

RICE UNIVERSITY

**Knowledge-Based Prediction of Chemical Shift
and Recognition of Protein Native Structure**

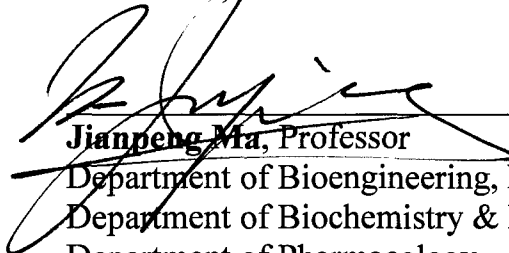
By

Zhao Ge

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

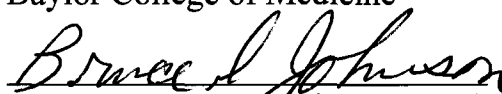
Master of Science

APPROVED, THESIS COMMITTEE:



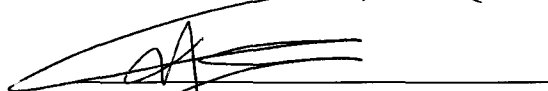
Jianpeng Ma, Professor

Department of Bioengineering, RICE University
Department of Biochemistry & Molecular Biology,
Department of Pharmacology
Baylor College of Medicine



Bruce R. Johnson, Distinguished Faculty Fellow

Department of Chemistry
Executive Director, Rice Quantum Institute



Yizhi Tao, Assistant Professor

Department of Biochemistry & Cell Biology
RICE University

HOUSTON, TEXAS

OCTOBER 2009

UMI Number: 1486032

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

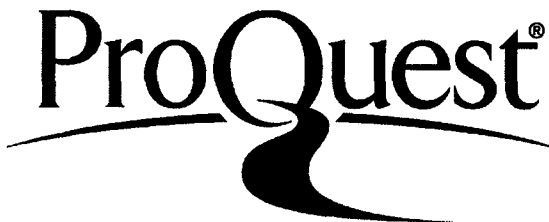
In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 1486032

Copyright 2010 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

This research work is sponsored and supported by a training fellowship from the Pharmacoinformatics Training Program of the Keck Center of Gulf Coast Consortia (NIH Grant No. R90 DK071505).

Keck Center of the Gulf Coast Consortia
Houston, Texas 77005

ABSTRACT

Knowledge-Based Prediction of Chemical Shift and Recognition of Protein Native Structure

by

Zhao Ge

We designed and implemented a suite of program which is able to accurately and automatically predict chemical shift of protein C-alpha nuclei on the simple basis of protein sequence and low-resolution C-alpha trace conformation. We applied this knowledge-based prediction approach on a group of C-alpha structures generated by computational modeling methods, and successfully identify the native structure by comparing the predicted and unassigned observed NMR data.

We begin the automatic prediction with construction of a knowledge-based protein structural profile library, which aims at capturing the most significant structural features affecting chemical shifts, even from a highly coarse-grained C-alpha model. The library is populated by more than 5000 non-homologous proteins, with publicly accessible structures from Protein Data Bank and more than 1.5 million pre-calculated chemical shifts by a widely used NMR predictive program SHIFTX. Fed with the minimum sequential and structural information, the program is able predict highly consistent chemical shifts comparing with experimental observed data from an NMR spectroscopy database BioMagResBank(BMRB). Overall, the proposed program achieves a correlation

coefficient of 0.937 and RMSD of 1.702 ppm towards observed chemical shifts. These results are slightly lower than those from achieved by the benchmark program SHIFTX, which utilizes semi-empirical hypersurfaces and semi-classical equations. On the same test sets, SHIFTX achieved a correlation coefficient of 0.945 and RMSD of 1.599 against experimental observations. In compensation, like most other predictive methods, SHIFTX requires high-resolution protein structures with three-dimensional all-atom coordinates, its accuracy of prediction will be highly compromised unless fed with all-atom high-resolution structure, which is normally exceedingly difficult to obtain. Combined with an optimization matching system using Monte Carlo method, we compared the predicted C-alpha chemical shifts with unassigned NMR data from BMRB, and successfully identify the native fold topology by the resemblance between two sets of chemical shifts.

In summary, the proposed program is one of the only methods which are capable to predict accurate chemical shifts, even on low-resolution C-alpha protein structures, which are far more accessible and readily obtained by currently available protein modeling methods. Based on the understanding that the similar pattern of chemical shifts reflects resemblance of two structures, we approved that prediction-recognition approach not only fundamentally improve the way of the NMR-assisted computational protein modeling, but is effective in accelerating the traditional protein structure determination and validation by NMR.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Dr. Jianpeng Ma, for being the best mentor I could ask for. Without his guidance, encouragement, support and correction, I wouldn't be here in Rice University to complete my degree and studies. Dr. Ma is the best scientist I have ever known, for his motivation, enthusiasm, hard work, and full of great ideas.

I would also like to thank my committee members - Dr. Bruce R. Johnson and Dr. Yizhi Tao, for providing me invaluable guidance and insight.

I would like thank Keck Center of Gulf Coast Consortia, and the training program of Pharmacoinformatics, which provide me with not only the three-year fellowship, but more importantly, the unique opportunity to continue my research in state-of-art edge of combined expertise from top institutions, professors and coworkers.

I would like to extent my thanks to previous and current members of the Ma group: Dr. Yifei Kong, Dr. Billy Poon, Dr. Yinghao Wu, Dr. Mingzhi Chen, Jun Shen, and Mingyang Lu, Jialin Li, Nasos Dousis, Brien Kirk, Chaoping Qin, Yuan Mei, Xiaorui Chen, Cheng Zhang for their thoughtful scientific discussion and help.

I would like to give my special thanks to my grand-parents. My grandpa taught me letters, mathematics and chess since I was 3; he is my first mentor into sciences and into the life. My Mom and Dad, who have been supporting and holding me whenever and wherever I am. I will say thanks to my dearest wife,

Kun Huang, who is always standing by me in my happiest time, who is always standing at my back at my hardest time.

Finally I will thank my life-time friend Milton Xiaomeng Yu and, Elaine Yiyang Xie, for their valuable discussion and friendship.

TABLE OF CONTENTS

Chapter	Page
1. Introduction and Background	1
1.1. Introduction	1
1.2. Background	4
1.3. Organization of Content.....	11
1.4. References.....	13
2. Accurate Prediction of Chemical Shift based on Protein C-alpha Conformation	21
2.1. Introduction.....	21
2.2. Materials and Methods.....	25
2.3. Results.....	33
2.4. Concluding Discussion	53
2.5. References.....	55
3. Determination of Protein Structure Assisted by Unassigned NMR Data	60
3.1. Introduction.....	60
3.2. Methods.....	64
3.3. Results.....	62
3.4. Concluding Discussion	74
3.5. References.....	75
4. Summary and Future Goals	92

LIST OF FIGURES

Figure	Page
Figure 2.1: Observed and predicted C-alpha chemical shifts of 1A6K	36
Figure 2.2: Observed and predicted C-alpha chemical shifts of 1BFK	37
Figure 2.3: Observed and predicted C-alpha chemical shifts of 1CEX	38
Figure 2.4: Observed and predicted C-alpha chemical shifts of 1CLL	39
Figure 2.5: Observed and predicted C-alpha chemical shifts of 1DMB	40
Figure 2.6: Observed and predicted C-alpha chemical shifts of 1HCB	41
Figure 2.7: Observed and predicted C-alpha chemical shifts of 1HFC	42
Figure 2.8: Observed and predicted C-alpha chemical shifts of 1HKA	43
Figure 2.9: Observed and predicted C-alpha chemical shifts of 1ONC	44
Figure 2.10: Observed and predicted C-alpha chemical shifts of 1RGE	45
Figure 2.11: Observed and predicted C-alpha chemical shifts of 1ROP	46
Figure 2.12: Observed and predicted C-alpha chemical shifts of 1RUV	47
Figure 2.13: Observed and predicted C-alpha chemical shifts of 1TOP	48
Figure 2.14: Observed and predicted C-alpha chemical shifts of 3LZT	49
Figure 2.15: Observed and predicted C-alpha chemical shifts of 4FGF	50
Figure 2.16: Observed and predicted C-alpha chemical shifts of 4I1B	51
Figure 2.17: Observed and predicted C-alpha chemical shifts of 5PTI	52
Figure 3.1: Calculation of agreement score with complete assignment of experimental NMR data	66
Figure 3.2: Alignment and calculation of agreement score with unassigned NMR data	67
Figure 3.3: Structure of Ubiquitin (PDB ID 1G6J, BMRB entry 5387)	68
Figure 3.4: Agreement between predicted and observed chemical shifts with complete assignment	69
Figure 3.5: Agreement between predicted and observed chemical shifts without assignment	70
Figure 3.6: Convergence of agreement score during alignment optimization with different initial assignment condition	72
Figure 3.7: Agreement scores predicted and observed chemical shifts at different number of optimization steps	73

LIST OF TABLES

Table	Page
Table 2.1: Comparison of chemical shift prediction methods	24
Table 2.2: PDB file format.....	27
Table 2.3: Elements of 30-dimensional structural profile	29
Table 2.4: Comparison of prediction from SHIFTX, C-alpha-based prediction and observed chemical shifts.....	33
Table 2.5: Correlation coefficient and RMSD among observed data, SHIFTX prediction and C-alpha-based prediction for 17 testing proteins	35

CHAPTER 1

Introduction and Background

1.1 Introduction

Proteins are important biological molecules that present and play essential roles in every biological process in all known biological organisms. For example, many proteins participate as enzymes that catalyze different biochemical reactions in cell. Immunoglobulin and antibodies are vitally involved in immune response of organism towards infections. Many other proteins are responsible to cell signal detection, processing and cell cycle. Still, there are many other so-called structural or mechanical proteins, which are capable of maintaining structure and mechanical functions of biological components¹⁻³.

The complex biological functionalities of protein result from their capability to fold into complicated conformations uniquely determined by their primary sequence¹. The folding of specific structure of one protein is also driven by various non-covalent interactions such as di-sulfur bonding, hydrogen bonding, hydrophobic packing, and Van der Waal's forces. In turn, the 3-D conformations of proteins are critical to their particular functionality within living organism. For instance, receptor proteins in cell signaling system must recognize particular target molecules. The study of proteins is aiming at the understanding of the fundamental interdependency among the sequence, the structure and the function⁴⁻¹⁰.

There are four hierarchical levels of protein structure. (1) The primary structure. All proteins are represented of unique sequences of combination of twenty common amino acids. The peptide chain is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. (2) The secondary structures are highly regular sub-structures, alpha helix, beta strand and sheet or irregular loop, which are defined by their patterns of hydrogen bonds between the main-chain peptide groups. (3) Tertiary structures of proteins are collective spatial arrangement of the secondary structures and (4) quaternary structure is the complex of several protein molecules or polypeptide chains, which usually called as protein subunits in this context and function as part of the larger assembly or protein complex.

***In silico* Protein Structure Prediction**

Due to the crucial relationship between structure and functionality, it is often essential to determine the three dimensional structure of proteins in order to understand their functions in a sense of molecular interactions. *In silico* protein structure prediction is to computationally determine the three-dimensional topology of proteins from primary sequence. The accurate modeling of protein structure, especially for those that are difficult to be accessed experimentally, will contribute to reveal their functional identity in important processes. On the other hand, computational studies often significantly improve the common experimental methods of structure determination, including NMR spectroscopy, X-ray crystallography, fiber diffraction, etc, which can produce information at

atomic resolution. Furthermore, the reverse problem of protein structure prediction is the sequence and structure designing of novel proteins, which is the ultimate challenge faced by therapeutic research and drug discovery. The major experimental and computational approaches for protein structure determination are introduced in the following paragraph.

1.2 Backgrounds

NMR in biological macro molecule structure determination

Nuclear Magnetic resonance is a phenomenon that magnetic nuclei, such as isotopic ^1H , ^{13}C , ^{15}N and ^{31}P , under an applied magnetic field, absorbing and radiating energy at a certain resonance frequency. The frequency depends on the strength of external magnetic field and a number of physical and chemical factors. The observation from NMR spectroscopy allows insight study on the quantum mechanical properties of atomic nucleus, in solution or in solid state. Furthermore, NMR is widely applied in medical imaging techniques, such as magnetic resonance imaging (MRI).

Nuclear magnetic resonance spectroscopy is unique technique among the others available for three-dimensional structure determination of biological macro molecules, such as proteins and nucleic acid at atomic resolution, since NMR data can be observed in solution. Most proteins maintain their structure and perform mechanical functions in organism fluids such as blood, and saliva. The solution condition such as temperature, pH and salt concentration of these physiological fluids can be accurately replicated in NMR experiments. In specific, NMR also excels X-ray crystallography in that it bypasses the routine of molecule crystallization.

Chemical shifts

The chemical shift is a numerical description of an atomic nucleus dependency of magnetic energy level on external magnetic field and electronic environments in molecules. The four most important and frequently studied nuclei are from hydrogen-1 (^1H), carbon-13 (^{13}C), nitrogen-15 (^{15}N), and phosphorus-31 (^{31}P). The chemical shift is also expressed as the variations of nuclear magnetic resonance frequencies of the same kind of nucleus, due to variations in the electron distribution.

Mathematically, chemical shift is usually defined as the ratio between the difference in precession frequency between two nuclei and the operating frequency of the magnet, and expressed by frequency in parts per million (ppm). The frequency is proportion to the strength of externally applied field, while ratio (chemical shift) is independent to it. With the increases of the applied field, the deviation of chemical shift changes significantly, which improves the resolution of NMR. Chemical Shifts are both important spectral indicators, and dependent upon complex electronic and geometric factors, therefore it potential provides rich resources of structural information. On the other hand, these sensitive dependencies make the interpretation and accurate prediction of chemical shifts extremely difficult. Great efforts have been extensively exerted during the past half century, to computationally predict of chemical shift based on resources, such as primary sequence, three-dimensional structure, all-atom coordinates, and through various algorithms, such as artificial neural networks, empirical potential functions or hyper-surfaces, classical calculations, statistical principle component analysis²⁰⁻³¹. These works will be covered in Chapter 2.

Other experimental tools

In comparison to NMR, X-ray crystallography is a suite of advanced experimental tools to investigate the arrangement of atoms within a crystal based on the fundamental understanding that X-rays can be deflected by the crystal in certain manner. X-ray crystallography begins with growing a pure crystal of the material whose structure is to be determined. A beam of x-rays is then passed through the crystal. The regular and repeating arrangement of atoms in the crystal gives rise to a complex pattern of spots, which originally were recorded on a receptor. The information about the positions of the atoms in the crystal is recorded originally in frequency domain. After a considerable amount of mathematical procedure, majorly the Fourier transformation, the experiment output will be transformed into space domain, and a map of electron densities can be calculated and displayed as contour maps resembling topographic maps in geography. The peaks in the electron density map correspond to the atomic positions in the molecule. From that map, a 3-D model of the molecule can be constructed. Biological X-ray crystallography is, to date, the most prolific discipline within the area of structural biology; out of the ~42000 protein structures solved, X-ray crystallography is responsible for ~36000, according to the Protein Data Bank (PDB).

Beside the NMR and X-ray, which potentially provide access to high and intermediate resolution three-dimensional structures of biological macromolecules. Other experimental methods, Cryo-electron microscopy (EM) and

small angle scattering (SAXS) are used to produce lower-resolution structural information in certain situations¹¹⁻¹⁴.

Computational approaches

Structural bioinformatics uses computational techniques and bioinformatics tools to model or mimic the 3D structures of bio-molecules⁸⁻¹⁰. In current stage of structural biology, the mission of tackling increasingly complicated cellular systems has led to a reality that structures of many bio-molecules, at least at early stages, can be obtained only at low to intermediate resolutions, therefore only incomplete structural information of these molecules can be obtained by experimental tools. Typical examples are seen in the measurements of cryo-electron microscopy (cryo-EM) and low resolution protein crystallography. One goal of the advanced structural bioinformatics methods is therefore to aid in interpretation of structural information at intermediate or higher resolutions. Moreover, it is a big challenge in structural biology that the conventional methods of building atomic model are not applicable to the intermediate resolution data. Therefore novel structural informatics tools are in high demand to bridge the missing link between the intermediate resolution structures and the conventional structural studies, which require at least atomic or C-alpha atom models.

Monte Carlo method is a widely used class of computational algorithms for simulating the behavior of various physical and biological systems. A Monte Carlo simulation attempts to overcome local energy barriers and find global low-

energy conformations⁶. Before the simulation begins, a set of conformational moves is selected. Beginning with the initial conformation, each subsequent conformation is mutated by a random move. If the change in energy ΔE is negative (i.e., the new conformation has a lower energy), the move is automatically accepted. If $\Delta E \geq 0$, the move will be accepted according to the Metropolis criterion. The simulation terminates when the ratio of accepted to attempted mutations (the acceptance ratio) falls below some threshold. The potential acceptance of higher energy conformations allows Monte Carlo simulations to overcome energy barriers and find globally low-energy conformations. Monte Carlo simulations are widely applied in modern computational biology in conformational sampling of bio-molecules, protein structure prediction and drug discovery.

We show in this thesis how these state-of-the-art computational methods can improve the study of protein structures in assistant of raw NMR data. The practical role of computational biophysics is now more important than ever. Because the output of community-wide efforts in structural genomics, typically by time-consuming and relatively expensive X-ray crystallography or traditional NMR spectroscopy, is lagging far behind the output of large-scale DNA sequencing efforts such as the Human Genome Project, computational modeling and prediction of protein structures can offer an efficient and fast alternative which will be very valuable to tasks as rational drug design. Furthermore, current structural biology methods, such as X-ray crystallography and Cryo-EM have a big limitation that they can only provide static structures of bio-molecules in

crystal state, while most biological events are dynamic processes in solution^{2,3},
On the contrary, NMR is capable of observing protein structures and dynamics in
solution, the vivid example of SAR-by-NMR in pharmaceutical researches. That
is why computational biophysics is so significant in extending the structural
information to fully understand the functional mechanism of biological targets. It
can not only capture the dynamic features, but sometimes also reveal the physics
and chemistry underneath. Moreover, it opens the door of biology for the well-
developed theoretical and computational methods in chemistry, physics,
mathematics and computer science, which greatly broadens the approaches to
understand the fast developing biological field.

