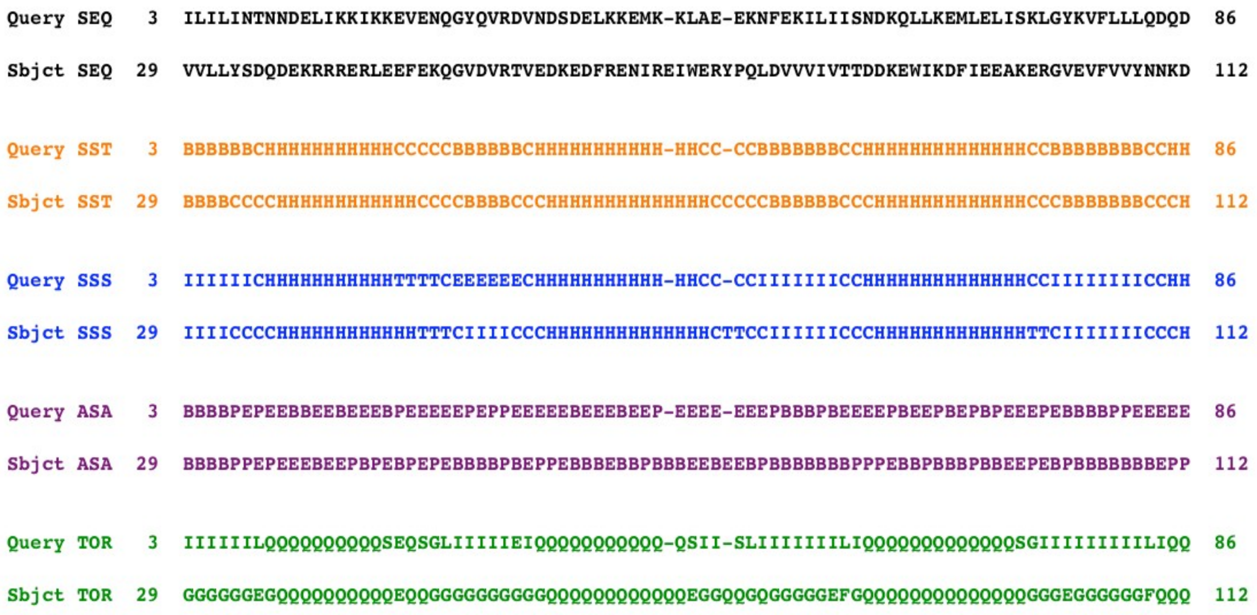
In E-Thrifty, the protein sequences with experimentally measured chemical shifts (i.e. the query sequence) are aligned against the known structures in our annotated nrPDB (i.e. the database sequences) using a modified version of the Smith-Waterman local alignment algorithm (Smith et al. 1981). A similarity matrix S of M×N dimensions is constructed, where M is the length of the query protein and N is the length of the database protein. Each element in the scoring matrix $S(i,j)$ indicates the substitution score for the query residue *i* with the database residue *j*. Once the scoring matrix is constructed, the optimal alignment between query and subject sequence is found by calculating an alignment matrix (H) using a dynamic programming, traceback procedure. This traceback protocol involves finding the maximum element in the alignment matrix and tracing back through the matrix from the maximum element to zero. Each element *H(i,j)* in the H matrix is calculated with the following recursive dynamic programming equation.

$$H(i,j) = max \left\{ \begin{matrix} H(i-1,j-1) + S(i,j) \\ H(i-1,j) - VGP(i,j,i-1,j) \\ H(i,j-1) - VGP(i,j,i,j-1) \\ 0 \end{matrix} \right\} \tag{5.10}$$

The initial conditions for the recursive algorithm are,

$$\begin{cases} H(i,0) = 0;\ 0 \le i \le M \\ H(0,j) = 0;\ 0 \le j\ \le N \end{cases}$$

In Eq. 5.10, $S(i,j)$ is the substitution score for the query residue i with the database residue *j*. VGP is the gap penalty function applied when there is a gap opening or extension between the i and i' positions in the sequence block or the j and j' positions in the structure block. For the assignment of each element $H(i,j)$ in the H matrix, the diagonal i.e. upper-left ($H(i-1,j-1)$), upper ($H(i-1,j)$) and left ($H(i,j-1)$) neighbor elements are compared and the maximum value among these three elements (if the maximum value is non-positive, then a 0.0 value) is assigned to the current element as score $H(i,j)$. After calculating all the elements of the H matrix as described in Eq. 5.10, the largest element in the H matrix (H_{max}) represents the optimal alignment score. The residue equivalence assignments can then be obtained by tracing back through the maximum element, H_{max} to the zero value in the H matrix, which is also the optimal sub-alignment between query and the subject sequences. Example local alignments between sequences of the query protein 2LCI and subject protein 2L82 (chain A) are shown in Figure 5.4.



```
Query SEQ   3  ILILINTNNDELIKKIKKEVENQGYQVRDVNDSDELKKEMK-KLAE-EKNFEKILIISNDKQLLKEMLELISKLGYKVFLLLQDQD  86

Sbjct SEQ  29  VVLLYSDQDEKRRRERLEEFEKQGVDVRTVEDKEDFRENIREIWERYPQLDVVVIVTTDDKEWIKDFIEEAKERGVEVFVVYNNKD  112

Query SST   3  BBBBBBCHHHHHHHHHHHCCCCCBBBBBBCHHHHHHHHHHH-HHCC-CCBBBBBBBCCHHHHHHHHHHHHHHCCBBBBBBBBCCHH  86

Sbjct SST  29  BBBBCCCCHHHHHHHHHHHHHCCCCBBBBCCCHHHHHHHHHHHHHCCCCCBBBBBBCCCHHHHHHHHHHHHHCCCBBBBBBBCCCH  112

Query SSS   3  IIIIIICHHHHHHHHHHHTTTTCEEEEEECHHHHHHHHHHH-HHCC-CCIIIIIIICCHHHHHHHHHHHHHHCCIIIIIIIICCHH  86

Sbjct SSS  29  IIIICCCCHHHHHHHHHHHHHTTTCIIIICCCHHHHHHHHHHHHHCTTCCIIIIIICCCHHHHHHHHHHHHHTTCIIIIIIICCCH  112

Query ASA   3  BBBBPEPEEBBEEBEEEBPEEEEEPEPPEEEEEBEEEBEEP-EEEE-EEEPBBBPBEEEEPBEEPBEPBPEEEPEBBBBPPEEEEE  86

Sbjct ASA  29  BBBBPPEPEEEBEEPBPEBPEPEBBBBPBEPPEBBBEBBPBBBEEBEEBPBBBBBBBPPPEBBPBBBPBBEEPEBPBBBBBBBEPP  112

Query TOR   3  IIIIIILQQQQQQQQQQSEQSGLIIIIIEIQQQQQQQQQQQ-QSII-SLIIIIIIILIQQQQQQQQQQQQQSGIIIIIIIIILIQQ  86

Sbjct TOR  29  GGGGGGEGQQQQQQQQQQEQQGGGGGGGGGGGQQQQQQQQQQQQQEGGQQGQGGGGGEFGQQQQQQQQQQQQGGGEGGGGGGFQQQ  112
```

Figure 5.4: Example sequence-structure alignments between query 2LCI and subject 2L82 (chain A) sequences are shown. Here SEQ, SST, SSS, ASA and TOR represent the sequence, secondary structure, structural motif, fractional ASA and torsion angle sequences respectively.

5.2.4 Chemical Shift Scoring and the Alignment Ranking

To further improve the alignment scoring, we implemented a backbone secondary chemical shift fitness score similar to SimShift (Ginzinger et al. 2006). Specifically, a secondary chemical shift fitness score is calculated for equivalent residue assignments in the alignment. Secondary chemical shifts can be defined as the difference between the observed experimental chemical shift (δ_{obs}) and the corresponding random coil shift (δ_{rc}) value for a specific atom (Wishart et al. 2011).

$$\Delta\delta = \delta_{obs} - \delta_{rc} \tag{5.11}$$

Secondary chemical shifts contain important structural and dynamic information about proteins (Wishart et al. 2001, Mielke et al. 2009). The backbone chemical shifts for the database structures are calculated using ShiftX2 (Han et al. 2012) and the secondary shift values are obtained using the formula mentioned above using the neighbor adjusted random coil values extracted from Schwarzinger et al. 2001. The secondary shift fitness score is then calculated as.

$$SC_{score} = \sum_a w_a \times \; corr\left(\delta_a^{obs}, \delta_a^{pred}\right) \tag{5.12}$$

where δ_a^{pred} represents the backbone chemical shift predicted by ShiftX2 (Han et al. 2012) for a specific atom a (a = $^{13}C_{\alpha}$, $^{13}C_{\beta}$, ^{13}C, $^{1}H_{N}$, $^{1}H_{\alpha}$, ^{15}N) for a set of database residues that are aligned with the query residues with experimental chemical shift referred to as δ_a^{obs}. The w_a's are the corresponding weighting coefficients for six backbone atoms. The function $corr\left(\delta_a^{obs}, \delta_a^{pred}\right)$ measures the correlation between the observed (query) and the predicted (subject) secondary chemical shifts over all the aligned residues for a specific backbone atom a. Therefore chemical shift fitness score is a weighted combination of chemical shift correlations of six backbone atoms over all the aligned residues.

The secondary chemical shift fitness score is then combined with the optimal sub-alignment score H_{max} with a scaling factor w_{corr} to produce the total score for each alignment.

$$S_{total} = \; H_{max} + \; w_{corr} \; \times \; SC_{score} \tag{5.13}$$

The final ranking of the alignments is performed according to this total score, S_{total}.

5.2.5 Optimization of E-Thrifty Parameters

To optimize the parameters described in Eq. 5.1, 5.8, 5.9, 5.12 and 5.13, a set of 30 proteins with complete experimental chemical shift information and available high-resolution X-ray structures were chosen. The training proteins had ~90% of their complete (^{1}H, ^{13}C and ^{15}N) backbone chemical shifts assigned. A set of homologs for the training proteins spanning a sequence identity range 20-40% was retrieved using a PSI-BLAST search. Once the training set was obtained, a sequence-structure alignment between the queries and the corresponding homolog proteins was performed, which produced a total of 1,777 alignment pairs. There were a total of 18 parameters to optimize (train) and the parameter set could be divided into three different groups in terms of their behavior during optimization; the scoring matrix weighting parameters, the gap spanning parameters, and the chemical shift weighting parameters.

The scoring matrix parameters ($w_{AA}, w_{Torsion}, w_{SM}, w_{ASA}$) were optimized by a grid search using the training alignments. We trained one parameter at a time and kept the other parameters constant at their initial values or the previously optimized values. Parameter optimization was terminated on the convergence of the average alignment score observed against the Cα rmsd (the higher the alignment score, the lower the Cα rmsd) for the training set of alignments. For the gap spanning parameters (u, v, d0, γ, w_{HBS}, w_d) in the VGP function, the initial values were chosen from a previous study by Shen et al. (2015). We then attempted to further optimize the parameter values through a grid search. However no significant improvement was observed (data not shown). Hence, we used the Shen et al. values. The chemical shift weighting parameters defined in Eq. 5.12 were optimized (trained) using a linear regression analysis. The training data for linear regression comprises of chemical shift correlation coefficients between the six backbone atoms of the query and the database equivalent residues (~X) and the Cα-rmsd (~Y) of the aligned region of the training proteins. A linear regression model was then fit to the training data. A similar regression analysis was performed to search an optimal value for w_{corr} described in Eq. 5.13.

The optimized parameter values determined from this study are: w_{ASA} = 4.25, w_{AA} = 1.0, w_{SM} = 5.11, $w_{Torsion}$ = 3.97, w_{SS} = 4.35, u = 3.0, v = 0.3, d0 = 6.5, γ = 2.0, w_{HBS} = 1.0, w_d = 2.0, w_{CO}= 3.75, w_{CA}= 4.5, w_{CB}= 4.75, w_N= 2.5, w_{HN}= 4.25, w_{HA}= 4.5, w_{corr} = 2.0.

5.2.6 Statistical Significance of E-Thrifty Alignments

In any database alignment protocol, it is important to properly assess the significance of an alignment that results from the comparison of a protein of a certain length to a database containing many different proteins of variable length. Hence, E-Thrifty alignments were evaluated using a BLAST-like e-value or expect-value (Altschul et al. 1996). We were particularly interested in seeing how high an alignment score can be expected to occur by chance by calculating an e-value associated with each optimal alignment. The e-value of an alignment having the score *S* can be calculated using the function described in Altschul et al. (1996).

$$E = K \times m' \times n' \times \mathrm{e}^{-\lambda \mathrm{S}} \quad (5.14)$$

Here K and λ are statistical parameters, $m' \times n'$ is the effective search space size and S is the alignment score of an optimal sub-alignment. The effective search space $m'n'$ is calculated using the equation below:

$$m' = m - l,\ n' = n - N \times l \quad (5.15)$$

where m is the number of residues in query protein, n is the total number of residues in the protein database, N is the total number of database proteins and l is the edge correction factor, which is used to calculate an "effective length" for a sequence. It eliminates the "edge effect" problem ensuring that a high-scoring alignment has a non-zero length and does not begin near to the end of either of two sequences being compared. K and λ values are taken from Altschul et al. (1996). The l value depends on the length of the database protein being compared with and is chosen from a set of empirical values depending on the $ln(m \times n)$ values described in the same study by Altschul et al. (1996).

