COMPUTATIONAL DEVELOPMENT TOWARDS HIGH-THROUGHPUT NMR-BASED PROTEIN STRUCTURE DETERMINATION

By

Woonghee Lee

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

Doctor of Philosophy

(Biochemistry)

at the

UNIVERSITY OF WISCONSIN-MADISON

2013

Date of final oral examination: 04/05/13

The dissertation is approved by the following members of the Final Oral Committee:

John L. Markley, Professor, Biochemistry
Ann C. Palmenberg, Professor, Biochemistry
Julie C. Mitchell, Professor, Biochemistry
Samuel E. Butcher, Professor, Biochemistry
Silvia Cavagnero, Professor, Chemistry

This thesis is dedicated to my wife, Jihyun Park, who has supported me in a long journey seemed endless.

TABLE OF CONTENTS……………………………………………………………………….ii

ACKNOWLEDGEMENTS……………………………………………………………………v

ABSTRACT…………………………………………………………………………………..vi

LIST OF ABBREVIATIONS……………………………………………………………...viii

Chapter 1: Introduction to protein structure determination by NMR spectroscopy………………1

1.1 Introduction to protein 3D structure determination……………………………...……2

1.2 NMR spectroscopy v.s. X-ray crystallography…………………………………………3

1.3 Current high-throughput NMR for protein 3D structures………………………………4

1.4 Contribution to high-throughput NMR…………………………………………………...8

Figures and tables………………………………………………………………………....10

Chapter 2: PINE-SPARKY: graphical interface for evaluating automated probabilistic peak assignments in protein NMR spectroscopy………………………………………………………15

2.1 Abstract…………………………………………………………………………………16

2.2 Introduction……………………………………………………………………………...17

2.3 Implementation…………………………………………………………………………..17

2.3.1 PINE2SPARKY converter………………………………………………………...18

2.3.2 SPARKY extensions……………………………………………………………….18

2.4 Results and conclusion…………………………………………………………………19

2.5 Worldwide dissemination………………………………………………………………20

Figures and tables………………………………………………………………………….22

Chapter 3: PONDEROSA, an automated 3D-NOESY peak picking program, enables automated protein structure determination……………………………………………………………..27

3.1 Abstract…………………………………………………………………………………28

3.2 Introduction……………………………………………………………………………...29

3.3 Implementation……………………………………………………………………..29

3.4 Results and conclusion……………………………………………………………..32

3.5 Application to CASD-NMR…………………………………………………………..33

Figures and tables………………………………………………………………………35

Chapter 4: ADAPT-NMR for Bruker spectrometers and ADAPT-NMR enhancer for visualization ....……………………………………………………………………………………45

4.1 Introduction to the reduced dimensionality and ADAPT-NMR…………………………46

4.2 ADAPT-NMR for Bruker spectrometers………………………………………………...47

4.2.1 Development of Pulse Sequences…………………………………………………47

4.2.2 Preparation of NMR parameter files for TopSpin………………………………...50

4.2.3 Development of the *ADAPT_ORTHO_run* AU program……….………………..51

4.2.4 Development of the *ADAPT_NMR_run* AU program………….………………...52

4.2.5 Preparation of NMR test samples and spectrometers……………………………..53

4.2.6 Results and discussion…………………………………………………………….54

4.2.7 Conclusion………………………………………………………………………...56

4.3 ADAPT-NMR Enhancer………………………………………………………………….57

4.3.1 Abstract…………………………………………………………………………...57

4.3.2 Introduction……………………………………………………………………….58

4.3.3 Implementation……………………………………………………………………59

4.3.4 Results and conclusion……………………………………………………………60

Figures and tables………………………………………………………………………62

Chapter 5: Structure determination and analysis of RV-C02 2A proteinase from strain W12 by NMR spectroscopy…………………………………………………………………………73

5.1 Introduction……………………………………………………………………………...74

5.2 Materials and Methods……………………………………………………………..75

5.2.1 Protein sample preparation………………………………………………..75

5.2.2 NMR spectroscopy…………………………………………………….…76

5.2.3 Structure determination of RV-C02 $2A^{pro}$……………………………………77

5.2.4 Hydrophobic scale measurement of the aromatic residues…………………………78

5.2.5 Spin relaxation and amid exchange experiment……………………………..78

5.2.6 Database accession number………………………………………………79

5.3 Results and discussion……………………………………………………………..79

Figures and tables……………………………………………………………………83

Chapter 6: PACSY, a relational database management system for protein structure and chemical shift analysis……………………………………………………………………………...91

6.1 Introduction………………………………………………………………………..92

6.2 Materials and Methods……………………………………………………………94

6.2.1 Database design……………………………………………………………...94

6.2.2 Software design……………………………………………………………...96

6.3 Results……………………………………………………………………………100

6.3.1 Database build………………………………………………………...…100

6.3.2 Database composition……………………………………………………101

6.3.3 Nomenclature……………………………………………………………101

6.3.4 PACSY statistics………………………………………………………...…101

6.3.5 Practical example……………………………………………………...103

6.4 Conclusions……………………………………………………………………105

Figures and tables…………………………………………………………………..108

Concluding remarks………………………………………………………………………125

Bibliography………………………………………………………………………………127

ACKNOWLEDGEMENTS

I cannot believe how fast time flies! It was 5 years ago that I came to Madison to be a research intern at NMRFAM. I had never before experienced a more beautiful summer. Everything was so memorable, and I remember the sights and sounds of the summer of 2008 as if I were still there. Life in Madison has been wonderful and has brought me the happiest moments of my life. Madison has been a perfect locale for my research; there has been easy access to well maintained, state-of-the art equipment at NMRFAM, which has always been well organized, has been an excellent motivational source for me. The expert facility staff and Markley group members are friendly and willing-to-help. My dissertation would not be so successful without the cooperation of these gracious people, I would like to thank them for their continuous and everlasting help.

I would like to express my sincerest gratitude to Dr. John Markley for his generous guidance, great kindness, and patience. If it were not for his excellent mentoring, as a foreign student in this country, I do not think that I would have been able to finish this Doctorate program. I wish to express my sincere thanks for the support of my knowledgeable committee members; Dr. Ann C. Palmenberg, Dr. Julie C. Mitchell, Dr. Samuel E. Butcher, and Dr. Silvia Cavagnero. I would like to also thank to Dr. Milo Westler for his support as the NMRFAM director. I cannot express my gratitude enough to Dr. Arash Bahrami's for his tremendous help and advice as a senior colleague. Dr. Marco Tonelli and Dr. Jin Hae Kim have always been available to discuss science and other issues when I have needed help. Also, I want to express my gratitude to Dr. Ronnie Frederick, Darius Chow, and Dr. Kaifeng Hu; it was a great experience and pleasure to work with them. Finally, I would like to thank all the people in NMRFAM, BMRB, CESG, and the Markley Lab.

ABSTRACT

Three-dimensional structures of proteins determined in solution by NMR spectroscopy have the unique advantage of revealing details of molecular structure and dynamics in a physiologically relevant state; however, the many tedious steps needed to solve and validate a structure make this method challenging. The barriers to NMR structure determination become higher for larger proteins whose spectra are harder to resolve. It is clear that advances need to be made in automating protein structure determination by NMR spectroscopy. The goal of my research has been to use computational methods to advance the development of high-throughput NMR spectroscopy. Accelerating and streamlining the structure determination process will enable investigators to spend less time solving structures and more time investigating challenging biomolecular systems. My goals have been to develop an automation protocol that integrates multiple steps, ensures the robustness of each step, incorporates iterative corrections, and includes visualization tools to validate and extend the results. I developed PINE-SPARKY as a graphical interface for checking and extending automated assignments made by the PINE-NMR server. ADAPT-NMR directs fast data collection by reduced dimensionality on the basis of ongoing NMR assignments. I helped develop a version of ADAPT-NMR (originally only for Varian spectrometers) for Bruker spectrometers, and I created ADAPT-NMR Enhancer as a visualization tool for validating and extending assignments made by ADAPT-NMR on either spectrometer system. I developed the PONDEROSA package to automate the next steps. PONDEROSA carries out automatic picking of 3D-NOESY peaks and iterative structure determinations with the protein sequence and the assignments as inputs. These automation and visualization tools cover almost all of the steps involved in protein structure determination by NMR spectroscopy. As a practical test of this technology, I solved the structure of the 2A proteinase from the human rhinovirus. As a side project, I built a relational database (PACSY DB) that combines information from the Protein Data Bank (PDB) and the Biological Magnetic Resonance data Bank (BMRB) and incorporates tools for structure analysis. PACSY DB can carry out complex queries that combine atomic coordinates, NMR parameters, and structural features of proteins.

LIST OF ABBREVIATIONS

ADAPT-NMR	Assignment-directed Data collection Algorithm utilizing a Probabilistic Toolkit in NMR
Agilent	NMR spectrometer manufacturer (previously Varian)
ANN	Artificial Neural Network
APSY	Automated Projection Spectroscopy
ARIA	Ambiguous Restraints for Iterative Assignment
ATNOS	Automated NOESY peak picking algorithm
AU	TopSpin macro programming language
AutoAssign	Automated backbone resonance assignments using triple-resonance NMR spectra
AUTOPSY	Automated Peak picking for NMR Spectroscopy
BMRB	Biological Magnetic Resonance Bank
Bruker	NMR spectrometer manufacturer
CANDID	Automated NOE assignment algorithm
CASD-NMR	Critical Assessment of automated Structure Determination by NMR
CASP	Critical Assessment of protein Structure Prediction
CCPN	Collaborative Computing Project for NMR
CCPNmr Analysis	A series of programs for macromolecular NMR spectroscopy integrated with the CCP data model from CCPN
CESG	Center for Eukaryotic Structural Genomics
CHIFIT	A computer program that carries out maximum likelihood parameter estimation in the D-1 indirectly detected dimensions of a D-dimensional NMR data set.
CING	Common Interface for NMR structure Generation
CNS	Crystallography and NMR system is a software library for computational

	structural biology
CS-Rosetta	System for chemical shifts based protein structure prediction using ROSETTA
CYANA	Combined Assignment and Dynamics Algorithm for NMR Applications
DNA	Deoxyribonucleic acid
FLYA	Fully automated protein NMR structure determination algorithm
GARANT	A General Algorithm for Resonance Assignment of Multidimensional Nuclear Magnetic Resonance Spectra
GFT-NMR	G-matrix Fourier Transform (GFT) Projection NMR Spectroscopy
HIFI-NMR	High-resolution Iterative Frequency Identification for NMR
HGP	Human Genome Project
HSQC	Heteronuclear Single Quantum Coherence
MARS	A program for robust automatic backbone assignment of 13C/15N labeled proteins
MATCH	Memetic Algorithm and Combinatorial Optimization Heuristics
MolProbity	All-atom structure validation for macromolecular crystallography
MUNIN	Multidimensional NMR Spectra Interpretation
NESG	Northeast Structural Genomics Consortium
NMR	Nuclear Magnetic Resonance
NMRFAM	The National Magnetic Resonance Facility at Madison
NMRPIPE	A program for Fourier processing of spectra in one to four dimensions as well as a variety of facilities for spectral display and analysis.
NMRVIEW	A program for the visualization and analysis of NMR datasets
NOE	Nuclear Overhauser Effect
NOESY	NOE Spectroscopy
PDB	Protein Data Bank

PICKY	A SVD-based NMR spectra peak picking method
PINE	Probabilistic Inference Network of Evidence
PINE-SPARKY	A graphical interface for evaluating automated probabilistic peak assignments in protein NMR spectroscopy.
PONDEROSA	Peak picking Of Noe Data Enabled by Restriction of Shift Assignments
PROCHECK	A program to check the stereochemical quality of protein structures
PSVS	Protein Structure Validation Software suite
RMSD	Root-Mean-Square Deviation
ROSETTA	Full-chain protein structure prediction server
SPARKY	A graphical NMR assignment and integration program for proteins nucleic acids and other polymers
STRIDE	An algorithm for the assignment of protein secondary structure elements given the atomic coordinates of the protein.
SUMO	Small Ubiquitin-like Modifier
TALOS	Torsion Angle Likelihood Obtained from Shift and sequence similarity
TARA	Topological Assembly of Regions Algorithm
TOPSPIN	Bruker's software package for acquiring processing and analyzing NMR data
TPPI	Time-Proportional Phase Incrementation method
Varian	NMR spectrometer manufacturer
VNMRJ	Varian's Java-based tool for data acquisition and processing
XEASY	A program for interactive computer-supported NMR spectrum analysis
XPLOR-NIH	A structure determination program which builds on the CNS program including additional tools developed at the NIH

CHAPTER 1

Introduction to Protein Structure Determination by NMR Spectroscopy

1.1 Introduction to protein 3D structure determination

We are living in the midst of biological phenomena; the air we breathe, the food we eat, the streets we walk; everything we do is related to biological phenomena. Scientific study has determined that proteins are of major importance in biological systems. The central dogma of molecular biology as stated by Francis Crick (Crick, 1970) reflects the essence of Biology.

"The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid"

Much scientific research of biological phenomena has been developed based on this statement. The results of the Human Genome Project (HGP), carried out from 1990 to 2003, provided information about genomic DNA (deoxyribonucleic acid), which serves as the fundamental library for all biological function in the human body. However, the molecules that actually perform the work in biological systems are the proteins, which are generated from DNA. It is known that the 3D structures of proteins are tightly coupled to biological functions. Even though epigenetics are currently enjoying popularity, there is no argument that proteins are the key actors in biological phenomena.

The study of protein 3D structure determination is an extremely important field in that the 3D structure of a protein is the key to revealing function and biological relevance. DNA and small biologically active organic compounds are usually rigid, fixed structures. Their functions are more related to the interactions that they make with other biological molecules rather than the alteration of their own structures. These structures are relatively easy to predict from their sequences or atom compositions as compared to the relative unpredictability of protein, or RNA structures from their sequence.

Despite the difficulty of solving protein structures, many researchers continue to work on it for the benefits it gives. However, despite their continuous and devoted efforts, the number of 3D structures deposited in the PDB (Protein Data Bank, Berman et al., 2000) ($<8 \times 10^4$) is dwarfed by the DNA sequence entry counts in GenBank (Benson et al., 1993) ($>1 \times 10^8$ entries) as shown in Fig. 1.

1.2 NMR spectroscopy v.s. X-ray crystallography

X-ray crystallography and NMR spectroscopy are popular methods to solve structures according to the composition of the structures in the PDB. About 8×10^4 structures determined by X-ray crystallography are deposited in the PDB, while about 1×10^4 are determined by NMR spectroscopy (by 2012 data deposition). The sum of the two methods is more than 99% of all depositions, while the remainder has been determined by structure prediction, electron microscopy and hybrid methods. The gap between the structures solved by NMR compared to X-ray is mostly because of the limitations in the NMR method. Too many peaks in the limited space cause difficulties in the assignment process because of overlaps, and current available resolutions are unable to distinguish all overlapped peaks. This problem has caused scientists to focus on smaller proteins (<20 kDa) by NMR spectroscopic methods (Fig. 2). Furthermore, automation in NMR spectroscopy is less advanced than in X-ray crystallography. Despite obvious limitations of NMR such as size, resolution, and relatively less automation, it has several strengths over X-ray (Wagner et al., 1992, Table 1); First, protein structures determined by NMR reflect molecular dynamics as experiments are performed in solution, which is more similar to the biological environment. The dynamic features of NMR data offer the ability to study the time scales of intramolecular motions and protein-ligand interactions. NMR can also be need to analyze thermodynamic states. Second, NMR structures do not have any artifacts from crystallization. When a protein crystallizes for X-ray structure determination, some of the surface residues can be perturbed by intermolecular contacts in the crystal, and this effect will not happen for the NMR structure because the protein was in solution. Another special feature of NMR structure is that various constraints can be used for the structure determination such as angle constraints, distance restraints, coupling constants, chemical shifts, and so on. They can be applied to further computational and modeling studies of protein structures as they are all molecular parameters. However, NMR structures can be biased because they are calculated with intensive use of constraint sets, and also require high concentrated protein samples. Nonetheless, NMR structures offer many benefits, and to maximize these benefits, high-throughput NMR is required.

1.3 Current high-throughput NMR for protein 3D structures

Fast and robust NMR structures obtained with the least effort from human intervention, can be achieved by improving the current conventional protocol used for structure determination. The typical steps of structure determination are (Wüthrich et al., 1990), 1) protein sample preparation, 2) NMR data acquisition and processing, 3) peak picking, 4) resonance assignments, 5) conformation restraint collection, 6) structure calculation, refinement and validation (Fig. 3). NMR spectroscopy integrated with computer science has improved these steps except for the protein sample preparation. As a result, many computer programs are used for the individual steps, and some parts are currently automated as well.

TopSpin by Bruker (http://www.bruker-biospin.com/topspin3.html) and VNMR by Varian (http://www.chem.agilent.com) are the two main programs used for NMR data acquisition. NMRpipe from the Bax group (Delaglio et al., 1995) has been the most popular program for NMR spectra processing, while SPARKY (Goddard,T.D. and Kneller,D.G. SPARKY 3, University of California, San Francisco), NMRView (Johnson et al., 1994), CCPNmr Analysis (Chignola et al., 2011) and XEASY (Bartels et al., 1995) have been used for peak picking and assignment. The flexible and extendable features of SPARKY by Python extension customizing make SPARKY unique and popular among users, and it is the most used assignment program according to the BMRB (Biological Magnetic Resonance Bank) deposition (Ulrich et al., 2007).