1.3 Organization of Content

The content of this thesis is organized as follows. In Chapter 2, we describe a computational procedure for predicting chemical shift based on C-alpha trace conformation, generated by low or mediate resolution experimental methods or any computational protein folding approaches. As stated in the introduction, NMR served as one of the major experimental approaches in structure determination for biological macromolecules, suffers from the exceedingly difficult chemical shift interpretation. Therefore, how to derive the connectivity among numerical signals to structural signatures of nuclei becomes the final missing link. Based on a knowledge-based structural profile library and automated prediction protocol, we could rapidly and accurately calculated chemical shift from C-alpha conformation with a high consistency compared at observed data from experiment. More importantly, it was revealed that, despite the highly-coarse grained C-alpha conformation and lack of all-atom coordinates, which are usually necessary for calculation, our knowledge-based approach is able to capture most essential principle factors in the molecule that affecting chemical shift values. The result is verified by the high correlation coefficient and low root mean square deviations (RMSD) between theoretically calculation and experimental observation.

In Chapter 3, the result from chapter 2 is applied in recognition of protein native structure from a number of generated models, by incorporating raw data observed directly from conventional NMR experiment. Furthermore, a new Monte Carlo protocol is presented, to bypass the labor-intensive procedure of

manual assignment between signal and nuclei. This implementation fundamentally boosts this solution-phase method of NMR to play more and more promising role in nowadays structure biology. Meanwhile, this work takes great advantage of different scaled models generated by computational protein structure modeling approaches. The minimum requirement on model resolution basically avoids the inevitable fallacy and difficulty in modeling small atoms and long side chain from those approaches. In other word, our method combines both raw experimental data and inexpensive computational model, in successfully identifying correct protein structures.

Finally, in Chapter 4, the entire thesis work is summarized and some important issues for future investigation are discussed.

1.4 References

1. Anfinsen C (1972). "The formation and stabilization of protein structure". *Biochem. J.* 128 (4): 737-49.
2. Brooks III, C. L., Karplus, M., and Pettitt, B. M., *Adv. Chem. Phys.*, 71, 1 (1988).
3. McCammon, J. A., and Harvey, S., *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge (1987).
4. Bahar I, Atilgan A R and Erman B 1997 Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* 2 173-81.
5. Leach, A. R. (2001) *Molecular Modelling: Principles and Applications*. Prentice Hall, Harlow, England; New York, 2nd edn.
6. Beniac, Daniel R, Andonov, Anton, Grudeski, Elsie, et al. Architecture of the SARS coronavirus prefusion spike. *Nat Struct Mol Biol* 13 (8): 751-2 Aug 2006.
7. Chiu, W, Baker, ML, Jiang, W, et al. "Electron cryomicroscopy of biological machines at subnanometer resolution". *STRUCTURE* 13 (3): 363-372 MAR 2005.
8. Jiang, W., Baker, M. L., Ludtke, S. J. & Chiu, W. (2001). "Bridging the information gap: computational tools for intermediate resolution structure interpretation". *J. Mol. Biol.* 308, 1033-1044.

9. Kong, Y. & Ma, J. (2003). "A structural-informatics approach for mining b- sheets: locating sheets in intermediate-resolution density maps". *J. Mol. Biol.* 332, 399-413.
10. Choi, Jung Min, Kang, Sung Yun, Bae, Won Jin, et al. "Probing the roles of active site residues in the 3'-5' exonuclease of the werner syndrome protein". *J Biol Chem* 282 (13): 9941-51 Mar 2007.
11. Svergun, D.I., Petoukhov, M.V., and Koch, M.H. (2001). "Determination of domain structure of proteins from X-ray solution scattering". *Biophys. J.* 80, 2946-2953.
12. Svergun, D.I., and Koch, M.H. (2002). "Advances in structure analysis using small-angle scattering in solution". *Curr. Opin. Struct. Biol.* 12, 654-660.
13. Buki, A, Okonkwo, DO, Wang, KKW, et al. "Cytochrome c release and caspase activation in traumatic axonal injury". *J. NEUROSCI* 20 (8): 2825-2834 APR 15 2000.
14. Baker, D., Prediction and design of macromolecular structures and interactions. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 2006. **361**(1467): p. 459-463.
15. Bowers, P.M., C.E.M. Strauss, and D. Baker, De novo protein structure determination using sparse NMR data. *Journal of Biomolecular Nmr*, 2000. **18**(4): p. 311-318.
16. Kihara, D., et al., TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints.

- Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(18): p. 10125-10130.
17. Kim, D.E., D. Chivian, and D. Baker, Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, 2004. **32**: p. W526-W531.
 18. Li, W., Y. Zhang, and J. Skolnick, Application of sparse NMR restraints to large-scale protein structure prediction. *Biophysical Journal*, 2004. **87**(2): p. 1241-1248.
 19. Lu, H. and J. Skolnick, Application of statistical potentials to protein structure refinement from low resolution Ab initio models. *Biopolymers*, 2003. **70**(4): p. 575-584.
 20. Meiler, J. and D. Baker, Rapid protein fold determination using unassigned NMR data. *Proceedings of the National Academy of Sciences of the United States of America*, 2003. **100**(26): p. 15404-15409.
 21. Meiler, J. and D. Baker, The fumarate sensor DcuS: progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy. *Journal of Magnetic Resonance*, 2005. **173**(2): p. 310-316.
 22. Rohl, C.A., Protein structure estimation from minimal restraints using rosetta. *Nuclear Magnetic Resonance of Biological Macromolecules, Part C*, 2005. **394**: p. 244-260.

23. Rohl, C.A. and D. Baker, De novo determination of protein backbone structure from residual dipolar couplings using rosetta. *Journal of the American Chemical Society*, 2002. **124**(11): p. 2723-2729.
24. Skolnick, J., et al., TOUCHSTONE: A unified approach to protein structure prediction. *Proteins-Structure Function and Genetics*, 2003. **53**(6): p. 469-479.
25. Wishart, D.S., B.D. Sykes, and F.M. Richards, The Chemical-Shift Index - a Fast and Simple Method for the Assignment of Protein Secondary Structure through Nmr-Spectroscopy. *Biochemistry*, 1992. **31**(6): p. 1647-1651.
26. Zhang, Y., A. Kolinski, and J. Skolnick, TOUCHSTONE II: A new approach to *ab initio* protein structure prediction. *Biophysical Journal*, 2003. **85**(2): p. 1145-1164.
27. Vranken, W.F. and W. Rieping, Relationship between chemical shift value and accessible surface area for all amino acid atoms. *Bmc Structural Biology*, 2009. **9**: p. -.
28. He, X., B. Wang, and K.M. Merz, Protein NMR Chemical Shift Calculations Based on the Automated Fragmentation QM/MM Approach. *Journal of Physical Chemistry B*, 2009. **113**(30): p. 10380-10388.
29. Ginzinger, S.W. and M. Coles, SimShiftDB; local conformational restraints derived from chemical shift similarity searches on a large synthetic database. *Journal of Biomolecular Nmr*, 2009. **43**(3): p. 179-185.

30. Elyashberg, M.E., K.A. Blinov, and A.J. Williams, The application of empirical methods of C-13 NMR chemical shift prediction as a filter for determining possible relative stereochemistry. *Magnetic Resonance in Chemistry*, 2009. **47**(4): p. 333-341.
31. Blinov, K.A., et al., Development of a fast and accurate method of C-13 NMR chemical shift prediction. *Chemometrics and Intelligent Laboratory Systems*, 2009. **97**(1): p. 91-97.
32. Vila, J.A., et al., Predicting C-13(alpha) chemical shifts for validation of protein structures. *Journal of Biomolecular Nmr*, 2007. **38**(3): p. 221-235.
33. Shen, Y. and A. Bax, Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *Journal of Biomolecular Nmr*, 2007. **38**(4): p. 289-302.
34. Kaur, J. and A.S. Brar, An approach to predict the C-13 NMR chemical shifts of acrylonitrile copolymers using artificial neural network. *European Polymer Journal*, 2007. **43**(1): p. 156-163.
35. Williamson, M.P. and C.J. Craven, Automated protein structure calculation from NMR data. *Journal of Biomolecular Nmr*, 2009. **43**(3): p. 131-143.
36. Wang, J.B., et al., Determination of Multicomponent Protein Structures in Solution Using Global Orientation and Shape Restraints. *Journal of the American Chemical Society*, 2009. **131**(30): p. 10507-10515.
37. Szymczyzna, B.R., et al., Synergy of NMR, Computation, and X-Ray Crystallography for Structural Biology. *Structure*, 2009. **17**(4): p. 499-507.

38. Shen, Y., et al., De novo protein structure generation from incomplete chemical shift assignments. *Journal of Biomolecular Nmr*, 2009. **43**(2): p. 63-78.
39. Sakakibara, D., et al., Protein structure determination in living cells by in-cell NMR spectroscopy. *Nature*, 2009. **458**(7234): p. 102-U10.
40. Guntert, P., Automated structure determination from NMR spectra. *European Biophysics Journal with Biophysics Letters*, 2009. **38**(2): p. 129-143.
41. Donald, B.R. and J. Martin, Automated NMR assignment and protein structure determination using sparse dipolar coupling constraints. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 2009. **55**(2): p. 101-127.
42. Berjanskii, M., et al., GeNMR: a web server for rapid NMR-based protein structure determination. *Nucleic Acids Research*, 2009. **37**: p. W670-W677.
43. Vogeli, B., L.S. Yao, and A. Bax, Protein backbone motions viewed by intraresidue and sequential H-N-H-alpha residual dipolar couplings. *Journal of Biomolecular Nmr*, 2008. **41**(1): p. 17-28.
44. Vila, J.A. and H.A. Scheraga, Factors affecting the use of C-13(alpha) chemical shifts to determine, refine, and validate protein structures. *Proteins-Structure Function and Bioinformatics*, 2008. **71**(2): p. 641-654.
45. Vila, J.A., Y.A. Arnautova, and H.A. Scheraga, Use of C-13(alpha) chemical shifts for accurate determination of beta-sheet structures in

- solution. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(6): p. 1891-1896.
46. Vila, J.A., et al., Quantum chemical C-13(alpha) chemical shift calculations for protein NMR structure determination, refinement, and validation. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(38): p. 14389-14394.
47. Ulrich, E.L., et al., BioMagResBank. Nucleic Acids Research, 2008. **36**: p. D402-D408.
48. Shin, J., W. Lee, and W. Lee, Structural proteomics by NMR spectroscopy. Expert Review of Proteomics, 2008. **5**(4): p. 589-601.
49. Shen, Y., et al., Consistent blind protein structure generation from NMR chemical shift data. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(12): p. 4685-4690.
50. He, Y., et al., NMR structures of two designed proteins with high sequence identity but different fold and function. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(38): p. 14412-14417.
51. Rapp, C.S., et al., Prediction of protein loop geometries in solution. Proteins-Structure Function and Bioinformatics, 2007. **69**(1): p. 69-74.
52. Latek, D., D. Ekonomiuk, and A. Kolinski, Protein structure prediction: Combining de novo modeling with sparse experimental data. Journal of Computational Chemistry, 2007. **28**(10): p. 1668-1676.

53. Cavalli, A., et al., *Protein structure determination from NMR chemical shifts*. Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(23): p. 9615-9620.
54. Campbell, I.D. and B. Sheard, *Protein-Structure Determination by Nmr*. Trends in Biotechnology, 1987. **5**(11): p. 302-306.

CHAPTER 2.

Accurate Prediction of Chemical Shift Based on Protein C-alpha Conformation

2.1 Introduction

In the past half century, Nuclear Magnetic Resonance has been recognized as one of the most important techniques, in biological macro-molecule structure determination for its advantage in lower sample requirement than X-ray Crystallography and its ability to mimic the solution condition in which most proteins performs particular physiological functions. Chemical shift, is the most important observable markers from NMR spectroscopy, which is critical for the fact that it potentially provides essential three-dimensional structural information about proteins, such as inter-atom distances, side-chain orientations, secondary structure, and di-sulfur bonding. However, these multiple dependencies upon geometric and electronic factors make both the interpretation and prediction exceedingly difficult.

Despite the fast development of instrumentation and software updates for NMR data processing in the past decade, the signal sequential assignment and structural calculation based on distance constraints are time-consuming, human effort intensive, and hence still severe prevention to high-throughput and accurate structure determination. Meanwhile, lots of efforts have been devoted in the prediction of chemical shift based on empirical data of known protein sequence

and structure¹²⁻²⁹. Accurate chemical shift prediction will not only dramatically accelerate the signal assignment procedure, but also benefit the verification of the proposed structures. SHIFTX, SHIFTY, PROSHIFT and CheShift are four current major prediction methods, though designed to use various approaches and different input data. SHIFTY predicts protein ¹H, ¹³C, and ¹⁵N chemical shifts on the basis of sequence homology, and requires solely the amino acid sequence for query protein. The algorithm utilized dynamic programming to detect sequence homologies between query sequence and hundreds of previously assigned proteins in the BioMagResBank¹. For given amino acid and atom type, the SHIFTY calculated the averaged chemical shift value over the similar sequence in the existing library. The accuracy of SHIFTY is fundamentally restricted by its over-simplified resource and the lack of three-dimensional conformational knowledge. PROSHIFT²¹ by Jen Meiler trained an artificial neural network (ANN) to predict the ¹H, ¹³C, and ¹⁵N using all-atom three-dimensional protein structure as well as the experimental conditions, totally 350 input units are fed into this three-layer fuzzy logic network, including the parameters describing the atom in focus as well as its spatial and covalent neighbors. PROSHIFT achieves the root mean square deviations of 0.3 ppm, 1.3 ppm, and 2.6 ppm for hydrogen, carbon, nitrogen chemical shifts respectively on test set. *CheShift*^{30,31} has been developed to predict ¹³C^α chemical shifts of protein structures. It is based on the generation of 696,916 conformations as a function of the ϕ , ψ , ω , χ_1 and χ_2 torsional angles for all 20 naturally occurring amino acids. Their ¹³C-alpha chemical shifts were computed at the DFT level of theory with a small basis set

and extrapolated, with an empirically-determined linear regression formula, to reproduce the values obtained with a larger basis set. Last, SHIFTX²⁰ is a hybrid predictive approach that employs empirically derived chemical shift hyper surfaces in combination with classical equations (for ring current, electric field, hydrogen bond and solvent effects) to calculate ¹H, ¹³C, and ¹⁵N chemical shifts from the coordinates for both backbone and side chain atoms. The chemical shift hyper surfaces contain dihedral angle, side chain orientation, secondary structure, and nearest neighbor effects that cannot be explicitly translated to analytical formulae. SHIFTX is acknowledged as one of the most accurate approaches in existence.

In contrast to all existing methods, our proposed approach reported in this chapter uses protein amino acid sequence and the coordinates for C-alpha atoms as only input, and predicts C-alpha chemical shift. We begin the prediction with the construction of C-alpha based structural profile library. The library consists of structural information of 5014 non-homologous proteins combined with pre-calculated chemical shifts from SHIFTX. By searching against this knowledge-based library, this program is able to simulate the C-alpha chemical shift from a low-resolution C-alpha trace. The result is nearly equally accurate as those calculated by SHIFTX, with much higher requirement of all-atom coordinates of the protein. In consideration of the importance of protein global structural properties over atomic-level details in many biological issues, and a general difficulties in obtaining the fine high-resolution structures, our interest is more

focused on the predictive ability of low-resolution models or C-alpha conformations rather than high-resolution, all-atomic structure.

Method	Requirements	Algorithm	Results
SHIFTY	Sequence	Sequence homology	Fairly
PROSHIFT	All-atom 3D coordinates	Artificial neural network	Moderate Accurate
CheShift	All-atom 3D coordinates	Quantum Mechanics Calculation	Accurate
SHIFTX	All-atom 3D coordinates	Semi-empirical semi-classical calculation	Accurate
Structural Profile Library Prediction	Sequence and C-alpha trace	Knowledge-based	Accurate

Table 2.1 Comparison of chemical shift prediction methods

2.2 Materials and Methods

Construction of Structural Profile Library

To construct the knowledge-based structural profile library, we select 5,014 non-homologous protein entries from Protein Data Bank (<http://www.rcsb.org>) using PISCES program, in purpose to reduce sequence redundancy and maintain maximum structural diversity. The full atomic coordinate files (.pdb) are downloaded from PDB server and screened through the culling procedures, with criteria:

- a) Sequence percentage identity $\leq 25\%$;
- b) Structure resolution $< 3.0 \text{ \AA}$;
- c) R-factor < 0.3 ;
- d) Individual chain length > 20 ;
- e) Non X-ray crystallography entries excluded;
- f) Entries with C-alpha only structures excluded.

In order to benchmark the prediction, 804 entries are randomly chosen into a “testing pool”, the rest of entries are separated in the “training pool”. Therefore, the prediction of any proteins in “testing pool” is generated with the knowledge solely from “training pool”. The excluded entries will reenter the training pool for general prediction purpose.

Profile Reader

A profile reader program written in PERL will scan the input pdb file for each protein, and generated the C-alpha level structural profile library in the next three steps.

Step 1: Only C-alpha coordinates are identified and recorded in a new “C-alpha trace” file, all other information about backbone Nitrogen, Oxygen, Hydrogen and atoms and all side chain particles such as C-beta, C-gamma and so on are ignored. Certain computational protein folding programs can generate only C-alpha trace coordinates or backbone model, with low confidence in side chain orientations. Except for the lack of side chain and main chain details, these “C-alpha trace” files follow PDB file format (ATOM Coordinate Section) strictly. Each record starts with “ATOM”, followed by its coordinates and sequential information, a sample record is expressed as

```
ATOM      10  CA  ARG A   2      63.313  35.100  82.885
1.00 51.84      A
```

Name	Description	Example
Record Name	Record header indicating the content	ATOM
Serial ID	Atom serial number.	10
Atom Name	Atom role name, such as "CA", "CB", "CG1", "N", "NH1", "O", "OD1"...	CA
Residue Name	Residue name in 3-letter Amino acid abbreviation, such as "ALA", "MET"	ARG
Chain ID	Usually starting from "A", "B" for multiple chain proteins, may be empty for single chain entries.	A
Sequential Number	Residue sequential number,	2
X	Atom X coordinate	63.313
Y	Atom Y coordinate	35.100
Z	Atom Z coordinate	82.885
Occupancy	Atom occupancy	1.00
Temperature Factor	Temperature factor or B value, Considering the atom position within a protein, and the interactions and forces it experiences, this factor describes the relative degree of freedom of one atom movement.	51.84
Record ID	Record identification field	A

Table 2.2 PDB file format

Step 2: Totally 1,475,237 residues are scanned and stored in library. The three-dimensional C-alpha trace by each protein is reconstructed by the profile reader, in order to calculate the geometric distances of covalent bond for each C-alpha atom and in-space interactions. The sequential order, amino acid types, and major properties for each C-alpha residue are summarized into a 30-dimensional vector, as described in table 3.