5.2.7 Generation of 3D Structures via MODELLER

As part of the E-Thrifty pipeline, a 3D structure of the query protein is generated via the MODELLER (Sali et al. 2003) software package. The sequence-structure alignment generated by E-Thrifty is converted into the required PIR format and used as input for the comparative modeling function in MODELLER. MODELLER then generates the 3D coordinates of a number of possible models. The generated models are further assessed using MODELLERS's core evaluation functions (GA341 and DOPE). The 3D structure that has the lowest energy after the assessment is chosen as the final 3D model. MODELLER was chosen for structure generation purposes as it proved to be the most suitable program for spatial-restraint based modeling. E-Thrifty has two structure generation output options. The default option generates a comparative model (with 3D coordinate data) of the query protein using the E-Thrifty sequence-structure alignment of the top template only. The other option offered by E-Thrifty employs Clustal Omega (Sievers et al. 2011) program to perform a multiple alignment between the query and several template sequences. This multiple alignment is then used to build the 3D structure of the query protein.

5.2.8 Assessment of E-Thrifty Generated Structures

To assess the quality of the templates/structures generated using E-Thrifty, as well as other threading programs, we used DALI (Holm et al. 2010), which is a web server designed for performing 3D coordinate comparisons. It is particularly useful for identifying proteins with 3D structure similarities that may not have any obvious sequence similarity. As a result DALI often serves as a "gold standard" for identifying remote structural homologs. For this component of the study we assessed the structural accuracy and fold similarity achieved by structures generated by E-Thrifty, POMONA and HHpred (with DALI as a control) using two different scoring functions: 1) the Cα rmsd and 2) the Template Modeling or TM-score (Zhang et al. 2004). The assessment process involved building a full-length model of the query protein based on the template structure and using scoring functions to evaluate the quality of the models. The rmsd at the Cα–atom level measures the distances between the 3D-coordinates of main chain α-carbon atoms. In this study the Cα-rmsd is used to evaluate the aligned residues between the query and the template (i.e. the quality of the alignment). Generally the higher the Cα-rmsd, the more dissimilar two structures are. On the other hand, a near-zero Cα-rmsd corresponds to two structurally similar or structurally identical structures. However, in our study Cα-rmsd was not used to assess the quality of full-length models generated using MODELLER. Indeed, as many other authors (Siew et al. 2000, Ortiz et al. 2002, Betancourt et al. 2001) have noted, Cα-rmsd is not a perfect measure of model quality as it fails to identify well-predicted sub-structures in the presence of large prediction errors (i.e. disordered loops) in other parts of the model. Instead we used the TM-score to measure model quality. Unlike other popular scoring functions such as the MaxSub score (Siew et al. 2000), the TM-score uses a size-dependent scale to eliminate the protein length dependence. It also considers all alignments or modeling residue pairs in its assessment rather than arbitrarily setting specific distance cutoffs and calculating only the fraction of residues with errors below a certain cutoff distance (Zhang et al. 2004). A TM-score typically falls in the range from 0 to 1.0, with a TM-score of 1.0 indicating a perfect match between two structures and a TM-score below 0.17 generally indicating a randomly chosen unrelated fold. A quantitative study by Xu et al. (2010) showed that proteins with a TM-score equal to 0.5 have a probability of 37% of being in the same CATH (Greene et al. 2007) topology family and a probability of 13% of being in the same SCOP (Murzin et al. 1995) fold family.

5.2.9 Clustering and Selection of E-Thrifty Templates

Among the candidate templates identified by E-Thrifty, many were found to score very similarly or very closely to one another for any given query protein. This is because multiple templates may share a similar fold, a similar sub-structure or a similar 3D structure. In some cases, multiple templates for a single query protein may increase the coverage when building a full-length model. Moreover, these templates can be used as a structural pool for comparative modeling purposes. Hence, it is useful to group the templates into a number of distinct clusters so that similar structures can be gathered together. In E-Thrifty, a hierarchical clustering algorithm that employed Cα-rmsd as a distance metric was used to group the set of candidate templates. The Cα-rmsd between two template proteins was calculated over a common set of residues that were aligned with the same set of query residues. Specifically the "complete linkage method" for hierarchical clustering was used to group the templates. At each stage, the cluster is formed when all the links (i.e. the Cα-rmsd) between pairs of objects in the cluster are less than a particular level.

E-Thrifty will generate up to 50 template hits, all of which are ranked according to their alignment scores and are given a cluster membership. To generate a 3D structure via MODELLER, a user can select either the top template from this list (the default single-template modeling option) or a variable number (default is 2) of top representative templates from the first five resultant clusters (the multiple-template modeling option).

5.2.10 CS-GAMDy Refinement of E-Thrifty Models

CS-GAMDy (Berjanskii et al. 2015) is a newly developed NMR chemical shift-based protein structure refinement method. It uses a knowledge-based scoring function and structural information derived from chemical shift information through a combination of molecular dynamics and a multi-criterion genetic algorithm to perform structure refinement. The software is able to effectively refine and improve a wide range of approximate or even erroneous models. In our study, we used CS-GAMDy to refine the full-length models generated by MODELLER using the E-Thrifty identified templates. Examples of full-length models refined by CS-GAMDy for eight query proteins extracted from TEST1 and TEST2 are shown in Figure 5.5 with the corresponding Cα-rmsd and TM-scores. Note that the reported Cα-rmsd is calculated only over the defined secondary structure regions, whereas the TM-score is calculated over all aligned residue pairs of the superposed proteins (query and template).

5.2.11 The E-Thrifty Web server and Stand-alone Program

E-Thrifty has been implemented as both a web server and a standalone program, both of which can be accessed at http://ethrifty.ca. E-Thrifty's modified nrPDB database is divided into ten equal subsets and searches against each database subset can be run independently of each other. A multicore system with a minimum memory of 4GB in each core is recommended to install and run the E-Thrifty program. In this multi-core environment, the average parallel run time for E-Thrifty on medium sized proteins (<200 residues) is ~90-120 minutes, whereas it generally takes ~120-180 minutes for larger proteins (>200 residues). The E-Thrifty method was written using several programming and scientific languages including Python, Perl and R. E-Thrifty accepts BMRB (NMR–Star 2.1 or NMR-Star 3.1) or SHIFTY-formatted chemical shift files and produces multiple output files. These output files include: 1) an alignment file showing the sequence – structure alignments for the top 50 hits; 2) a summary file containing alignment scores, chemical shift scores, total scores, e-value and the cluster membership associated with each hit and 3) a 3D structure (PDB coordinates) of the query protein using the top scoring template (or multiple templates from the clusters) generated via MODELLER. The E-Thrifty server supports a number of user-selectable options related to comparative modeling which includes the sequence identity threshold for template selection, the model building mode (either "single" or "multiple") and exclusion of flexible terminal regions in the modeling process.

5.3 Results

The performance of E-Thrifty was evaluated on two independent test sets. The first test set (TEST1) consisted of 15 proteins extracted from "CS-Rosetta structure generation on existing entries" page located at the BMRB (Ulrich et al. 2008) web server ("CS-ROSETTA test cases", 2016). The second data set (TEST2) consisted of nine blind targets from the recent Critical Assessment of Automated Structure Determination of Proteins by NMR (CASD-NMR-2013) (Rosato et al. 2015). The structure alignment server DALI (Holm et al. 2010) was used to identify the "correct" homologs for test proteins in each data set. For the TEST1 and TEST2 proteins, the best templates found by E-Thrifty are listed in Tables 5.1-5.3. The performance of E-Thrifty was compared against several well-regarded threading, *ab initio* structure generation and chemical shift threading methods, including HHpred (Söding et al. 2005), POMONA (Shen et al. 2015), CS23D (Wishart et al. 2008) and CS-Rosetta (Shen et al. 2008). More specifically, POMONA (Shen et al. 2015) is a threading/homology search program that uses only chemical shift generated structural information to obtain highly probable alignments for query proteins. HHpred (Söding et al. 2005) detects homologs by using Hidden Markov Model (HMM) profile alignments and predicted secondary structures. CS23D (Wishart et al. 2008) detects structural homologs via chemical shift threading (torsion angle matching and secondary structure matching) as well as via sequence comparison. CS-Rosetta (Shen et al. 2008) predicts the protein structures via chemical shift-based fragment matching, in conjunction with *ab-initio* protein modeling algorithms.

The performance of E-Thrifty and the other methods was evaluated for the TEST1 proteins with two sequence identity cutoffs (≤30% and ≤95%.) whereas for the TEST2 proteins, only a ≤30% sequence identity cutoff was used (as most TEST2 proteins exhibit very low sequence identity to known structures). For evaluation consistency, the same protein structure modeling software, MODELLER (Sali et al. 1993) was used to build full-length models using the templates identified by the different threading methods assessed in this study. The quality of the template models was then evaluated using the TM-score (Zhang et al. 2004), a widely used metric to assess the folding similarity between two proteins.

The results of the TM-score evaluation on TEST1 and TEST2 proteins are shown in Tables 5.1-5.3. These tables also describe the performance of HHpred (Söding et al. 2005), CS23D (Wishart et al. 2008), CS-Rosetta (Shen et al. 2008), DALI (Holm et al. 2010) and POMONA (Shen et al. 2015). DALI (Holm et al. 2010) was used to calculate the "gold standard" corresponding to the "true" or best structural homologs identified through 3D coordinate-based structural superposition against all PDB structures. All of the programs chosen here for comparison were run though the corresponding web service or the corresponding program on the same test proteins. The POMONA and the HHpred templates with the highest alignment scores (within two sequence identity cutoffs) were identified as the optimal threading results. The lowest energy structure produced by CS-Rosetta was considered as the best template. As the TEST2 proteins are blind targets from the CASD-NMR-2013 competition (Rosato et al. 2015), we decided that in order to make the comparison fair and unbiased, all database templates that were solved or deposited into the PDB after 2013 would be excluded. The last column in each table includes the results from the DALI server (Holm et al. 2010), which essentially indicates the “true” answer or the “gold standard”. This result was included to help assess each program’s performance. The inclusion of the DALI data also helps to define the upper limit regarding how well a given threading program can practically perform. The database proteins for which DALI produces the highest Z-scores (within two sequence identity cutoffs) were selected as the DALI outputs. Note that in selecting the templates from the different programs we relied on the reported sequence identity as measured by the respective alignments. Note that if one of the programs had no answer for any particular test case, we report the TM-score as 0.0.

Query			E-Thrifty		POMONA		HHpred		DALI
Protein Name	PDB/ BMRB	Length/ Fold	PDB	TM-score	PDB	TM-score	PDB	TM-score	TM-score
KaiA	1M2F/ 5031	135/ (α/β)	1YS7A	0.67	2WRZA	0.34	2PL1A	0.69	0.72
NEDTH	1F3Y/ 4448	165/ (α/β)	2O1CA	0.66	3A6UA	0.62	4ICKA	0.67	0.66
NCS-1	2LCP/ 4378	190/ (α)	2L4HA	0.60	2TN4A	0.37	2L4HA	0.60	0.61
Sortage	1IJA/ 4879	148/ (β)	2XWGA	0.70	3FN5A	0.68	4O8TA	0.57	0.70
PyJ	1FAF/ 4403	79/ (α)	1BMTA	0.51	3QPPA	0.38	2DN9A	0.56	0.60
ERp18	2K8V/ 15964	157/ (α/β)	2LNSA	0.55	3VWWB	0.51	3GNJA	0.45	0.58
ApolPBP1A	2JPO/ 15256	142/ (α)	2WC5A	0.63	3TNWD	0.29	1OOHA	0.57	0.63
Pru Av 1	1E09/ 4671	159/ (α/β)	4Q0KA	0.84	3US7A	0.83	2I9YA	0.61	0.83
Ets-1	2JV3/ 4205	110/ (α)	2DKXA	0.61	2Y1IA	0.28	2DKXA	0.61	0.61
cg2496	2KPT/ 16569	148/ (α/β)	2KW7A	0.63	3PVHA	0.64	3PVHA	0.64	0.64
NCAM	1EPFA/ 4162	191/ (β)	2YD1A	0.79	3QP3C	0.79	2KKQA	0.66	0.78
PG	2HZE/ 4113	108/ (α/β)	3L4NA	0.76	4I2UA	0.78	4I2UA	0.78	0.81
AT5g22580	1RJJ/ 6011	111/ (α/β)	1TR0A	0.72	1TR0C	0.72	1TR0A	0.72	0.70
N-WASP	1MKE/ 5554	144/ (α/β)	3SYXA	0.57	2XQNM	0.62	1XODA	0.56	0.61
Grx2	1G7O/ 4318	215/ (α)	1EEMA	0.63	2WRTG	0.63	1PN9A	0.65	0.72
Average TM-score			0.66		0.57		0.62		0.68

Table 5.1: Template recognition performances of four threading programs on TEST1 proteins using sequence identity cutoff as ≤ 30%. The E-Thrifty column shows the top template identified by E-Thrifty, whereas the next two columns show the top templates identified by HHpred and POMONA. The DALI answers for TEST1 proteins are displayed in the last column. Template information includes the PDB ID of the template and the TM-score of full-length model generated using the corresponding template.