While the SPARKY type of visual tool provides manual assignment and semi-automated peak picking functions with a simple local maxima algorithm, AUTOPSY (Koradi et al., 1998), PICKY (Alipanahi et al., 2009), MUNIN (Orekhov et al., 2001) and ChiFit (Chylla et al., 1995) offer more advanced automated peak picking methods. A local maxima algorithm, which searches for the local maxima of the intensities in the frequency domain data, is used for all semi-automated and fully-automated peak picking programs except for ChiFit because it is easy to implement; however, it cannot recognize overlapped peaks or discriminate between real peaks and residual peaks. ChiFit models peaks from time domain data, and has the ability to discriminate between a real peak and a noise peak. ChiFit shows better resolution compared to the other approaches, but there are high computing resource requirements, and 3D spectra are not fully supported.

Fast NMR spectrum acquisition has become a challenging field of study because not only are NMR spectrometers costly to purchase and maintain but also spectra from some proteins need to be collected in a short period of time because of their unreliable stability. The reduced dimensionality (RD) approach is the most successful and highly developed method for that. The Szyperski group introduced this approach first (Kim et al., 2003). Their G-matrix Fourier transform (GFT) employs quadrature detection of all simultaneously evolving signals, while the time-proportional phase incrementation (TPPI) method by the Natterer group (Schulte-Herbrüggen et al., 1999) does sequential quadrature detection on the signals of each of the simultaneously evolving types of nuclei. Both methods collect a 45° 2D tilted plane along with two 0°, 90° 2D orthogonal planes and construct 3D peaks by their own methods. However, collecting only one more 2D tilted plane which is fixed to 45° is not enough so the Wüthrich group proposed APSY-NMR (Hiller et al., 2005), which collects more optimal tilted planes, which are optimized before the data collection starts. HIFI-NMR (Eghbalnia et al., 2005) adopts an artificial intelligence approach by predicting the optimal tilt planes to be collected during the experiment interactively by processing spectra and picking peaks. In addition, unlike other approaches, which do not provide visualization tools for investigating the result, the HIFI enhancer has been developed for visual validation and modification of results from HIFI-NMR.

Automation of resonance assignments is another challenging field of study. Many programs such as GARANT (Bartels et al., 1996), MATCH (Volk et al., 2008), AutoAssign (Moseley et al., 2004), MARS (Jung et al., 2004) and PINE-NMR (Bahrami et al., 2009) are publicly available, however, PINE-NMR appears to be the most successful assignment package used so far because it gives not only backbone assignments but also side chain assignments as well as providing a web service so that users do not have to install any program or do any manual parameter settings. Fig. 4 shows how it has been successfully disseminated to the bio-NMR field. PINE-NMR's citation record shows outstanding numbers compared to other programs. Furthermore, PINE's probabilistic approach offers multiple probable assignment choices for a peak, and the feature is very effective and useful along with the PINE-SPARKY visualization tool for fast and robust resonance assignment (Lee et al., 2009).

Torsion angle constraints are used for structure calculation. Some of the most popular programs for deriving constraints are TALOS (Cornilescu et al., 1999) and its descendant, TALOS+ (Shen et al., 2009). A reference database derived from chosen PDB and BMRB entries for coordinates and chemical shifts, respectively, is searched by TALOS for tripeptide sequences, and TALOS provides mean and deviation values of φ and ψ angles for each residue. Other restraints, such as residual dipolar couplings and hydrogen bonds, have not been used as intensively for structure calculation. Some groups (Bax and others) use them for refined structure calculation, and there is a clear need for automated prediction programs for these types of constraints.

Various programs are used for the calculation, validation and refinement of structures. CNS (Brunger et al., 2008) and XPLOR-NIH (Schwieters et al., 2003) are the traditional programs used for both X-ray and NMR, whereas ARIA (Linge et al., 2003), CANDID (Hermann et al., 2002) and CYANA (Güntert 2004) are optimized for NMR structures and offer easier and more convenient set up for users. The internal algorithms are similar, and differences depend on their scoring systems, computation methods, and libraries for simulated annealing. The scoring system adopts preliminary validation functions for the refinements; however, individual validation tools are developed for checking the reliabilities. PROCHECK (Laskowski et al., 1996), included in most calculation programs, is no longer considered the standard. MolProbity (Chen et al., 2010) is recognized as a more faithful validation tool, so that the validation web servers such as PSVS (Bhattacharya et al., 2007) and iCING (Vuister et al., http://nmr.cmbi.ru.nl/icing/) are adopting MolProbity as well as PROCHECK. Validation before deposition into PDB and BMRB is now a required routine step.

In spite of these developments in automation, users still need to understand all the details of each step and what the programs accept and give. In other words, users should be experts at data analysis and interpretation. There is no actual fully automated program that accepts raw spectra as inputs and provides final structure as the output. FLYA (Lopez-Mendez et al., 2006), proposed by the Güntert group, is the only fully-automated protocol leading from frequency domain spectra to structures. However, as the authors mentioned in the paper, false accumulation prevents it from building reliable structures. For example, if one step is not perfect, the next step accepts erroneous information and results in more errors. FLYA tries to relieve this problem by analysis and iteration; however, it seems that it is still not widely accepted according to the PDB deposition history. Only three structure entries (2DCP, 2DCQ, 2DCR) can be found in PDB determined by FLYA and they are also part of the publication.

1.4 Contribution to high-throughput NMR

As described above, a high-quality computer programs for high-throughput NMR are needed. NMRFAM (National Magnetic Resonance Facility at Madison) has contributed to the field for many years. Analysis of the weaknesses of the NMRFAM software suggested the need for supportive programs to work in conjunction with their pre-existing programs. Even though PINE-NMR and HIFI-NMR represented very innovative and cutting-edge technology, their results were difficult to analyze and realize. In addition, some important steps of the process, for example, structure determination, had not been touched by NMRFAM. Thus, my dissertation has focused on promoting high-throughput NMR for protein structure studies. How the contributions have been made will be covered in the following chapters.

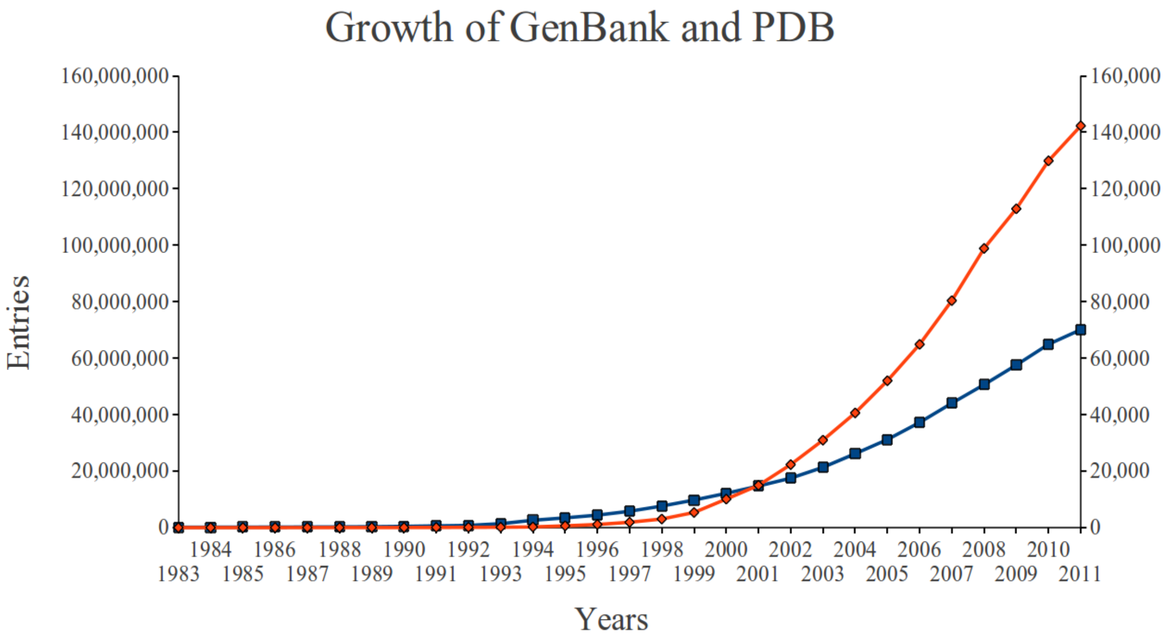


Figure 1. Double y-axis line chart illustrates deposition numbers of GenBank and PDB. Red line and left y-axis shows an annual trend of GenBank deposition, while blue line and right y-axis shows that of PDB.

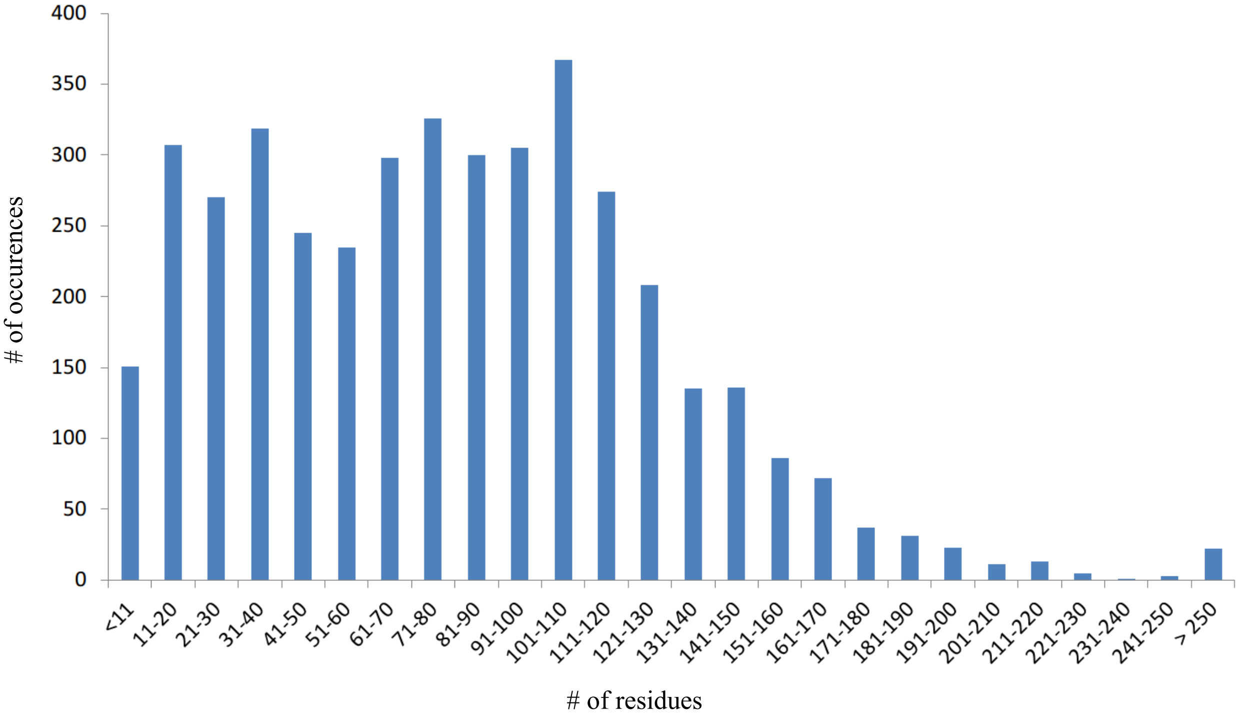


Figure 2. Bars in this histogram indicate the numbers of occurrences in each category of residue counts found in SEQ_DB table of PACSY database. As it can be seen, NMR spectroscopic methods have been used for smaller proteins (>90%) rather than proteins larger than 200 amino acids.

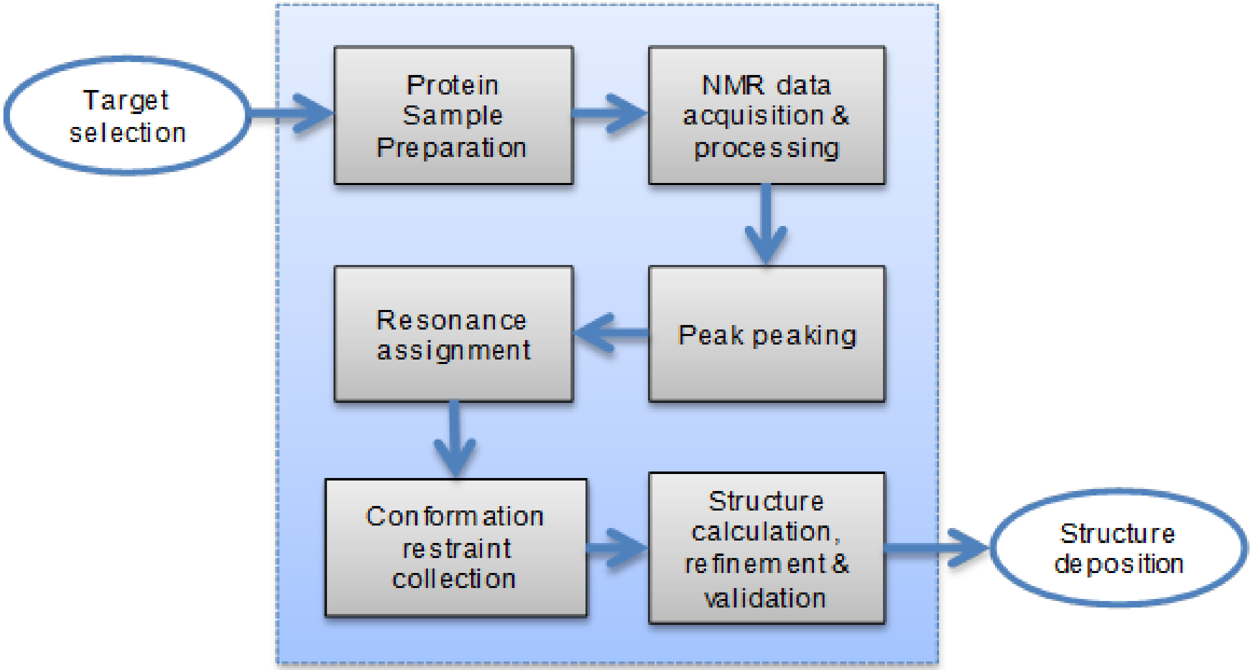

Target selection
Protein Sample Preparation
NMR data acquisition & processing
Peak peaking
Resonance assignment
Conformation restraint collection
Structure calculation, refinement & validation
Structure deposition

Figure 3. The protocol of NMR protein 3D structure determination suggested by the Wüthrich group. Each step inside the dashed-line box requires frequent repetition.

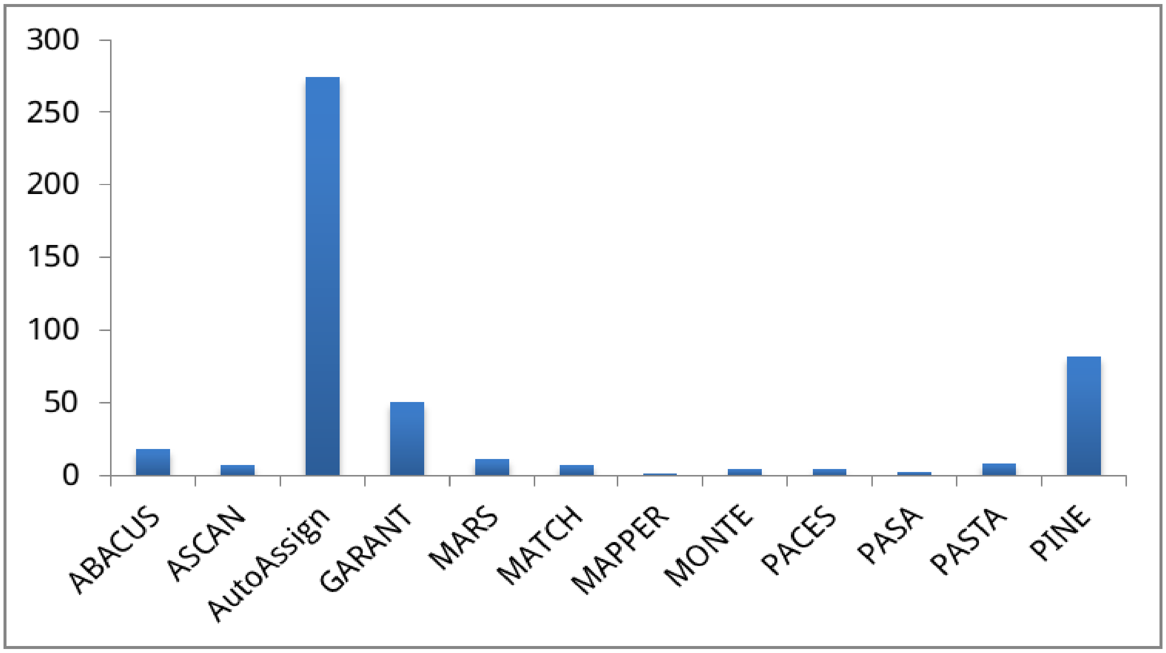


Figure 4. Histogram representing the number of citations from BMRB depositions for each automated resonance assignment program as of Jan.31, 2013. The most widely used program is AutoAssign published in 2004. PINE published in 2009 shows rapid growth in its use.