We use three adjacent amino acid residues, or tri-peptide, as basic storage unit. The amino acid types and secondary structure types for the tri-peptide are the most principle components in the 30 dimensional profiles. They are used primarily in identifying structural unit with similar electronic, magnetic and geometric properties, while they often share close chemical shift. The next 23 components in the profile numerically describe the constitution of neighborhood with the center C-alpha atom, and possible in space-interactions and forces it may experience. The last component is the angle among by the three residue C-alpha atoms, which is essential in determining the secondary structure identity. The last 24 components contribute to the sensitive deviation in chemical shift for tripeptides within identical secondary structure environment.

Index	Description	Elements in profile	Explanation	Format
1-3	The amino acids type of and center residue (i th) and its adjacent residues ($i-1$ th and $i+1$ th)	AA_i, AA_{i-1}, AA_{i+1}	Derived from given sequence, one-letter 20 common amino acid type	3 alphabets (A-Y)
4-6	Secondary structure of center residue (i th) and its adjacent residues	SS_{i-1}, SS_i, SS_{i+1}	Calculated and defined from given C-alpha coordinates. (A = helix, B = strand, T = turn)	3 alphabets (A,B,T)
7-26	The numbers of the twenty different amino acids located within the sphere with a radius of 15 Å	$neibor_j, j = 1...20$	1 = Ala, 2 = Cys, ... 19 = Trp, 20 = Tyr	20 integers
27-29	The numbers of the neighbors within the sphere having their secondary structures as α -helix, β -strand, or turn	$neibor_ss_A, neibor_ss_B, neibor_ss_T$	First number stands for number in helix, second number for strand, and third number for turn	3 integers
30	backbone angle among three adjacent C-alpha atoms	$Ca_{i-1} Ca_i Ca_{i+1}$	0-360 (degree)	1 real number

Table 2.3 Elements of 30-dimensional structural profile

Step 3: Parallel to the profile reading, the original PDB file with all-atom coordinates for each protein is fed to SHIFTX server, by which a theoretical C-alpha chemical shift values is calculated. The library is then accomplished with the one and half million records with unique structural profiles and a registered chemical shifts correspondingly Grouped by the index 1-3 in alphabet order, the library is designed for fast search and compare during the prediction.

Chemical shift prediction based on C-alpha trace conformation

The prediction of chemical shift for any given C-alpha model is performed “weighted profile matching” system. The target C-alpha trace file, (generated by any protein structure modeling program, or simplified from all-atom coordinate pdb file), is first scanned by profile reader similar to the procedure used in preparing the structural profile library. Each target C-alpha atom is registered with a specific 30-dimensional profile, called target profile.

The types of three amino acid residues and their secondary structure environment (first six components) of target profile are used as *keywords* in searching against entire library. Any profile with identical *keywords* is marked as “matched files”, and their pre-calculated chemical shifts will contribute the prediction in certain degree according to the structural resemblance between the matched profile and the target profile. This contribution system is expressed explicitly as

$$CS_{prediction} = \frac{\sum_{matched} CS_i \times weight_i}{\sum_{matched} weight_i}$$

The predicted value of target C-alpha $CS_{prediction}$ is defined as a weighted average of all the CS_i . CS_i is the pre-calculated chemical shift registered to the i th matched profile located in the library. The weight for specific profile is determined by its resemblance to the target profile and is defined as

$$weight_i = \left(const_1 \times (ratio \times \vec{v}_i \bullet \vec{v}_{target})^{-1} + const_2 \times \sum_{k=A,B,T} Diff_neighbor_k + const_3 \times Diff_CaCaCa \right)^{-1}$$

$$\text{where } ratio = \frac{|\vec{v}_{target}|}{|\vec{v}_i|} \text{ or } ratio = \frac{|\vec{v}_i|}{|\vec{v}_{target}|} \text{ and } ratio \leq 1.0$$

The weighted average scheme is embedded in the fact that the resemble structure profiles reflects a similar environment which in turn yield close chemical shifts.

The higher the resemblance of the 30 components between two profiles, the larger is the weight to reflect the major contribution from a closer chemical shift value.

Specifically in the equation, the first element describes the deviation of neighborhood constitutions between matched profile and target, both the crowdedness of the neighborhood and identities of neighbor residues are taken into consider by the dot product of two vectors. For example, residue A and B share

same number of neighbor residues, but one is constituted of most highly hydrophobic, the other is of most hydrophilic. The neighborhood information reflects the reverse property of A and B, which may produce significant difference in prediction chemical shift. In another case, residue C and D may have similar properties and share the neighbors of similar amino acid types. However, C is buried inside a protein, while D is exposed on the surface and may have much fewer neighbors around, which may as well results in serious change of chemical shift. The second and third components in the equation are designed to reflect the similarity secondary structure environments and topological configuration. The constant parameter for each component in the equation is individually optimized by experience and testing.

$$\text{const}_1 = 50$$

$$\text{const}_2 = 0.1$$

$$\text{const}_3 = 1.0$$

2.3 Results

Table 3 shows the comparison between the predictions of the C-alpha library method and all-atom SHIFTX on the test set of 806 protein chains. Totally 87,025 C-alpha chemical shifts are predicted, among which 218 “outliners” were found. The outliners have out-ranged experimental data from BMRB server. For example, C-alpha chemical shift was smaller than 10 ppm, which is far from the average and obviously experimental a typo from the BMRB database. We also recalculate the correlation coefficient for the clean set of 86,807 residues by excluding the outliners.

Data set		Between BMRB and SHIFTX	Between BMRB and C-alpha based prediction	Between SHIFTX and C-alpha based prediction
Test set with 87,025 residues	Correlation	0.9306	0.9231	0.9604
	Coefficient			
	RMSD (Å)	1.8047	1.8897	1.3148
Clean set with 86,807 residues	Correlation	0.9450	0.9369	0.9604
	Coefficient			
	RMSD (Å)	1.5985	1.7018	1.3141

Table 2.4 Comparison of prediction from SHIFTX, C-alpha-based prediction and observed chemical shifts

In order to closely investigate the performance of c-alpha based prediction, we choose the same testing set as those used in verification of SHIFTX. Among 38 proteins structure mentioned in SHIFTX works, only 17 have both atomic resolution structure and experiment observed chemical shift publicly accessible from PDB and BMRB. It was claimed by SHIFTX authors that they achieved a correlation coefficient of 0.98 and RMSD of 0.98 ppm of C-alpha chemical shift against experimental data. We will show in Table 5 the comparison of correlation coefficient and RMSD from SHIFTX and proposed approach for 17 individual proteins, ranging from 49 to 340 amino acid residues in size.

Specifically, for two sets of data with same number (n) of elements,

$$X = (x_1, x_2, \dots, x_n) \text{ and } Y = (y_1, y_2, \dots, y_n)$$

the correlation coefficient is calculated as

$$\text{corr}(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}$$

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

And root mean square deviation is calculated as

$$\text{RMSD}(X, Y) = \sqrt{E((X - Y)^2)} = \sqrt{\frac{\sum_{i=1}^n (x_i - y_i)^2}{n}}$$

PDB ID/BMRB Accession	Number of Residues	Resolution (Å)	BMRB vs. SHIFTX	BMRB vs. C-alpha based Prediction
			Correlation Coefficient / RMSD(ppm)	Correlation Coefficient / RMSD(ppm)
1A6K/4061	139	1.10	0.97036/1.09797	0.95616/1.28586
1BKF/4077	100	1.60	0.97722/1.08713	0.94755/1.56961
1CEX/4101	188	1.00	0.97922/1.07413	0.95159/1.61877
1CLL/547	137	1.70	0.96511/1.26681	0.96198/1.27295
1DMB/7354	340	1.80	0.97837/1.01397	0.94483/1.56025
1HCB/4022	241	1.60	0.97285/1.03273	0.93638/1.52448
1HFC/4064	146	1.56	0.96442/1.23993	0.93878/1.61688
1HKA/4299	119	1.50	0.97373/1.08903	0.94649/1.52207
1ONC/4371	89	1.70	0.96579/1.0995	0.91952/1.59997
1RGE/4259	90	1.15	0.97656/1.14588	0.95165/1.51193
1ROP/4072	49	1.70	0.96673/0.81666	0.94378/1.05236
1RUV/4031	109	1.30	0.95218/1.07515	0.89290/1.59904
1TOP/4401	88	1.78	0.96515/1.23703	0.95253/1.40106
3LZT/4562	114	0.92	0.96857/1.17786	0.92067/1.76936
4FGF/4091	112	1.60	0.97453/1.01452	0.94361/1.47203
4I1B/1061	144	2.00	0.95309/1.21731	0.92589/1.46858
5PTI/46	49	1.00	0.97698/1.09177	0.95329/1.51675

Table 2.5 Correlation coefficient and RMSD among observed data, SHIFTX prediction