Query			E-Thrifty		POMONA		HHpred		CS23D	CS-ROSETTA	DALI
Protein Name	**PDB/BMRB**	**Length / Fold**	**PDB ID**	**TM-score**	**PDB ID**	**TM-score**	**PDB ID**	**TM-score**	**TM-score**	**TM-score**	**TM-score**
KaiA	1M2F/5031	135/(α/β)	1R8JA	0.85	1R8JA	0.85	2PL1A	0.69	0.50	0.55	0.85
NEDTH	1F3Y/4448	165/(α/β)	2KDWA	0.74	2KDVA	0.75	4S2XA	0.75	0.72	0.34	0.75
NCS-1	2LCP/4378	190/(α)	1FPWA	0.73	1S1EA	0.71	1FPWA	0.73	0.0	0.63	0.71
Sortage	1IJA/4879	148/(β)	1T2OA	0.87	1T2OA	0.87	4O8TA	0.57	0.0	0.45	0.87
PyJ	1FAF/4403	79/(α)	1BMTA	0.51	1QDBB	0.38	2PF4E	0.60	0.30	0.77	0.64
ERp18	2K8V/15964	157/(α/β)	3PH9A	0.63	3PH9B	0.63	3PH9A	0.63	0.53	0.33	0.63
ApolPBP1A	2JPO/15256	142/(α)	1DQEA	0.66	2FJYB	0.86	4INWA	0.67	0.83	0.40	0.87
Pru Av 1	1E09/4671	159/(α/β)	4BK6A	0.88	4C9CA	0.86	4BK7A	0.89	0.89	0.86	0.91
Ets-1	2JV3/4205	110/(α)	4MHVA	0.83	4MHVB	0.83	4MHVA	0.83	0.75	0.71	0.79
cg2496	2KPT/16569	148/(α/β)	2KW7A	0.63	3PVHA	0.64	3PVHA	0.64	0.36	0.77	0.64
NCAM	1EPFA/4162	191/(β)	2VAJA	0.86	2XY2A	0.86	2XY2A	0.86	0.76	0.42	0.88
PG	2HZE/4113	108/(α/β)	1JHBA	0.83	1KTEA	0.85	4RQRA	0.88	0.91	0.88	0.89
AT5g22580	1RJJ/6011	111/(α/β)	1TR0A	0.72	1TR0C	0.72	1Q4RA	0.67	0.66	0.41	0.68
N-WASP	1MKE/5554	144/(α/β)	2IFSA	0.74	2XQNM	0.62	2IFSA	0.74	0.41	0.62	0.74
Grx2	1G7O/4318	215/(α)	3IR4A	0.95	3IR4A	0.95	3IR4A	0.95	0.92	0.28	0.95
Average TM-score			0.76		0.75		0.74		0.66	0.56	0.79

Table 5.2: Template recognition performances of four threading and two protein structure prediction programs on TEST1 proteins using sequence identity cutoff as ≤ 95%. The result includes E-Thrifty, POMONA, DALI identified template information and structure prediction by CS23D and CS-Rosetta. Template information includes the PDB ID of the template and the TM-score of full-length model generated using the corresponding template. A TM-score evaluation of predicted structures by CS23D and CS-Rosetta is also presented.

CASD-NMR Targets			**E-Thrifty**		**POMONA**		**HHpred**		**CS23D**	**CS-ROSETTA**	**DALI**
Protein Name	**PDB /BMRB**	**Length / Fold**	**PDB ID**	**TM-score**	**PDB ID**	**TM-score**	**PDB ID**	**TM-score**	**TM-score**	**TM-score**	**TM-score**
NTPASE	2LCI/ 17613	134/ (α/β)	2L82A	0.79	2L69A	0.76	2MR6A	0.79	0.91	0.78	0.79
BUB1	2LAH/ 17524	160/ (α)	3ESLA	0.80	3ESLB	0.80	3ESLA	0.80	0.85	0.63	0.80
FUS	2LA6/ 17508	99/ (α/β)	1A9NB	0.71	1RK8A	0.72	2CQCA	0.68	0.67	0.66	0.71
NFU1	2M5O/ 19068	97/ (α/β)	3W63A	0.57	3R5GA	0.34	1TH5A	0.61	0.63	0.73	0.61
DNAJC2	2M2E/ 18909	73/ (α)	4UUSA	0.55	3ZNVA	0.53	2DIMA	0.56	0.59	0.69	0.62
NKX 3.1	2L9R/ 17484	69/ (α)	1YRNA	0.71	2R5YB	0.66	2DA2A	0.71	0.71	0.52	0.68
NFU1	2LTM/ 18489	107/ (α/β)	2M8WA	0.66	2M8WA	0.66	1PQXA	0.59	0.48	0.53	0.66
TSTM	2LOJ/ 18214	63/ (β)	2JRAA	0.37	NA	0.0	2JRAA	0.37	0.36	0.40	0.37
IF3-like fold	2LN3/ 18145	83/ (α/β)	3P04A	0.66	3PP7A	0.43	4PWWA	0.48	0.61	0.27	0.68
Average TM- score			0.65		0.50		0.62		0.62	0.58	0.65

Table 5.3: Template recognition performances of E-Thrifty, POMONA, HHpred, CS23D and CS-Rosetta protocols on TEST2 (nine blind targets from the CASD-NMR-2013 competition) proteins are shown. Note that CS23D and CS-Rosetta were run without any homology threshold on these proteins. The final column describes the highest possible alignment quality within the specified sequence identity threshold (≤30%) for these nine proteins. Each threading program column includes the PDB ID of the identified template and the TM-score of full-length model

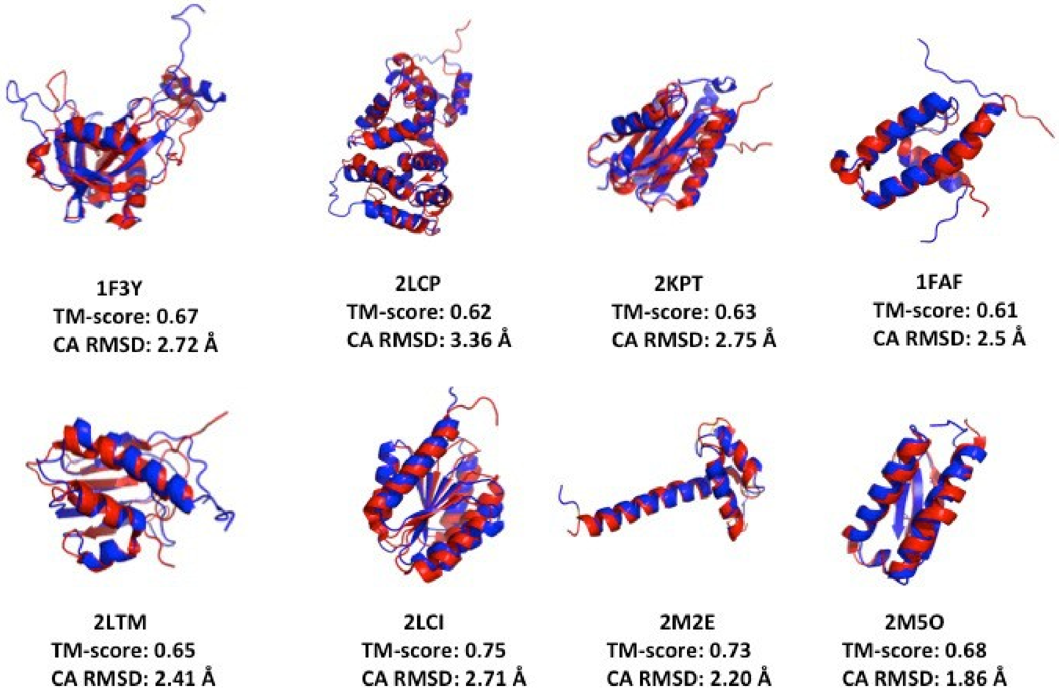


Figure 5.5: E-Thrifty template models after the CS-GAMDy refinement for eight query proteins are shown. Query proteins (red) are shown superposed with the template models (blue) using PyMOL. The TM-score and the Cα rmsd of the secondary structure regions between query and template models are displayed below each model

5.4 Discussion

E-Thrifty's performance

E-Thrifty's performance on the TEST1 and TEST2 data set is described in detail in Tables 5.1-5.3. In terms of TM-score evaluation of the full-length template (or database) models, E-Thrifty consistently gave a better or comparable performance when evaluating against other prediction programs for both test sets. In particular, E-Thrifty achieved an average TM-score of 0.66 for proteins with ≤ 30% sequence identity in the TEST1 data set, as opposed to 0.57 achieved by POMONA (Shen et al. 2015) and 0.62 achieved by HHpred (Söding et al. 2005) respectively (Table 5.1). In Table 5.2, we can see that all three programs perform almost equally well in terms of their average TM-score (E-thrifty=0.76, POMONA=0.75, HHpred=0.74). This was expected given that a ≤95% sequence identify cutoff allows near identical homologs to be used in modeling. We also tested the performance of CS23D (Wishart et al. 2008) and CS-Rosetta (Shen et al. 2008) on these 15 proteins listed in Table 5.2. Both of these programs perform fairly well for these test cases. In particular, CS23D has an average TM-score of 0.66 whereas CS-Rosetta has an average TM-score of 0.56. The purpose of this evaluation is to show that the threading algorithms are applicable to a wide range of applications and to assure users that chemical shift threading methods can attain similar performances as sequence-only threading or comparative modeling methods.

As a second test for E-Thrifty, we selected a number of recent CASD-NMR-2013 (Rosato et al. 2015) targets which we called TEST2. The TEST2 proteins consist of nine targets for which the majority of these proteins are structurally dissimilar to most (or even all) proteins in the PDB. For this "difficult" data set, E-Thrifty achieved an average TM-score of 0.65 with ≤30% sequence identity (Table 5.3). This performance is identical to the average TM-score achieved by the "gold standard" structure superposition program DALI (Holm et al. 2010). Moreover, most of the top 50 hits obtained by E-Thrifty for the nine targets were also identified by DALI through 3D structural superposition, which confirms that E-Thrifty generally finds the correct answer for most (if not all) cases. The average TM-scores for POMONA (Shen et al. 2015) and HHpred (Söding et al. 2005) were 0.50 and 0.62. CS-Rosetta and CS23D are also run on the same set of proteins yielding average TM-scores of 0.58 and 0.62 respectively.

E-Thrifty's sensitivity to missing chemical shifts

It is notable that E-Thrifty failed to find a high quality, structurally similar homolog in the first 100 alignments for one of the CASD-NMR-2013 (Rosato et al. 2015) targets, "YR313A" (2LTL). The best templates for "YR313A" as identified by DALI (Holm et al. 2010) had comparatively low E-Thrifty alignment scores with very modest chemical shift correlations. Further investigation revealed that "YR313A" actually had no ^{13}CO backbone chemical shift assignments. This adversely affected the chemical shift-based structural parameter prediction and consequently led to poorer alignments with the nrPDB database proteins. It is also notable that POMONA (Shen et al. 2015) also failed to detect a good quality template (a template with a TM-score≥0.5) for "YR313A" for the same reason. The missing chemical shift issue led us to exclude this particular query protein from the comparison study. The stand-alone version of the E-Thrifty program now performs a check of the level of completeness of the query protein chemical shift assignments and provides a warning if a significant number of chemical shifts are missing.

TM-score distribution of E-Thrifty versus other templates

The quality (TM-score) of E-Thrifty identified structure templates was also compared with the actual structural homologs found by DALI (Holm et al. 2010) as well as the top alignment hits generated by POMONA (Shen et al. 2015) and HHpred (Söding et al. 2005). For this comparison, DALI homologs with a Z-score ≥ 2, HHpred homologs with a probability ≥ 10 and the top 500 alignments from POMONA and E-Thrifty were chosen. The TM-score distribution of the templates identified by the three threading programs (HHpred, POMONA and E-Thrifty) and DALI (a structure matching program) for two proteins, the P-LOOP NTPase fold (PDB: 2LCI, BMRB: 17613) and the Homeobox domain of the Nkx 3.1 protein (PDB: 2L9R, BMRB: 17484) derived from the TEST2 set is shown in Figure 5.6. Note that 974 DALI alignments, 196 HHpred alignments, 1000 POMONA and 1000 E-Thrifty alignments were used in this comparison. Relative to other programs, E-Thrifty shows somewhat better performance in TM-score≥0.5 sub-region, except for the (0.60-0.70) bin. However, it is notable that E-Thrifty was able to retrieve a higher number of similar folds (representative TM-score ≥0.50) compared to its chemical shift-only counterpart POMONA (265 vs. 218 in 1000 alignments). These high-quality templates consist of more than 25% of the total templates identified by E-Thrifty. In the case of random and unrelated fold rejection (TM-score≤0.3), E-Thrifty showed the second best performance (only 8% of total alignments) after HHpred (which had just 6% of total alignments).