Table 1. Pros and cons of NMR spectroscopy over X-ray crystallography in protein 3D structure determination.

Pros	Cons
● Mobile proteins in solution	● Size limitation (less than 30kDa)
● More similar to biological environment	● Less automated procedure
● Dynamics study capability	● Requires expertised researcher to analyse the data
● No artifacts from crystallization	● Less structure deposited in PDB ($1\text{x}10^4$ v.s. $8\text{x}10^4$)[1]
● Various restraints and factors can be applied to further computational and modeling studies for prediction	

[1] Data deposition numbers from the PDB and BMRB in 2012.

CHAPTER 2

PINE-SPARKY: Graphical Interface for Evaluating Automated Probabilistic Peak Assignments in Protein NMR Spectroscopy.

Adapted in part from Lee W, Westler WM, Bahrami A, Eghbalnia HR, Markley JL (2009) PINE-SPARKY: graphical interface for evaluating automated probabilistic peak assignments in protein NMR spectroscopy. *Bioinformatics* 25:2085

2.1 Abstract

Summary

PINE-SPARKY supports the rapid, user-friendly and efficient visualization of probabilistic assignments of NMR chemical shifts to specific atoms in the covalent structure of a protein in the context of experimental NMR spectra. PINE-SPARKY is based on the very popular SPARKY package for visualizing multidimensional NMR spectra (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco). PINE-SPARKY consists of a converter (PINE2SPARKY), which takes the output from an automated PINE-NMR analysis and transforms it into SPARKY input, plus a number of SPARKY extensions. Assignments and their probabilities obtained in the PINE-NMR step are visualized as labels in SPARKY's spectrum view. Three SPARKY extensions (PINE Assigner, PINE Graph Assigner, and Assign the Best by PINE) serve to manipulate the labels that signify the assignments and their probabilities. PINE Assigner lists all possible assignments for a peak selected in the dialog box and enables the user to choose among these. A window in PINE Graph Assigner shows all atoms in a selected residue along with all atoms in its adjacent residues; in addition, it displays a ranked list of PINE-derived connectivity assignments to any selected atom. Assign the Best-by-PINE allows the user to choose a probability threshold and to automatically accept as "fixed" all assignments above that threshold; following this operation, only the less certain assignments need to be examined visually. Once assignments are fixed, the output files generated by PINE-SPARKY can be used as input to PINE-NMR for further refinements.

Availability

The program, in the form of source code and binary code along with tutorials and reference manuals, is available at http://pine.nmrfam.wisc.edu/PINE-SPARKY.

2.2 Introduction

Despite rapid progress toward automating many facets of research in structural biology, visualization and expert verification of computational results continue to be required. PINE-NMR (Bahrami et al., 2009) is an automated protein NMR assignment package that accepts, as input, the amino acid sequence of a protein and peak lists associated with defined NMR experiments and provides, as output, probabilistic backbone and side chain assignments and an analysis of the secondary structure. PINE-NMR can accommodate prior information about assignments or stable isotope labeling schemes. PINE-NMR achieves robust and consistent results that have been shown to be effective in subsequent steps of NMR structure determination. In cases where the input data do not support unequivocal assignments (because of weak signals or too many missing signals) PINE-NMR provides multiple ranked possibilities that need to be evaluated. The PINE-SPARKY software package described here provides a graphical interface for reviewing possible assignments in the context of their experimental basis (peaks in multidimensional NMR spectra) and for choosing among them. The software enables the expert to inject additional knowledge into the assignment process in an efficient and straightforward manner.

2.3 Implementation

We selected SPARKY as the viewing and verification tool, because currently it is the most popular NMR visualization and assignment program according to software citations in BMRB

(Ulrich et al., 2007). Another benefit is that SPARKY enables programmers to utilize its internal classes to write Python extensions. PINE-SPARKY consists of two parts: 1) PINE2SPARKY, which converts PINE-NMR assignments and their associated probabilities to SPARKY inputs, and 2) PINE. SPARKY extensions, which support intuitive interfaces that enable various visualization and assignment tasks.

2.3.1 PINE2SPARKY converter

Multiple assignments and their probabilities (output from PINE-NMR) are converted into labeled objects (Fig. 2A), and these objects are incorporated into SPARKY save files by the PINE2SPARKY converter (Fig. 1A). After the user chooses which assignment is correct, the incorrect labels can be removed. Colors of the labels are associated with the level probability. These can be configured by the user, but the default spectrum is blue for the highest probability and red for the lowest. We developed PINE2SPARKY under Lazarus, an IDE of Free Pascal, and the software is compatible with multiple operating systems (MS Windows, MacOSX, and Linux).

2.3.2 SPARKY extensions

PINE Assigner is a dialog box. The peak to be analyzed is selected prior to opening the dialog box. The dialog box lists all possible assignments for that peak (Fig. 2A) and contains buttons that simplify the assignment selection process. Each buttons is labeled with its function (Update, Assign, Best probability, Unassign, Floating labels, Graph, Stop, Close).

PINE Graph Assigner is graphical window consisting of four parts: the covalent structural representation of a tripeptide, a list of spectra associated with different NMR experiments that PINE-NMR used for the assignment (Fig. 2B), buttons with defined functions (Previous residue, Next residue, Update, Assign, Unassign, Close), and list of labels. When the user chooses a residue from the protein sequence, the graphical window displays all the atoms in that residue as well as the atoms in the residues sequentially to either side. Atoms with assignments are color coded (yellow for 1H, red for 13C, blue for 15N); gray denotes atoms that PINE-NMR was unable to assign. Chemical shifts and their standard deviations associated with the assignments are displayed below and to the right of each assigned atom. When the user clicks on an individual atom and a spectrum, PINE Graph Assigner displays a ranked list of PINE-derived assignment connectivities to that atom from that spectrum. By going to the spectrum view, the user sees a list of available peak labels associated with the chosen atom. One can assign or unassign peaks with a few mouse clicks. The list of spectra includes only those currently loaded into PINE-SPARKY.

Assign the Best by PINE enables the user to bypass the manual steps needed to fix assignments. The user can choose a threshold, such as 90%, and Assign the Best by PINE will fix all assignments with probabilities greater than or equal to this value (Fig. 2C).

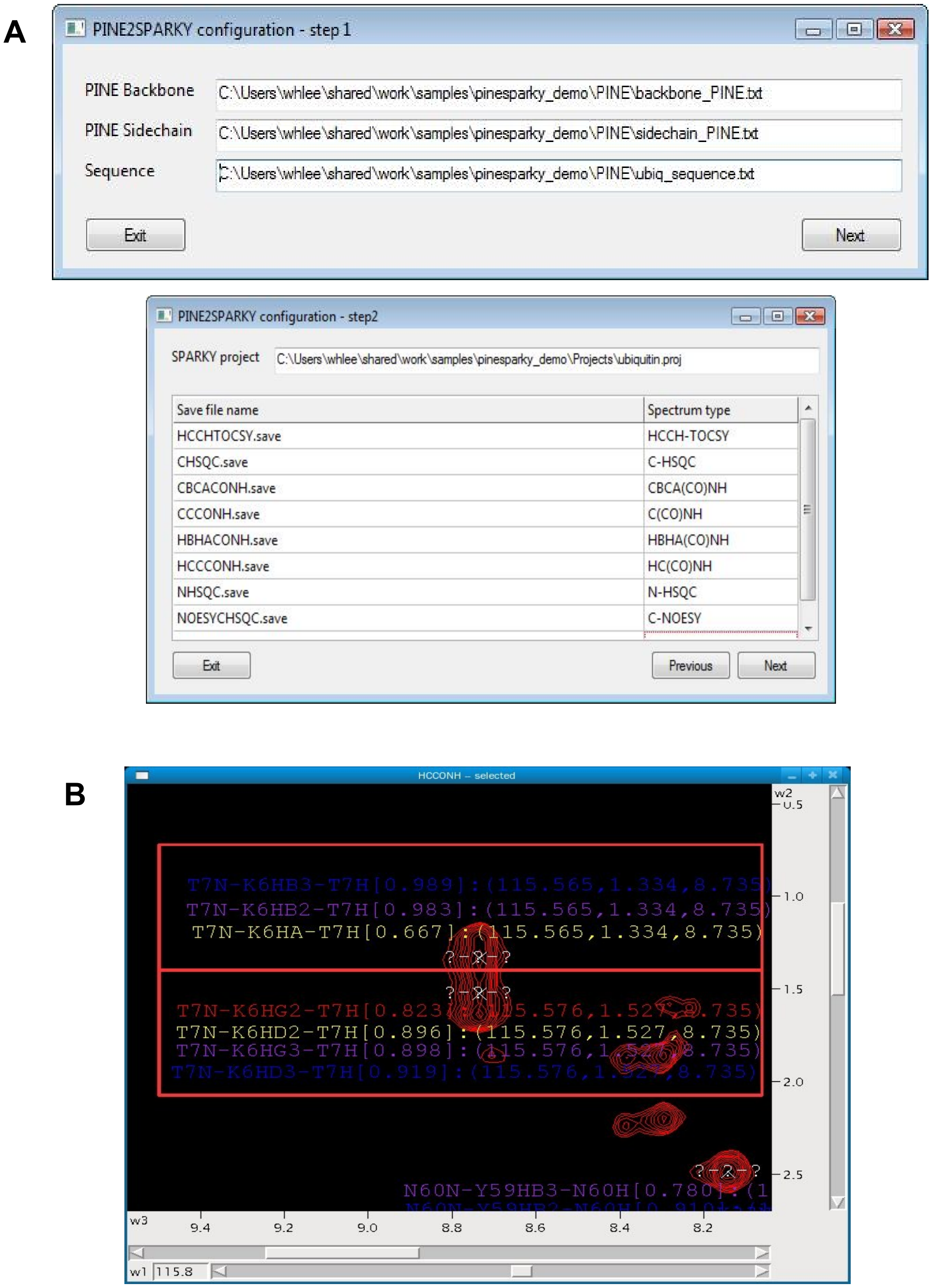
2.4 Results and Conclusion

We used NMR data from the 76-residue protein, human ubiquitin, to illustrate the use of PINE-SPARKY in a structure determination project. ^{1}H-^{15}N HSQC, ^{1}H-^{13}C HSQC, CBCA(CO)NH, and HBHA(CO)NH data sets were collected to support backbone assignments, and (H)CC(CO)NH, H(CC)(CO)NH, and HCCH-TOCSY data sets were collected to support sidechain assignments. ^{15}N-edited NOESY and ^{13}C-edited NOESY data sets were used in a subsequent structure determination. NMRpipe (Delaglio et al., 1999) was used to process all

NMR spectra, and NMRdraw (Delaglio et al., 1999) was used to pick peaks in all but the NOESY data sets. ATNOS (Herrmann et al., 2002) was used to pick NOESY peaks. We generated a SPARKY project and save files with the processed spectra. PINE-NMR was used to generate probabilistic assignments, and these were uploaded via the PINE2SPARKY converter. Tolerances for ^{13}C and ^{15}N were set at 0.4 ppm, and that for 1H was set to 0.03 ppm. Overall assignment quality assessed has been presented in Table 1. Assign the Best by PINE was performed with a threshold of 0.9 (90%) with all (non NOESY) NMR spectra. Peaks that remained unassigned after that process were assigned with PINE Graph Assigner and PINE Assigner. Assign the Best by PINE with 0.9 threshold assigned more than 90% of the peaks automatically. After this the procedure, it was possible to quickly assign the remaining peaks with small number of clicks using PINE Graph Assigner. TALOS (Cornilescu et al., 1999) was used to determine torsion angle constraints from the assigned chemical shifts: 106 torsion angles involving 53 residues were judged to be "good" by TALOS, and these were used constraints along with the NOESY data in 3D structure calculations by CYANA (Güntert, 2004). In the resulting 20 best structures, the root mean standard deviation was 0.46 Å for backbone atoms and 1.22 Å for all heavy atoms in the structured regions (Fig. 2D). The following is an analysis of the time required to determine the structure following initial data collection: PINE-NMR run (~1h), PINE-SPARKY analysis (30m), TALOS analysis (20m), CYANA structure determination (7m) with 16 CPUs.

2.5 Worldwide dissemination

The PINE-SPARKY package has been available to download since 2009 from webpage (http://pine.nmrfam.wisc.edu/pine-sparky). Quick start tutorial, reference manual, and sample test results are also available from the webpage. In order to understand how the program in being used, the IP addresses where the download has been made have been back-traced. The locations are plotted on the world map in Fig. 3. It can be seen that the locations wide-spread. A survey on PINE-SPARKY has been sent to users via email, and the results show how users feel about the effectiveness of PINE-SPARKY in practical applications (Table 2). 15 out of 18 responders (83%) thought PINE-SPARKY was helpful and the remainder (17%) did not know how to use the package (Table 2A). All participants who could figure out how to use the PINE-SPARKY package responded that the assignment job was completed in fewer than 30 days. Two of them (13%) responded they could finish the job in a day (Table 2B). Most of the participants (81%) thought PINE-SPARKY package outperformed SPARKY alone or was superior to other packages (Table 2C). In addition, all of them would like to use PINE-SPARKY in the future for their projects (Table 2D). As it can be seen in the survey, PINE-SPARKY has been found to be an efficient assignment procedure for structure determination by NMR spectroscopy. The increasing citation number of PINE-SPARKY also supports this.


A
B
Exit
Next
Previous
HCCONH – selected
w2
0.5
1.0
1.5
2.0
2.5
w3
9.4
9.2
9.0
8.8
8.6
8.4
8.2
w1 115.8
T7N-K6HB3-T7H[0.989]:(115.565,1.334,8.735)
T7N-K6HB2-T7H[0.983]:(115.565,1.334,8.735)
T7N-K6HA-T7H[0.667]:(115.565,1.334,8.735)
?-?-?
?-?-?
T7N-K6HG2-T7H[0.823]:(115.576,1.527,8.735)
T7N-K6HD2-T7H[0.896]:(115.576,1.527,8.735)
T7N-K6HG3-T7H[0.898]:(115.576,1.527,8.735)
T7N-K6HD3-T7H[0.919]:(115.576,1.527,8.735)
?-?-?
N60N-Y59HB3-N60H[0.780]:(1

PINE2SPARKY configuration - step 1

PINE Backbone C:\Users\whlee\shared\work\samples\pinesparky_demo\PINE\backbone_PINE.txt

PINE Sidechain C:\Users\whlee\shared\work\samples\pinesparky_demo\PINE\sidechain_PINE.txt

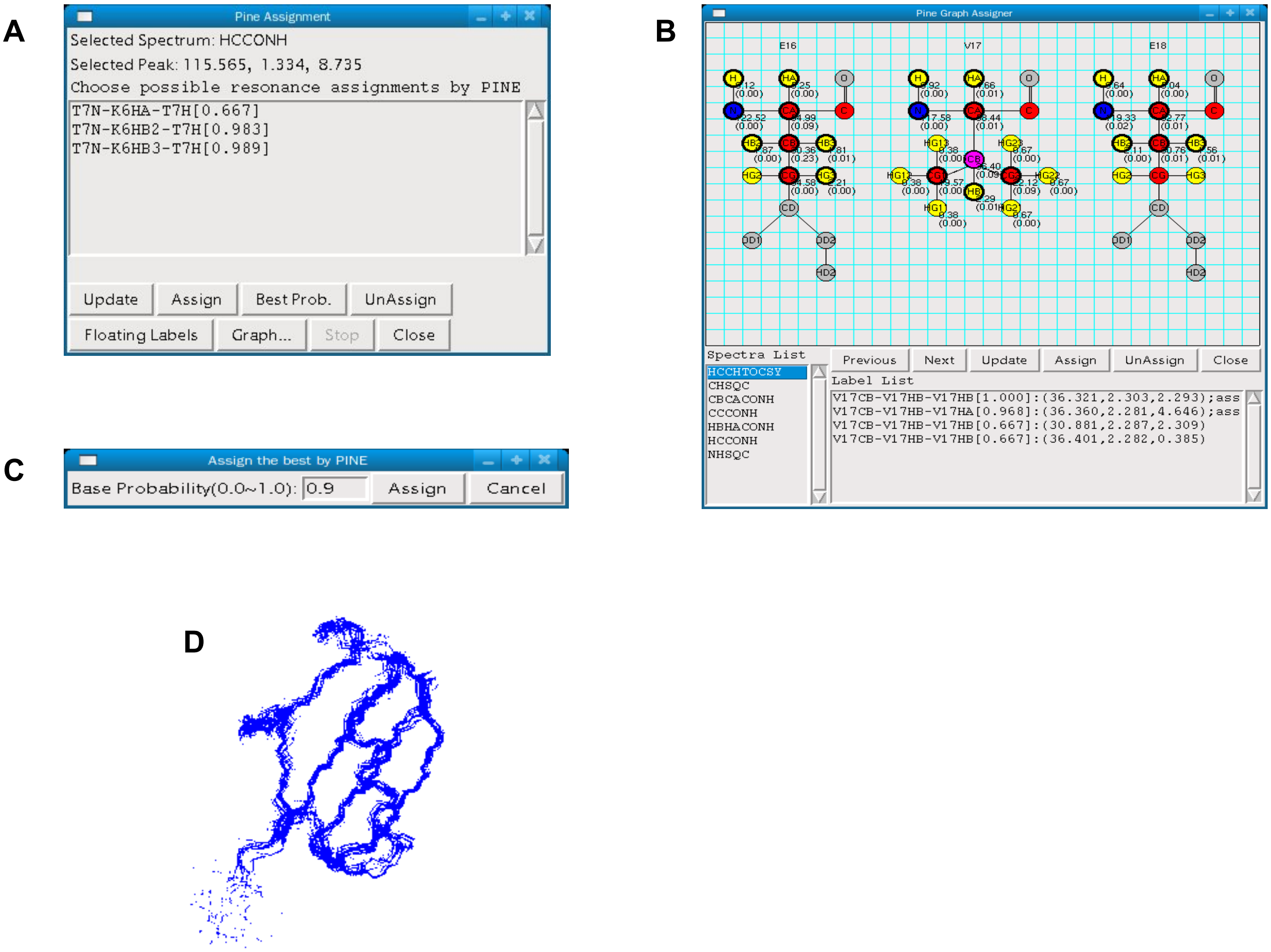
Sequence C:\Users\whlee\shared\work\samples\pinesparky_demo\PINE\ubiq_sequence.txt

PINE2SPARKY configuration - step2

SPARKY project C:\Users\whlee\shared\work\samples\pinesparky_demo\Projects\ubiquitin.proj

Save file name	Spectrum type
HCCHTOCSY.save	HCCH-TOCSY
CHSQC.save	C-HSQC
CBCACONH.save	CBCA(CO)NH
CCCONH.save	C(CO)NH
HBHACONH.save	HBHA(CO)NH
HCCCONH.save	HC(CO)NH
NHSQC.save	N-HSQC
NOESYCHSQC.save	C-NOESY

Figure 1. PINE2SPARKY converter incorporates PINE probabilistic assignment results into the SPARKY projects to make them useable by SPARKY extensions. (A) screen shots of PINE2SPARKY converter user interface. (B) a screen shot of SPARKY with the incorporated PINE probabilistic assignments.