and C-alpha-based prediction for 17 testing proteins

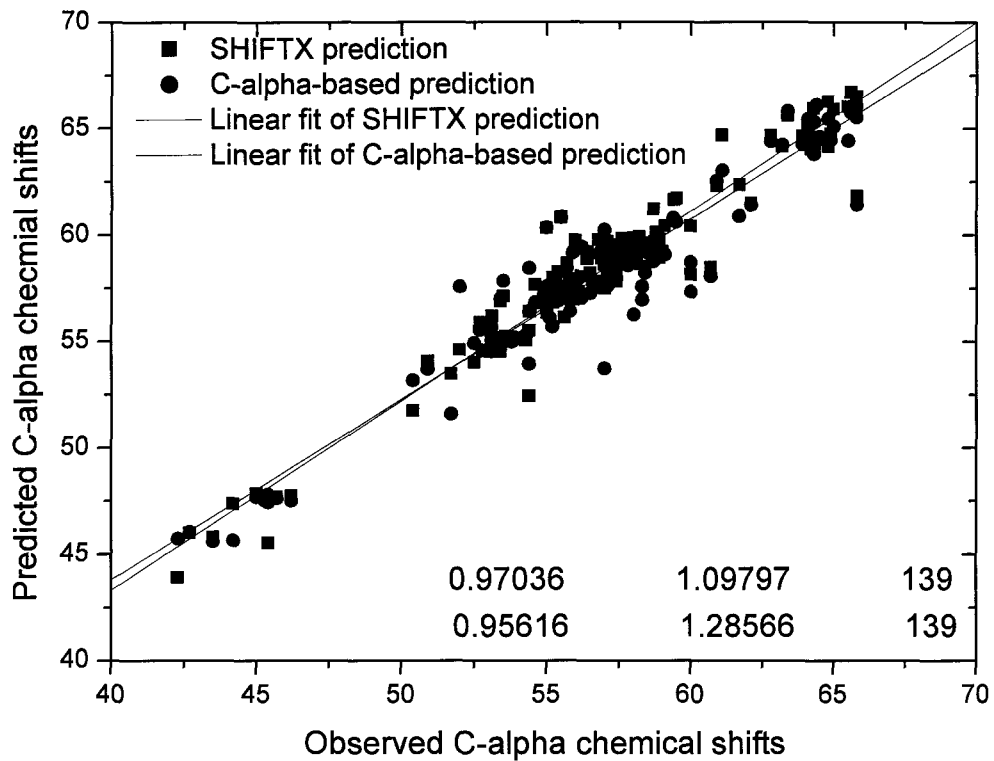


Figure 2.1 Comparison of observed and predicted C-alpha chemical shifts of 1AK6

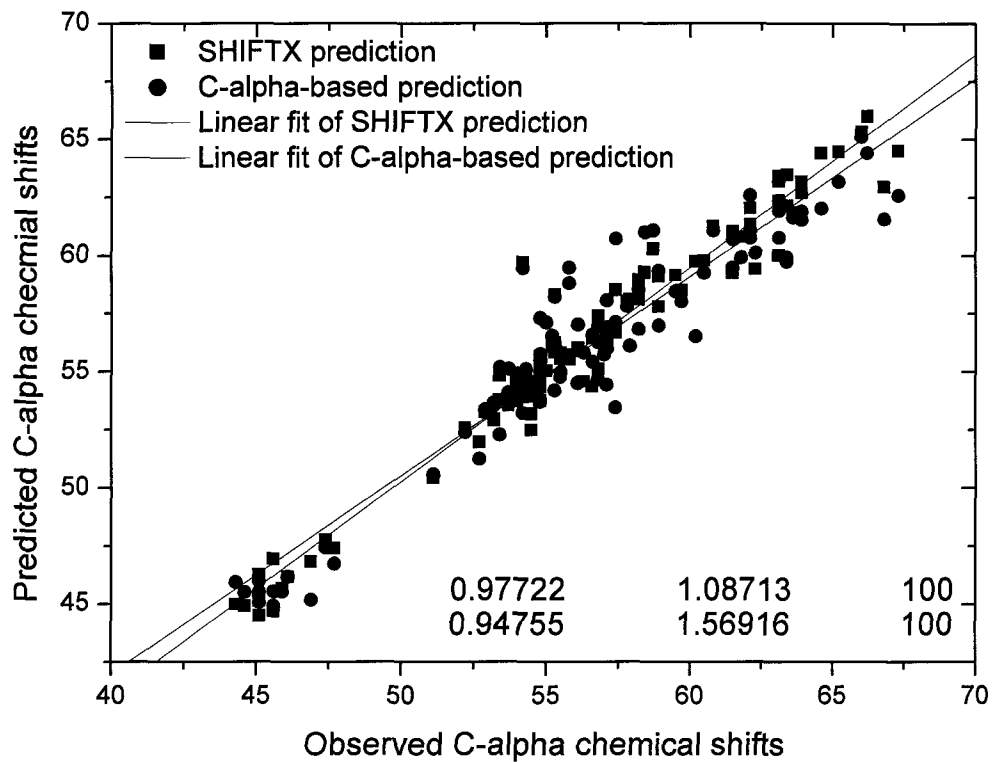


Figure 2.2 Comparison of observed and predicted C-alpha chemical shifts of 1BFK

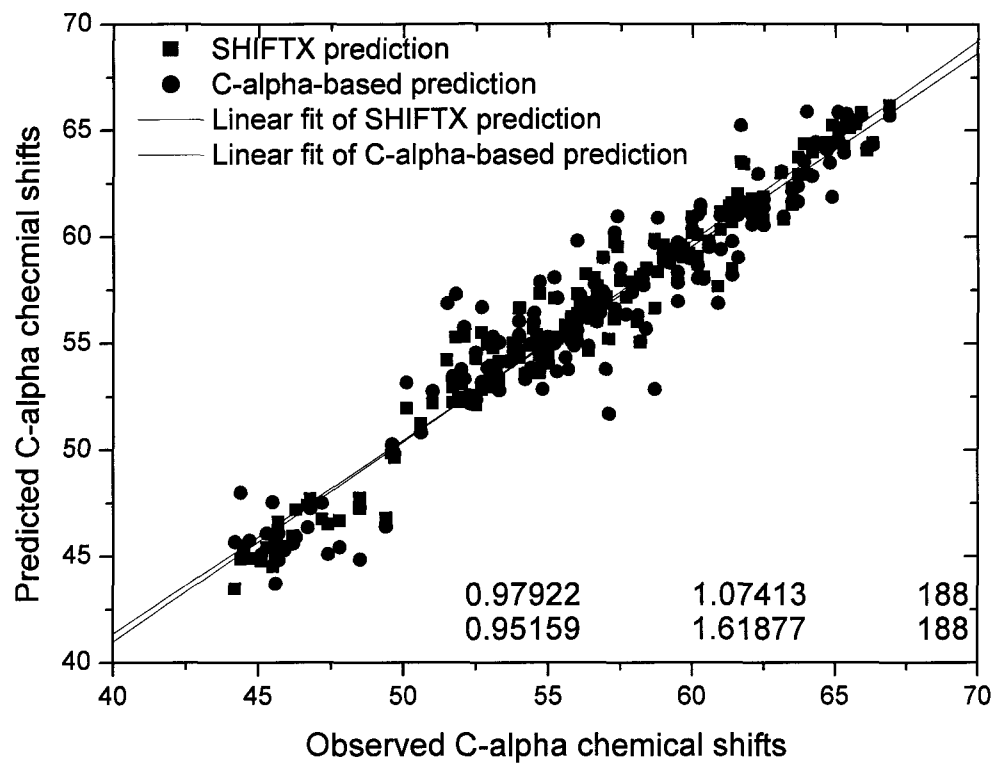


Figure 2.3 Comparison of observed and predicted C-alpha chemical shifts of 1CEX

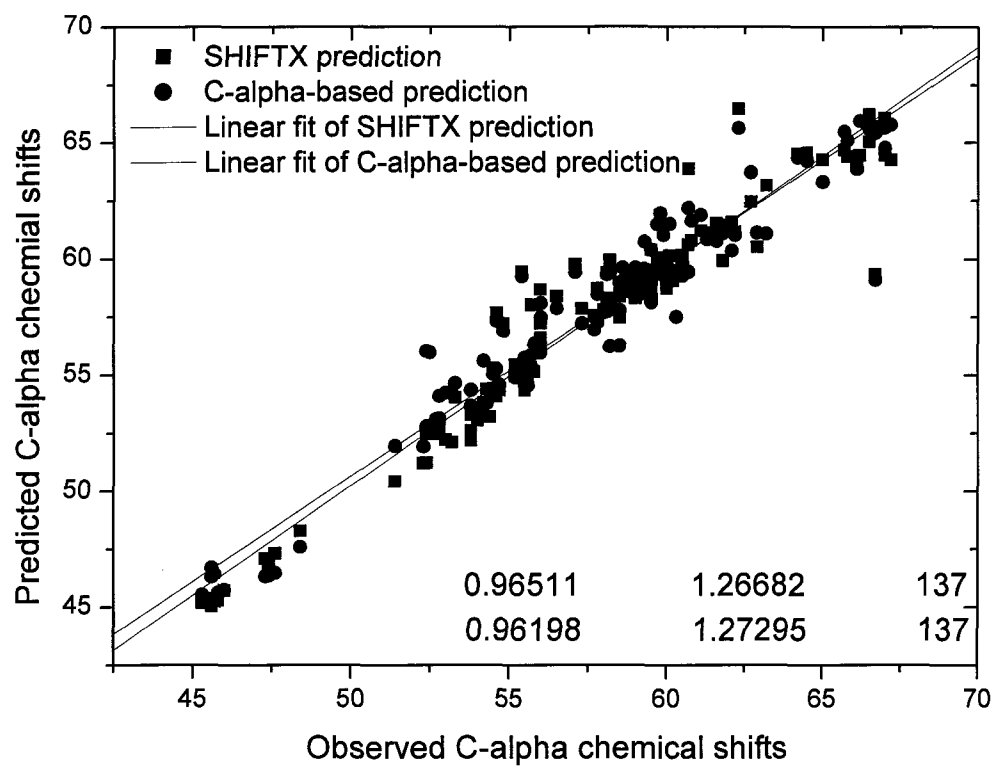


Figure 2.4 Comparison of observed and predicted C-alpha chemical shifts of 1CLL

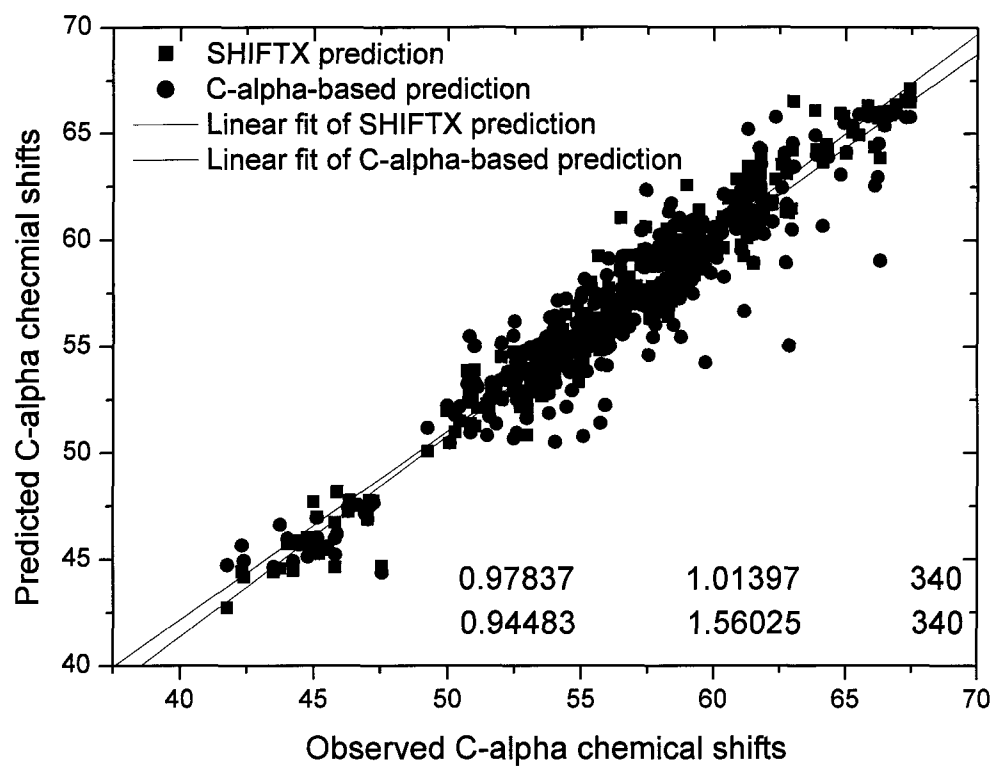


Figure 2.5 Comparison of observed and predicted C-alpha chemical shifts of 1DMB

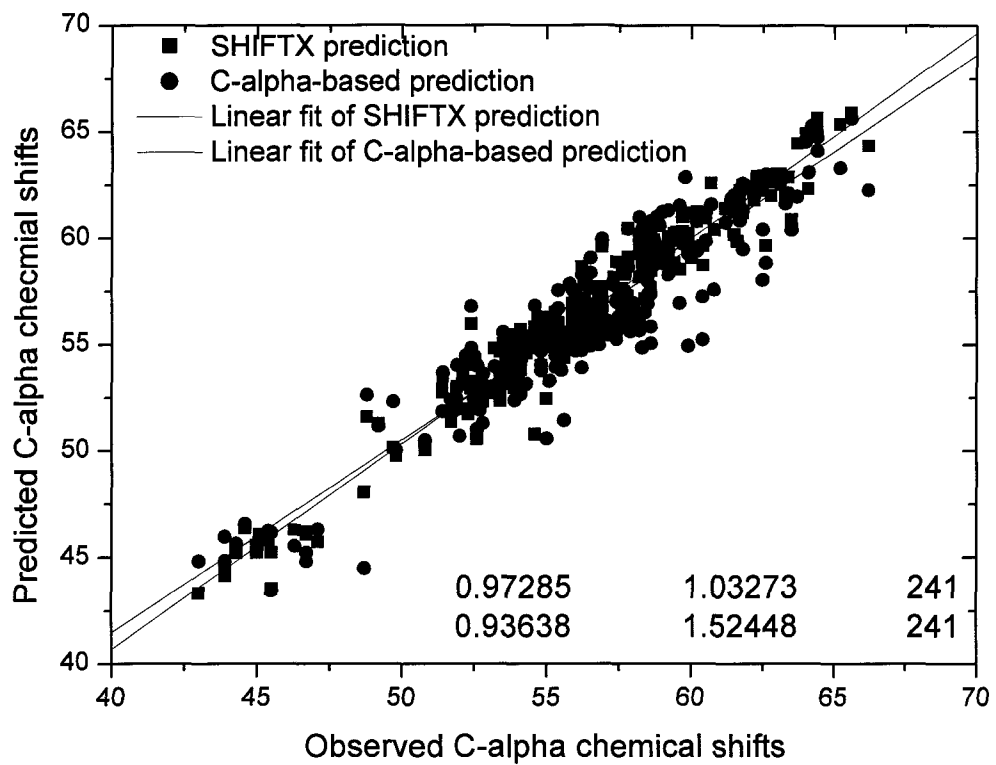


Figure 2.6 Comparison of observed and predicted C-alpha chemical shifts of 1HCB

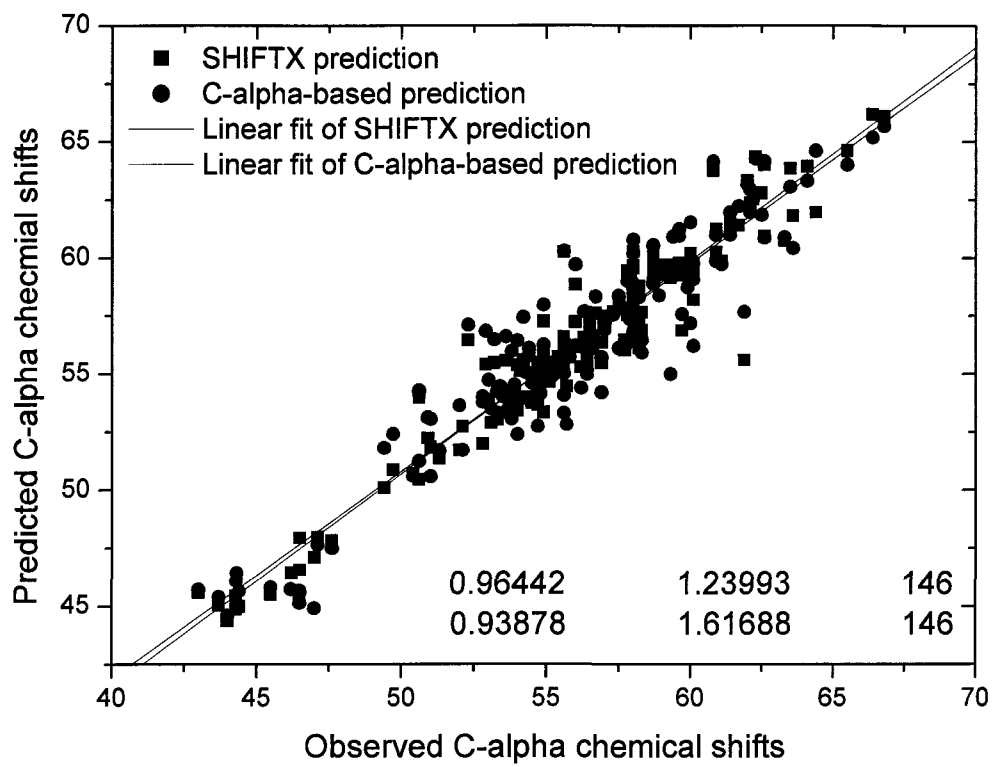


Figure 2.7 Comparison of observed and predicted C-alpha chemical shifts of 1HFC

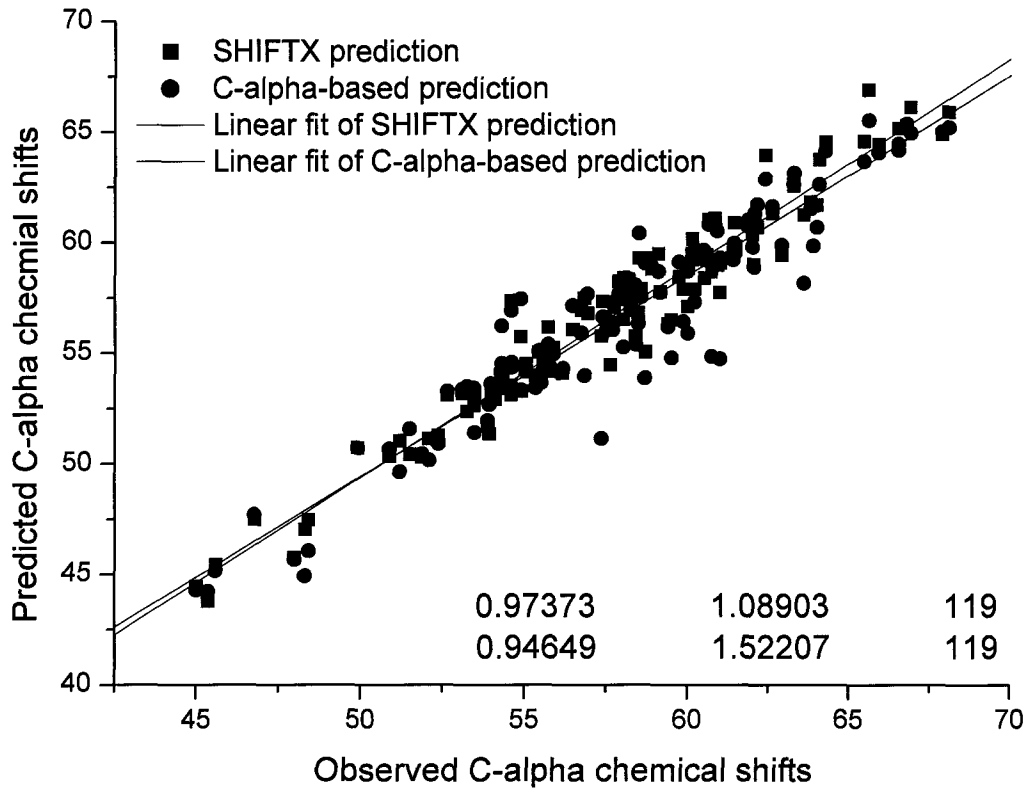


Figure 2.8 Comparison of observed and predicted C-alpha chemical shifts of 1HKA

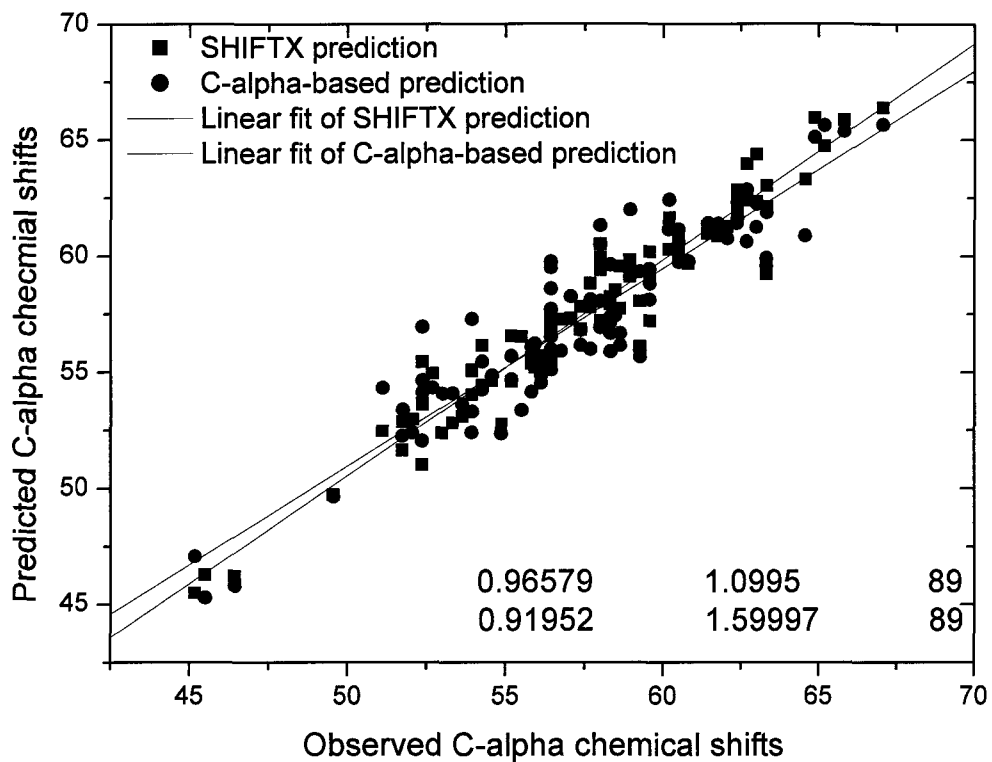


Figure 2.9 Comparison of observed and predicted C-alpha chemical shifts of 10NC

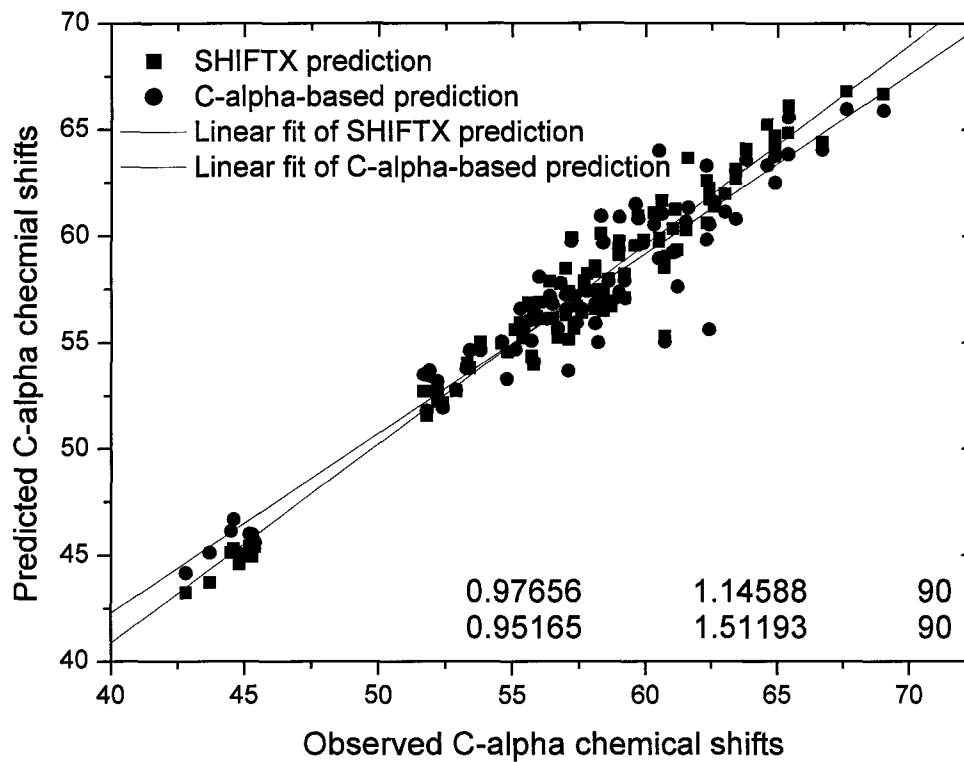


Figure 2.10 Comparison of observed and predicted C-alpha chemical shifts of 1RGE

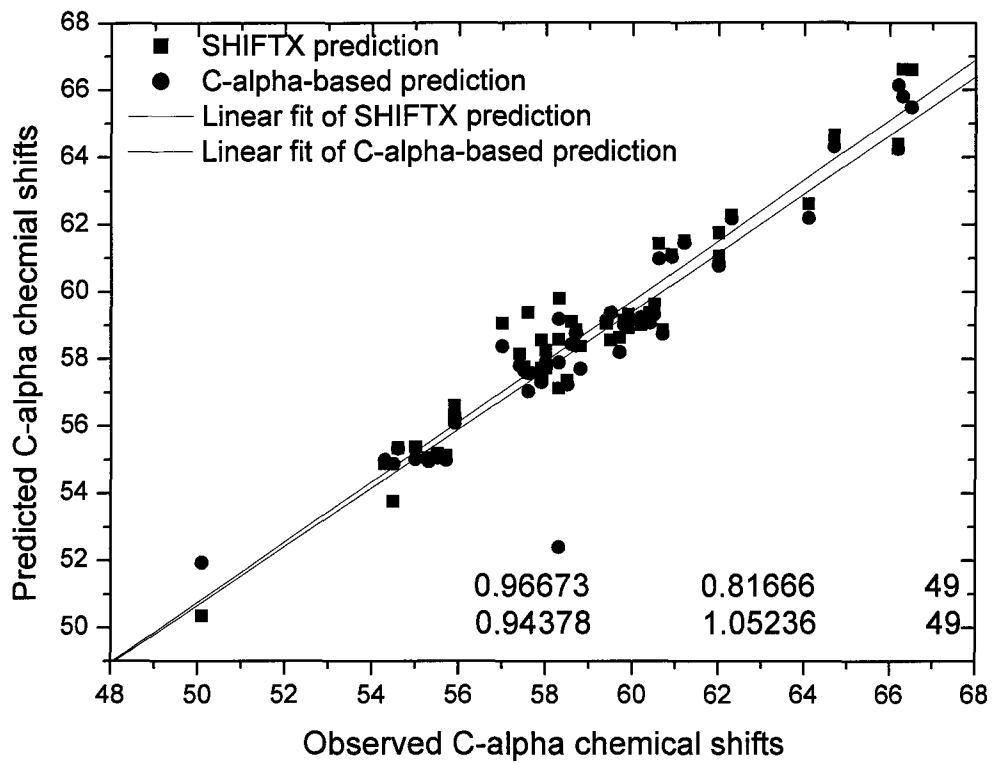


Figure 2.11 Comparison of observed and predicted C-alpha chemical shifts of 1ROP

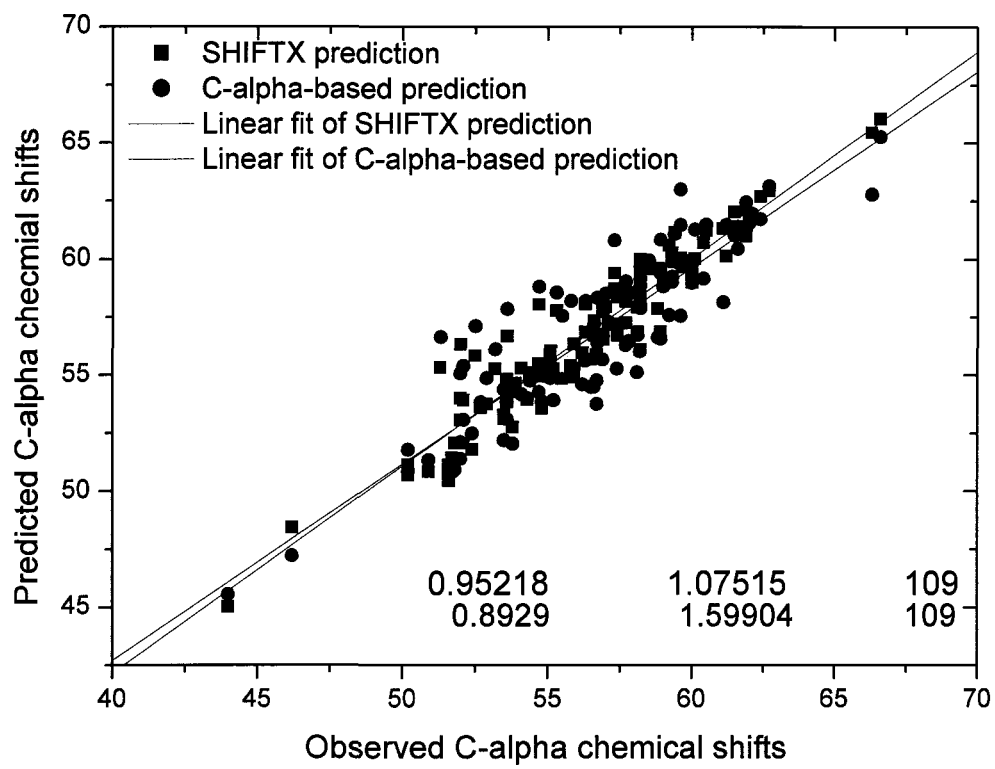


Figure 2.12 Comparison of observed and predicted C-alpha chemical shifts of 1RUV

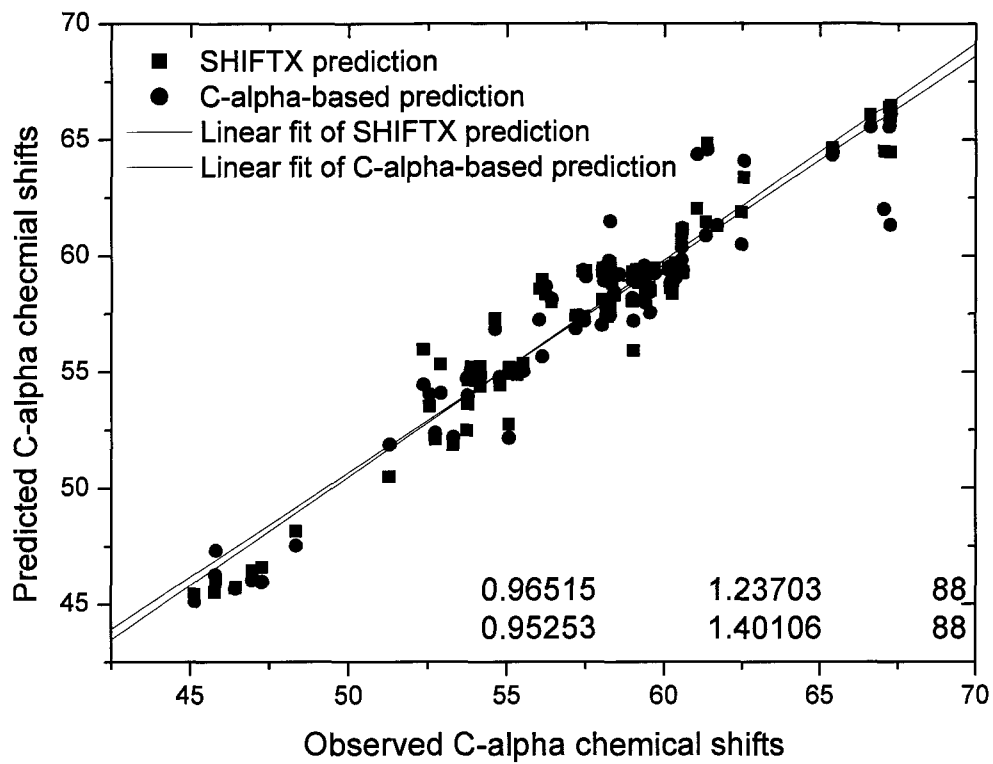


Figure 2.13 Comparison of observed and predicted C-alpha chemical shifts of 1TOP

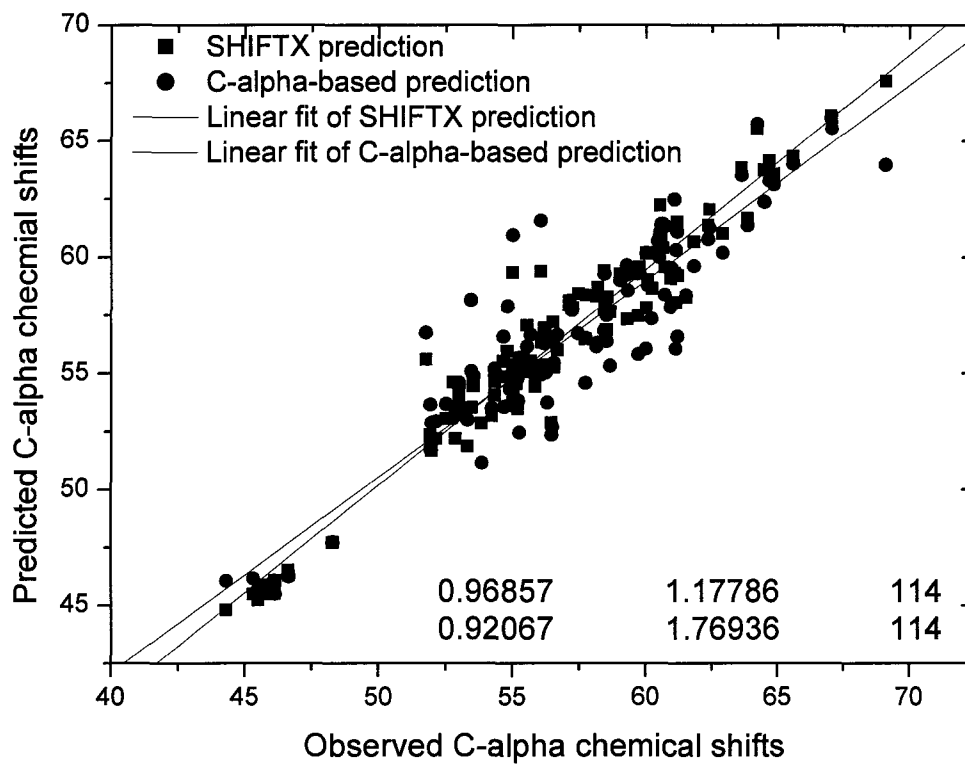


Figure 2.14 Comparison of observed and predicted C-alpha chemical shifts of 3LZT

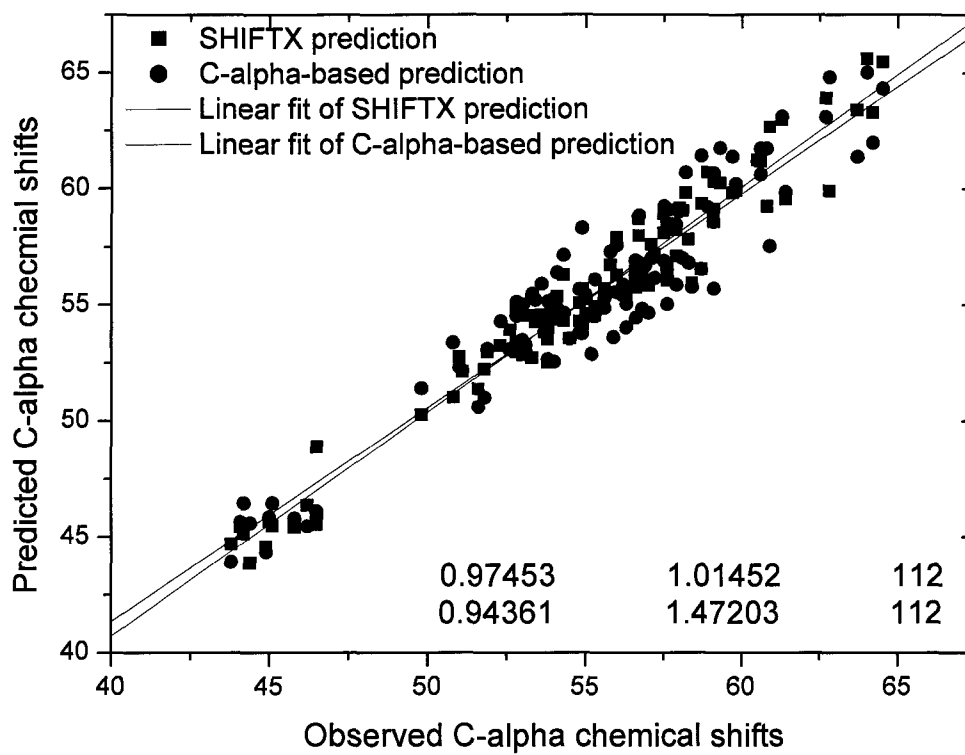


Figure 2.15 Comparison of observed and predicted C-alpha chemical shifts of 4FGF

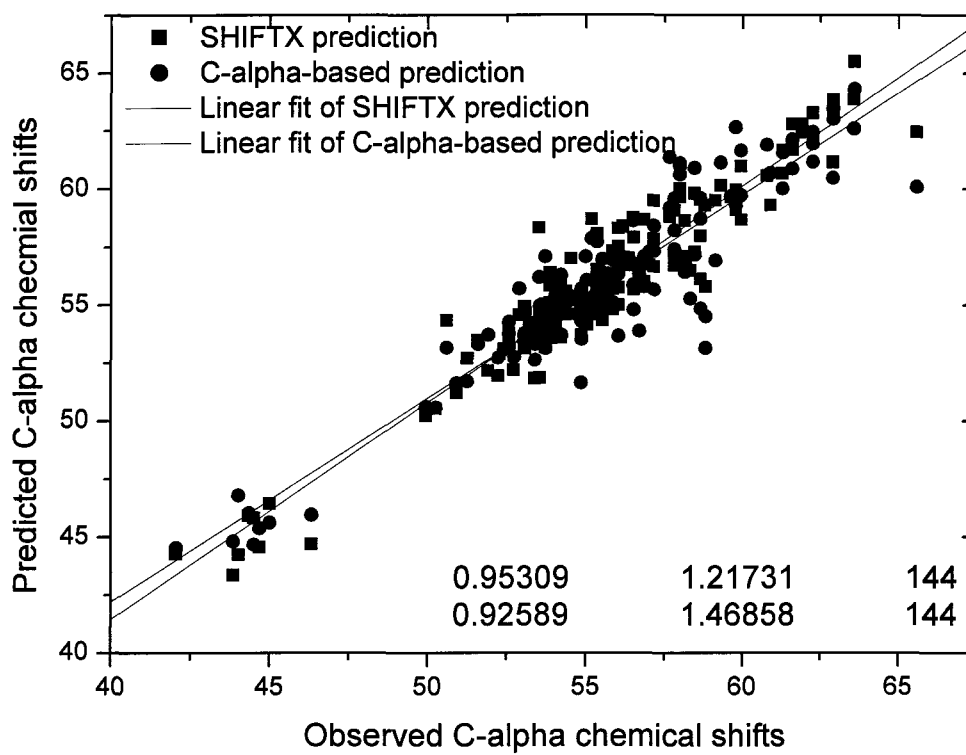


Figure 2.16 Comparison of observed and predicted C-alpha chemical shifts of 4I1B

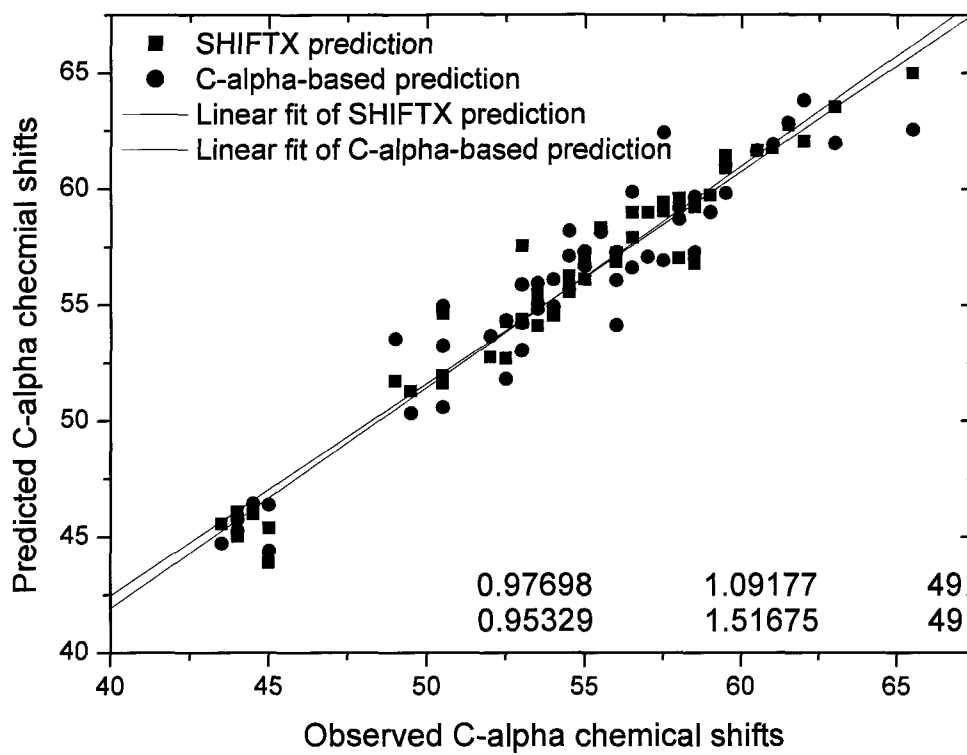


Figure 2.17 Comparison of observed and predicted C-alpha chemical shifts of 5PTI

2.4 Conclusion and Discussion

We reported a suite of program that successfully predicts accurate C-alpha chemical shift from a knowledge-based structural profile library. The program is fed with low-resolution models with C-alpha atoms only, which are readily constructed by any computational modeling approaches. This study reveals the fact that current computation method can capture the essence of NMR parameters from even such a coarse-grained level of protein structure. Specifically, the overall quality (measured by correlation coefficient and RMSD) of this C-alpha based prediction is in the same level as high-resolution all-atom structure required predictive program SHIFTX.