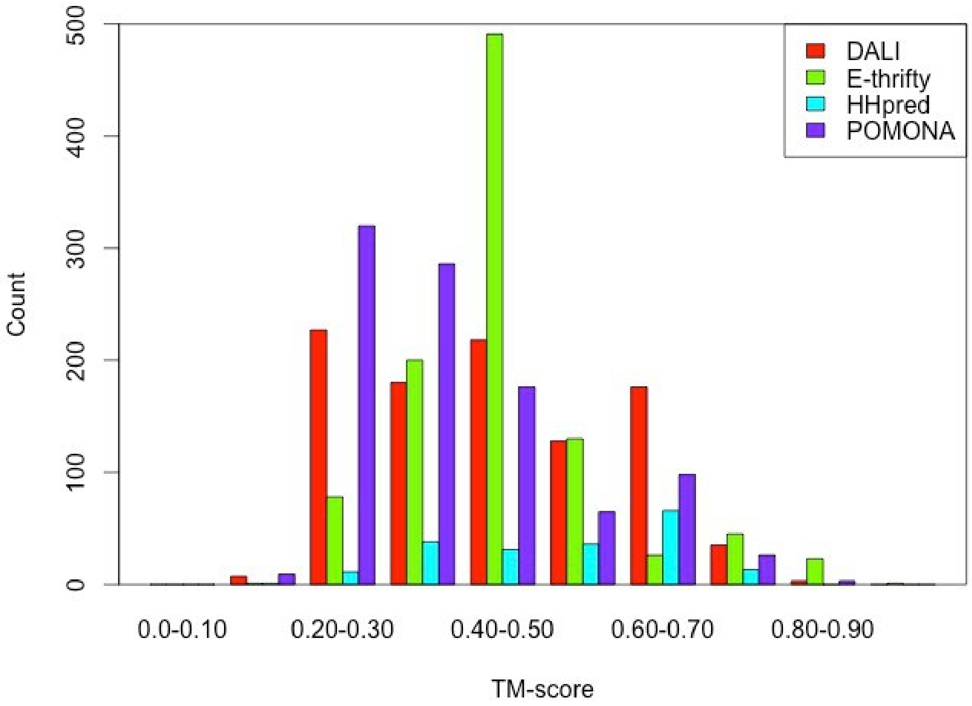


Figure 5.6: TM-score distribution of full-length models using input templates identified by DALI, E-Thrifty, HHpred and POMONA for P-LOOP NTPase fold (PDB: 2LCI, BMRB: 17613) and Homeobox domain of Nkx 3.1 protein (PDB: 2L9R, BMRB: 17484).

Correlation between template quality and E-Thrifty alignment scores

We also examined the correlation between E-Thrifty alignment scores and the overall alignment quality. The alignment scores versus the Cα-rmsd of equivalent residues between the P-LOOP NTPase fold (PDB: 2LCI, BMRB: 17613) and the top 400 database alignments are shown in Figure 5.7. The red and blue dots in the figure indicate the alignments spanning the ≥30% and the ≤30% sequence identity ranges.

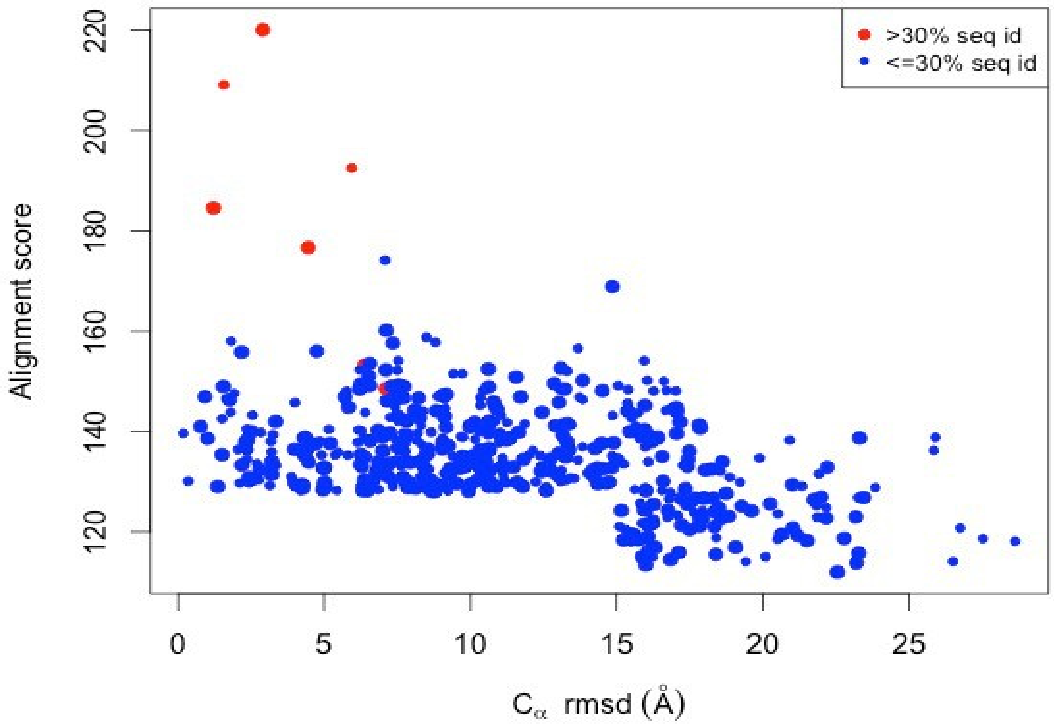


Figure 5.7: Alignment score vs. Cα – rmsd between equivalent residues of the P-loop NTPase protein (PDB: 2LCI, BMRB: 17613) and template proteins is shown. The red dots represent the templates with ≤95% (or >30%) sequence identity and blue dots indicate the templates with ≤ 30% sequence identity.

As can be seen in Figure 5.7, the Cα-rmsd of the E-Thrifty alignments weakly correlate (Pearson's correlation coefficient = -0.40) with the corresponding alignment scores. In particular, we found that the top 25% of hits found by E-Thrifty had a high TM-score range (~0.5-0.9) for the P-loop NTpase protein. However, the distribution of Cα-rmsd values appears to be quite sparse. This is probably due to the fact that Cα-rmsd measures are fairly sensitive to small structural defects in the aligned regions. On the other hand, the TM-score measured for the full-length template models shows a better correlation (Pearson's correlation coefficient = 0.54) with the alignment scores for the same set of templates (Figure 5.8). In this figure, the green dots represent the positive templates identified by both DALI (Holm et al. 2010) and E-Thrifty. These positive templates accounted for almost 20% of the total number of templates.

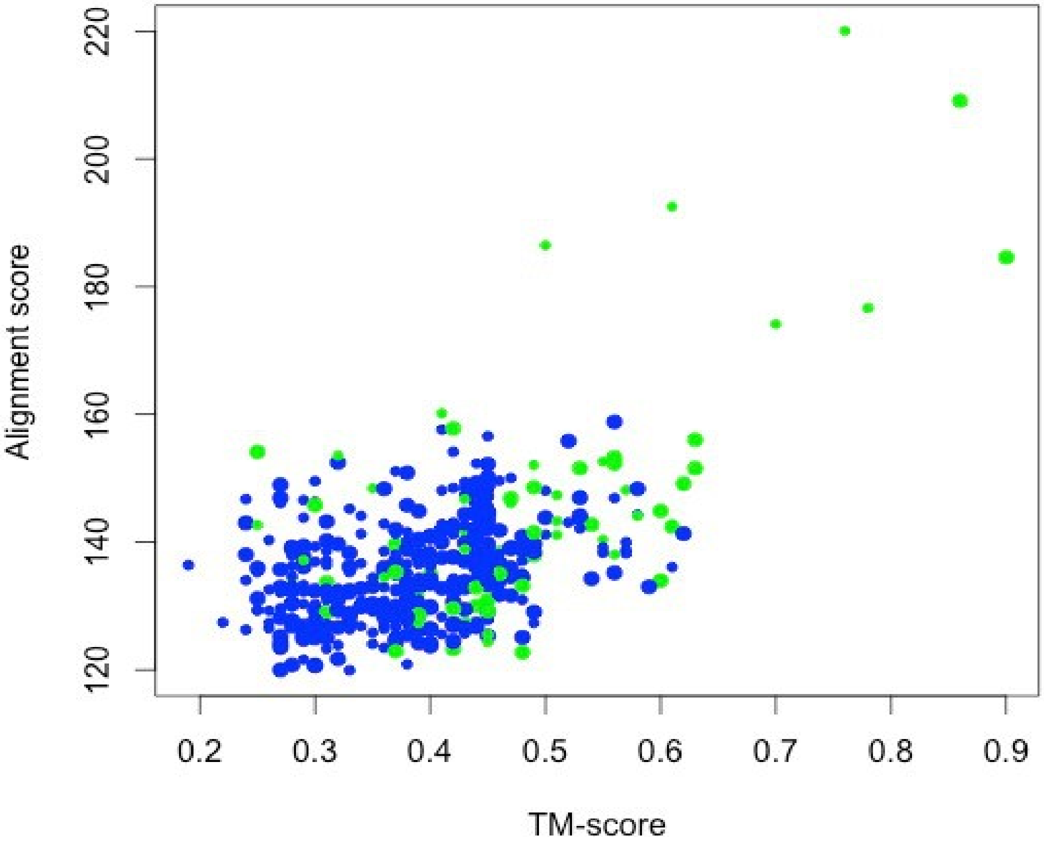


Figure 5.8: E-Thrifty alignment score vs. TM-score is shown for top 500 templates identified by E-Thrifty for the P-loop NTPase protein (PDB: 2LCI, BMRB: 17613). The green dots indicate the positive templates identified by both E-Thrifty and DALI, which is approximately 20% of the template population.

Detecting remote homologs

All nine targets of CASD-NMR-2013 (the TEST2 data set) are proteins with very low sequence identity to any known structure. Two proteins, 2LOJ (BMRB: 18214) and 2LN3 (BMRB: 18145), proved to be particularly challenging for almost all of the programs we tested. Both of these proteins have low sequence identity homologs in the PDB. E-Thrifty was able to correctly identify the most likely template for 2LN3 (3P04A, TM-score=0.66) according to DALI (Holm et al. 2010). For 2LOJ, E-Thrifty as well as

HHpred (Söding et al. 2005) identified 2JRAA as the best template with a relatively modest TM-score of 0.37. DALI also failed to find a template with a better TM-score (i.e. TM-score>0.37) within its threshold Z-score≥2. Further searches through the PDB to identify other structural homologs for 2LOJ revealed only one other homolog 4YNX (with >50% sequence identity), which was solved in 2015 (after our exclusion date). Therefore at the time of its deposition in 2013, 2LOJ appears to be one of those truly novel protein folds that now comes along only very rarely.

Query		Seq only	Seq+ Sec. Struct.	Seq+Sec. Struct. +Torsion angles
PDB	**BMRB**			
2M5O	19068	0.31	0.20	0.31
2LTM	18489	0.25	0.30	0.53
2M2E	18909	0.40	0.30	0.37
2LCI	17613	0.46	0.79	0.79
1E09	4671	0.80	0.76	0.84
2JV3	4205	0.30	0.30	0.34
2JPO	15256	0.30	0.30	0.30
1IJA	4879	0.20	0.68	0.69
2K8V	15964	0.20	0.30	0.30
Average TM-score		**0.36**	**0.44**	**0.50**

Table 5.4: E-Thrifty performances for nine proteins randomly chosen from TEST1 and TEST2 sets using sequence only; sequence and shift-derived secondary structures (HHpred features) and sequence, shift-derived secondary structures and torsion angles (POMONA features) with a sequence identity threshold of ≤30% are shown.

E-Thrifty performance using different combinations of sequence/structure features

We also analyzed E-Thrifty's performance using different sequence/structure feature combinations with a sequence identity threshold of ≤30%. This was done to assess which properties (sequence, secondary structure, chemical shifts, torsion angles, etc.) were most important to E-Thrifty's overall performance. For this assessment, we randomly chose nine proteins from both the TEST1 and TEST2 sets. As can be seen in Table 5.4 using sequence as the only input, E-Thrifty produces an average TM-score of 0.36 (compared to a TM-score of 0.65 for the full E-Thrifty package). Using sequence data alone, only 1/9 of these proteins generated high quality (TM-score>0.50) matches or found the correct "gold standard" structural homolog. This reduced level of performance was certainly expected. Indeed, this highlights the fact that sequence data alone is not often sufficient to identify high quality structural homologs with low sequence identity.