Selected Spectrum: HCCONH
Selected Peak: 115.565, 1.334, 8.735
Choose possible resonance assignments by PINE

```
T7N-K6HA-T7H[0.667]
T7N-K6HB2-T7H[0.983]
T7N-K6HB3-T7H[0.989]
```

Spectra List

```
HCCHTOCSY
CHSQC
CBCACONH
CCCONH
HBHACONH
HCCONH
NHSQC
```

Label List

```
V17CB-V17HB-V17HB[1.000]:(36.321,2.303,2.293);ass
V17CB-V17HB-V17HA[0.968]:(36.360,2.281,4.646);ass
V17CB-V17HB-V17HB[0.667]:(30.881,2.287,2.309)
V17CB-V17HB-V17HB[0.667]:(36.401,2.282,0.385)
```

Base Probability(0.0~1.0): 0.9

Figure 2. SPARKY extensions of PINE-SPARKY package and 3D ubiquitin structure calculated by using PINE-SPARKY. (A) PINE Assigner dialog. (B) PINE Graph Assigner dialog. (C) Assign the Best by PINE threshold dialog. (D) Calculated protein 3D structures of ubiquitin using PINE-SPARKY.

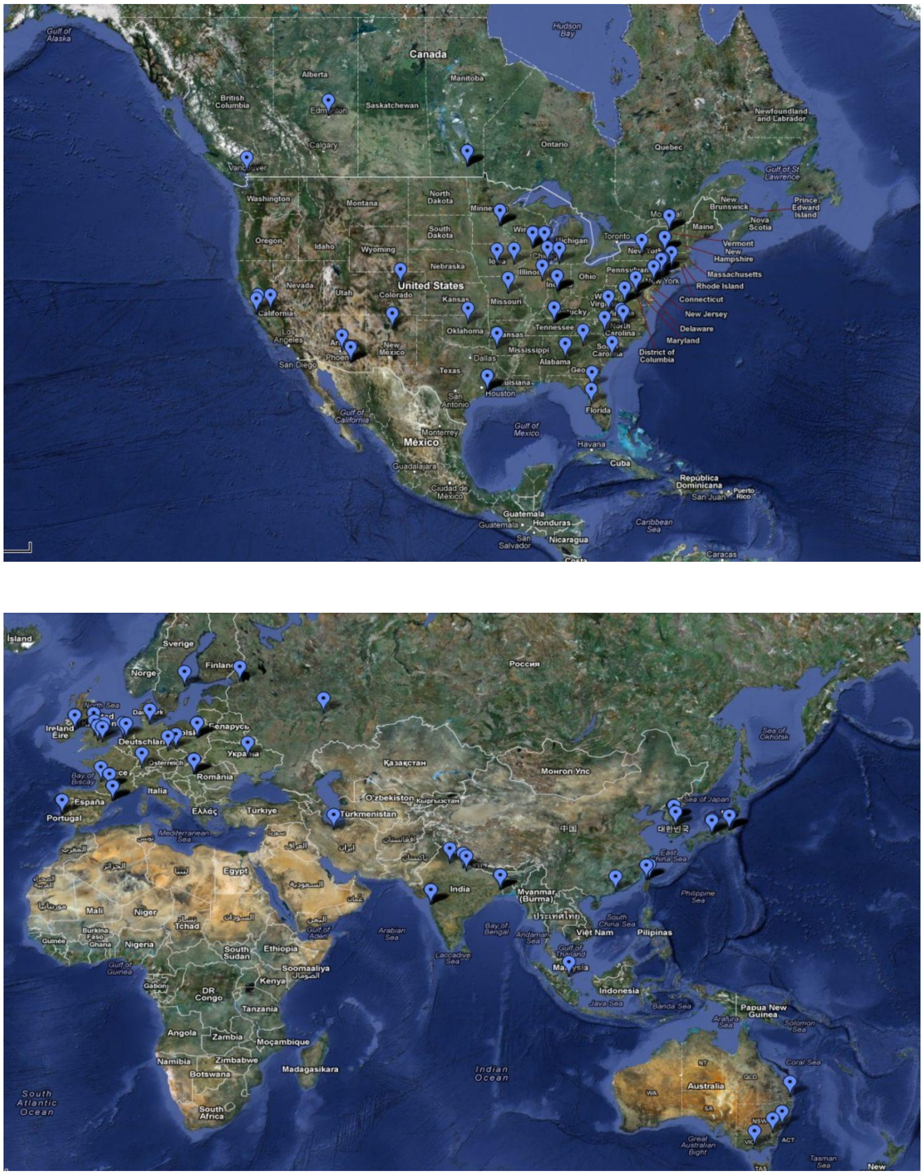


Figure 3. The locations where the PINE-SPARKY package has been downloaded. As it shows, the download map to locations where NMR facilities exist. There has not been any download made from Africa, South America and New Zealand.

Table 1. Quality assessment of PINE-SPARKY assignment used for ubiquitin. 88% assignment completeness could have been reached by PINE-SPARKY without any manual intervention, while 95% could have been reached by using one-hour of manual intervention with PINE-SPARKY tools.

	Ideal assignments(%)	Automatically selected(%)[1]	Manually selected(%)[2]	Auto/Manual(%)
All assignments	590(100)	520(88.136)	562(95.254)	520/562(92.527)
Backbone assignments	287(100)	272(94.774)	276(96.167)	272/276(98.551)
Sidechain assignments	303(100)	248(81.848)	286(94.389)	248/286(86.713)

[1] 90% of probability is the standard cut-off for the Assign-the-best-by-Pine function in the PINE-SPARKY plug-in.

[2] Additional one-hour work of PINE Graph Assigner and PINE Assigner was conducted for the validation of the automatic assignment.

Table 2. A part of the survey results from PINE-SPARKY users (*Jan 20, 2011 – Nov 09, 2012. Total participants: 18 people*)

A. Which of the following best describes your most recent (or typical) PINE-SPARKY experience?

1	PINE-SPARKY really helped me by reducing assignment time and it is easy to use.	11	61%
2	It took me a while to learn PINE-SPARKY, but I found it helpful afterwards.	4	22%
3	Installing PINE-SPARKY was successful, but I have no idea how to use it.	3	17%
4	I can't even install it.	0	0%
5	Others, please specify.	0	0%
	Total	18	100%

B. If you succeeded on installing and using PINE-SPARKY, how long did it take you to assign all the spectra?

1	Over 30 days.	0	0%
2	7-30 days.	7	47%
3	1-7 days.	4	27%
4	less than one day.	2	13%
5	Others, please specify.	2	13%
	Total	15	100%

C. How does PINE-SPARKY compare to SPARKY alone or combination of SPARKY with other assignment packages?

1	Better	13	81%
2	Same	2	13%
3	Worse	0	0%
4	I have not tried other assignment packages.	1	6%
	Total	16	100%

D. Do you plan on using PINE-SPARKY again in the future?

1	Yes	16	100%
2	No	0	0%
	Total	16	100%

CHAPTER 3

PONDEROSA, an Automated 3D-NOESY Peak Picking Program, Enables Automated Protein Structure Determination.

Adapted in part from Lee W, Kim JH, Westler WM, Markley JL (2011) PONDEROSA, an automated 3D-NOESY peak picking program, enables automated protein structure determination. *Bioinformatics* 27:1727

3.1 Abstract

Summary

PONDEROSA (Peak-picking Of Noe Data Enabled by Restriction of Shift Assignments) accepts input information consisting of a protein sequence, backbone and sidechain NMR resonance assignments, and 3D-NOESY (^{13}C-edited and/or ^{15}N-edited) spectra, and returns assignments of NOESY crosspeaks, distance and angle constraints, and a reliable NMR structure represented by a family of conformers. PONDEROSA incorporates and integrates external software packages (TALOS+, STRIDE and CYANA) to carry out different steps in the structure determination. PONDEROSA implements internal functions that identify and validate NOESY peak assignments and assess the quality of the calculated three-dimensional structure of the protein. The robustness of the analysis results from PONDEROSA's hierarchical processing steps that involve iterative interaction among the internal and external modules. PONDEROSA supports a variety of input formats: SPARKY assignment table (.shifts) and spectrum file formats (.ucsf), XEASY proton file format (.prot), and NMR-STAR format (.star). To demonstrate the utility of PONDEROSA, we used the package to determine 3D structures of two proteins: human ubiquitin and Escherichia coli iron-sulfur scaffold protein variant IscU(D39A). The automatically generated structural constraints and ensembles of conformers were as good as or better than those determined previously by much less automated means.

Availability

The program, in the form of binary code along with tutorials and reference manuals, is available at http://ponderosa.nmrfam.wisc.edu/.

3.2 Introduction

A major challenge of structural biology is to close the gap between known sequences of proteins [>1× 10^8 in GenBank (Benson et al., 2008)] and their 3D structures (~ 1× 10^5 in PDB; Berman et al., 2000). Automation now plays a key role in speeding up the determination of protein structures by X-ray crystallography. However, the determination of protein structures by NMR spectroscopy includes a larger number of steps that present greater challenges for automation. The steps basically are sequential; however, some of them may need to be iterated in order to yield a satisfactory protein structure. Software packages have been developed to automate individual steps, and in some cases to pipeline several steps (Bahrami et al., 2009; Lopez-Mendez and Güntert, 2006). One of the challenges has been to automate the final steps beyond backbone and sidechain peak assignment, including the determination of torsion angle constraints, the assignment of NOESY cross peaks and the determination of distance constraints, the analysis of secondary structure, and the calculation of a validated 3D protein structure. The PONDEROSA (Peak-picking Of Noe Data Enabled by Restriction of Shift Assignments) software package described here bridges this gap and is meant to be used with an automated resonance assignment package such as PINE-NMR introduced earlier by our group (Bahrami et al., 2009).

3.3 Implementation

PONDEROSA (Fig. 1A) accepts resonance assignments in popular file formats (SPARKY; T.D. Goddard and D. G. Kneller, SPARKY 3; University of California, San Francisco, XEASY; Bartels et al., 1995 or NMR-STAR; http://www.bmrb.wisc.edu/dictionary/), an amino acid sequence file in either one- or three-letter code, and ^{13}C-NOESY and/or ^{15}N-NOESY datasets in SPARKY (.ucsf) format. By integrating internal functions and external programs, PONDEROSA provides as output NOE peak lists, NOE assignments, structural constraints and a family of conformers representing the 3D structure.

Internal functions: The major internal functions of PONDEROSA simulate and validate NOESY peaks and manage interactions among the internal and external software routines. PONDEROSA uses available resonance assignments to simulate all possible short, medium- and long-range peaks (Fig. 2 and 3). The NOESY simulation starts with the ^{1}H-^{15}N assignments and ^{1}H-^{13}C assignments. For example, ^{15}N-NOESY, on the basis of a ^{1}H-^{15}N assignment, a region is identified in the ^{15}N plane of the 3D ^{15}N-NOESY spectrum that should contain the ^{1}H diagonal (circle, Fig.2A). The peak within the circle in Fig.2A is identified on the basis of a local maximum (x in box), and positions of possible intra-residue NOESY cross peaks are identified from the list resonances assigned ^{1}H to that residue (circles, Fig.2B). Local maxima within the cross peak regions are identified (x's). 3D ^{13}C-NOESY data are analyzed in a similar way to identify intra-residue NOE peaks (Fig.2C). The identification of inter-residue NOE peaks follows (Fig.3). At the position of each diagonal peak in a selected 3D plane (small black circles), the positions of all possible cross peaks are identified from the list of ^{1}H peak assignments. Those previously identified as intra-residue NOEs (blue circles) are differentiated from all others (yellow circles). The yellow circles represent a simulation of all possible inter-residue NOE peaks (Fig.3A). Each of the possible peaks (yellow circles) is validated, first, by determining if it contains a local maximum above a given threshold, and second, if it has a local maximum by determining whether a diagonal peak exists within a given tolerance above a given threshold. If these criteria are met, PONDEROSA considers the peak to represent an intra-residue NOE; otherwise it is discarded. For example, peak 1 (blue circle) is confirmed by two matching diagonal peaks (blue boxes) whereas peak 2 lacks a horizontal diagonal peak (red box). The small gray peaks in each panel indicate the region of the diagonal (Fig.3B). Members of the set of simulated peaks are validated by comparing them to peaks detected in the experimental NOESY datasets under different threshold levels. The sets of validated peak lists are provided to the external programs that determine torsion angle restraints, assign NOESY peaks, calculate structures and analyze secondary structure. The results from these programs are recycled to PONDEROSA for the next iteration (Fig. 4).

PONDEROSA examines the effect of the threshold level on a structural quality score that incorporates the root mean standard deviation (RMSD) of backbone atoms in structured regions as determined by STRIDE, the number of constraint and van der Waals violations, and number of residues in favored and disallowed Ramachandran regions. If both ^{13}C- and ^{15}N-edited NOESY data are present, PONDEROSA interactively determines optimal thresholds for each.

External programs: PONDEROSA interacts with TALOS+ (Shen et al., 2009) for identifying structured regions and for determining torsion angle restraints from assigned chemical shifts, STRIDE (Frishman et al., 1995) for analyzing secondary structure, and CYANA (Güntert, 2004) for assigning NOESY cross peaks and calculating 3D structures.

Graphical User Interface: An intuitive graphical user interface (Fig. 1B) enables specification of the number of CPU nodes, steps and cycles to be used in CYANA iterations, the limit on the number of NOESY peaks to be searched for on the basis of local peak maxima, and the weighting factors for RMSD distance violations and torsion angle dispersions.