To further improve the quality of the prediction, several approaches are taken into consideration. Firstly, without raising the difficulty in modeling the structure, we can still take more comprehensive geometric information from the C-alpha traces, such as the penta-peptide sequence and structure, $C(i-2)C(i-1)C(i)C(i+1)C(i+2)$, the secondary structure from this broader structural unit, $SS(i-1)SS(i)SS(i+1)SS(i+2)$. This five-element unit may contains much more structural signatures than the current tripeptide in that the secondary structure of the unit is more definitive and can be sophisticated categorized. Two more Ca bond angles can be obtained $C(i-2)C(i-1)C(i)$, $C(i-1)C(i)C(i+1)$, $C(i)C(i+1)C(i+2)$ in closely describing the topology of the structural unit. Two dihedral angles $C(i-2)C(i-1)C(i)C(i+1)$, $C(i-1)C(i)C(i+1)C(i+2)$ are added to the structural library as well. These dihedral angles has been approved by previous work to reflects protein secondary structure and even tertiary topological

signature, in turn will adjust to the chemical shift prediction from the effect of long-distant forces and global shape of protein. The structural library is more focused on the short-range interactions and lack of such long-distance contributions. Secondly, the library describes the inter-residue network by simply the counting of neighbor residues and the distribution of their types. However, the relative distance and position are highly correlated the effect on chemical shifts from those neighbors. In an effort to include every possible effect on the chemical shift prediction, we plan to construct an even comprehensive structural profile, with all geometric properties of penta-peptide, and a fully description of the three-dimensional sphere of target C-alpha neighborhood, which depict the number, type, distance and orientation about each neighbor residue.

2.5 References

1. Doreleijers, J. F. et al. BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J Biomol NMR* 26, 139-46 (2003).
2. Meiler, J. PROSHIFT: protein chemical shift prediction using artificial neural networks. *J Biomol NMR* 26, 25-37 (2003).
3. Bailey-Kellogg, C. et al. The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J Comput Biol* 7, 537-58 (2000).
4. Li, W. et al. TOUCHSTONEX: protein structure prediction with sparse NMR data. *Proteins* 53, 290-306 (2003).
5. Meiler, J. & Baker, D. Rapid protein fold determination using unassigned NMR data. *Proc Natl Acad Sci U S A* 100, 15404-9 (2003).
6. Meiler, J. & Baker, D. The fumarate sensor DcuS: progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy. *J Magn Reson* 173, 310-6 (2005).
7. Rohl, C. A. & Baker, D. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J Am Chem Soc* 124, 2723-9 (2002).
8. Grishaev, A. & Llinas, M. CLOUDS, a protocol for deriving a molecular proton density via NMR. *Proc Natl Acad Sci U S A* 99, 6707-12 (2002).