Using the combination of sequence and shift-derived secondary structure, E-Thrifty showed a modestly improved average TM score of 0.44 (compared to 0.65 for the full package). Using these two features, a total of 3/9 of these proteins generated high quality (TM-score>0.50) matches or found the correct "gold standard" structural homolog. Interestingly, HHpred (Söding et al. 2005), which also uses just these two features, was able to achieve an impressive TM-score of 0.59. Obviously for this test, E-Thrifty was not optimized to work with just two input features (i.e. sequence and predicted secondary structure) and certainly if we had optimized its training it may have performed somewhat better. It is also important to note that HHpred is a web server that uses PSIPRED (Jones 1999) for its secondary structure prediction. PSIPRED uses a fully up-to-date PDB-derived database of assigned secondary structures to assist with its secondary structure prediction routine. So in this case, HHpred likely identified exact matches to all nine query proteins in its PDB database, thereby giving it a significant advantage over E-Thrifty, which had all of the query proteins removed from its database. In other words, HHpred may have had all the answers in hand already.

Using the combination of sequence, shift-derived secondary structures and shift-derived torsion angles, E-Thrifty shows an average TM-score of 0.50 (compared to 0.65 for the full package). Using these three features, 4/9 of these proteins generated high quality (TM-score>0.50) matches or found the correct "gold standard" structural homolog. Interestingly, POMONA (Shen et al. 2015), which also uses these three features, was able to achieve a TM-score of 0.54. As noted previously, E-Thrifty was not optimized or re-parameterized to work only with these three features and certainly if some optimization was done, E-Thrifty might have performed much better. Overall, the primary intent of these experiments was to illustrate the role that: 1) sequence alone; 2) sequence + secondary structure and 3) sequence + secondary structure + torsion angles had in the performance of E-Thrifty. These data clearly show that more information is better. The secondary objective was to highlight how important parameter optimization can be in getting a threading program to work optimally. Clearly parameter optimization is important. The third objective was to illustrate why the inclusion of additional features (ASA, chemical shift matching, super-secondary structure) were needed to boost the overall threading performance of E-Thrifty.

Limitations and potential improvements of E-Thrifty

One of the limitations of E-Thrifty is that it relies on categorical, somewhat imprecise character-based representations to describe both the query structures and the corresponding database structures. For example, real-valued φ/ψ angles are converted into a discretized 9-letter torsion alphabet; while real valued fractional ASA (fASA) values are classified into an even simpler 3-letter alphabet. Approaches that use numeric torsion angles and numeric fASA values might be expected to further improve E-Thrifty's performance. This is because numeric values would be far more precise and would capture far more subtle information about these torsion angle and fASA features. While E-Thrifty makes use of a number of high quality dynamic programming alignment routines, a further improvement in its sensitivity for detecting remote homologs could potentially be achieved by including more powerful Hidden Markov Model profile alignments. These alignment methods have consistently proven to be very effective in detecting distant homologs (Krogh et al. 1994, Eddy 1998, Karplus et al. 1998) and appear to play a key role in the high level of performance achieved with HHpred (Söding et al. 2005). While improved alignment methods could be particularly beneficial, improved scoring functions may prove to be equally useful. Indeed, we suspect further improvements could be achieved by designing a suitable Z-score value for a more effective assessment of E-Thrifty sequence-structure alignment quality.

Next steps for E-Thrifty

As a chemical shift threading method E-Thrifty is particularly good at automatically generating "approximate" or initial 3D protein models. However, to obtain truly high-quality, atomic resolution structures it will be necessary to couple E-Thrifty to other kinds of programs that can perform true structural refinement. As highlighted in Figure 5.6, E-Thrifty can be easily coupled to CS-GAMDy (Berjanskii et al. 2015) to perform chemical shift refinement. These refinements certainly improved the quality and accuracy of the starting structures. While these refinement calculations can take several hours, it is not unreasonable to imagine having E-Thrifty tightly coupled to CS-GAMDy (either as a stand-alone program or as a web server) in the near future. Further enhancements to E-Thrifty will likely include the addition of other structure "massaging" or refinement options such as XPLOR-NIH (Schwieters et al. 2003), AMBER (Pearlman et al. 2005) or DYANA (Güntert et al. 1997). Adding these tools to the pipeline would also allow E-Thrifty to incorporate other experimental measures such as NOEs, J-couplings and residual dipolar couplings into its structure generation and refinement protocols. Finally, in the rare situations where no structural homolog can be found, it may be possible to consider blending CS23D (Wishart et al. 2008), Cheshire (Cavalli et al. 2007), CS-Rosetta (Shen et al. 2008) or other *ab initio* structure predictors with E-Thrifty. This would lead to the creation of a much more fail-safe and far more comprehensive chemical shift-based structure generation pipeline. Indeed, our long-term plan is for E-Thrifty to become fully integrated into the next release of CS23D.

5.5 Conclusion

In this study, we have described a new and particularly powerful protein fold recognition method called E-Thrifty that uses chemical shift threading to generate high quality coordinate models for proteins having little or no sequence identity to any protein in the PDB. We believe this represents a significant step towards "solving" protein structures using only chemical shift information. As outlined above, E-Thrifty uses chemical shift derived secondary structures, chemical shift derived fASA values and chemical shift derived torsion angles to perform a comprehensive alignment between the query sequence (with experimentally determined chemical shift assignments) and a large database of proteins with known structures and predicted chemical shifts. E-Thrifty exploits a number of recently developed chemical shift analysis tools (CSI 2.0 (Hafsa et al. 2014), CSI 3.0 (Hafsa et al. 2015a), ShiftASA (Hafsa et al. 2015b), TALOS-N (Shen et al. 2013), ShiftX2 (Han et al. 2011)) to generate chemical shifts or chemical-shift derived information for both the query and the database proteins. A Smith-Waterman local alignment algorithm with a variable gap penalty function was found to be the best tool for performing the sequence-structure alignment. The weighting coefficients and fitness scores used to evaluate the alignments were optimized through both a parameter grid search and a linear regression analysis. E-Thrifty includes a chemical shift fitness score and an e-value scoring system to fully evaluate the alignments between the query and the database proteins. In addition, E-Thrifty performs a cluster analysis step for all identified folding templates to group them according to their structural similarity. The templates identified by E-Thrifty can be subsequently used for chemical shift-based structure refinement.

In terms of performance, E-Thrifty achieved an average TM-score of 0.66 for query sequences having ≤30% sequence identity (as measured on an independent test set of 15 proteins). E-Thrifty's performance was found to be comparable to the "gold standard" DALI (Holm et al. 2010) which had an average TM-score of 0.68. In contrast to E-Thrifty or other structure prediction routines, DALI uses experimentally derived coordinate data to identify structural homologs of proteins by structural superposition (i.e. DALI knows the answer, whereas E-Thrifty predicts the answer). E-Thrifty was also evaluated on a number of recent CASD-NMR-2013 targets and achieved an average TM-score performance of 0.65 on nine test proteins with ≤30% sequence identity. The performance of E-Thrifty clearly demonstrates its ability to "predict" a 3D fold by using only chemical shift information. With its exceptional performance, we believe that E-Thrifty could be a very useful contribution towards the goal of rapid and automated protein structure generation and refinement by NMR chemical shifts.

Chapter 6

Contributions and Future Prospects

6.1 Introduction

In this chapter, I will summarize the main contributions of this thesis and suggest a number of possible areas for further exploration as well as some ideas for future research directions. In discussing my major research contributions, I believe it is important to emphasize that this is an applied computing thesis. It is not a traditional "pure computing science" thesis that is focused on developing a novel computing technique or devising breakthrough machine-learning algorithms. Instead, I focused on applying state-of-the-art computing science knowledge to solve a number of "real world" and extremely difficult problems in structural biology. Working on challenging inter-disciplinary problems such as computational protein folding or NMR chemical shift analysis requires in-depth knowledge of several very different domains, including computing science, NMR spectroscopy, structural biology and biochemistry. Solving inter-disciplinary problems requires not only a good knowledge of the individual disciplines themselves, but an ability to identify which problems are solvable with which techniques.

My long-term career goal is to help solve one of the most vexing and computationally challenging problems in biology – the protein folding problem. Simply stated, the protein folding problem attempts to answer the question: How does a protein sequence determine a protein's three-dimensional shape? From a machine learning perspective, one of the best ways of solving this problem is to gather a large body of data on protein structures and protein sequences and to "learn" the rules for protein folding. In fact, this was the original motivation behind the Protein Structure Initiative (PSI) (Smith et al. 2007) and for the establishment of the Protein Data Bank (PDB). Indeed, since 1959 a total of 115,000 protein structures have been solved and deposited into the PDB. While this may seem like a large number, it is still insufficient to cover all of protein folding space and for computers (or humans) to "learn" the rules of protein folding. Adding more structures would seem to be a trivial solution to this problem, but determining protein structures is non-trivial and expensive. If we assume that each protein structure costs

$50,000-$75,000 to complete and each structure takes 3-4 person years of work, one can calculate that the total amount of money and time spent on experimentally solving all the structures now in the PDB is somewhere between $6-8 billion and nearly 500,000 person years. Given these huge costs and the vast amount of time already spent on solving the existing set of structures, there is considerable interest (and pressure from funding bodies) to find much more efficient and automated computational approaches to solve protein structures.

This "automation" challenge is what motivated my research for this thesis. Over the next few pages, I will highlight the contributions of this thesis towards facilitating automated or near-automated protein structure determination via NMR spectroscopy.

6.2 Contributions of this thesis

As mentioned in the introductory chapter, this thesis is sub-divided into four sub projects. I will summarize the key findings in these chapters and briefly describe the methods used in each sub-project in the following paragraphs.

In chapter 2, I described a significantly improved secondary structure identification method called CSI 2.0 (Hafsa et al. 2014) that can be used to accurately determine the secondary structure of proteins using NMR chemical shifts, alone. CSI 2.0 was designed to be a successor to the much older and somewhat simplistic Chemical Shift Index or CSI algorithm (Wishart et al. 1992 & 1994). In particular, CSI 2.0 replaces the simple digital chemical shift filter used in the original CSI with a much more powerful "feature filter" that uses machine learning techniques to locate secondary structures along the protein chain. CSI 2.0 exploits all six backbone chemical shifts ($^{13}C_{\alpha}$, $^{13}C_{\beta}$, ^{13}C, ^{15}N, ^{1}HN, $^{1}H_{\alpha}$) with sequence-derived features instead of using only backbone ^{13}C and ^{1}H chemical shifts like its predecessor. With an extended feature set and sophisticated machine-learning algorithms, CSI 2.0 is able to obtain a much more accurate 3-state secondary structure (α-helix, β-strand and coil) assignment. More specifically, a multi-class "kernelized" SVM classifier is used to train the prediction model in which a RBF kernel performs feature dimension reduction. The training model parameters (the "regularization" or "cost" parameter in the SVM classifier and "kernel width parameter" in the Gaussian RBF kernel) were optimized using a repeated 10-fold cross-validation. A multi-residue Markov Model was designed for post-assignment filtering to remove any "confusing" or "scrambled" assignments in contiguous stretches of helices, strands or coil structures along the protein chain (e.g. *CCBHH* or *BBHCC or HCHCH).* Our tests indicate that CSI 2.0 achieves an average Q3 accuracy (a standard evaluation metric for 3-state secondary structure identification) of 90.56% for a training set of 181 proteins in a repeated 10-fold cross-validation and 89.35% for a test set of 59 proteins. The average Segment Overlap or SOV score (another common evaluation metric) for the test proteins was found to be 88.45. This performance represents a substantial improvement over the original CSI (Q3 of 89% vs. 79%) as well as over the other state-of-the-art chemical shift-based methods for secondary structure identification (Q3 of 89% vs. ~86%). This performance improvement was statistically significant when using three of the most widely used secondary structure assignment protocols, DSSP (Kabsch et al. 1983), STRIDE (Frishman et al. 1995) and VADAR (Willard et al. 2003). Interestingly, the standard 3D coordinate based secondary structure identification methods, such as DSSP and STRIDE appear to perform no better than CSI 2.0, which uses no coordinate data. Based on data presented in chapter 2 and the level of agreement measured between different secondary structure identification methods (NMR vs. X-ray vs. different programs), it appears that CSI 2.0 reaches the maximum performance that secondary structure assignment methods can achieve.

CSI 2.0 not only shows an improved performance for identifying secondary structure locations, but it also successfully detects different secondary structures in structurally dissimilar proteins sharing high sequence identity - something that misleads other programs. CSI 2.0 was also able to identify the absence of secondary structures in proteins that had been unfolded in urea or other denaturants. This is something that would fool programs that place too much emphasis on sequence data and too little emphasis on chemical shift data. CSI 2.0 is currently implemented as a web server (http://csi.wishartlab.com). The server accepts protein NMR chemical shift assignments in most standard formats and generates a comprehensive, colorful report describing the protein's secondary structure assignments. I believe that CSI 2.0, with its superior performance, will be a useful contribution to the field of biomolecular NMR. In particular, it could be used in providing accurate NMR constraint data (such as torsion and distance restraints) in the early stages of protein structure determination as well as in defining secondary structure locations in the final protein model(s). CSI 2.0 could also serve as a robust alternative to standard coordinate-based methods for secondary structure identification.