3.4 Results and conclusion

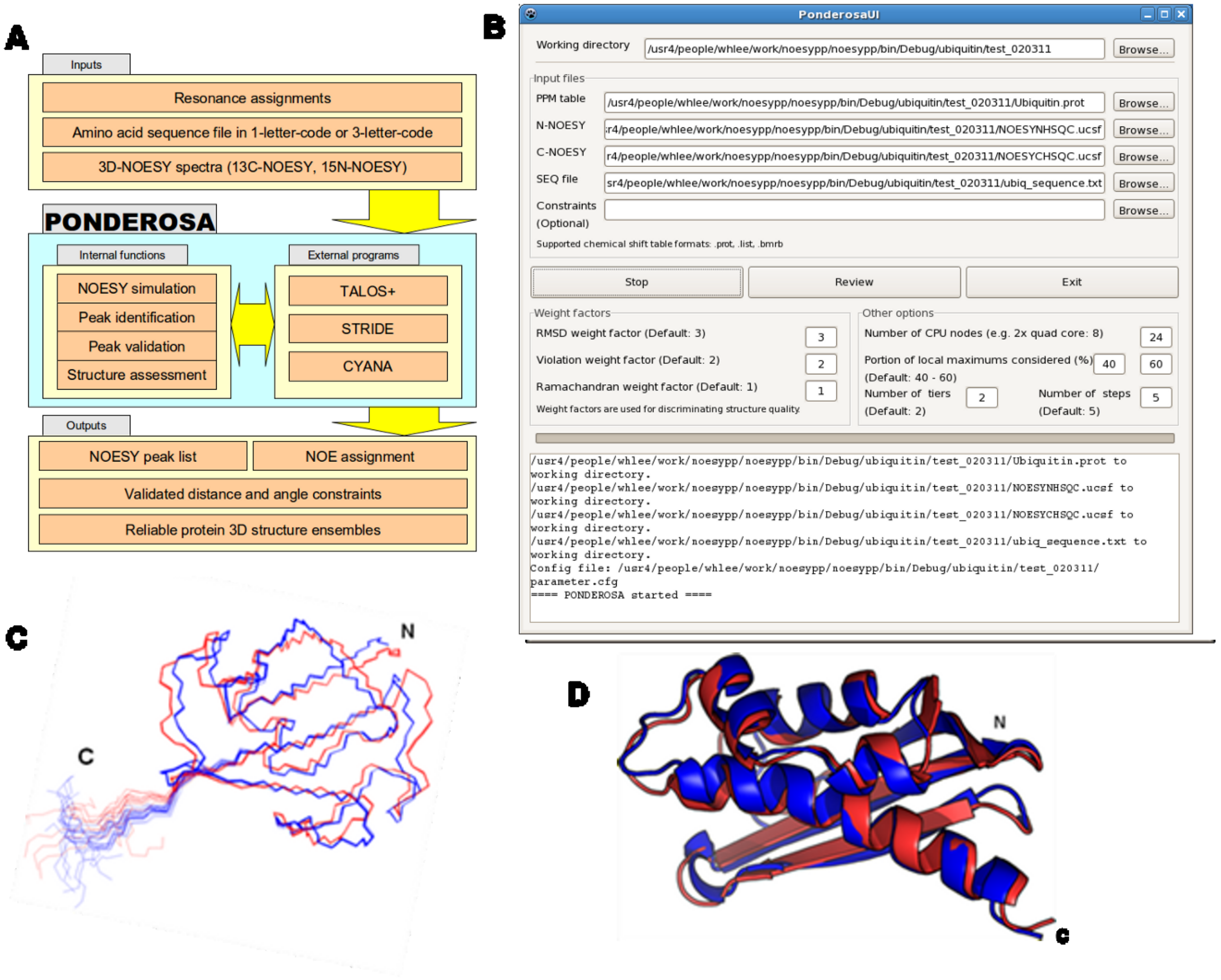
We selected two proteins to illustrate the use of PONDEROSA for NOESY peak picking and automated structure determination: human ubiquitin (76 residues) and Escherichia coli iron-sulfur scaffold protein variant IscU(D39A) (128 residues). We chose human ubiquitin because it is a well-known test sample for protein NMR technology development with 3D structures deposited in the Protein Data Bank (PDB), e.g. 1D3Z (Cornilescu et al., 1998). We chose IscU(D39A) (Kim et al., 2009) because it is a larger protein with a recently deposited non-automatically derived NMR structure (PDB 2KQK) that exhibited variation in the position of secondary structural elements within the family of 20 conformers. In determining the structures of both proteins, we used ^{1}H-^{15}N HSQC, ^{1}H-^{13}C HSQC, CBCA(CO)NH, HNCACB and HBHA(CO)NH datasets for backbone assignments, and (H)CC(CO)NH, H(CC)(CO)NH and HCCH-TOCSY datasets for sidechain assignments. We used NMRpipe (Delaglio et al., 1995) to process all spectra and then converted the spectra to SPARKY (.ucsf) files. We used PINE-NMR and PINE-SPARKY (Lee et al., 2009) to assign the spectra of human ubiquitin, but assigned IscU(D39A) by a manual assignment strategy. We processed 3D ^{13}C-NOESY and ^{15}N-NOESY datasets with NMRPipe and converted the spectra to .ucsf files for input to PONDEROSA. The total times required for the structure determinations with 24 CPUs were 9 h for human ubiquitin and 15 h for IscU(D39A).

The 20 best conformers of human ubiquitin determined by PONDEROSA (Fig. 1C) had a RMSD of 0.09 Å for backbone atoms and 0.48 Å for all heavy atoms in structured regions. The 20 best conformers of IscU(D39A) determined by PONDEROSA had an RMSD of 0.20 Å for backbone atoms and 0.61 Å for all heavy atoms in structured regions. The structures determined by PONDEROSA were very similar to those determined earlier by more manual approaches: 1.15 Å RMSD for the backbone atoms of human ubiquitin (PONDEROSA versus 1D3Z) and 1.30 Å for structured backbone atoms IscU(D39A) (PONDEROSA versus 2KQK) (Fig. 1D). Analysis by two standard validation suites, PSVS (Bhattacharya et al., 2007) and iCing (http://nmr.cmbi.ru.nl/icing/#welcome), revealed that the PONDEROSA-derived structures were of equivalent quality to the structures of the same proteins in the Protein Data Bank (1D3Z and 2KQK) determined by less automated means.

3.5 Application to CASD-NMR

CASD-NMR (Critical Assessment of Automated Structure Determination of Proteins from NMR Data) is a community-wide experiment comparing their automated methods for experimental NMR data analysis (Antonio et al., 2009). This project has been funded by the European Commision within the e-NMR project (www.e-nmr.eu). The goal of this project is to improve each participant's automated method to deploy similar protein structure as close as possible compare to manually refined structure using the same experimental data (Table 3). CASD-NMR committee provides chemical shift assignments, unrefined NOE peak lists and NOESY spectra of determined structures not yet publicly released. Participants show their ability to generate structures by their own automated methods. The first run of CASD-NMR comprising 2010's results has been published (Rosato et al., 2012). However, PONDEROSA has participated in CASD-NMR since 2011 and has deposited four proteins in 2011 (HR6470A, HR6430A, HR5460A, OR36) and another four proteins in 2012 (OR135, StT322, YR313A, HR2876B). The structures from PONDEROSA and PDB-deposited structure have been superimposed to show the difference by using PyMOL program (Fig. 5). Except for HR5460A, the RMSDs between the

PONDEROSA and published structures were less than 3 Å. Even for HR5460A, the fold achieved by PONDEROSA is similar to that published.



Working directory /usr4/people/whlee/work/noesypp/noesypp/bin/Debug/ubiquitin/test_020311

Input files

PPM table /usr4/people/whlee/work/noesypp/noesypp/bin/Debug/ubiquitin/test_020311/Ubiquitin.prot

N-NOESY :r4/people/whlee/work/noesypp/noesypp/bin/Debug/ubiquitin/test_020311/NOESYNHSQC.ucsf

C-NOESY r4/people/whlee/work/noesypp/noesypp/bin/Debug/ubiquitin/test_020311/NOESYCHSQC.ucsf

SEQ file sr4/people/whlee/work/noesypp/noesypp/bin/Debug/ubiquitin/test_020311/ubiq_sequence.txt

Constraints (Optional)

Supported chemical shift table formats: .prot, .list, .bmrb

Weight factors

RMSD weight factor (Default: 3) 3

Violation weight factor (Default: 2) 2

Ramachandran weight factor (Default: 1) 1

Weight factors are used for discriminating structure quality.

Other options

Number of CPU nodes (e.g. 2x quad core: 8) 24

Portion of local maximums considered (%) (Default: 40 - 60) 40 60

Number of tiers (Default: 2) 2

Number of steps (Default: 5) 5

```
/usr4/people/whlee/work/noesypp/noesypp/bin/Debug/ubiquitin/test_020311/Ubiquitin.prot to
working directory.
/usr4/people/whlee/work/noesypp/noesypp/bin/Debug/ubiquitin/test_020311/NOESYNHSQC.ucsf to
working directory.
/usr4/people/whlee/work/noesypp/noesypp/bin/Debug/ubiquitin/test_020311/NOESYCHSQC.ucsf to
working directory.
/usr4/people/whlee/work/noesypp/noesypp/bin/Debug/ubiquitin/test_020311/ubiq_sequence.txt to
working directory.
Config file: /usr4/people/whlee/work/noesypp/noesypp/bin/Debug/ubiquitin/test_020311/
parameter.cfg
==== PONDEROSA started ====
```

Figure 1. Organization of PONDEROSA and its use in automated determination of three-dimensional structures of proteins. (A) Inputs, internal functions and external programs, and outputs. (B) Screen shot of the front end of PONDEROSA. The current version of the front end accepts a single 15N-NOESY and/or a single ^{13}C NOESY data set. However, additional data sets can be entered by editing a configuration file in the command line as described in the PONDEROSA web page. (C) Comparison of the family of 20 conformers representing the structure of human ubiquitin: (blue) PONDEROSA-derived structure; (red) previously deposited structure (1D3Z). MOLMOL software (Koradi et al., 1996) was used to superimpose the families of conformers. (D) Comparison of ribbon diagrams representing the structure of IscU(D39A):

(blue) PONDEROSA-derived structure; (red) previously deposited structure (2KQK). PyMOL software (Schrödinger et al., http://www.pymol.org) was used to create the figure.

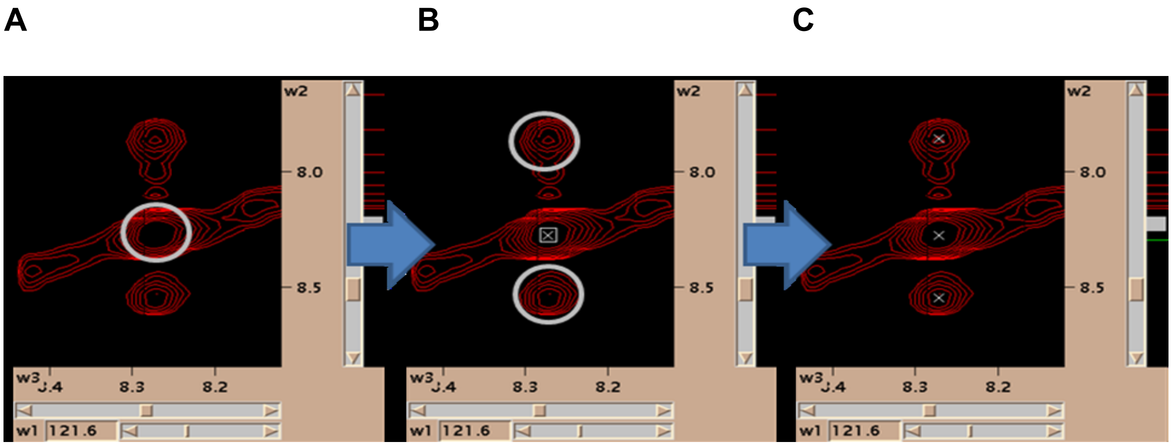


Figure 2. Example showing how PONDEROSA identifies and filters NOESY data to identify intra-residue peaks. (A) On the basis of a ^{1}H-^{15}N assignment, a region is identified in the ^{15}N plane of the 3D 15N NOESY spectrum that should contain the ^{1}H diagonal (circle). (B) The peak within the circle in A is identified on the basis of a local max-imum (× in box), and positions of possible intra-residue NOESY cross peaks are identified from the list resonances assigned ^{1}H to that residue (circles). (C) Local maxima within the cross peak regions are identified (×'s). 3D ^{13}C-NOESY data are analyzed in a similar way to identify intra-residue NOE peaks.

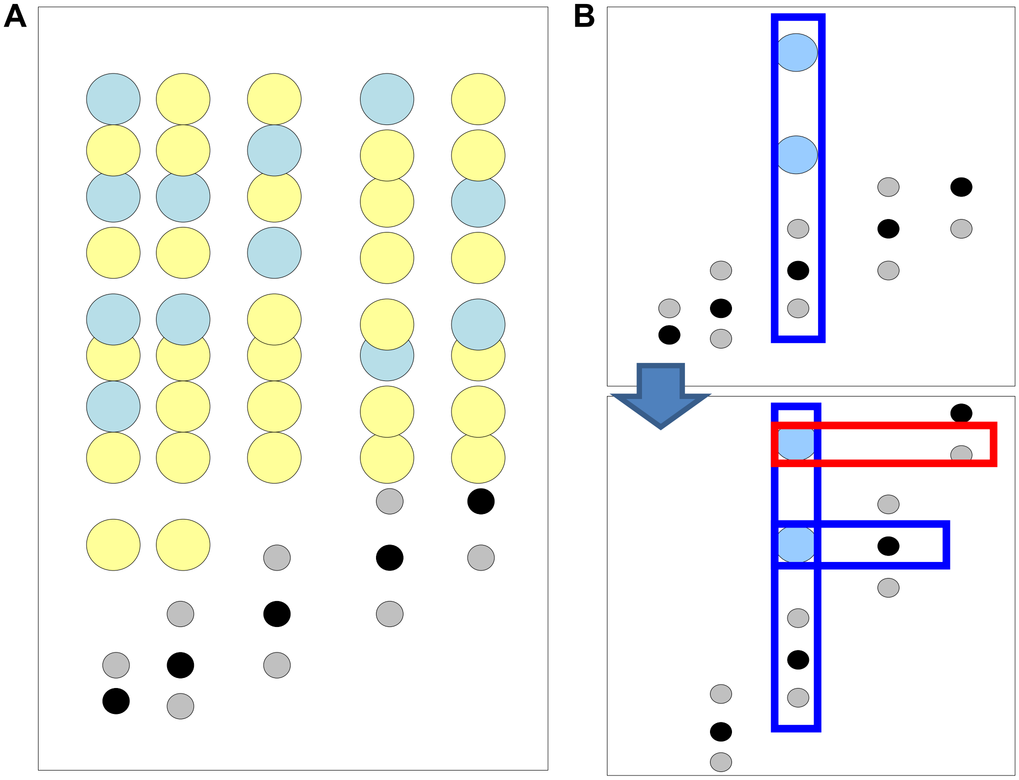


Figure 3. Identification of inter-residue NOE peaks. (A) At the position of each diagonal peak in a selected 3D plane (small black circles), the positions of all possible cross peaks are identified from the list of 1H peak assignments. Those previously identified as intra-residue NOEs (blue circles) are differentiated from all others (yellow circles). The yellow circles represent a simulation of all possible inter-residue NOE peaks. (B) Each of the possible peaks (yellow circles) is validated, first, by determining if it contains a local maximum above a given threshold, and second, if it has a local maximum by determining whether a diagonal peak exists within a given tolerance above a given threshold. If these criteria are met, PONDEROSA considers the peak to represent an intra-residue NOE; otherwise it is discarded. For example, peak 1 (blue circle) is confirmed by two matching diagonal peaks (blue boxes) whereas peak 2 lacks a

horizontal diagonal peak (red box). The small gray peaks in each panel indicate the region of the diagonal.

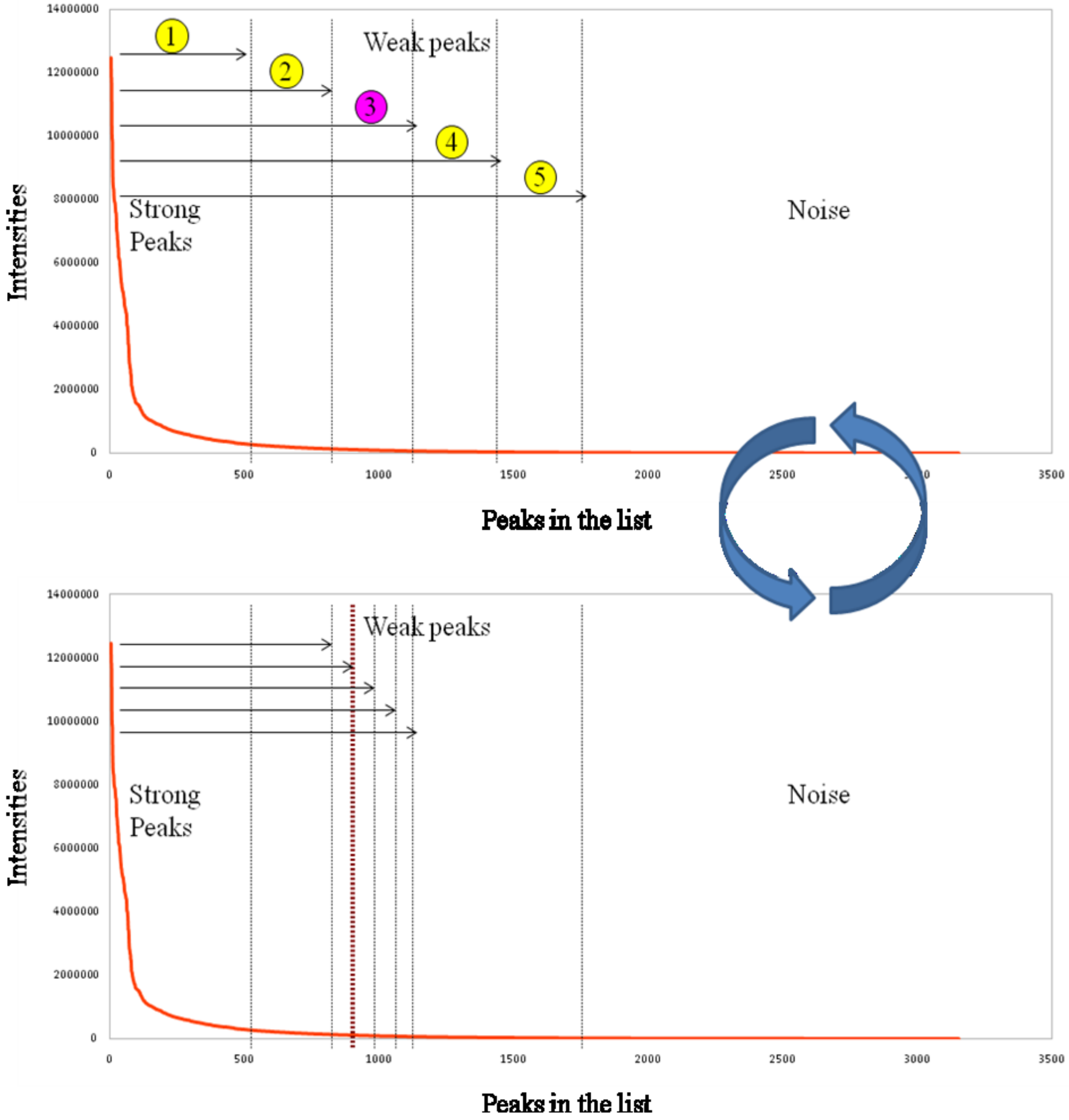


Figure 4. PONDEROSA creates NOESY peak lists at several threshold levels as input for CYANA executions and compares the quality of the resulting structures. (A) Illustration of five widely separated thresholds chosen and the number of NOESY peaks in the list submitted to CYANA. In this case, threshold “3” yielded the highest quality structure. (B) For the next run, PONDEROSA chooses more narrowly spaced thresholds above and below “3” to generate the input peak lists. This process is repeated until the desired quality structure is obtained.