9. Brooks III, C. L., M. Karplus, and B. M. Pettitt. 1988. Proteins: a theoretical perspective of dynamics, structure, and thermodynamics. *Adv. Chem. Phys.* 71:1-249.
10. Ming, D., Y. Kong, Y. Wu, and J. Ma. 2003b. Substructure synthesis method for simulating large molecular complexes. *Proc. Natl. Acad. Sci. USA.* 100:104-109.
11. Numerical Recipes: The Art of Scientific Computing. Cambridge University Press, Cambridge, UK.
12. Xu, X.P. and D.A. Case, *Probing multiple effects on N-15, C-13 alpha, C-13 beta, and C-13 ' chemical shifts in peptides using density functional theory.* *Biopolymers*, 2002. **65**(6): p. 408-423.
13. Xu, X.P. and D.A. Case, *Automated prediction of N-15, C-13(alpha), C-13(beta) and C-13 ' chemical shifts in proteins using a density functional database.* *Journal of Biomolecular Nmr*, 2001. **21**(4): p. 321-333.
14. Wishart, D.S., et al., *Automated H-1 and C-13 chemical shift prediction using the BioMagResBank.* *Journal of Biomolecular Nmr*, 1997. **10**(4): p. 329-336.
15. Wishart, D.S. and A.M. Nip, *Protein chemical shift analysis: a practical guide.* *Biochemistry and Cell Biology-Biochimie Et Biologie Cellulaire*, 1998. **76**(2-3): p. 153-163.
16. Wang, Y.J. and O. Jardetzky, *Predicting N-15 chemical shifts in proteins using the preceding residue-specific individual shielding surfaces from phi,*

- psi(i-1), and chi(1) torsion angles*. Journal of Biomolecular Nmr, 2004. **28**(4): p. 327-340.
17. Wang, Y.J. and O. Jardetzky, *Investigation of the neighboring residue effects on protein chemical shifts*. Journal of the American Chemical Society, 2002. **124**(47): p. 14075-14084.
 18. Wang, Y.J., *Secondary structural effects on protein NMR chemical shifts'*. Journal of Biomolecular Nmr, 2004. **30**(3): p. 233-244.
 19. Pearson, J.G., et al., *Predicting chemical shifts in proteins: Structure refinement of valine residues by using ab initio and empirical geometry optimizations*. Journal of the American Chemical Society, 1997. **119**(49): p. 11941-11950.
 20. Neal, S., et al., *Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts*. Journal of Biomolecular Nmr, 2003. **26**(3): p. 215-240.
 21. Meiler, J., *PROSHIFT: Protein chemical shift prediction using artificial neural networks*. Journal of Biomolecular Nmr, 2003. **26**(1): p. 25-37.
 22. Luman, N.R., M.P. King, and J.D. Augspurger, *Predicting N-15 amide chemical shifts in proteins. I. An additive model for the backbone contribution*. Journal of Computational Chemistry, 2001. **22**(3): p. 366-372.
 23. Gronwald, W., et al., *ORB, a homology-based program for the prediction of protein NMR chemical shifts*. Journal of Biomolecular Nmr, 1997. **10**(2): p. 165-179.

24. Benod, C., M.A. Delsuc, and J.L. Pons, *CRAACK: Consensus program for NMR amino acid type assignment*. Journal of Chemical Information and Modeling, 2006. **46**(3): p. 1517-1522.
25. Vranken, W.F. and W. Rieping, *Relationship between chemical shift value and accessible surface area for all amino acid atoms*. BMC Structural Biology, 2009. **9**: p. -.
26. He, X., B. Wang, and K.M. Merz, *Protein NMR Chemical Shift Calculations Based on the Automated Fragmentation QM/MM Approach*. Journal of Physical Chemistry B, 2009. **113**(30): p. 10380-10388.
27. Ginzinger, S.W. and M. Coles, *SimShiftDB; local conformational restraints derived from chemical shift similarity searches on a large synthetic database*. Journal of Biomolecular Nmr, 2009. **43**(3): p. 179-185.
28. Elyashberg, M.E., K.A. Blinov, and A.J. Williams, *The application of empirical methods of C-13 NMR chemical shift prediction as a filter for determining possible relative stereochemistry*. Magnetic Resonance in Chemistry, 2009. **47**(4): p. 333-341.
29. Blinov, K.A., et al., *Development of a fast and accurate method of C-13 NMR chemical shift prediction*. Chemometrics and Intelligent Laboratory Systems, 2009. **97**(1): p. 91-97.
30. Vila, J.A., et al., *Predicting C-13(alpha) chemical shifts for validation of protein structures*. Journal of Biomolecular Nmr, 2007. **38**(3): p. 221-235.
31. Vila, J.A., et al., *Quantum chemical C-13(alpha) chemical shift calculations for protein NMR structure determination, refinement, and*

- validation*. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(38): p. 14389-14394.
32. Shen, Y. and A. Bax, *Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology*. Journal of Biomolecular Nmr, 2007. **38**(4): p. 289-302.
33. Kaur, J. and A.S. Brar, *An approach to predict the C-13 NMR chemical shifts of acrylonitrile copolymers using artificial neural network*. European Polymer Journal, 2007. **43**(1): p. 156-163.

CHAPTER 3.

Determination of Protein Native Structure Assisted by Unassigned NMR Data

3.1 Introduction

Proteins play essential roles in organisms and participate in almost all processes in the cells. Their versatile functionalities are greatly owing to the ability to fold into specific three-dimensional shapes determined uniquely by the amino acid sequences. The major goals of current protein folding study are to determine the three-dimensional tertiary structure from one-dimensional primary sequences. From decades of extensive study, substantial progress has been made in terms of understanding of folding mechanisms and of actual prediction of three-dimensional structures¹²⁻²¹. Currently, it is still a critical challenge faced by structural biologists in the exploration of life, to understand how to reliably *determine overall topology for most proteins how proteins adopt the stable structures to allow specific functions.*

With the dramatic advancement in biological experimental techniques and facilitation in the past few decades, structural biology benefits a boost of experimental data. Meanwhile, there is an unprecedented demand of computational simulations of biological systems to incorporate the explosion of real world data for two reasons. First, computational modeling of proteins allow fast and efficient alternative methods to explain the experimental results, that will

ultimately contribute to reveal the insight into the most difficult tasks in molecular structural biology, such as structural-based drug design. Second to assisting the experimental data interpretation, *in silico* modeling approaches are capable of revealing the dynamic processes that cannot be directly detected by common experimental techniques in structural biology. In summary, computational simulations serves as irreplaceable tools in linking the classical theories from mathematics, physics, and chemistry with developing studies in structural biology.

Protein structure determination from NMR data

Traditional approaches provide us valuable insight on the utility of NMR data in protein structure determination¹⁻³⁷, meanwhile, several projects are launched to determine protein native structures from unassigned or sparse NMR data⁹⁶⁻¹¹⁴. With the efforts exerted in automatic signal assignment of NOESY and other spectroscopy^{77,88,93,95,112}, some group tried to derive geometric distance information between atoms from unassigned NOESY experiments (CLOUDS⁴⁵⁻⁴⁷). Others incorporate *de novo* fold prediction algorithms to generate candidate structures. Sparse and unassigned NMR data³⁷⁻⁷⁶, often including residual dipolar couplings (RDCs), NOE and backbone chemical shifts, are served as restrictions in filtering incorrect conformations¹⁻³⁶, such as TOUCHSTONEX⁴ and RosettaNMR^{5,6}

Recently, our group developed a *de novo* protein C-alpha-based structure prediction approach¹¹⁵⁻¹¹⁷. Based on sequence information alone this novel

protocol can rapidly generate a group of possible structures with only C-alpha atoms (C-alpha traces). The lack of other backbone atoms and entire side chain potentially enables this approach to handle large proteins (> 200 amino acids). In order to select correct conformations or single out native structures candidate models generated, other physical potential functions and experimental data are necessary. When NMR experimental chemical shift and its assignment are available, the calculated chemical shifts for candidate models are compared with actual experimental data. An agreement score is calculated as sum of the deviations between corresponding chemical shifts in two sets, and represents the resemblance of between predicted and observed values. The candidate models are ranked according to their consistency with the NMR data.

When the assignment of experimental chemical shift is unavailable or incomplete, for each candidate model, an optimal assignment is identified by Monte Carlo/ Simulated annealing search in need of the comparison. The combined approaches were tested on a 76 residue protein Ubiquitin (PDB: 1G6J), among 84 generated candidates with RMSD from 4~15 angstroms. After the assignment optimization, the native structure is recognized without any assignment information.

In this chapter, we report the result from the recent study in recognizing native structure of proteins or domains under constraints from NMR data. Since we used C-alpha-based model, we are focusing on distinguishing the overall correct topology, instead of delivering atomic resolution structure. This success attempt could help to increase the effective resolution of NMR method since a

reasonably correct three-dimensional topology conveys more structural information than, for example, a 10 Å density map by other technique such as cryogenic electron microscopy (cryo-EM), from which no detailed structural model could be derived. For small non-crystallizable proteins, therefore, our method serves as a bridge connecting cryo-EM, which works better with very large systems, and x-ray crystallography, which must have quality crystals.

3.2 Methods

As described in Chapter II, for given C-alpha traces of target protein, we are able to predict accurate chemical shifts of all residues from the target model, except the first and last residues, based on the knowledge-based structural profile library. Parallel with the prediction, the target protein is gone through the conventional NMR experiments. Needless of further signal assignment, A set of C-alpha chemical shift can be derived conveniently from the NMR spectra HN(CO)CA and HNCA. Both spectra can be performed within hours and in an automatic format. In our research, we download the C-alpha chemical shifts of testing proteins from BMRB database ¹. The native structure recognition with the help from experimental data is discussed in the following two cases respectively.

When the assignment is available

With a strong assumption that complete assignment is available between specific experimental chemical shift signals for each C-alpha atom in a protein, an agreement score is designed to describe the resemblance of two sets of chemical shifts, both in the same order as the sequence, one set is directly observed from experiment and the other is predicted by the knowledge-based structural profile library. It is calculated by comparing the experimental chemical shift with predicted ones registered to the same atoms. The score is obtained by

$$score = \sum_{i=1}^{N_{residue}} \delta_i$$

$$\delta_i = \begin{cases} 1 & \text{if } |CS_{\text{exp}} - CS_{\text{predicted}}| < \text{low} \\ \frac{|CS_{\text{exp}} - CS_{\text{predicted}}|}{(\text{high} - \text{low})} & \text{if } \text{high} > |CS_{\text{exp}} - CS_{\text{predicted}}| > \text{low} \\ 0 & \text{if } |CS_{\text{exp}} - CS_{\text{predicted}}| > \text{high} \end{cases}$$

high and *low* are determined upon further analysis of predicted chemical shift, specifically in this study, *high* = 3.0 ppm, *low* = 1.0 ppm.

Considering the ideal situation that every single pair of assigned predicted and observed chemical shifts are close enough that their difference falls into the range of 0 to *low*, then the agreement score is 1 for each pair. Therefore, the highest possible sum of agreement score could be $(N_{\text{residue}} - 2)$, as the first and last residues chemical shift prediction are ignored due to missing prediction values. This highest agreement score reflect a complete match between two sets of chemical shift and in turn indicating that the model resemblance the native structure to certain extent. On the contrary, the sum of agreement score could be 0 in the worst case, in which the deviation of either pair is smaller than value *high*.

Predicted CS			Exp. CS assigned		
i	aa	CS_pre	i	aa	CS_exp
2	Q	53.4742	2	Q	55.0
3	I	56.3627	3	I	59.6
4	F	56.2557	4	F	55.2
5	V	60.9527	5	V	60.8
6	K	55.1466	6	K	54.5
...

Figure 3.1 Calculation of agreement score with complete assignment of experimental NMR data

When the assignment is not available or incomplete

In a realistic case, and for the purpose to use unassigned NMR data, we assume the sequential assignment information is incomplete or not available at all, where signals from HN(CO)CA, and HNCA spectra have no correlation with sequential order. Initially, we linked the individual chemical shift from the two lists randomly; a pseudo score can be calculated based on the random pairs. We then performed a Monte Carlo and simulated annealing procedure to maximize the score. At each step, two pairs are randomly picked, their partners exchanged, and new links are accepted if the total agreement score increases after the exchange, otherwise, the pairs remain untouched. Typically for structure with 100 residues, 15000 MC steps are performed until an optimal assignment is found and the agreement scores converge to a static value.

In order to minimize the biases from MC/SA process, totally 20 trajectories are performed in parallel and independently for each candidate model.

The agreement scores are taken at multiple steps during MC iterations; an average over these twenty trajectories is recorded as an overall native quality for the model. The native structure is to be recognized by its overall high agreement scores and fast convergence.

Predicted CS			Exp. CS unassigned		
i	aa	CS_pre	i	aa	CS_exp
2	Q	53.4742	?	?	54.5
3	I	56.3627	?	?	55.0
4	F	56.2557	?	?	55.2
5	V	60.9527	?	?	59.6
6	K	55.1466	?	?	60.8
...

Figure 3.2 Alignment and calculation of the agreement score with unassigned experimental NMR data

Recognition of native structure of Ubiquitin

We test the approaches against a real a real protein Ubiquitin (PDB ID 1G6J, BMRB entry 5387). Ubiquitin is small mediator proteins in a number biological process, which was firstly discovered in eukaryotic organism and was intensively studied for its function of tagging other proteins for ATP dependent degradation. Ubiquitin consists of 76 amino acids residues, which are grouped into one anti-parallel beta-strand sheet, with four strands, and two alpha-helices.

Meanwhile 84 simulated C-alpha structures are generated by a *de novo* modeling approaches by Wu and Ma based on sequence only. Each model is aligned in three-dimensional space with the known native structure, a C-alpha

root mean square difference between simulated and native C-alpha traces are calculated, the RMSDs are ranged from 4.5 to 14.5 Angstrom.



Figure 3.3 Structure of Ubiquitin (PDB ID 1G6J, BMRB entry 5387)

In both cases, with complete known assignment or without any assignment information, the native structure is identified from 84 candidate models. The correct structure has highest agreement scores among all the models, calculated by the deviation between the predicted and observed chemical shifts. In the latter case, ten thousand steps of Monte Carlo assignment optimization are performed to achieve an optimal assignment matching.

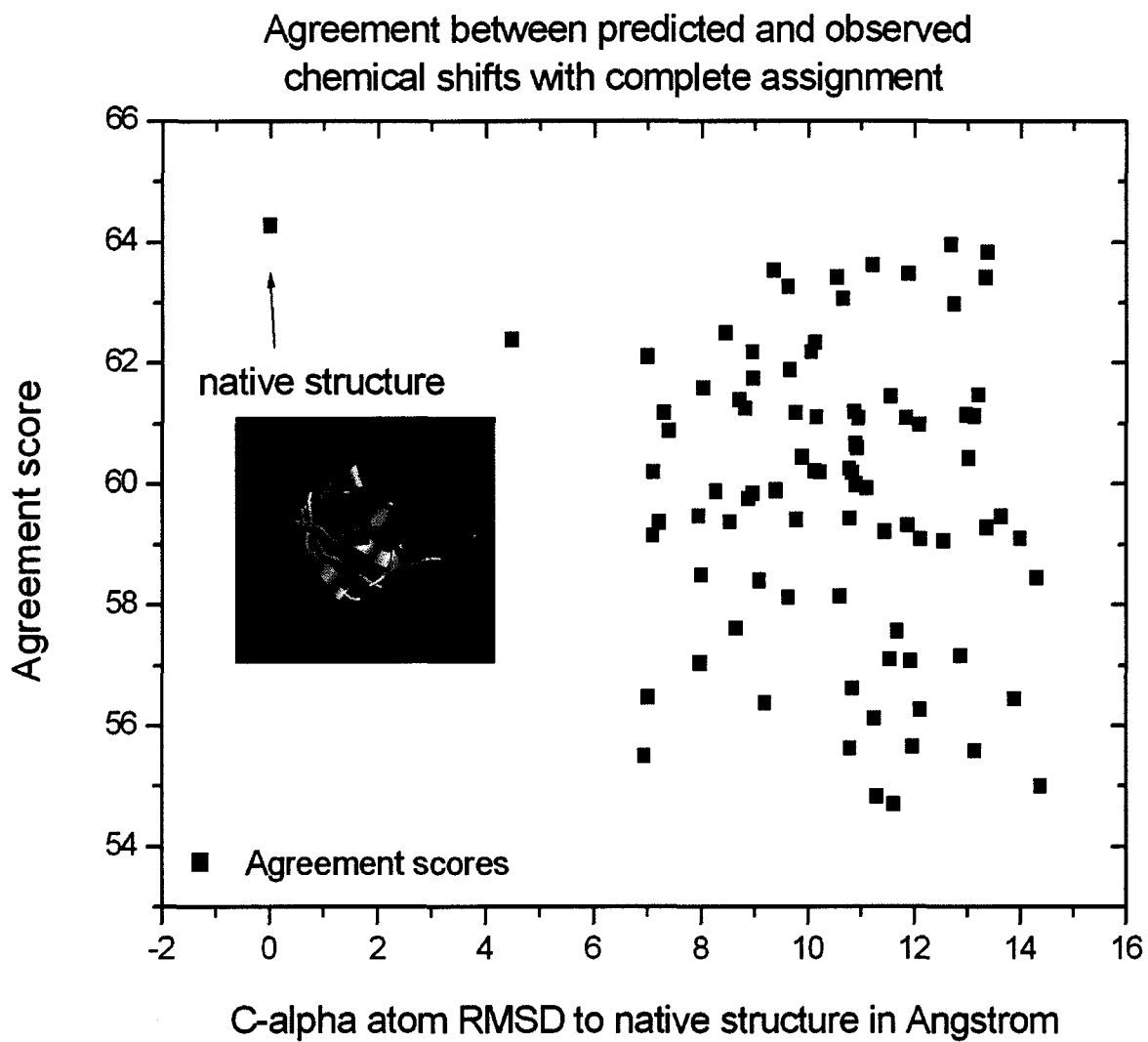


Figure 3.4 Agreement between predicted and observed chemical shifts with complete assignment