In chapter 3, I described a successor to CSI 2.0 and named it CSI 3.0 (Hafsa et al. 2015a). CSI 3.0 greatly extends the concept of chemical shift assignment of secondary structure over what was described in CSI (Wishart et al. 1992 & 1994) and CSI 2.0 (Hafsa et al. 2014). In particular, it uses a pipeline of several well-tested, previously published programs, including its predecessor CSI 2.0 to identify the locations of secondary and super-secondary structures. More specifically, CSI 3.0 accurately identifies a total of 11 types of secondary and super-secondary structures, including 8 types of local secondary structures (helices, β-strands, coils) and 5 common β-turns (type I, II, I', II' and VIII)) as well as 3 types of super-secondary structures or topological features (β-hairpins, edge beta-strands and interior beta-strands). The increased number of secondary and supersecondary structures identified by CSI 3.0 makes it both distinct and superior to any other chemical shift-analysis tool.

CSI 3.0 uses a combination of heuristic and machine-learning algorithms to locate the secondary and super-secondary structure locations along the protein chain. A binary SVM classifier with an RBF kernel is used for interior and edge β-strand classification, whereas a simple rule-based algorithm was developed to classify different β-turn types. CSI 3.0 was tested on a set of 13 high-resolution protein structures that were solved by X-ray and NMR and spanning a broad range of secondary structure content and folds. It exhibited >90% average agreement over all 11 types of secondary and super-secondary assignments when compared to those made via standard coordinate analysis programs such as DSSP, STRIDE and VADAR. CSI 3.0 has been converted to a web server (http://csi3.wishartlab.com) and accepts most standard formats of chemical shift assignment input. It outputs colourful CSI plots (bar graphs and secondary structure icons) along with secondary/super-secondary structure textual assignments, which can be readily used as constraints for structure determination and refinement. The images generated by CSI 3.0 may be used for presentations and publications. I believe that the additional secondary/supersecondary structure data, along with the useful topological information and colourful graphical output generated by CSI 3.0 will not only improve the quality of preliminary protein structure descriptions (often obtained shortly after chemical shift assignments are completed) but also facilitate protein structure determination by NMR. In particular, with the recent trends towards protein structure determination and refinement using chemical shifts (only), chemical shift threading or minimal numbers of NOEs, this added information could prove to be particularly useful to a growing number of NMR spectroscopists.

Chapter 4 describes an improved fractional Accessible Surface Area (fASA) estimation method named as ShiftASA (Hafsa et al. 2015b). The method is based on machine learning techniques (specifically a boosted tree regression model) that combines chemical-shift and sequence derived features to accurately estimate per-residue fractional ASA (fASA) values of amino acid residues in water-soluble proteins. ShiftASA exploits the well-known chemical shift sensitivity to ASA as well as the "pseudo-ASA" information content in sequence data. The fASA real value estimation problem was mapped as a Stochastic Gradient Boosting Regression Tree model and model parameters were optimized using a repeated 10-fold cross validation on a training set of 344 data points. Each data point in the training set consisted of 15 chemical shift and sequence-derived features spanning a three amino acid residue window. This approach showed a correlation coefficient between predicted and experimental fASA values of 0.79 when evaluated on a set of 65 independent test proteins. This represents an 8.2% improvement over the next best performing (sequence-only) method. On a separate test set of 92 proteins, ShiftASA reported a mean correlation coefficient of 0.82, which was 12.3% better than the next best performing method. The mean absolute error in ASA values was found to drop from 0.19 to 0.14 Å^2 and the root mean squared error fell from 0.26 to 0.19 Å^2 compared to its sequence-only and chemical shift-only counterparts. On the second test set (TEST2), ShiftASA attained a mean correlation coefficient of 0.82, a clear improvement over correlation coefficients of 0.67 and 0.73 reported by the best performing sequence-only and chemical-shift-only methods, respectively. ShiftASA is also available as a web server (http://shiftasa.wishartlab.com). It accepts most standard chemical shift assignment formats as input and outputs real-value fASA predictions. ShiftASA also allows flexible, categorical prediction of binary or ternary ASA states.

Overall, I believe that ShiftASA, with its significantly improved prediction of ASA parameters, will not only facilitate protein fold recognition and *de novo* protein structure prediction by providing useful structural constraints, but it will contribute to the generation and refinement of protein structures by NMR chemical shifts (only). In addition, an accurate estimation of fASA values can be used to calculate useful thermodynamic parameters from chemical shift data.

In chapter 5, I described a new and particularly effective protein fold recognition method called E-Thrifty that uses chemical shift threading to generate high quality 3D coordinate models for proteins having little or no sequence identity to any protein in the PDB. E-Thrifty uses chemical shift derived secondary structures, chemical shift derived fASA values and chemical shift derived torsion angles to perform a comprehensive alignment between the query sequence (with experimentally determined chemical shift assignments) and a database of proteins with known structures and predicted chemical shifts. E-Thrifty exploits a number of existing tools (CSI 2.0 (Hafsa et al. 2014), CSI 3.0 (Hafsa et al. 2015a), ShiftASA (Hafsa et al. 2015b), TALOS-N (Shen et al. 2013), and ShiftX2 (Han et al. 2011)) to generate chemical shifts or chemical-shift predicted data for both the query and the database proteins. E-Thrifty includes a chemical shift fitness score and an e-value scoring system to evaluate the alignments between the query and the database proteins. In addition, E-Thrifty performs a cluster analysis of identified folding templates to group them according to structural similarity. The templates identified by E-Thrifty can be subsequently used for structure modeling and chemical shift based structure refinement. A Smith-Waterman local alignment algorithm with a variable gap penalty function was implemented to achieve optimal sequence-structure alignment. The weighting coefficients and fitness scores used to evaluate the alignments were optimized through either a grid search or by fitting via linear regression. E-Thrifty achieves an average TM-score of 0.66 for query sequences having ≤30% sequence identity (as measured on an independent test set of 15 proteins). E-Thrifty's performance was found to be comparable to the "gold standard" DALI (Holm et al. 2010) which had an average TM-score of 0.68. Unlike E-Thrifty or other structure prediction routines, DALI uses experimentally derived coordinate data to identify structural homologs of proteins by structural superposition (i.e. DALI knows the answer, whereas E-thrifty is predicting the answer). E-Thrifty was also evaluated on a number of recent CASD-NMR-2013 (Critical Assessment of Automated Structure Determination by NMR) (Rosato et al. 2015) targets and an average TM-score performance of 0.65 was achieved on nine test proteins with ≤30% sequence identity. The performance of E-Thrifty clearly demonstrates its ability to "predict" 3D structure by using only chemical shift information in the presence of suitable templates in the structure database. Given its exceptional performance, I believe that E-Thrifty will be a very useful contribution towards the goal of rapid and automated structure generation and refinement by NMR chemical shifts.

6.3 Future prospects

In this section, I will suggest a number of possible areas for further development along with several ideas for future research directions. As mentioned in the previous section, most my thesis work focused on structural parameter calculators and chemical shift threading methods that could be used to automate or facilitate initial protein structure determination or structure characterization. The next obvious step in the process is to integrate these methods into structure refinement methods.

CS23D 2.0, ShiftRefiner and NMRrefiner

Many of the tools I have developed will be integrated into three different programs and/or web servers that are being developed in Dr. Wishart's laboratory for NMR-based protein structure determination and refinement. These include CS23D 2.0, ShiftRefiner and NMRrefiner. CS23D 2.0 is a successor of CS23D (Wishart et al. 2008), a web-based protein structure determination pipeline that uses NMR chemical shifts to determine and refine 3D protein structures. CS23D 2.0 is currently in the developmental stage. Some of the existing weaknesses in CS23D, namely its poor performance in threading and chemical shift refinement, will be corrected in the new version. In particular, CS23D 2.0 will incorporate my improved chemical shift threading method, E-thrifty, to more accurately identify structural similar templates. It will also use a more advanced homology modeling program called MODELLER (Sali et al. 1993, Webb et al. 2014) and a recent chemical shift refinement protocol called CS-GAMDy (Berjanskii et al. 2015) to significantly improve its overall performance and accuracy.

ShiftRefiner is a web service program, currently being developed that greatly extends the capabilities of CS-GAMDy. ShiftRefiner will use predicted secondary structures and fASA estimations from CSI 2.0 (Hafsa et al. 2014) and ShiftASA (Hafsa et al. 2015b). These predictions will serve as target functions for CS-GAMDy's biased molecular dynamics (MD) and genetic algorithm steps. In particular, CSI 2.0's predicted secondary structure will be used as a part of a chemical shift fitness score while ShiftASA's fASA predictions will be used as the second scoring function to bias the MD protocol during GAMDy refinement. Shift-predicted secondary structure "scores" and fASA "scores" will be among four structural fitness scores randomly selected to assess the quality of the protein models during different stages of refinement in ShiftRefiner. Upon completion, ShiftRefiner will have two different functions. The first will be a "fast" 2-3 hour computational protocol that allows good quality NMR structures to be refined so that they improve their structural and chemical shift quality/agreement by at least 50% over the starting state. The second will be a "slow" protocol that is capable of refining starting NMR structures that are as much as 5 Angstroms RMSD away from the correct structure to a near correct structure (<2 Angstroms RMSD) within 24 CPU hours. Ultimately ShiftRefiner will be incorporated into an improved protein folding suite (tentatively called CS23D 3.0) to assist with protein structure generation from raw NMR chemical shift data.

Another web service program called NMR-Refiner for chemical shift and NOE based structure refinement is also under development. It will include options to add NOE and residual dipolar coupling (RDC) restraints during the biased MD and genetic algorithm refinement steps. NMR-Refiner will also incorporate other functionalities to optimize protein-protein and protein-ligand complexes.

Ab-initio and Sparse Data Protein Structure Prediction

Predicting protein structures from sequence and very limited experimental information is referred to as *ab-initio* structure prediction. This problem is considered extremely difficult because it requires extensive conformation searches through a vast and complex multidimensional hyperspace. CS-ROSETTA (Shen et al. 2008) is currently the best state-of-the art *ab-intio* or sparse-data protein structure prediction tool available. It uses a coarse-grained chemical shift fragment-based search through conformational space using a knowledge-based scoring function that favors protein-like features. The peptide fragments (defined by chemical shift similarity) used by CS-ROSETTA are simply three and nine residue fragments that exhibit maximal chemical shift similarity. This kind of fragment selection procedure uses little or no local structural or topological information. As a result, CS- ROSETTA frequently fails to generate structures for non-homologous or "hard-to-predict" protein structures. The methods proposed in this thesis could be used to generate much more useful structural and topological restraints and much more precise peptide fragments that could greatly improve ab-initio methods such as CS- ROSETTA. In particular using my improved protocols for identifying secondary structures (CSI 2.0), calculating accessible surface area (ShiftASA), identifying interior and edge β-strands, and β-turns (CSI 3.0) or selecting better chemical shift threaded templates (E-thrifty) could be bundled into an improved fragment selection protocol. An *ab-inito* structure generation functionality with these proposed fragment selection features is now being planned for a new version of CS23D (CS23D 3.0) over the next year or two.

Concluding Remarks

Overall, I believe my work has advanced the field of NMR-based protein structure determination and refinement. It has also laid a solid foundation for future work in my supervisor's laboratory that should make protein structure determination by NMR spectroscopy much faster, much more accurate and much more robust. I also believe that this work nicely illustrates how the application of advanced computing science techniques can lead to significant and useful advances in scientific disciplines that have not previously exploited these ideas or methods.