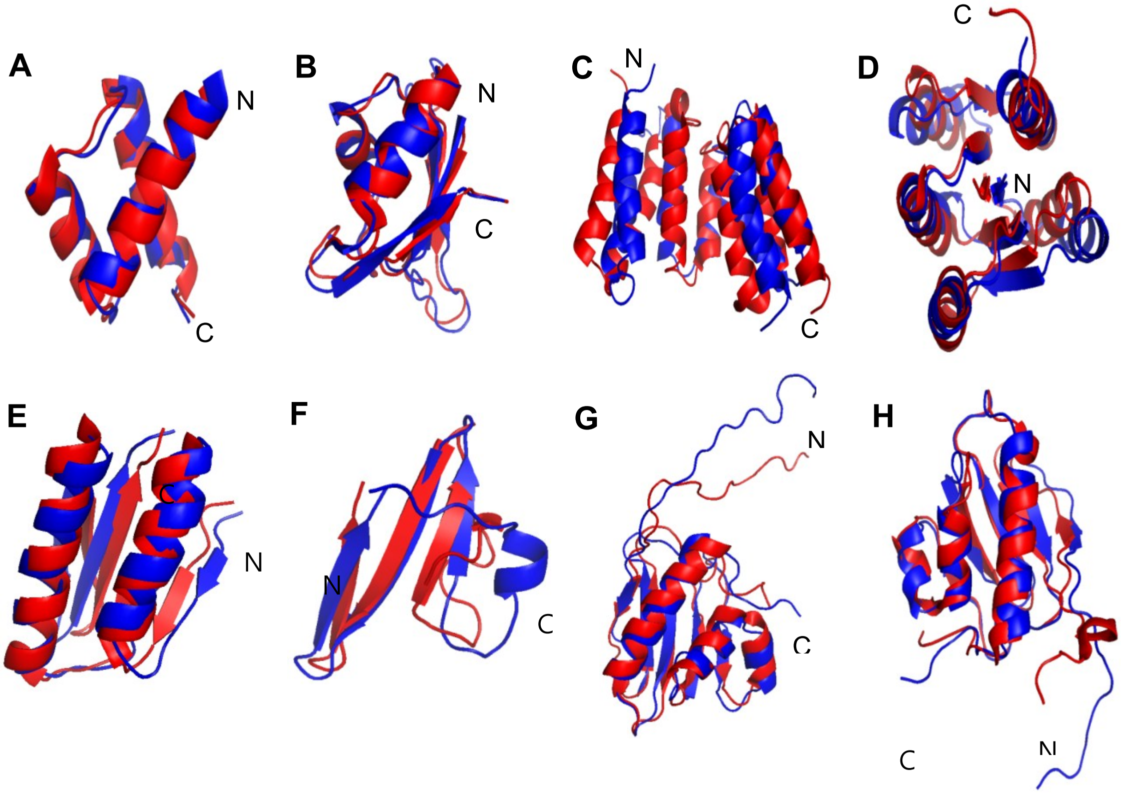


Figure 5. Superimposed structures of CASD-NMR targets and PDB-deposited structures. PDB-deposited structures are illustrated in red, while PONDEROSA calculated structures are illustrated in blue. (A-D) 2011 CASD-NMR targets (E-H) 2012 CASD-NMR targets. (A) HR6470A. Backbone r.m.s.d. (vs 2L9R): 0.8 Å. (B) HR6430A. Backbone r.m.s.d. (vs 2LA6): 0.9 Å. (C) HR5460A. Backbone r.m.s.d. (v.s. 2LAH): 7.2 Å. (D) OR36. Backbone r.m.s.d. (v.s. 2LCI): 1.8 Å. (E) OR135. Backbone r.m.s.d. (v.s. 2LN3): 1.4 Å. (F) StT322. Backbone r.m.s.d. (v.s. 2LOJ): 2.6 Å. (G) YR313A. Backbone r.m.s.d. (v.s. 2LTL): 1.5 Å. (H) HR2876B. Backbone r.m.s.d. (v.s. 2LTM): 1.0 Å.

TABLE 1. Statistics for the NMR structure of human ubiquitin determined by PONDEROSA

Conformationally restricting distance constraints		
Intraresidue [$i = j$]		415
Sequential [$(i- j) = 1$]		463
Medium Range [$1 < (i - j) \leq 5$]		223
Long Range [$(i - j) > 5$]		454
Total		1555
Dihedral angle constraints		
φ		54
ψ		55
Hydrogen-bond constraints		27
CYANA target function [Å]		2.09
Average rmsd to the mean CYANA coordinates [Å]		
Regular secondary structure elements, backbone heavy		0.09
Regular secondary structure elements, all heavy atoms		0.48
Backbone heavy atoms N, Cα, C$'$	(2–71)	0.26
All heavy atoms	(2–71)	0.86
PROCHECK Z-scores (φ and Ψ/all dihedral angles)		-0.35/-3.49
MOLPROBITY Mean score/Z-score		25.79/-2.90
Ramachandran plot summary ordered residue ranges [%]		
Most favored regions		95.7
Additionally allowed regions		4.3
Generously allowed regions		0
Disallowed regions		0
Average number of distance constraint violations per CYANA conformer		
0.2 – 0.5 Å		4
> 0.5 Å		1
Average number of angle constraint violations per CYANA conformer		
> 10°		6

TABLE 2. Statistics for the NMR structure of IscU(D39A) determined by PONDEROSA

Conformationally restricting distance constraints		
Intraresidue [i = j]		417
Sequential [(i– j) = 1]		484
Medium Range [1 < (i – j) ≤ 5]		352
Long Range [(i – j) > 5]		449
Total		1702
Dihedral angle constraints		
φ		101
ψ		101
Hydrogen-bond constraints		56
CYANA target function [Å]		2.38
Average rmsd to the mean CYANA coordinates [Å]		
Regular secondary structure elements, backbone heavy		0.20
Regular secondary structure elements, all heavy atoms		0.61
Backbone heavy atoms N, Cα, C′	(27–126)	0.73
All heavy atoms	(27–126)	1.14
PROCHECK Z-scores (φ and Ψ/all dihedral angles)		0.47/-2.07
MOLPROBITY Mean score/Z-score		28.20/-3.31
Ramachandran plot summary ordered residue ranges [%]		
Most favored regions		99.2
Additionally allowed regions		0.8
Generously allowed regions		0
Disallowed regions		0
Average number of distance constraint violations per CYANA conformer		
0.2 – 0.5 Å		0
> 0.5 Å		4
Average number of angle constraint violations per CYANA conformer		
> 10°		7

TABLE 3. Registered participant lists in CASD-NMR (*as of March 2013*).

Software	Group
csRosetta	Baker and Lange group
CYANA	Güntert group
UNIO	Herrmann group
PONDEROSA	Markley group
AutoStructure	Montellione group
ARIA	Nilges group
CheShire	Vendruscolo group
The WeNMR csRosetta web portal	Bonvin group
I-Tasser	Zhang group

CHAPTER 4

ADAPT-NMR for Bruker Spectrometers and ADAPT-NMR Enhancer for Visualization

Adapted in part from Lee W, Bahrami A, Markley JL (2013) ADAPT-NMR Enhancer: complete package for reduced dimensionality in protein NMR spectroscopy. *Bioinformatics.* 29(4):515-7

Adapted in part from Lee W, Hu K, Tonelli M, Bahrami A, Neuhardt E, Glass KC, Markley JL Fast automated NMR spectra collection and assignment by ADAPT-NMR on Bruker spectrometers. *in preparation*

4.1 Introduction to the reduced dimensionality and ADAPT-NMR

Functional study of proteins has been carried out along with 3D structure determination because of their relevance. Protein structure determination by NMR spectroscopy has been vital for biological science since Kurt Wüthrich and colleagues introduced a serial set of phases to achieve this goal (Wüthrich 1990). The benefits given by NMR structures are obvious such as structure in solution, in short time scale, dynamics research, and application to drug discovery. However, there is still significant room for methods development in biological NMR spectroscopy as NMR is an insensitive technique and currently lacks accessible automated methods to study proteins. Routine protein structure determination by NMR currently is incredibly expensive due to the instrument and human time necessary to solve a structure.

HIFI-NMR (Eghbalnia et al., 2005), introduced by NMRFAM, has been shown to be a successful approach to automating the reduced dimensionality (RD) method for rapid NMR data collection; however, it still misses overlapped or weak peaks especially when applied to larger proteins (>20 kDa). HIFI-Enhancer was developed for the visual validation and modification of constructed peaks and spectra from HIFI-NMR, however the improvement made by HIFI-Enhancer could not be fed back into HIFI-NMR for iterative run to improve the results. These limitations have prevented HIFI-NMR from becoming more popular.

ADAPT-NMR (Bahrami et al., 2012) has been designed to overcome the problems of HIFI-NMR by integrating fast data collection with automated resonance assignment, utilizing methods from HIFI-NMR and PINE-NMR (Bahrami et al., 2011). ADAPT-NMR has been implemented on Varian (Agilent) spectrometers by Arash Bahrami and Marco Tonelli since 2012. Figure 1 illustrates how HIFI-NMR and PINE-NMR are designed to be integrated into ADAPT-NMR.

The initial data collection in ADAPT-NMR is the same as in the HIFI-NMR except for the probabilistic approach made for peak identification. ADAPT-NMR applied the PINE-NMR algorithm to evaluate the quality of the data and to derive resonance assignments. The assessment of the assignment determines which experiment type and tilted planes need to be collected for better quality peak identification and assignment. The initial report for ADAPT-NMR included the results of its application to six selected proteins on Varian (Agilent) spectrometers (Table 1).

4.2 ADAPT-NMR for Bruker spectrometers

The results shows that ADAPT-NMR performed better than sequential use of HIFI-NMR and PINE-NMR led to rapid and robust assignments. It was important to implement ADAPT-NMR for Bruker spectrometers because there are more Bruker than Varian (Agilent) spectrometer users in the world. The MATLAB part of ADAPT-NMR from Varian version which does 2D peak picking, 3D peak generation, tilt angle and experiment type prediction could be reused as it requires NMRpipe frequency domain data generated by NMRPipe. However, in order to support Bruker spectrometers, many aspects of ADAPT-NMR had to be modified because Bruker pulse programming and consoles are different from Varian. For example, Bruker utilizes AU programs as console programs while Varian (Agilent) uses VNMRJ scripts. These differences necessitated the rewriting of both the programs and the way that they are structured.

4.2.1 Development of Pulse Sequences

Dr. Kaifeng Hu from NMRFAM participated in developing pulse sequences for reduced dimensionality on Bruker spectrometers. To accelerate the NMR data acquisition, the 3D experiments HNCO, HN(CA)CB, HNCA, HN(CO)CA, HN(CA)CO, CBCA(CO)NH and C(CCO)NH were modified for reduced dimensionality (2D) through the simultaneous co-evolution of both indirect dimensions, i.e. ^{15}N and ^{13}C. All experiments together with the appropriate acquisition parameter setting are included on our website (http://pine.nmrfam.wisc.edu/ADAPT-NMR/). As an example, here we show in more detailed way how we adapt the conventional HNCO (parameter setting *HNCOGPWG3D*) to an ADAPT-NMR version with a semi-constant time of N co-evolving together with the C dimension with the following modification of MC acquisition:

```
#       ifdef HIFI

        F1PH(calph(ph4, +90), caldel(d0, +in0) & caldel(d10, +in10) & caldel(d29, +in29) &
caldel(d30, -in30) & caldel(d31, +in31))

        F2PH(calph(ph5, +90), caldel(d60, +in60))

#       else

        F1PH(calph(ph4, +90), caldel(d0, +in0))

        F2PH(calph(ph5, +90), caldel(d10, +in10) & caldel(d29, +in29) & caldel(d30, -in30) &
caldel(d31, +in31))

#       endif     /*HIFI*/
```

aqseq is set to 321 in the pulse sequence with N set to be at the 2nd dimension (inner loop). When the *ZGOPTNS* flag *HIFI* is turned on, the real and imaginary part of N is acquired but without any independent time evolution by using dummy delay *d60* and by setting *TD2* to 2. When the *ZGOPTNS* flag is set to *HIFI*, that is, the time for the chemical shift of ^{15}N will co-evolve with the time for the chemical shift of ^{13}C. The original constant time of ^{15}N dimension thus might limit the maximum available evolution time along ^{13}C. To match the possible requirement for high resolution along C dimension, time evolution of N dimension is adapted to a semi-constant time version as the following:

```
#         ifdef   HIFI

"FACTOR2=d30*10000000*2/td1"

"in30=FACTOR2/10000000"

#         else

"FACTOR2=d30*10000000*2/td2"

"in30=FACTOR2/10000000"

#         endif     /*HIFI*/

"if ( in30 > in10 ) { in31 = 0; } else { in31=in10-in30; }"

"if ( in30 > in10 ) { in30 = in10; }"
```

With *"in10=inf2/4"* and *"in29=in10"*, in which *inf2* is defined by the spectral width of ^{15}N dimension. The semi-constant time period is defined by setting *"d30=d23/2+p14/2+d31"*, in which *"d23=16m"* assuming J_{nco} of about 15 Hz. The pulse sequence is shown in Fig. 2., with the main modification highlighted in the broken box. To achieve fast and fully automated NMR data acquisition, processing and NMR signal assignment, all original 3D NMR experiments are adapted to an mode of reduced dimensionality through synchronizing the chemical shift evolution (see the *F1PH* line above with the definition of *HIFI*) of N and C by correlating the time incremental interval *inf2 (N) = 1/SW*$_N$ ** COS(ϑ)* and *inf1 (C) = 1/SWc * SIN(ϑ)*, in which the angle of is defined as the tilted angle between the two orthogonal planes (H-N and H-C planes).

4.2.2 Preparation of NMR parameter files for TopSpin

All these experiments are run at 25 °C on a 500 MHz Bruker AVANCE III spectrometer equipped with a z-gradient 5 mm TCI probe. TopSpin parameter sets for 3D experiments were used for the orthogonal plane data collection. Among all these experiments, the universal carrier position of ^{1}H, ^{15}N, C^{α} (shaped pulse), $C^{aliphatic}$ (C^{α} or C^{β}, shaped pulse), C^{O} (shaped pulse) were applied at 4.76 ppm (H_2O frequency), 118 ppm, 56 ppm, 45 ppm and 176 ppm respectively. 1024, 32, 64, 64 and 64 complex data points with spectral widths of 16 ppm, 36 ppm, 32 ppm, 70 ppm and 22 ppm, respectively, were collected along the ^{1}H, ^{15}N, C^{α}, $C^{aliphatic}$ and C^{O} dimensions. However, they were modified for the better results during actual the orthogonal plane collection as discussed in the section 4.2.6.

To successfully realize automated data acquisition and in-line (on-the-fly) data processing, the data acquisition are manually optimized for better water suppression to obtain optimal signal-to-noise ratio and the last INEPT delay, τ is universally set to 2.3 ms (*d26*, among all NMR experiments) with soft water selective pulse p11 (*sp1*, power level is manually optimized) to be

1ms. This will render a universal phase correction for the direct detected ^{1}H dimension among all different experiments. Furthermore, the receiver phase ph31 is adjusted (flipped) to achieve phasing agreement in data processing among all experiments.

4.2.3 Development of the *ADAPT_ORTHO_run* AU program

The AU program to collect orthogonal 2D NMR spectra, *ADAPT_ORTHO_run*, is designed to conduct automated spectrometer operation and Fourier transformation for orthogonal planes of experiment types stated in the experiment list file (Fig. 3A). Three input files are required to run *ADAPT_ORTHO_run*; *parameters.txt* (ADAPT_NMR parameter file), *ORTHO_list.txt* (experiment list file) and *nmrpipe.par* (NMRpipe parameter file). The experiment list file, *ORTHO_list.txt*, with modified or default parameters such as number of scans, number of increments, carrier positions, spectral widths are designed to be read in the AU program.