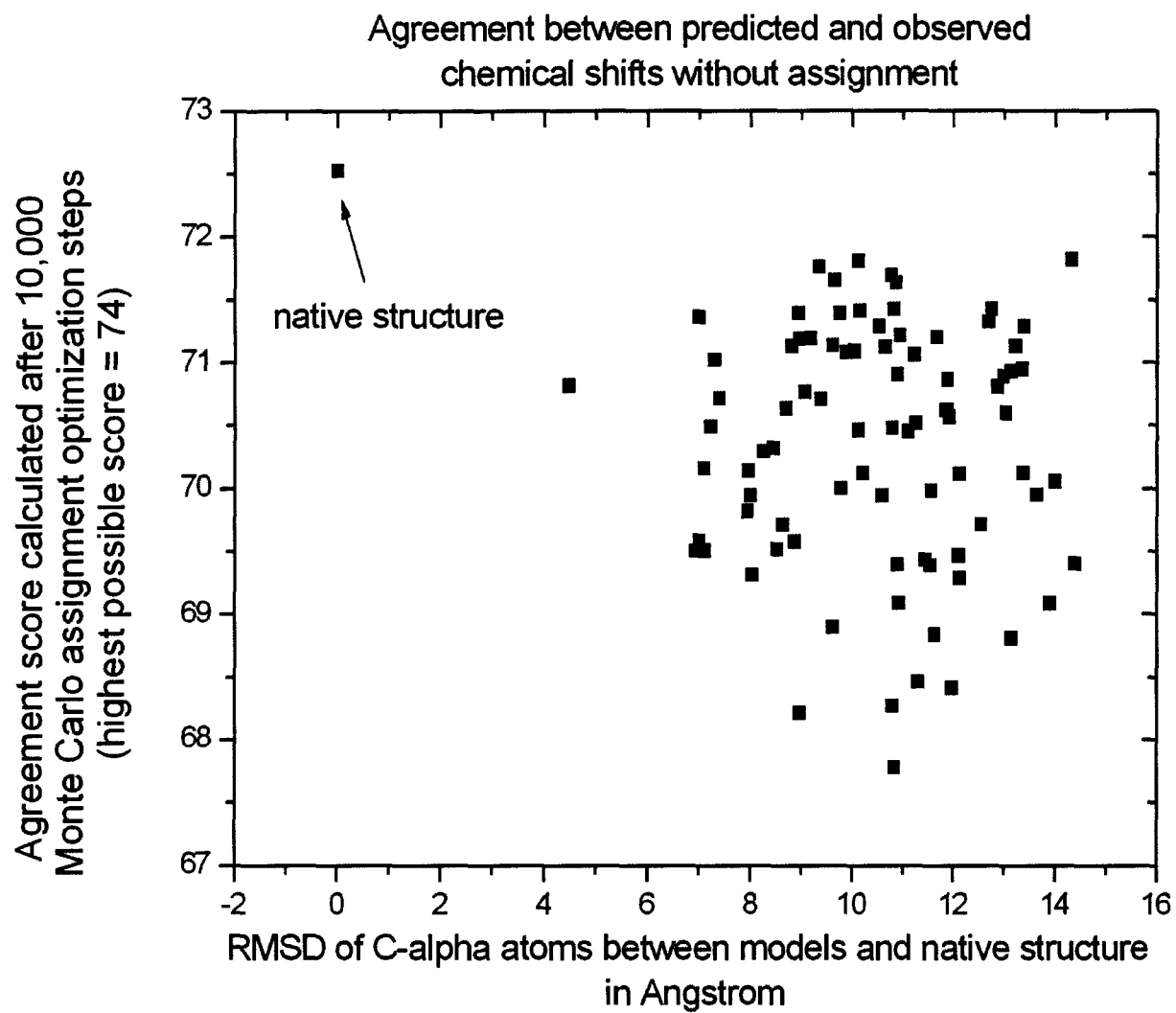


Figure 3.5 Agreement between predicted and observed chemical shifts without assignment

Agreement score convergence in optimization

We also studied the performance of agreement score during the optimization procedure. To simulate the actual condition for incomplete experimental data, we set five initial conditions, with 100% (full assignment), 80%, 50%, 20%, and 0% (complete random) assignment respectively. For native structure, the agreement scores in all five levels converged to the same and almost the highest value of 74 (Figure 3.6). However, even after quite long Monte Carlo optimization steps (20000), the chemical shifts predicted from denatured conformations cannot converge with such a high score, due to their inconsistent pattern from observed data (Figure 3.7).

We discovered that the final scores after optimization are even higher than the scores calculated from known assignment, meanwhile, the optimal assignment found by MC procedure is often different from actual assignment. We argue that the consistence between predicted and experimental data from native structure guarantees that it could obtain a higher agreement score comparing with those from simulated models. In other words, even after extensive long optimization, the inappropriate agreement from denatured structures prevents them from achieving good match in any circumstances.

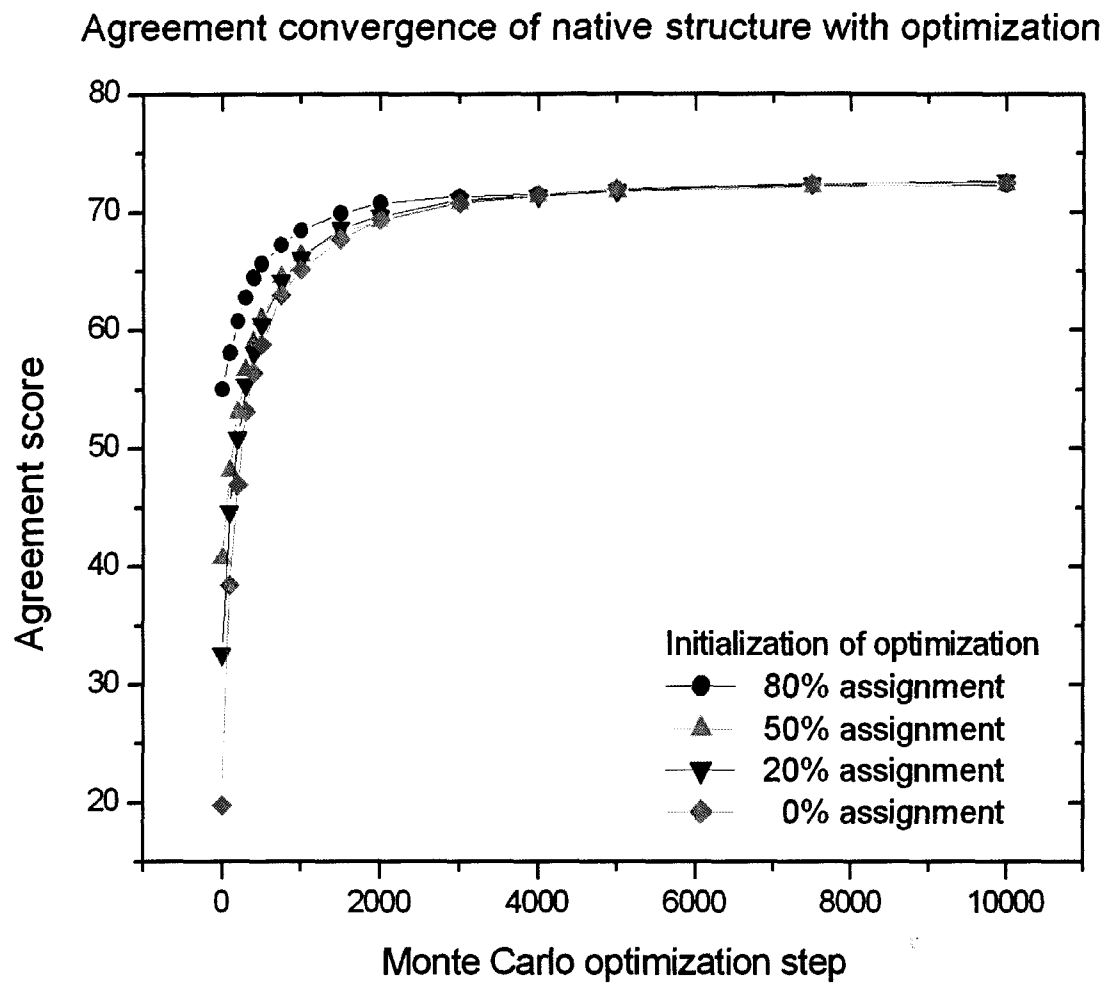


Figure 3.6 Convergence of agreement score during alignment optimization with different initial assignment condition

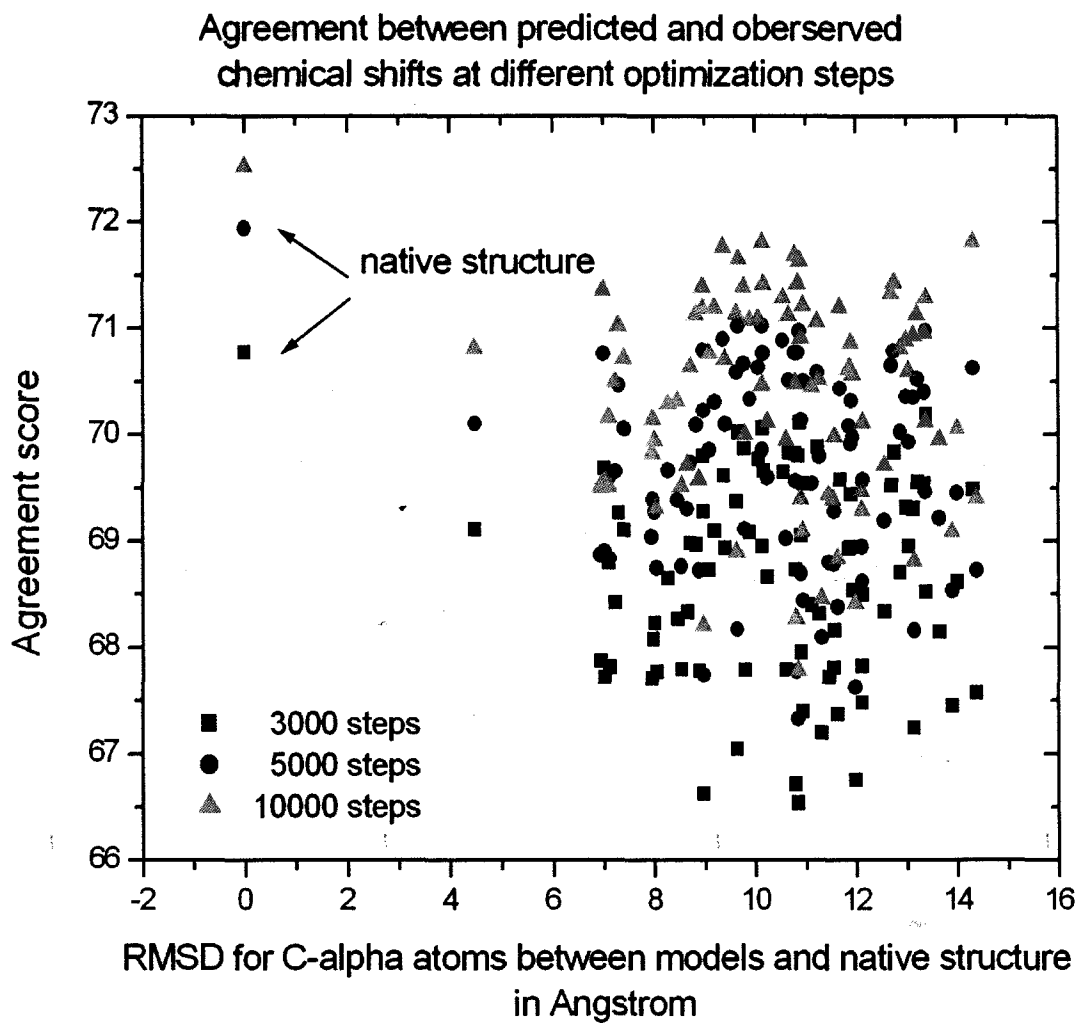


Figure 3.7 Agreement scores predicted and observed chemical shifts at different number of optimization steps

3.4 Concluding Discussion

In this chapter, we reported results of a computational study for recognition tertiary topology of proteins, or protein domains in the assistant from conveniently accessible NMR data. A knowledge-based structural profile and pre-calculated chemical shift library and coarse-grained protein models (C-alpha-traces) were used in the structure determination. One important feature in this study was that a novel protocol of Monte Carlo simulation was employed to overcome the time-consuming and often man power intensive signal nuclei assignment in regular NMR experiment procedure. The algorithm made significant contribution to native structure determination in that it potentially provides a link from low-resolution models (C-alpha traces) with raw NMR data.

Together with the recent development of other computational methods ⁴⁵, ⁴⁶ that simulated the overall shape of large molecular complexes with promising success ^{47,48}, we believe that these methods will eventually enable NMR to be a main stream experimental technique in the field of structural biology.

3.5 References

1. Ab, E., et al., *Direct use of unassigned resonances in NMR structure calculations with proxy residues*. Journal of the American Chemical Society, 2006. **128**(23): p. 7566-7571.
2. Altieri, A.S. and R.A. Byrd, *Automation of NMR structure determination of proteins*. Current Opinion in Structural Biology, 2004. **14**(5): p. 547-553.
3. Atkinson, R.A. and V. Saudek, *The direct determination of protein structure by NMR without assignment*. Febs Letters, 2002. **510**(1-2): p. 1-4.
4. Auguin, D., et al., *Superposition of chemical shifts in NMR spectra can be overcome to determine automatically the structure of a protein*. Spectroscopy-an International Journal, 2003. **17**(2-3): p. 559-568.
5. Ayers, D.J., et al., *Enhanced protein fold recognition using secondary structure information from NMR*. Protein Science, 1999. **8**(5): p. 1127-1133.
6. Billeter, M., P. Schultze, and K. Wuthrich, *Determination of Protein Secondary Structure from Patterns of Distance Constraints Observed by Nmr in Solution*. Experientia, 1984. **40**(6): p. 603-603.
7. Bowers, P.M., C.E.M. Strauss, and D. Baker, *De novo protein structure determination using sparse NMR data*. Journal of Biomolecular Nmr, 2000. **18**(4): p. 311-318.

8. Braun, W., *Distance Geometry and Related Methods for Protein-Structure Determination from Nmr Data*. Quarterly Reviews of Biophysics, 1987. **19**(3-4): p. 115-157.
9. Campbell, I.D. and B. Sheard, *Protein-Structure Determination by Nmr*. Trends in Biotechnology, 1987. **5**(11): p. 302-306.
10. Chen, H.L. and H.X. Zhou, *Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data*. Proteins-Structure Function and Bioinformatics, 2005. **61**(1): p. 21-35.
13. Clore, G.M. and A.M. Gronenborn, *Applications of 3-Dimensional and 4-Dimensional Heteronuclear Nmr-Spectroscopy to Protein-Structure Determination*. Progress in Nuclear Magnetic Resonance Spectroscopy, 1991. **23**: p. 43-92.
14. Clore, G.M. and A.M. Gronenborn, *NMR structure determination of proteins and protein complexes larger than 20 kDa*. Current Opinion in Chemical Biology, 1998. **2**(5): p. 564-570.
15. Linge, J.P., S.I. O'Donoghue, and M. Nilges, *Automated assignment of ambiguous nuclear overhauser effects with ARIA*. Nuclear Magnetic Resonance of Biological Macromolecules, Pt B, 2001. **339**: p. 71-90.
16. Williamson, M.P. and C.J. Craven, *Automated protein structure calculation from NMR data*. Journal of Biomolecular Nmr, 2009. **43**(3): p. 131-143.

17. Wang, J.B., et al., *Determination of Multicomponent Protein Structures in Solution Using Global Orientation and Shape Restraints*. Journal of the American Chemical Society, 2009. **131**(30): p. 10507-10515.
18. Szymczyna, B.R., et al., *Synergy of NMR, Computation, and X-Ray Crystallography for Structural Biology*. Structure, 2009. **17**(4): p. 499-507.
19. Shen, Y., et al., *De novo protein structure generation from incomplete chemical shift assignments*. Journal of Biomolecular Nmr, 2009. **43**(2): p. 63-78.
20. Sakakibara, D., et al., *Protein structure determination in living cells by in-cell NMR spectroscopy*. Nature, 2009. **458**(7234): p. 102-U10.
21. Guntert, P., *Automated structure determination from NMR spectra*. European Biophysics Journal with Biophysics Letters, 2009. **38**(2): p. 129-143.
22. Donald, B.R. and J. Martin, *Automated NMR assignment and protein structure determination using sparse dipolar coupling constraints*. Progress in Nuclear Magnetic Resonance Spectroscopy, 2009. **55**(2): p. 101-127.
23. Berjanskii, M., et al., *GeNMR: a web server for rapid NMR-based protein structure determination*. Nucleic Acids Research, 2009. **37**: p. W670-W677.
24. Vogeli, B., L.S. Yao, and A. Bax, *Protein backbone motions viewed by intraresidue and sequential H-N-H-alpha residual dipolar couplings*. Journal of Biomolecular Nmr, 2008. **41**(1): p. 17-28.

25. Vila, J.A. and H.A. Scheraga, *Factors affecting the use of C-13(alpha) chemical shifts to determine, refine, and validate protein structures*. Proteins-Structure Function and Bioinformatics, 2008. **71**(2): p. 641-654.
26. Vila, J.A., Y.A. Arnautova, and H.A. Scheraga, *Use of C-13(alpha) chemical shifts for accurate determination of beta-sheet structures in solution*. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(6): p. 1891-1896.
27. Vila, J.A., et al., *Quantum chemical C-13(alpha) chemical shift calculations for protein NMR structure determination, refinement, and validation*. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(38): p. 14389-14394.
28. Ulrich, E.L., et al., *BioMagResBank*. Nucleic Acids Research, 2008. **36**: p. D402-D408.
29. Shin, J., W. Lee, and W. Lee, *Structural proteomics by NMR spectroscopy*. Expert Review of Proteomics, 2008. **5**(4): p. 589-601.
30. Shen, Y., et al., *Consistent blind protein structure generation from NMR chemical shift data*. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(12): p. 4685-4690.
31. He, Y., et al., *NMR structures of two designed proteins with high sequence identity but different fold and function*. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(38): p. 14412-14417.