Bibliography

Adamczak R, Porollo A, Meller J (2004) Accurate prediction of solvent accessibility using neural networks–based regression. Proteins: Struct, Funct, Bioinf 56(4): 753-767

Adamczak R, Porollo A, Meller J (2005) Combining prediction of secondary structure and solvent accessibility in proteins. Proteins: Struct, Funct, Bionf 59(3): 467-475

Adams PD, Baker D, Brunger AT, Das R, DiMaio F, Read RJ, Richardson DC, Richardson JS, Terwilliger TC (2013) Advances, interactions, and future developments in the CNS, Phenix and Rosetta structural biology software systems. Annu Rev Biophys 43: 265-287

Ahmad S, Gromiha MM (2002) NETASA: neural network based prediction of solvent accessibility. Bioinformatics 18(6): 819-824

Ahmad S, Gromiha MM, Sarai A (2003) Real value prediction of solvent accessibility from amino acid sequence. Proteins: Struct, Funct, Bioinf 50(4): 629-635

Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2009) A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci 106(50): 21149-21154

Altschul SF, Gish W (1996) Local alignment statistics. Meth Enzymol 266:460-480

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17): 3389-3402

Andrec M, Snyder DA, Zhou Z, Young J, Montelione GT, Levy RM (2007) A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. Proteins: Struct, Funct, Bionf 69(3): 449-465

Anfinsen CB (1973) Principles that Govern the Folding of Protein Chains. Science 81: 223-230

Arnold JT, Dharmatti SS, Packard ME (1951) Chemical effects on nuclear induction signals from organic compounds. J Chem Phys 19(4): 507-507

Avbelj F, Kocjan D, Baldwin RL (2004) Protein chemical shifts arising from α-helices and β-sheets depend on solvent exposure. Proc Natl Acad Sci USA 101(50): 17394-17397

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. (2005) The universal protein resource (UniProt). Nucleic Acids Res 33(Database issue): D154-D159

Bax A (1989) Homonuclear Hartmann-Hahn experiments. Method Enzymol 176: 151-168

Becker ED (1999) High resolution NMR: theory and chemical applications. Academic Press ISBN: 0-12-084662-4

Benkert P, Tosatto SC, Schomburg D (2008) QMEAN: A comprehensive scoring function for model quality assessment. Proteins: Struct Funct Bioinf 71(1): 261-277

Berjanskii M, Arndt D, Liang Y, Wishart DS (2015) A robust algorithm for optimizing protein structures with NMR chemical shifts. J Biomol NMR 63(3):255-264

Berjanskii M, Tang P, Liang J, Cruz JA, Zhou J, Zhou Y, Bassett E, MacDonell C, Lu P, Wishart DS (2009) GeNMR: a web server for rapid NMR-based protein structure determination. Nucleic Acids Res 37(Web server issue):W670-W677

Berjanskii MV, Neal S, Wishart DS (2006) PREDITOR: a web server for predicting protein torsion angle restraints. Nucleic Acids Res 34(suppl 2): W63-W69

Berjanskii MV, Wishart DS (2005) A simple method to predict protein flexibility using secondary chemical shifts. J Am Chem Soc 127(43): 14970-14971

Berjanskii MV, Wishart DS (2013) A Simple Method to Measure Protein Side-Chain Mobility Using NMR Chemical Shifts. J Am Chem Soc 135(39): 14536-14539

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28(1): 235-242

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Betancourt MR, Skolnick J (2001) Universal similarity measure for comparing protein structures. Biopolymers 59(5):305–309

Biemann K (1989) Tables of spectral data for structure determination of organic compounds. Springer, Berlin

Biswas KM, DeVido DR, Dorsey JG (2003) Evaluation of methods for measuring amino acid hydrophobicities and interactions. J Chromatogr A 1000(1): 637-655

Blanco FJ, Rivas G, Serrano L (1994) A short linear peptide that folds into a native stable β-hairpin in aqueous solution. Nat Struct Mol Biol 1(9): 584-590

Bodenhausen G, Ruben DJ (1980) Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem Phys Lett, 69(1): 185-189

Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235-242

Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170

Brügel W (1979) Handbook of NMR spectral parameters. Heyden, London Vol. 1: 47

Brunger AT (1993) XPLOR Version 3.1 A System for X-ray Crystallography and NMR. Yale University Press, New Haven and London

Camilloni C, De Simone A, Vranken WF, Vendruscolo M (2012) Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. Biochemistry 51(11): 2224-2231

Castellani F, van Rossum B, Diehl A, Schubert M, Rehbein K, Oschkinat H (2002) Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. Nature 420(6911): 98-102

Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci 104(23): 9615-9620

Chan HS, Dill KA (1993) The protein folding problem. Phys today 46(2): 24-32
Chaudhuri TK, Paul S (2006) Protein-misfolding diseases and chaperone-based therapeutic approaches. FEBS J 273(7): 1331-1349

Chen H, Zhou HX (2005) Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res 33(10): 3193-3199

Chen L, Oughtred R, Berman HM, Westbrook J (2004) TargetDB: a target registration database for structural genomics projects. Bioinformatics 20(16): 2860-2862

Cheung MS, Maguire ML, Stevens TJ, Broadhurst RW (2010) DANGLE: a Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure. J Magn Reson 202(2): 223-233

Chothia C (1976) The nature of the accessible and buried surfaces in proteins. J Mol Biol 105(1): 1-12

Cieplak M, Niewieczerzał S (2009) Hydrodynamic interactions in protein folding. J Chem Phys 130(12): 124906 - 124906

Cole C, Barber JD, Barton GJ (2008) The Jpred 3 secondary structure prediction server. Nucleic Acids Res 36(suppl 2): W197-W201

Crammer K, Singer Y (2001) On the algorithmic implementation of multiclass kernel-based vector machines. J Mach Learn Res 2: 265–292

Crammer K, Singer Y (2002) On the learnability and design of output codes for multiclass problems. Mach Learn 47(2-3): 201-233

Crick F (1970) Central dogma of molecular biology. Nature 227(5258): 561-563

Croy CH, Koeppe JR, Bergqvist S, Komives EA (2004) Allosteric changes in solvent accessibility observed in thrombin upon active site occupation. Biochemistry 43(18): 5246-5255

CS-ROSETTA test cases (May 15, 2016) Retrieved from BMRB website: https://csrosetta.bmrb.wisc.edu/csrosetta?page=bmrb_entries

Dobson CM (1999) Protein misfolding, evolution and disease. Trends Biochem Sci 24: 329–332

Drenth J (1994) Principles of Protein X-ray Crystallography. Springer-Verlag, New York, NY and London

Durrett R (2010) Probability: theory and examples. Vol 3 Cambridge University Press, London

Eddy SR (1998) Profile hidden markov models. Bioinformatics 14:755–763

Eghbalnia HR, Wang L, Bahrami A, Assadi A, Markley JL (2005) Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements. J Biomol NMR 32(1): 71-81

Eisenberg D, Weiss RM, Terwilliger TC (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci 81(1): 140-144

Eisenhaber F, Argos P (1993) Improved strategy in analytic surface calculation for molecular systems: handling of singularities and computational efficiency. J Comput Chem 14(11): 1272-1280

Engelman DM, Steitz TA, Goldman A (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biomol Struct 15(1): 321-353

Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B (2001) EVA: Continuous automatic evaluation of protein structure prediction servers. Bioinformatics 17: 1242–1243

Fesinmeyer RM, Hudson FM, Olsen KA, White GW, Euser A, Andersen NH (2005) Chemical shifts provide fold populations and register of β-hairpins and β- sheets. J Biomol NMR 33(4): 213-231

Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins: Struct, Funct, Bionf 23(4): 566-579

Garg A, Kaur H, Raghava GPS (2005) Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins: Struct, Funct, Bioinf 61(2): 318-324

Ginzinger SW, Fischer J (2006) SimShift: identifying structural similarities from NMR chemical shifts. Bioinformatics 22(4):460-465

Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, …, Orengo CA (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35:D291–D297

Gronenborn AM, Clore GM (1994) Identification of N-terminal helix capping boxes by means of 13C chemical shifts. J Biomol NMR 4(3): 455-458

Güntert P (2004) Automated NMR structure calculation with CYANA. Methods Mol Biol 278:353-378

Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 273(1):283-298

Hafsa NE, Wishart DS (2014) CSI 2.0: a significantly improved version of the Chemical Shift Index. J Biomol NMR 60:131-146

Hafsa NE, Arndt D, Wishart DS (2015a) CSI 3.0: a web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts. Nucleic Acids Res 43(W1): W370-W377

Hafsa NE, Arndt D, Wishart DS (2015b) Accessible surface area from NMR chemical shifts. J Biomol NMR 62(3): 387-401

Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50(1): 43-57

Hansen DF, Neudecker P, Vallurupalli P, Mulder FA, Kay LE (2009) Determination of Leu side-chain conformations in excited protein states by NMR relaxation dispersion. J Am Chem Soc, 132(1): 42-43

Hausrath AC (2006) A kinetic theory of tertiary contact formation coupled to the helix-coil transition in polypeptides. J Chem Phys 125(8): 084909- 084909

He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK (2009) Predicting intrinsic disorder in proteins: an overview. Cell Res 19(8): 929-949

Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, ..., Zhou Y (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Sci Rep 5:1476

Heinemann U, Illing G, Oschkinat H (2001) High-throughput three-dimensional protein structure determination. Curr Opin Biotechnol 12(4): 348-354

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89(22):10915-10919

Holbrook SR, Muskal SM, Kim SH (1990) Predicting surface exposure of amino acids from protein sequence. Protein Eng 3(8): 659-665

Holm L, Rosenström P (2010) Dali server: conservation mapping in 3D. Nucleic Acids Res 38(suppl 2):W545-W549

Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci 78(6): 3824-3828

Hung LH, Samudrala R (2003) Accurate and automated classification of protein secondary structure with PsiCSI. Protein Sci 12(2): 288-295

Hutchinson EG, Thornton JM (1994) A revised set of potentials for β-turn formation in proteins. Protein Sci 3(12): 2207-2216

Huyghues-Despointes BM, Langhorst U, Steyaert J, Pace CN, Scholtz JM (1999) Hydrogen-exchange stabilities of RNase T1 and variants with buried and solvent-exposed Ala→ Gly mutations in the helix. Biochemistry 38(50): 16481-16490

Ippolito JA, Alexander RS, Christianson DW (1990) Hydrogen bond stereochemistry in protein structure and function. J Mol Biol 215(3): 457-471

Janin J (1979) Surface and inside volumes in globular proteins. Nature 277: 491-492

Jardetzky O, and Roberts GCK (1981) NMR in molecular biology. Academic Press, New York

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292(2): 195-202

Jones DT, Tress M, Bryson K, Hadley C (1999) Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. Proteins Suppl 3: 104-111

Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358:86–89

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12): 2577-2637

Kadokura H, Katzen F, Beckwith J (2003) Protein disulfide bond formation in prokaryotes. Annu Rev Biochem 72(1): 111-135

Karatzoglou A, Meyer D, Hornik K (2006) Support Vector Machines in R. J Stat Softw 15(9): 1-28

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) kernlab-an S4 package for kernel methods in R. J Stat Softw 11: 1-20

Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14(10):846-856

Kemmink J, Darby NJ, Dijkstra K, Nilges M, Creighton TE (1996) Structure determination of the N-terminal thioredoxin-like domain of protein disulfide isomerase using multidimensional heteronuclear 13C/15N NMR spectroscopy. Biochemistry 35(24)*:* 7684-7691

KK, Lemak A, et al. (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci 105:4685-4690

Krogh A, Brown M, Mian IS, Sjölander K, Haussler D (1994) Hidden markov models in computational biology - Applications to protein modeling. J Mol Biol 235:1501–1531

Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28(5): 1-26

Kumar A, Ernst RR, Wüthrich K (1980) A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. Biochem Bioph Res Comm 95(1): 1-6

Kuszewsk J, Gronenborn AM, Clore GM (1995) The impact of direct refinement against proton chemical shifts on protein structure determination by NMR. J Magn Reson 107:293–297

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157(1): 105-132

Labudde D, Leitner D, Krüger M, Oschkinat H (2003) Prediction algorithm for amino acid types with their secondary structure in proteins (PLATON) using chemical shifts. J Biomol NMR 25(1): 41-53

Lavigne P, Willard L, Sykes BD, Bagu JR, Boyko R, Holmes CE (2000) Structure-based thermodynamic analysis of the dissociation of protein phosphatase-1 catalytic subunit and microcystin-LR docked complexes. Protein Sci 9(2): 252-264

Lee B, Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55(3): 379-400

Lesk AM (2001) Introduction to protein architecture: the structural biology of proteins. Oxford University Press, New York

Levitt M (1978) Conformational preferences for globular proteins. Biochemistry 17(20): 4277-4284

Li X, Pan XM (2001) New method for accurate prediction of solvent accessibility from protein sequence. Proteins: Struct, Funct, Bioinf 42(1): 1-5

London RE, Wingad BD, Mueller GA (2008) Dependence of amino acid side chain 13C shifts on dihedral angle: application to conformational analysis. J Am Chem Soc, 130(33): 11097-11105

Madhusudhan MS, Marti-Renom MA, Sanchez R, Sali A (2006) Variable gap penalty for protein sequence–structure alignment. Protein Eng Des Sel 19(3): 129-133

Manavalan P, Ponnuswamy PK (1978) Hydrophobic character of amino acid residues in globular proteins. Nature 275: 673-674

Marsh JA (2013) Buried and accessible surface area control intrinsic protein flexibility. J Mol Biol 425: 3250-3263

Mayrose I, Graur D, Ben-Tal N, Pupko T (2004) Comparison of site-specific rate-inference methods: Bayesian methods are superior. Mol Biol Evol 21: 1781-1791

McCallister EL, Alm E, Baker D (2000) Critical role of β-hairpin formation in protein G folding. Nat Struct Mol Biol 7(8): 669-673

Medek A, Hajduk PJ, Mack J, Fesik SW (2000) The use of differential chemical shifts for determining the binding site location and orientation of protein-bound ligands. J Am Chem Soc 122(6):1241-1242

Mielke SP, Krishnan VV (2003) Protein structural class identification directly from NMR spectra using averaged chemical shifts. Bioinformatics 19(16): 2054-2064