As *ADAPT_ORTHO_run* does not really perform any analysis on the spectra such as peak picking or assignment rather than the Fourier transformation, only NMRpipe installation path is taken out from the ADAPT_NMR parameter file. Pre-installation of NMRpipe is necessary for transformation from time-domain data to frequency-domain data. When orthogonal planes are collected by the *ADAPT_ORTHO_run*, Fourier transformation by NMRpipe follows automatically. NMRpipe is called by *ADAPT_ORTHO_run*; thus, some NMRpipe parameters such as phasing, extracting, zero filling, and solvent filters should be set in the NMRpipe parameter file. As Fig. 3A illustrates, the AU program *ADAPT_ORTHO_run* runs on TopSpin (at least version 3.0 and patch level 4 are required) in iterative manner by listed experiments in the experiment list file. During parameter application stage for the acquisition in the AU program, parameters are read from presets installed, but they are updated regarding the experiment list file unless they are set to *DEFAULT* in the file. *ADAPT_ORTHO_run* generates transformation script files for each data directory and executes them for the automated orthogonal plane preparation.

4.2.4 Development of the *ADAPT_NMR_run* AU program

Another AU program, *ADAPT_NMR_run*, was designed to integrate with the ADAPT-NMR and magnet operation module to collect 2D tilted planes. Fig. 3B illustrates how this program works between the NMR spectrometer and ADAPT-NMR on TopSpin. Four input files are required to run this program; *parameters.txt* (ADAPT_NMR parameter file), *ADAPT_list.txt* (experiment list file), *nmrpipe.par* (NMRpipe parameter file), and a protein sequence file. Unlike *ADAPT_ORTHO_run*, all the information from the *parameters.txt* is read by the program for the refined ADAPT-NMR settings such as peak picking, assignment level, and digital resolution. The format of the experiment list is a little bit different from *ORTHO_list.txt*. The *ADAPT_list.txt* does not contain carrier positions and spectral widths for direct and indirect dimensions because *ADAPT_NMR_run* acquires them from the orthogonal planes experimented from *ADAPT_ORTHO_run*. However, number of increment (*ni*) and scan (*nt*) are adjustable separate from defaults for better the resolution. The most important feature of *ADAPT_NMR_run* is that it always runs the ADAPT-NMR engine before collecting data for discovering the best tilt angle and experiment type. For doing that, ADAPT-NMR picks 2D peaks from both orthogonal planes and tilt planes, and constructs peaks from them in the 3D space. If the number of the constructed 3D peaks count from a certain experiment is not sufficient, the ADAPT-NMR determines the best tilt angle to be collected for filling the gap. If the numbers are enough for all experiment types in the *ADAPT_list.txt*, ADAPT-NMR starts probabilistic assignment. The details of

ADAPT-NMR are well-explained in Bahrami's previous paper (Bahrami et al., 2012). Currently supported experiments are, ^{1}H,^{15}N HSQC, HNCO, HN(CA)CO, HNCA, HN(CO)CA, HN(CA)CB, CBCA(CO)NH, HBHA(CO)NH, C(CO)NH, and H(CCO)NH. However, the use of ADAPT-NMR for side chain experiments such as HBHA(CO)NH, C(CO)NH, and H(CCO)NH is only recommended for small proteins (less than 5 kD). If the completeness of the assignment does not exceed the *assignment_level* parameter defined in the *parameters.txt*, it will suggest the experiment types and tilt angles to fill the gap of the sequential assignment. When it achieves the specified level, *ADAPT_NMR_run* is designed to finish its process.

4.2.5 Preparation of NMR test samples and spectrometers

Two samples were selected to be tested for ADAPT-NMR on Bruker spectrometers; chlorella ubiquitin (76 residues) and BRPF1 bromodomain (117 residues). As a control, we used Chlorella ubiquitin as it was the same sample used from the previous ADAPT-NMR publication for Varian (Agilent) spectrometers. We also selected BRPF1 bromodomain as a new challenge for bigger protein expecting somewhat reasonable coverage of automated assignments. Marco Tonelli made the ubiquitin sample for the test, and Karen Glass' group provided the BRPF1 bromodomain sample. 1.1mM [U-^{13}C,^{15}N]-chlorella ubiquitin was prepared by *E. coli* cell-free synthesis in the NMR buffer (10 mM phosphate, 0.04% NaN_3, 90% H_2O and 10% D_2O at pH 6.6), whereas 1.0 mM [^{13}C, ^{15}N]- BRPF1 bromodomain was prepared in the NMR buffer (20 mM Tris-HCl, 150 mM NaCl, 10 mM DTT, 90% H_2O and 10% D_2O at pH 6.8).

We used TopSpin 3.0 with patch level 4 on a CentOS 5.5 workstation linked to the 500 MHz Bruker AVANCE III spectrometer. To install the ADAPT-NMR for Bruker package, the installation script file *install.py* was executed, which installs MATLAB libraries, ADAPT-NMR executables, pulse sequence and TopSpin parameters for the experiments, and AU programs (*ADAPT_ORTHO_run* and *ADAPT_NMR_run*).

4.2.6 Results and discussion

The ^{1}H carrier position in all NMR experiments for assignment in the version of ADAPT-NMR is set on-resonance water signal, which is determined from a 1D ^{1}H experiment by applying a very short (e.g. 0.5 ms), excitation pulse. The ^{1}H pulse width is manually calibrated and applied to all NMR experiments for assignment in the version of ADAPT-NMR using the command of *"getprosol"*.

Before running ADAPT-NMR, the orthogonal planes (H-N and H-C planes) are checked in the conventional version by turning off the *ZGOPTNS* flag *"HIFI"* and by setting TD1 or TD2 to be 1. Water suppression is manually optimized by adjusting the power level (sp1) of the soft water selective pulse p11. Then by turning on the *ZGOPTNS* flag *"HIFI"* and setting TD2 to be 2 (for real and imaginary part of N) and TD1 according to the desired resolution, parameters are stored for running ADAPT-NMR in the mode of reduced dimensionality.

The input tables for the automatic orthogonal and tilted 2D plane collection of both chlorella ubiquitin and BRPF1 bromodomain are shown in Table 2. As ubiquitin protein is known to provide sharp and well-dispersed peaks in NMR spectra, default numbers of scans for each experiment were used. For better resolution, spectral widths for the HNCO and HNCACO experiments were refined to 20.0 ppm from the default value of 22.0 ppm. For BRPF1 bromodomain, the numbers of scans were increased from the default parameters as shown in

Table 2B due to the water signal and the weak peak intensities in the HNCB and HNCACO experiments. Numbers of increments for both proteins were optimized for both speed and resolution.

Overall time spent by ADAPT-NMR was about 20 h for chlorella ubiquitin, and 45 h for BRPF1 bromodomain. The optimal parameters for the orthogonal 2D planes were found by repetitive runs of *ADAPT_ORTHO_run* which took about 3~4 hours. By subsequent *ADAPT_NMR_run* execution, tilted 2D planes were collected (Table 3). As the ADAPT-NMR is a descendant of HIFI-NMR and PINE-NMR, the selection of tilted angles for the initial recording of spectra is very similar to how HIFI-NMR works. At the first stage of collection, ADAPT-NMR collected a certain numbers of tilt angles and moved to the other experiment. This took a very short time because only a 3D construction had been made without any deep analysis on the quality of the 3D peaks and agreement between experiment types. The recorded 2D tilted planes through experiment list queue in this stage are shown in the table without parentheses. After recording sufficient numbers of 2D tilted planes to construct 3D spectra to identify enough peaks to assign, ADAPT-NMR started to run the torsion angle prediction module and resonance assignment module. By executing these modules, ADAPT-NMR determined what experiment needed to be further collected and what angle should be collected for the experiment. The planes collected by this procedure are written in the Table 3 within parentheses. The difference in time spent for suggesting tilt angles with and without the parentheses was huge. ADAPT-NMR spent approximately a minute to suggest a new angle at first when it did not run torsion angle prediction module and resonance assignment module. When it came to the final stage, it took about one hour.

Completeness of chemical shift assignment is illustrated in Fig. 4. The bar color in the diagram indicates the probability of the assignment robustness (green: >99%, cyan: >85%~99%, yellow: 50%~85%, red: <50%, gray: no possible assignment found). As we expected for the chlorella ubiquitin, the result was very good (Fig. 4A). The resonance assignments were nearly complete (98%) disregarding prolines and the first residue which could not be assigned with confidence. The result from the Bruker version of ADAPT-NMR was the same as that with the Varian (Agilent), confirming that Bruker version worked same (Bahrami et al., 2012). On the other hand, NMR spectra from BRPF1 bromodomain required more manual intervention (Fig. 4B) because it did not give the same extent of completeness as chlorella ubiquitin. Nevertheless, using ADAPT-NMR was worthy for this case because 86% of resonances were picked and assigned automatically in two days. Only 14% of resonances were left for manual assignment, which could be conducted easily by using the ADAPT-NMR Enhancer program developed for the visualization and the verification of the ADAPT-NMR results (Lee et al., 2012). Otherwise, the program could be used to enhance the completeness of automated resonance assignment by adding and separating some vague peaks by visualizing multiple 2D planes. The program is available for download from the ADAPT-NMR Enhancer official webpage at NMRFAM (http://pine.nmrfam.wisc.edu.edu/adapt-nmr-enhancer).

4.2.7 Conclusion

Protein structure determination by NMR spectroscopy is limited due to many factors including the supply of NMR spectrometers, NMR experts, a lack of automation, and stability of protein samples. ADAPT-NMR helps to overcome these limitations by a combination of rapid data collection and automated resonance assignment taking only one or two days. Previously, ADAPT NMR was only configured to run on Varian (Agilent) spectrometers. However, Bruker spectrometers are used more extensively in the biological NMR field. Here we report our homemade pulse programs and AU programs for adopting ADAPT-NMR running on Bruker spectrometers. To demonstrate ADAPT-NMR's abilities on Bruker instruments, we selected two proteins for the test; chlorella ubiquitin and BRPF1 bromodomain. As expected, the assignment quality of chlorella ubiquitin was excellent and equivalent to that with the Varian (Agilent) spectrometer version. We also showed the application of ADAPT-NMR to a larger, and more challenging protein as BRPF1 bromodomain. The results were 86% assignment without any human intervention.

4.3 ADAPT-NMR Enhancer

4.3.1 Abstract

Summary

ADAPT-NMR offers an automated approach to the concurrent acquisition and processing of protein NMR data with the goal of complete backbone and sidechain assignments. What the approach lacks is a useful graphical interface for reviewing results and for searching for missing peaks that may have prevented assignments or led to incorrect assignments. Because most of the data ADAPT-NMR collects are 2D tilted planes used to find peaks in 3D spectra, it would be helpful to have a tool that reconstructs the 3D spectra. The software package reported here, ADAPT-NMR Enhancer, supports the visualization of both 2D tilted planes and reconstructed 3D peaks on each tilted plane. ADAPT-NMR Enhancer can be used interactively with ADAPT-NMR to automatically assign selected peaks or it can be used to produce PINE-SPARKY-like

graphical dialogs that support atom-by-atom and peak-by-peak assignment strategies. Results can be exported in various formats including XEASY proton file (.prot), PINE pre-assignment file (.str), PINE probabilistic output file, SPARKY peak list file (.list), and TALOS+ input file (.tab). As an example, we show how ADAPT-NMR Enhancer was used to extend the automated data collection and assignment results for the protein Aedes aegypti sterol carrier protein 2.

Availability

The program, in the form of binary code along with tutorials and reference manuals, is available at http://pine.nmrfam.wisc.edu/adapt-nmr-enhancer

4.3.2 Introduction

One of the goals of protein NMR spectroscopy is to increase its throughput by automating the steps of data collection, spectral assignment, and structure determination. The latest approach toward this goal from our laboratory is ADAPT-NMR (Bahrami et al., 2012), a software package that interfaces with the NMR spectrometer and uses an algorithm for devising a pathway for optimal data collection to approach the goal of complete data assignment. As new data are collected, ADAPT-NMR analyzes the set of data collected up to that point and chooses the next step for data collection. Each data collection step involves choosing a 3D NMR experiment and a particular tilted plane that will identify peaks in the 3D spectrum. ADAPT-NMR incorporates an earlier approach to fast data collection, HIFI-NMR (Eghbalnia et al., 2005), and an algorithm for automated probabilistic assignment, PINE-NMR (Bahrami et al., 2009). The output from ADAPT-NMR is a probabilistic assignment table and analysis of secondary structure. As a means for visualizing the spectral data, picked peaks, and spin system assemblies underlying these assignments we have developed the standalone software package described here, ADAPT-NMR Enhancer (Fig. 5).

4.3.3 Implementation

ADAPT-NMR Enhancer is an SDI (Single Document Interface) application written in C++ with QT4 libraries (http://qt.nokia.com) for graphical user interface. The software supports multiple operating systems (MS Windows, MacOSX and Linux). ADAPT-NMR Enhancer offers three active dialog boxes: *Main Window Dialog*, *PINE Assignment Dialog*, and *Probable Assignment Dialog*. The *Main Window Dialog* (Fig. 6A) allows the visualization of peaks picked in 2D tilted planes and their positions in 3D space. 2D and 3D peak lists are located to the left of the dialog box; file I/O (input/output), visual manipulation, peak picking, linking and assignment tools are located at the top of the dialog box. A maximum of 6 synchronized 2D tilted planes can be viewed at once. The *x*-axis represents the ^{1}H chemical shift dimension, which is invariant with tilt angle. However, the *y*-axis is a combination of ^{13}C and ^{15}N chemical shifts as represented by the tilt angle. Thus, it is hard for users to judge the correctness of 3D peaks constructed from peaks in tilted 2D planes. ADAPT-NMR Enhancer offers two functions to resolve this problem. When one chooses a constructed 3D peak from the 3D peak list at the left side of the dialog box, circles appear in the displayed 2D tilted planes at positions where peaks are expected, and a lime-colored dot identifies peaks associated with the 3D reconstruction (Fig. 6A). Alternatively (not shown), the user can right-click and drag a 2D peak to give it the lime dot and identify the corresponding peak in the 3D peak list; again, regions in the displayed 2D planes where peaks are expected are circled. Tools located at the top of the *Main Window Dialog* can be used, not only to validate the automated peak picking and assignment, but also to add missing peaks, remove peaks picked in error, or correct assignments. PINE-SPARKY (Lee et al., 2009) tools have been incorporated into ADAPT-NMR Enhancer to assist with resonance assignments. The *PINE Assignment Dialog* (Fig. 6B) displays the peptide chain with atoms associated with assigned chemical shifts with their probabilities indicated by color coding. The candidate list box shows all 3D resonances for a given experiment that PINE considered as possible assignments for the selected atom. If the constructed 3D spectrum does not exhibit the predicted peak, the user can examine the linked 2D tilted planes for evidence of a peak. This examination is accomplished by double-clicking the candidates so as to view the corresponding 3D peak. The Probable Assignment Dialog box pops up when a 3D peak from the 3D list box or from the spectral view is selected. It lists possible assignments for a peak along with their probabilities. The PINE Assignment Dialog box is based on atoms, whereas the Probable Assignment Dialog box is based on peaks. The user can either confirm or modify the assignment for a 3D peak. The decision is stored in the confirmation list box, and the results can be exported in a variety of file formats (Fig. 7).

4.3.4 Results and conclusion

AeSCP-2 (110 residues) is the *Aedes aegypti* sterol carrier protein 2, which is involved in cellular lipid transport mechanisms related to lipid uptake and metabolism (Singarapu et al., 2010). This protein was previously used to test ADAPT-NMR (Bahrami et al., 2012). Although assignments were made to 510 atoms with >99% probability of correctness, the assignment probabilities of 24 atoms was 99% or lower, and no assignment were obtained for five atoms. We used ADAPT-NMR Enhancer to visualize and improve the quality of the assignments. We manually added peaks that had not been picked by the automated algorithm; we deleted picked peaks clearly arising from noise; and we modified the priority scores of the peaks on the basis of manual assessment. With the new peak set as input, ADAPT-NMR yielded improved scoring: of the 24 assignments initially scored at less than 99% probability, only 7 remained lower than 99% probability. We then used the manual features of ADAPT-NMR-Enhancer to determine why these 7 assignments were of lower probability. We found, for example, that because residue 80 is proline, the CBCA(CO)NH data set yielded no connectivities from P80 to the CA and CB of L79. However, we could easily confirm the assignment from HNCA(HNCB) data. Another atom with low assignment probability, A60CA, was found to have a low peak intensity that prevented its detection in the CBCA(CO)NH experiment. The missing peak is easily added by using the editing tool of ADAPT-NMR Enhancer so that ADAPT-NMR will recognize the peak when it is re-run. All backbone resonance assignments were confirmed or completed by means of a "sequential walk" through the 3D HNCA(HNCB) and CBCA(CO)NH data. The "Lock" tool in ADAPT-NMR Enhancer, which enables one to predict the position of a 3D peak by selecting two peaks from 2D tilted planes, was found to be useful in confirming assignments. In cases where a large number of noise peaks have been deleted ADAPT-NMR, will suggest another experiment and tilt angle for data collection. All the results we have gotten are shown in the Table 4. The detailed strategies used are documented at http://pine.nmrfam.wisc.edu/adapt-nmr-enhancer.