32. Rapp, C.S., et al., *Prediction of protein loop geometries in solution*. Proteins-Structure Function and Bioinformatics, 2007. **69**(1): p. 69-74.
33. Latek, D., D. Ekonomiuk, and A. Kolinski, *Protein structure prediction: Combining de novo modeling with sparse experimental data*. Journal of Computational Chemistry, 2007. **28**(10): p. 1668-1676.
34. Cavalli, A., et al., *Protein structure determination from NMR chemical shifts*. Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(23): p. 9615-9620.
35. Campbell, I.D. and B. Sheard, *Protein-Structure Determination by Nmr*. Trends in Biotechnology, 1987. **5**(11): p. 302-306.
36. Constantine, K.L., et al., *Protein-ligand NOE matching: A high-throughput method for binding pose evaluation that does not require protein NMR resonance assignments*. Journal of the American Chemical Society, 2006. **128**(22): p. 7252-7263.
37. Dancea, F. and U. Gunther, *Automated protein NMR structure determination using wavelet de-noised NOESY spectra*. Journal of Biomolecular Nmr, 2005. **33**(3): p. 139-152.
38. Delaglio, F., G. Kontaxis, and A. Bax, *Protein structure determination using molecular fragment replacement and NMR dipolar couplings*. Journal of the American Chemical Society, 2000. **122**(9): p. 2142-2143.
39. Devlieg, J., et al., *Restrained Molecular-Dynamics Procedure for Protein Tertiary Structure Determination from Nmr Data - a Lac Repressor Headpiece Structure Based on Information on J-Coupling and from*

- Presence and Absence of Noe*. Israel Journal of Chemistry, 1986. **27**(2): p. 181-188.
40. Ding, K.Y. and A.M. Gronenborn, *Protein backbone H-1(N)-C-13(alpha) and N-15-C-13(alpha) residual dipolar and J couplings: New constraints for NMR structure determination*. Journal of the American Chemical Society, 2004. **126**(20): p. 6232-6233.
41. Duggan, B.M., et al., *SANE (Structure assisted NOE evaluation): An automated model-based approach for NOE assignment*. Journal of Biomolecular Nmr, 2001. **19**(4): p. 321-329.
42. Fossi, M., et al., *Influence of chemical shift tolerances on NMR structure calculations using ARIA protocols for assigning NOE data*. Journal of Biomolecular Nmr, 2005. **31**(1): p. 21-34.
43. Fossi, M., et al., *Quantitative study of the effects of chemical shift tolerances and rates of SA cooling on structure calculation from automatically assigned NOE data*. Journal of Magnetic Resonance, 2005. **175**(1): p. 92-102.
44. Gao, G.H. and D.C. Wang, *Progress in the determination of protein solution structure by NMR*. Progress in Biochemistry and Biophysics, 1999. **26**(3): p. 228-233.
45. Grishaev, A. and M. Llinas, *CLOUDS, a protocol for deriving a molecular proton density via NMR*. Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(10): p. 6707-6712.

46. Grishaev, A. and M. Llinas, *Protein structure elucidation from minimal NMR data: The CLOUDS approach*. Nuclear Magnetic Resonance of Biological Macromolecules, Part C, 2005. **394**: p. 261-295.
47. Grishaev, A. and M. Llinas, *BACUS: A Bayesian protocol for the identification of protein NOESY spectra via unassigned spin systems*. Journal of Biomolecular Nmr, 2004. **28**(1): p. 1-10.
48. Gronenborn, A.M. and G.M. Clore, *Protein-Structure Determination by 2d-Nmr and 3d-Nmr*. Abstracts of Papers of the American Chemical Society, 1989. **198**: p. 65-Anyl.
49. Gronwald, W., et al., *Automated assignment of NOESY NMR spectra using a knowledge based method (KNOWNOE)*. Journal of Biomolecular Nmr, 2002. **23**(4): p. 271-287.
50. Gronwald, W., et al., *CAMRA: Chemical shift based computer aided protein NMR assignments*. Journal of Biomolecular Nmr, 1998. **12**(3): p. 395-405.
51. Guittet, E., et al., *Application of 3-Dimensional Nmr to Protein Studies - Recent Developments*. Journal De Chimie Physique Et De Physico-Chimie Biologique, 1992. **89**(2): p. 125-133.
52. Guntert, P., *Automated NMR protein structure calculation*. Progress in Nuclear Magnetic Resonance Spectroscopy, 2003. **43**(3-4): p. 105-125.
53. Harvey, T.S., S. Bagby, and M. Ikura, *Automated Assignment of Multidimensional Protein Noe Spectra*. Journal of Cellular Biochemistry, 1993: p. 253-253.

54. Heald, S.L., et al., *Sequential Nmr Resonance Assignment and Structure Determination of the Kunitz-Type Inhibitor Domain of the Alzheimers Beta-Amyloid Precursor Protein*. *Biochemistry*, 1991. **30**(43): p. 10467-10478.
55. Herrmann, T., P. Guntert, and K. Wuthrich, *Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS*. *Journal of Biomolecular Nmr*, 2002. **24**(3): p. 171-189.
56. Herrmann, T., P. Guntert, and K. Wuthrich, *Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA*. *Journal of Molecular Biology*, 2002. **319**(1): p. 209-227.
57. Holak, T.A., M. Nilges, and H. Oschkinat, *Improved Strategies for the Determination of Protein Structures from Nmr Data - the Solution Structure of Acyl Carrier Protein*. *Febs Letters*, 1989. **242**(2): p. 218-224.
58. Huang, X.M., F. Moy, and R. Powers, *Evaluation of the utility of NMR structures determined from minimal NOE-based restraints for structure-based drug design, using MMP-1 as an example*. *Biochemistry*, 2000. **39**(44): p. 13365-13375.
59. Huang, X.M. and R. Powers, *Validity of using the radius of gyration as a restraint in NMR protein structure determination*. *Journal of the American Chemical Society*, 2001. **123**(16): p. 3834-3835.

60. Hudaky, P. and A. Perczel, *Toward direct determination of conformations of protein building units from multidimensional NMR experiments VI. Chemical shift analysis of his to gain 3D structure and protonation state information*. Journal of Computational Chemistry, 2005. **26**(13): p. 1307-1317.
61. Hung, L.H. and R. Samudrala, *PROTINFO: secondary and tertiary protein structure prediction*. Nucleic Acids Research, 2003. **31**(13): p. 3296-3299.
62. Hung, L.H. and R. Samudrala, *Accurate and automated classification of protein secondary structure with PsiCSI*. Protein Science, 2003. **12**(2): p. 288-295.
63. Hus, J.C., D. Marion, and M. Blackledge, *De novo determination of protein structure by NMR using orientational and long-range order restraints*. Journal of Molecular Biology, 2000. **298**(5): p. 927-936.
64. Jefson, M.R., *Applications of Nmr-Spectroscopy to Protein-Structure Determination*. Annual Reports in Medicinal Chemistry, 1988. **23**: p. 275-283.
65. Jung, Y.S., M. Sharma, and M. Zweckstetter, *Simultaneous assignment and structure determination of protein backbones by using NMR dipolar couplings*. Angewandte Chemie-International Edition, 2004. **43**(26): p. 3479-3481.

66. Kanelis, V., J.D. Forman-Kay, and L.E. Kay, *Multidimensional NMR methods for protein structure determination*. Iubmb Life, 2001. **52**(6): p. 291-302.
67. Kirnarsky, L., O. Shats, and S. Sherman, *Improving the efficiency of protein structure determination from NMR*. Journal of Molecular Structure-Theochem, 1997. **419**: p. 213-220.
68. Klaus, W. and R. Moser, *Nuclear-Magnetic-Resonance Studies and Molecular-Dynamics Simulations of the Solution Conformation of a Designed, Alpha-Helical Peptide*. Protein Engineering, 1992. **5**(4): p. 333-341.
69. Korukottu, J., et al., *Fast high-resolution protein structure determination by using unassigned NMR data*. Angewandte Chemie-International Edition, 2007. **46**(7): p. 1176-1179.
70. Koharudin, L.M.I., et al., *Use of very long-distance NOEs in a fully deuterated protein: an approach for rapid protein fold determination*. Journal of Magnetic Resonance, 2003. **163**(2): p. 228-235.
71. Kozerski, L., et al., *Towards stereochemical and conformational assignment in flexible molecules using NOEs and molecular modelling*. Journal of the Chemical Society-Perkin Transactions 2, 1997(9): p. 1811-1818.
72. Kraulis, P.J., *Protein 3-Dimensional Structure Determination and Sequence-Specific Assignment of C-13-Separated and N-15-Separated*

- Noe Data - a Novel Real-Space Ab-Initio Approach*. Journal of Molecular Biology, 1994. **243**(4): p. 696-718.
73. Kuszewski, J., et al., *Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments*. Journal of the American Chemical Society, 2004. **126**(20): p. 6258-6273.
74. Labudde, D., et al., *Prediction algorithm for amino acid types with their secondary structure in proteins (PLATON) using chemical shifts*. Journal of Biomolecular Nmr, 2003. **25**(1): p. 41-53.
75. Latek, D., D. Ekonomiuk, and A. Kolinski, *Protein structure prediction: Combining de novo modeling with sparse experimental data*. Journal of Computational Chemistry, 2007. **28**(10): p. 1668-1676.
76. Liu, G.H., et al., *NMR data collection and analysis protocol for high-throughput protein structure determination*. Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(30): p. 10487-10492.
77. Lohr, F. and H. Ruterjans, *Unambiguous NOE assignments in proteins by a combination of through-bond and through-space correlations*. Journal of Biomolecular Nmr, 1997. **9**(4): p. 371-388.
78. Lopez-Mendez, B. and P. Guntert, *Automated protein structure determination from NMR spectra*. Journal of the American Chemical Society, 2006. **128**(40): p. 13112-13122.

79. Marassi, F.M. and S.J. Opella, *Simultaneous assignment and structure determination of a membrane protein from NMR orientational restraints*. Protein Science, 2003. **12**(3): p. 403-411.
80. Marin, A., et al., *From NMR chemical shifts to amino acid types: Investigation of the predictive power carried by nuclei*. Journal of Biomolecular Nmr, 2004. **30**(1): p. 47-60.
81. Mayer, K.L., et al., *Structure determination of a new protein from backbone-centered NMR data and NMR-assisted structure prediction*. Proteins-Structure Function and Bioinformatics, 2006. **65**(2): p. 480-489.
82. Meiler, J. and D. Baker, *Rapid protein fold determination using unassigned NMR data*. Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(26): p. 15404-15409.
83. Meiler, J. and D. Baker, *The fumarate sensor DcuS: progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy*. Journal of Magnetic Resonance, 2005. **173**(2): p. 310-316.
84. Mertens, H.D.T. and P.R. Gooley, *Validating the use of database potentials in protein structure determination by NMR*. Febs Letters, 2005. **579**(25): p. 5542-5548.
85. Moreau, V.H., A.P. Valente, and F.C.L. Almeida, *Prediction of the amount of secondary structure of proteins using unassigned NMR spectra: A tool for target selection in structural proteomics*. Genetics and Molecular Biology, 2006. **29**(4): p. 762-770.

86. Mueller, G.A., et al., *NMR assignment of protein side chains using residue-correlated labeling and NOE spectra*. Journal of Magnetic Resonance, 2003. **165**(2): p. 237-247.
87. Neal, S., et al., *Accurate prediction of protein torsion angles using chemical shifts and sequence homology*. Magnetic Resonance in Chemistry, 2006. **44**: p. S158-S167.
88. Nilges, M., *Calculation of Protein Structures with Ambiguous Distance Restraints - Automated Assignment of Ambiguous Noe Crosspeaks and Disulfide Connectivities*. Journal of Molecular Biology, 1995. **245**(5): p. 645-660.
89. Nilges, M. and S.I. O'Donoghue, *Ambiguous NOEs and automated NOE assignment*. Progress in Nuclear Magnetic Resonance Spectroscopy, 1998. **32**: p. 107-139.
90. Oschkinat, H., T. Muller, and T. Dieckmann, *Protein-Structure Determination with 3-Dimensional and 4-Dimensional Nmr-Spectroscopy*. Angewandte Chemie-International Edition, 1994. **33**(3): p. 277-293.
91. Parker, L.L., A.R. Houk, and J.H. Jensen, *Cooperative hydrogen bonding effects are key determinants of backbone amide proton chemical shifts in proteins*. Journal of the American Chemical Society, 2006. **128**(30): p. 9863-9872.
92. Podlogar, B.L., et al., *Protein structure determination using a combination of comparative modeling and NMR spectroscopy. Application to the*

- response regulator protein, Spo0F*. Journal of Medicinal Chemistry, 1997. **40**(21): p. 3453-3455.
93. Pristovsek, P. and L. Franzoni, *Stereospecific assignments of protein NMR resonances based on the tertiary structure and 2D/3D NOE data*. Journal of Computational Chemistry, 2006. **27**(6): p. 791-797.
94. Quine, J.R., et al., *Mathematical aspects of protein structure determination with NMR orientational restraints*. Bulletin of Mathematical Biology, 2004. **66**(6): p. 1705-1730.
95. Rieping, W., et al., *ARIA2: Automated NOE assignment and data integration in NMR structure calculation*. Bioinformatics, 2007. **23**(3): p. 381-382.
96. Rohl, C.A., *Protein structure estimation from minimal restraints using rosetta*. Nuclear Magnetic Resonance of Biological Macromolecules, Part C, 2005. **394**: p. 244-260.
97. Schieborr, U., et al., *How much NMR data is required to determine a protein-ligand complex structure?* Chembiochem, 2005. **6**(10): p. 1891-1898.
98. Schnell, J.R., et al., *Rapid and accurate structure determination of coiled-coil domains using NMR dipolar couplings: Application to cGMP-dependent protein kinase I alpha*. Protein Science, 2005. **14**(9): p. 2421-2428.

99. Sherman, S., et al., *Improvement in accuracy of protein local structure determination from NMR data*. *Theochem-Journal of Molecular Structure*, 1996. **368**: p. 153-161.
100. Spronk, C.A.E.M., et al., *Validation of protein structures derived by NMR spectroscopy*. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 2004. **45**(3-4): p. 315-337.
101. Turner, D.L., et al., *Determination of solution structures of paramagnetic proteins by NMR*. *European Biophysics Journal with Biophysics Letters*, 1998. **27**(4): p. 367-375.
102. Wang, B. and K.M. Merz, *A fast QM/MM (Quantum Mechanical/Molecular Mechanical) approach to calculate nuclear magnetic resonance chemical shifts for macromolecules*. *Journal of Chemical Theory and Computation*, 2006. **2**(1): p. 209-215.
103. Wang, C.C., et al., *2DCSi: identification of protein secondary structure and redox state using 2D cluster analysis of NMR chemical shifts*. *Journal of Biomolecular Nmr*, 2007. **38**(1): p. 57-63.
104. Wishart, D., *NMR spectroscopy and protein structure determination: Applications to drug discovery and development*. *Current Pharmaceutical Biotechnology*, 2005. **6**(2): p. 105-120.
105. Wishart, D.S., B.D. Sykes, and F.M. Richards, *The Chemical-Shift Index - a Fast and Simple Method for the Assignment of Protein Secondary Structure through Nmr-Spectroscopy*. *Biochemistry*, 1992. **31**(6): p. 1647-1651.

106. Wu, H.H., L.D. Finger, and J. Feigon, *Structure determination of protein/RNA complexes by NMR*. Nuclear Magnetic Resonance of Biological Macromolecules, Part C, 2005. **394**: p. 525-545.
107. Wuthrich, K., *The Method of Protein-Structure Determination by Nmr in Solution - Initial New Insights Relating to Molecular Mobility*. Biological Chemistry Hoppe-Seyler, 1988. **369**(4): p. 200-201.
108. Wuthrich, K., *Protein-Structure Determination in Solution by Nmr-Spectroscopy*. Journal of Biological Chemistry, 1990. **265**(36): p. 22059-22062.
109. Wuthrich, K., *The Ramachandran Plot and the Nmr Method for Protein-Structure Determination*. Current Science, 1990. **59**(17-18): p. 825-831.
110. Wuthrich, K., *6 Years of Protein-Structure Determination by Nmr-Spectroscopy - What Have We Learned*. Ciba Foundation Symposia, 1991. **161**: p. 136-149.
111. Xu, Y., et al., *A computational method for NMR-constrained protein threading*. Journal of Computational Biology, 2000. **7**(3-4): p. 449-467.
112. Xu, Y., et al., *Automatic assignment of NOESY cross peaks and determination of the protein structure of a new world scorpion neurotoxin using NOAH/DIAMOD*. Journal of Magnetic Resonance, 2001. **148**(1): p. 35-46.
113. Zheng, D.Y., et al., *Automated protein fold determination using a minimal NMR constraint strategy*. Protein Science, 2003. **12**(6): p. 1232-1246.

114. Zhukov, I. and A. Ejchart, *NMR spectroscopy in structural proteomics. NMR-based protein structure determination*. Polimery, 2003. **48**(1): p. 28-34.
115. Ma, J. (2005). Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. *Structure* 13, 373-380.
116. Wu, Y., Chen, M., Lu, M., Wang, Q., and Ma, J. (2005). Determining Protein Topology from Skeletons of Secondary Structures. *J. Mol. Biol.*
117. Lu, M., and Ma, J. (2005). The Role of Shape in Determining Molecular Motion. *Biophys. J.*, in press.

CHAPTER 4.

Summary and Future Goals

This thesis describes computational methods for predicting chemical shifts based on minimal protein structural information. Automated prediction program and comprehensive library are developed and integrated in the fields of structural bioinformatics, knowledge-based potential, and Monte-Carlo simulation in order to complement, improve or guide the experimental results in NMR. The main contributions of this thesis are as follows: (a) Structural profile library has been designed to capture significant protein geometrical and energetic properties from atom coordinates. (b) A new computational protocol combined with the profile library has been developed to accurately predict protein C-alpha chemical shifts based on coarse-grained protein three-dimensional structures from any experimental databanks or computational simulations methods. (c) Incorporated with experimentally observed NMR data and partial signal-nuclei assignment, an approach was implemented to recognize native models among group of simulated structures. (d) A novel Monte-Carlo sampling technique has been developed and applied when the assignment is incomplete or not available.

Further research to follow up the results of this thesis could take a number of different directions. In the study of structural profile library, in order to predict chemical shift more accurately, additional comprehensive geometric features may be taken into consider, such as dihedral bond, hydrogen bond, and topology of surrounding Carbon nuclei. These properties can still be readily obtained from

low-resolution models, without adding difficulties to experiments and computational cost to modeling approaches. In the study of native protein structure recognition, other information from biological experiments may be incorporated as supplement to chemical shifts to improve the determination.

In summary, the fundamental goals of research are aimed at the understanding the macro bio-molecule structures. The computational methods developed in this thesis enable such studies and are useful to take advantage of experimental data and to aid computational structure prediction that are beyond conventional means.