Mielke SP, Krishnan VV (2004) An evaluation of chemical shift index-based secondary structure determination in proteins: influence of random coil chemical shifts. J Biomol NMR 30(2): 143-153

Mielke SP, Krishnan VV (2009) Characterization of protein secondary structure from NMR chemical shifts. Prog Nucl Magn Reson Spectrosc 54(3-4): 141-165

Momen-Roknabadi A, Sadeghi M, Pezeshk H, Marashi SA (2008) Impact of residue accessible surface area on the prediction of protein secondary structures. BMC Bioinformatics 9(1): 357

Montgomerie S, Cruz JA, Shrivastava S, Arndt D, Berjanskii M, Wishart DS (2008) PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. Nucleic Acids Res 36(suppl 2):W202-W209

Montgomerie S, Sundraraj S, Gallin WJ, Wishart DS (2006) Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics 7: 301

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4):536-540

Myers JK, Nick PC, Martin SJ (1995) Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. Protein Sci 4(10): 2138-2148

Naderi-Manesh H, Sadeghi M, Arab S, Moosavi MAA (2001) Prediction of protein surface accessibility with information theory. Proteins: Struct, Funct, Bioinf 42(4): 452-459

Nguyen MN, Rajapakse JC (2005) Prediction of protein relative solvent accessibility with a two-stage SVM approach. Proteins: Struct, Funct, Bioinf 59(1): 30-37

Opitz CA, Kulke M, Leake MC, Neagoe C, Hinssen H, Hajjar RJ, Linke WA (2003) Damped elastic recoil of the titin spring in myofibrils of human myocardium. Proc Natl Acad Sci 100(22): 12688-12693

Ortiz AR, Strauss CE, Olmea O (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 11(11):2606 –2621

Ösapay K, Case DA (1991) A new analysis of proton chemical shifts in proteins. J Am Chem Soc 113(25):9436-9444

Ösapay K, Case DA (1994) Analysis of proton chemical shifts in regular secondary structure of proteins. J Biomol NMR 4:215-230

Ösapay K, Theriault Y, Wright PE, Case DA (1994) Solution structure of carbonmonoxy myoglobin determined from nuclear magnetic resonance distance and chemical shift constraints. J Mol Biol 244:183–197

Ozenne V, Bauer F, Salmon L, Huang JR, Jensen MR, Segard S, Blackledge M (2012) Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinf 28(11): 1463-1470

Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, DeBolt S, ..., Kollman P (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comput Phys Commun 91(1):1-41

Peng J, Xu J (2010) Low-homology protein threading. Bioinformatics 26(12):i294-i300

Perkins SJ, Johnson LN, Phillips DC, Dwek RA (1977) Conformational changes, dynamics and assignments in ^{1}H NMR studies of proteins using ring current calculations Hen egg white lysozyme. FEBS Lett 82(1): 17-22

Perkins SJ, Wüthrich K (1979) Ring current effects in the conformation dependent NMR chemical shifts of aliphatic protons in the basic pancreatic trypsin inhibitor. BBA-Protein Struct 576(2): 409-423

Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol 9(1): 51

Pimentel GC, McClellan AL (1971) Hydrogen bonding. Annu Rev Phys Chem *22*(1): 347-385

Pollastri G, Baldi P, Fariselli P, Casadio R (2002) Prediction of coordination number and relative solvent accessibility in proteins. Proteins: Struct, Funct, Bioinf 47(2): 142-153

Proctor WG, Yu FC (1950) The dependence of a nuclear magnetic resonance frequency upon chemical compound. Phys Rev 77(5): 717

R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, ISBN 3-900051-07-0, URL: http://www.R-project.org

Ratnaparkhi GS, Ramachandran S, Udgaonkar JB, Varadarajan R (1998) Discrepancies between the NMR and X-ray structures of uncomplexed barstar: analysis suggests that packing densities of protein structures determined by NMR are unreliable. Biochemistry 37(19): 6958-6966

Richards FM (1974) The interpretation of protein structures: total volume, group volume distributions and packing density. J Mol Biol 82(1): 1-14

Richards FM (1977) Areas, volumes, packing and protein structure. Annu Rev Biophys Bioeng 6: 151-176

Richards FM (1991) The protein folding problem. Sci Am 264(1): 54-63

Richardson JS (1981) The anatomy and taxonomy of protein structure. Adv Protein Chem 34:167-339

Ridgeway G (2007) Generalized Boosted Models: A guide to the gbm package. R package vignette, URL: http://CRAN.R-project.org/package=gbm

Rosato A, Vranken W, Fogh RH, Ragan TJ, Tejero R, Pederson K, Vuister GW (2015) The second round of critical assessment of automated structure determination of proteins by NMR: CASD-NMR-2013. J Biomol NMR 62(4):413-424

Rost B (1995) TOPITS: threading one-dimensional predictions into three-dimensional structures. Proc Int Conf Intell Syst Mol Biol 3:314-21

Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. Proteins: Struct, Funct, Bioinf 20(3): 216-226

Rost B, Sander C, Schneider R (1994) Redefining the goals of protein secondary structure prediction. J Mol Biol 235:13-26

Rost B, Schneider R, Sander C (1997) Protein fold recognition by prediction-based threading. J Mol Biol 270(3):471-480

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779-815

Schwarzinger S, Kroon GJ, Foss TR, Chung J, Wright PE, Dyson HJ (2001) Sequence-dependent correction of random coil NMR chemical shifts. J Am Chem Soc 123(13): 2970-2978

Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160(1):65-73

Serpa JJ, Makepeace KA, Borchers TH, Wishart DS, Petrotchenko EV, Borchers CH (2014) Using isotopically-coded hydrogen peroxide as a surface modification reagent for the structural characterization of prion protein aggregates. J Proteomics 100:160-166

Sharma D, Rajarathnam K (2000) 13C NMR chemical shifts can predict disulfide bond formation. J Biomol NMR 18(2): 165-171

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, ... , Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci, 105(12):4685-4690

Shen Y, Vernon R, Baker D, Bax A (2009a) De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 43(2): 63-78

Shen Y, Delaglio F, Cornilescu G, Bax A (2009b) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44(4): 213-223

Shen Y, Bryan PN, He Y, Orban J, Baker D, Bax A (2010) De novo structure generation using chemical shifts for proteins with high-sequence identity but different folds. Protein Sci 19(2): 349-356

Shen Y, Bax A (2012) Identification of helix capping and β-turn motifs from NMR chemical shifts. J Biomol NMR 52(3): 211-232

Shen Y, Bax A (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. J Biomol NMR, 56(3): 227-241

Shen Y, Bax A (2015) Homology modeling of larger proteins guided by chemical shifts. Nat Methods 12(8): 747-750

Siepen JA, Radford SE, Westhead DR (2003) β-edge strands in protein structure prediction and aggregation. Protein Sci 12:2348-2359

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, ..., Thompson JD (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7(1):539

Siew N, Elofsson A, Rychlewski L, Fischer D (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16(9):776-785

Singh H, Singh S, Raghava GP (2014) Evaluation of protein dihedral angle prediction methods. PloS one 9(8):5667

Smith JD, Clayton DA, Fields S, Hellinga HW, Kuriyan J, Levitt M, Taylor SS (2007) Report of the Protein Structure Initiative Assessment Panel. Retrieved from National Institute of General Medical Sciences website: https://www.nigms.nih.gov/News/reports/archivedreports2009-2007/Pages/PSIAssessmentPanel2007.aspx

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195-197

Söding J (2005) Protein homology detection by HMM–HMM comparison. Bioinformatics 21(7):951-960

Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244-W248

Söding J, Remmert M (2011) Protein sequence comparison and fold recognition: progress and good practice benchmarking. Curr Opin Struct Biol 21(3):404-411

Song Y, DiMaio F, Wang RYR, Kim D, Miles C, Brunette TJ, ..., Baker D (2013) High-resolution comparative modeling with RosettaCM. Structure 21(10):1735-1742

Steinbeck C (2004) Recent developments in automated structure elucidation of natural products. Nat Prod Rep 21(4): 512-518

Sternlicht H, Wilson D (1967) Magnetic Resonance Studies of Macromolecules. I. Aromatic-Methyl Interactions and Helical Structure Effects in Lysozyme. Biochemistry 6(9):2881-2892

Thompson MJ, Goldstein RA (1996) Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins: Struct, Funct, Genet 25(1):38-47

Tigelaar HL, Flygare WH (1972) Molecular zeeman effect in formamide and the α proton chemical shift in poly (L-alanine). J Am Chem Soc 94(2):343-346

Trevor H, Robert T, Friedman JJH (2001) The Elements of Statistical Learning. New York: Springer (Vol. 1)

Tyagi M, Bornot A, Offmann B, de Brevern AG (2009) Analysis of loop boundaries using different local structure assignment methods. Prot Sci 18(9):1869-1881

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin, ..., Markley JL (2008) BioMagResBank. Nucleic Acids Res 36(suppl 1):D402-D408

UniProt Consortium (2010) The universal protein resource (UniProt) in 2010. Nucleic Acids Res 38(suppl 1):D142-D148

Valdar WSJ (2002) Scoring residue conservation. Proteins: Struct, Funct, Bioinf 48(2): 227-241

Viera AJ, Garrett JM (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37(5):360-363

Voet D, Voet JG (1995) Biochemistry. J Wiley and Sons, New York

Vranken WF, Rieping W (2009) Relationship between chemical shift value and accessible surface area for all amino acid atoms. BMC Struct Biol 9(1):20

Wagner M, Adamczak R, Porollo A, Meller J (2005) Linear regression models for solvent accessibility prediction in proteins. J Comput Biol 12(3):355-369

Wang CC, Chen JH, Lai WC, Chuang WJ (2007) 2DCSi: identification of protein secondary structure and redox state using 2D cluster analysis of NMR chemical shifts. J Biomol NMR 38(1):57-63

Wang CC, Chen JH, Yin SH, Chuang WJ (2006) Predicting the redox state and secondary structure of cysteine residues in proteins using NMR chemical shifts. Proteins: Struct, Funct, Bionf 63(1): 219-226

Wang G, Dunbrack RLJ (2003) PISCES: a protein sequence culling server. Bioinformatics 19(12):1589-1591

Wang L, Eghbalnia HR, Markley JL (2007) Nearest-neighbor effects on backbone α and β carbon chemical shifts in proteins. J Biomol NMR 39(3): 247-257

Wang Y, Jardetzky O (2002a) Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 11(4):852-861

Wang Y, Jardetzky O (2002b) Investigation of the neighboring residue effects on protein chemical shifts. J Am Chem Soc 124(47):14075-14084

Webb B, Sali A (2014) Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics 5-6

Weston J, Watkins C (1998) Multi-class support vector machines. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London

Willard L, Ranjan A, Zhang H, Monzavi H, Boyko RF, Sykes BD, Wishart DS (2003) VADAR: a web server for quantitative evaluation of protein structure quality. Nucleic Acids Res 31(13): 3316-3319

Wishart DS (2011) Interpreting protein chemical shift data. Prog Nucl Magn Reson Spectrosc 58(1):62-87

Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res 36(W2): W496-W502

Wishart DS, Case DA (2001) Use of chemical shifts in macromolecular structure determination. Methods Enzymol 338:3-34

Wishart DS, Nip AM (1998) Protein chemical shift analysis: a practical guide. Biochem Cell Biol 76:153–163

Wishart DS, Sykes BD (1994a) Chemical shifts as a tool for structure determination. Methods Enzymol 239:363-392

Wishart DS, Sykes BD (1994b) The 13C chemical shift index: a simple method for the identification of protein secondary structure using 13C chemical shift data. J Biomol NMR 4(2): 171-180

Wishart DS, Sykes BD, Richards FM (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry 31(6): 1647-1651

Wittekind M, Mueller L (1993) HNCACB, a high-sensitivity 3D NMR experiment to correlate amide-proton and nitrogen resonances with the α-and β-carbon resonances in proteins. J Magn Reson 101(2): 201-205

Wüthrich K (1986) NMR of Proteins and Nucleic Acids. John Wiley and Sons, New York, NY

Wüthrich K (1990) Protein structure determination in solution by NMR spectroscopy. J Bio Chem 265(36): 22059-22062

Xu J, Zhang Y (2010) How significant is a protein structure similarity with TM-score= 0.5? Bioinformatics 26(7):889-895

Yang S, Valas R, Bourne PE (2009) Evolution studied using protein structure. Structural Bioinformatics, Wiley-Blackwell 561-573

Yuan Z, Huang B (2004) Prediction of protein accessible surface areas by support vector regression. Proteins: Struct, Funct, Bioinf 57(3): 558-564

Zemla A, Venclovas C, Fidelis K, Rost B (1999) A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. Proteins 34: 220-223

Zhang H, Neal S, Wishat DS (2003) RefDB: A database of uniformly referenced protein chemical shifts. J Biomol NMR 25: 173-195

Zhang W, Dunker AK, Zhou Y (2008) Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. Proteins: Struct, Funct, Bioinf 71(1): 61-67

Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57:702-710