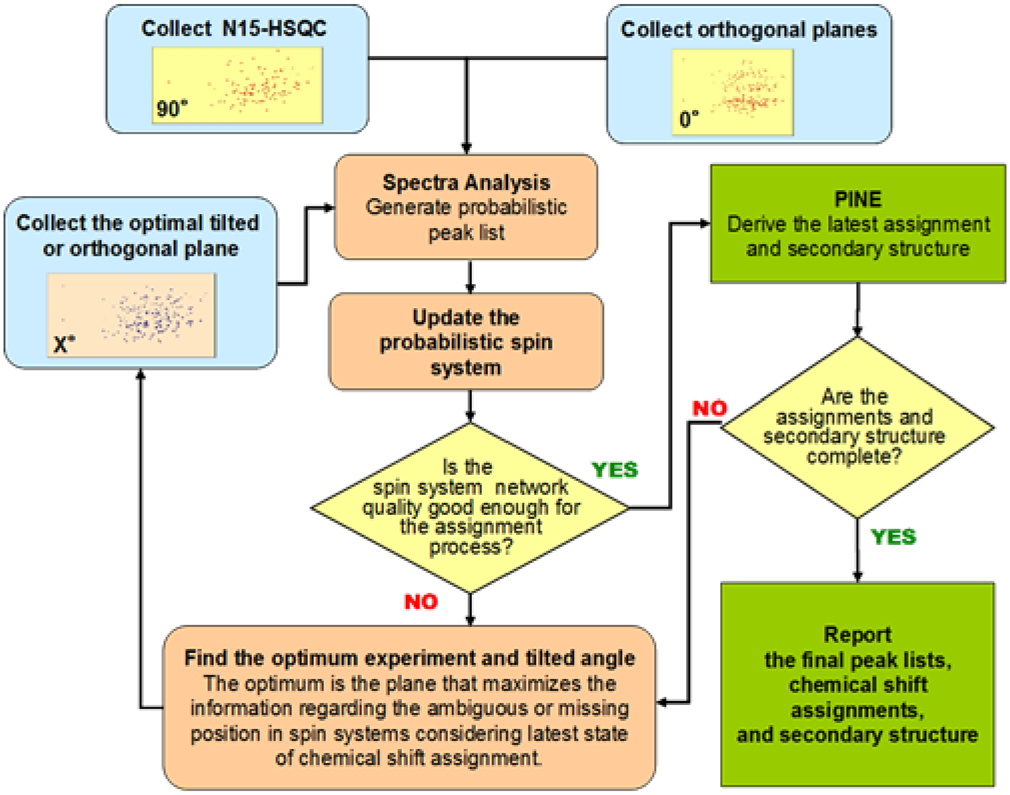

Collect N15-HSQC
90°
Collect orthogonal planes
0°
Spectra Analysis
Generate probabilistic peak list
Collect the optimal tilted or orthogonal plane
X°
PINE
Derive the latest assignment and secondary structure
Update the probabilistic spin system
Are the assignments and secondary structure complete?
NO
YES
Is the spin system network quality good enough for the assignment process?
YES
NO
Find the optimum experiment and tilted angle
The optimum is the plane that maximizes the information regarding the ambiguous or missing position in spin systems considering latest state of chemical shift assignment.
Report the final peak lists, chemical shift assignments, and secondary structure

Figure 1. The improvement of reduced dimensionality through ADAPT-NMR. A flowchart illustrates the design of ADAPT-NMR integrating HIFI-NMR and PINE-NMR. HIFI portions are in sky blue boxes, PINE-NMR portions are in green boxes, while others in neither green nor sky blue boxes are newly developed MATLAB portion for integration between HIFI-NMR and PINE-NMR. *Adapted from Bahrami et al., 2011*

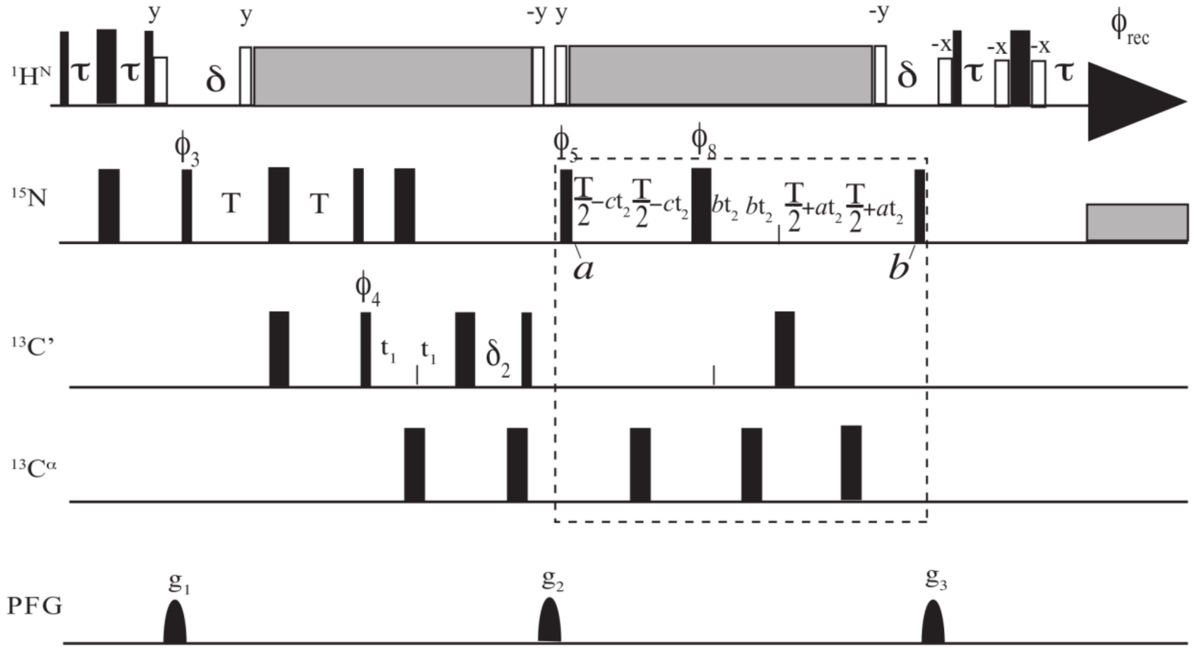


Figure 2. Pulse sequence of HNCO in ADAPT-NMR version. The time constants are; t (d26) = 2.3ms, T (d23) = 16ms, d (d21) = 5.5ms. The evolution of chemical shift of ^{15}N is modified to a semi-constant time version, as shown in the box, in order to match the possible requirement for high resolution along C dimension. The chemical shift evolution of N and C is synchronized by correlating the time incremental interval inf2 (N) and inf1 (C) to a common tilted angle between the two orthogonal planes.

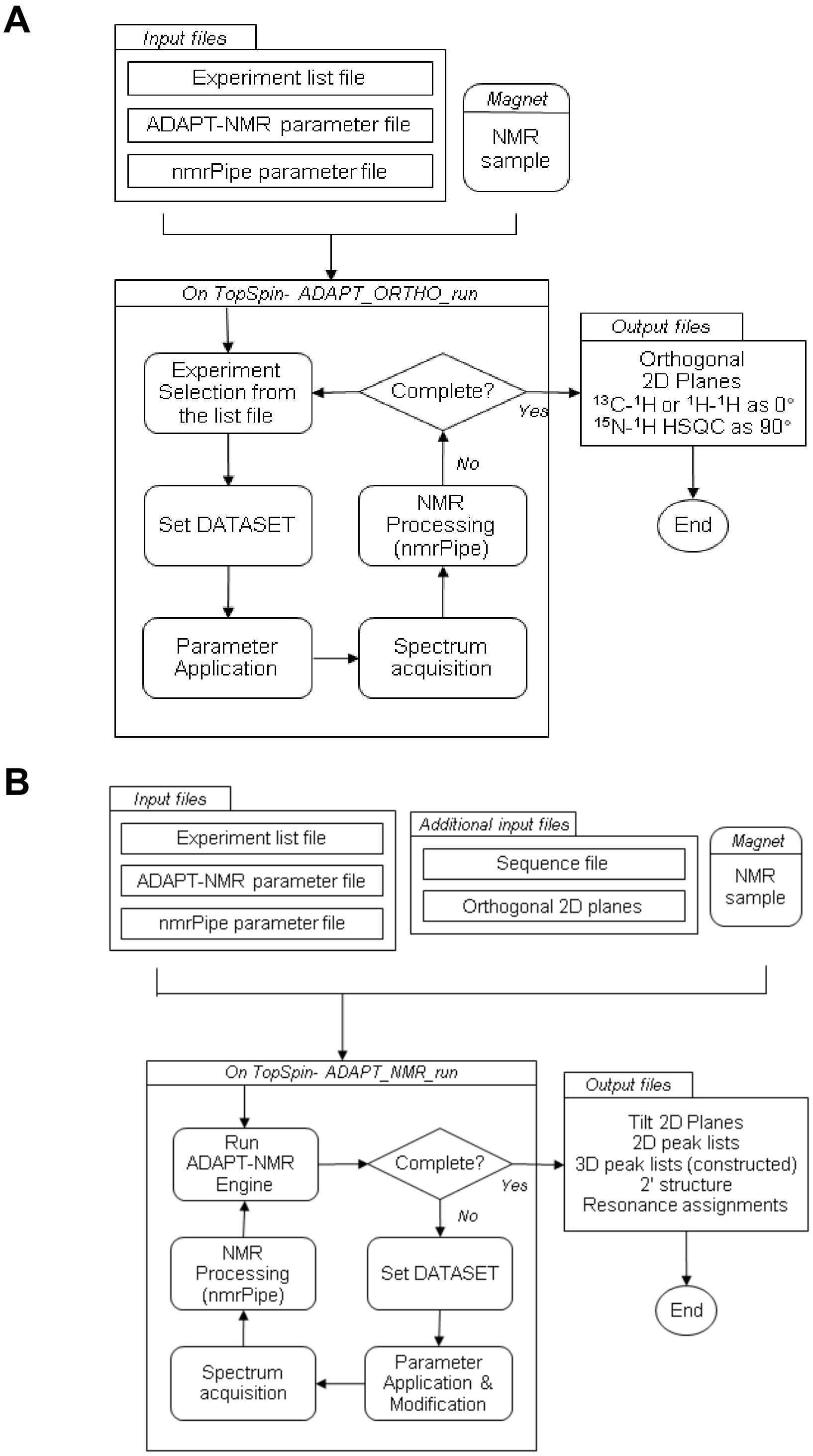


Figure 3. Overall flowchart of ADAPT-NMR. (A) Flowchart of ADAPT_ORTHO_run, (B) Flowchart of ADAPT_NMR_run.

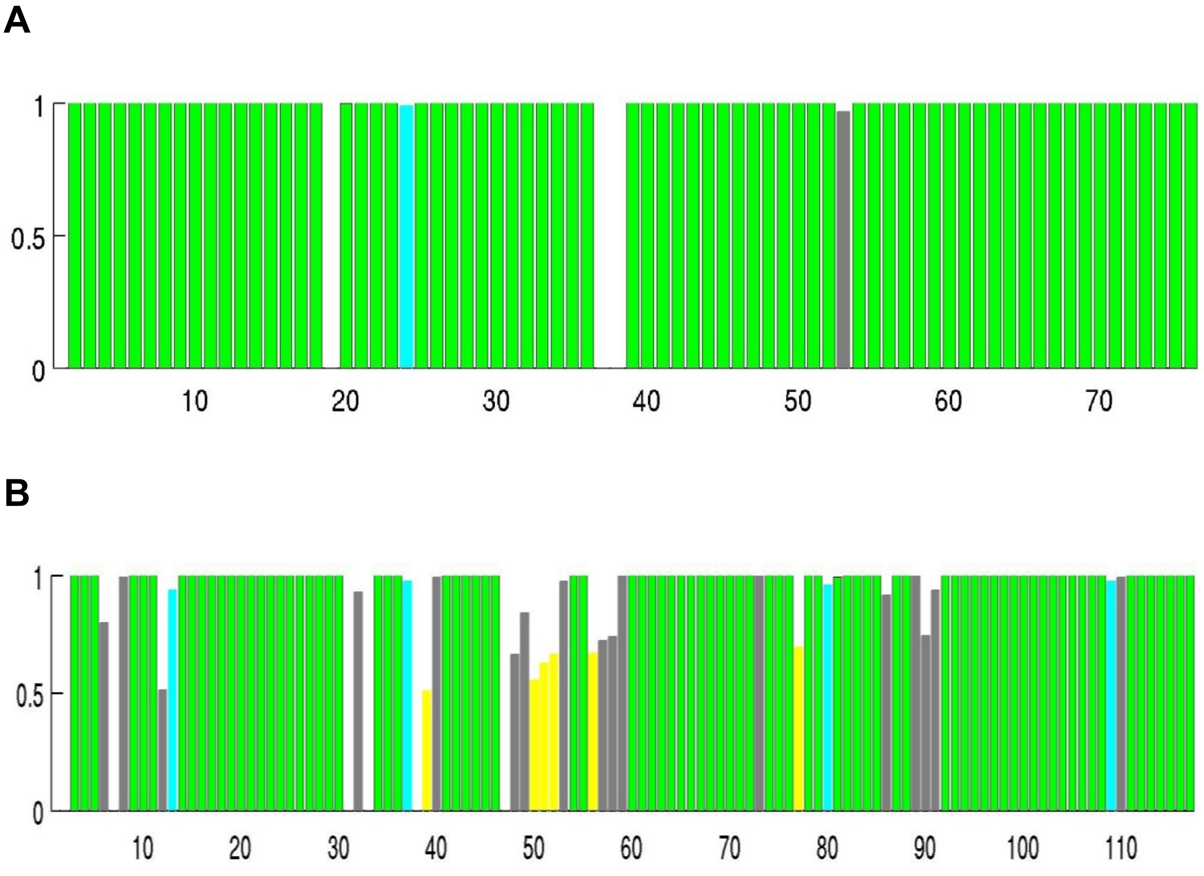


Figure 4. Bar diagram graphs of ADAPT-NMR for Bruker results. (A) Assignment diagram for Ubiquitin. (B) Assignment diagram for BRPF1 bromodomain.

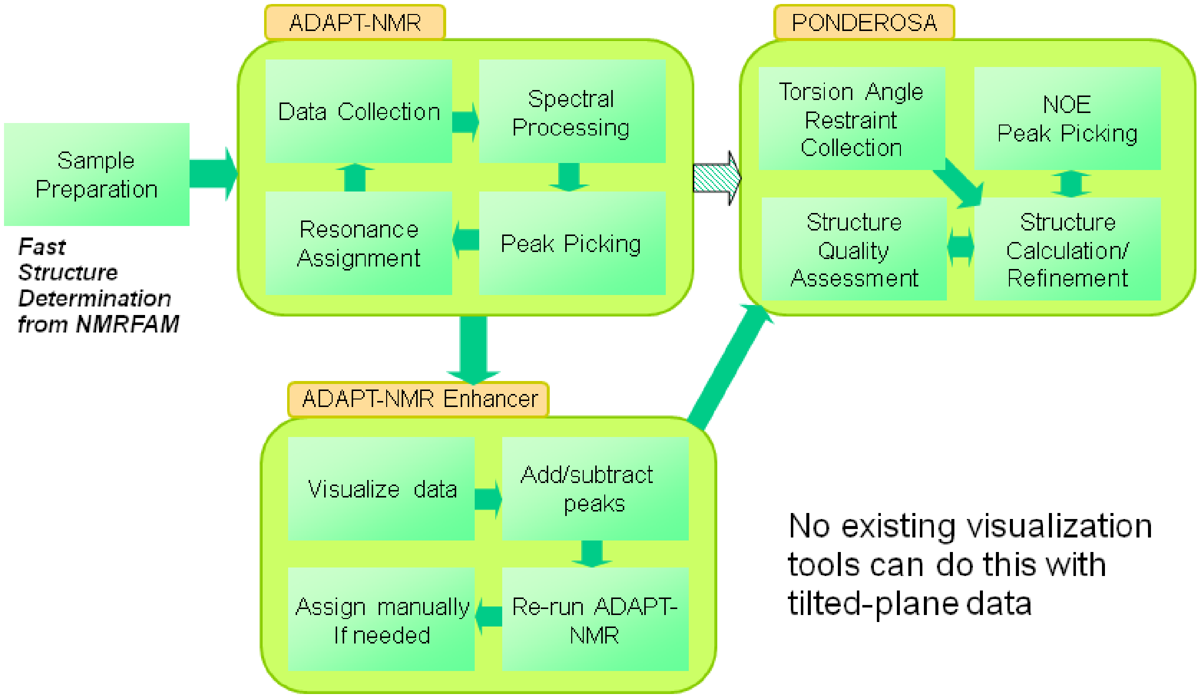
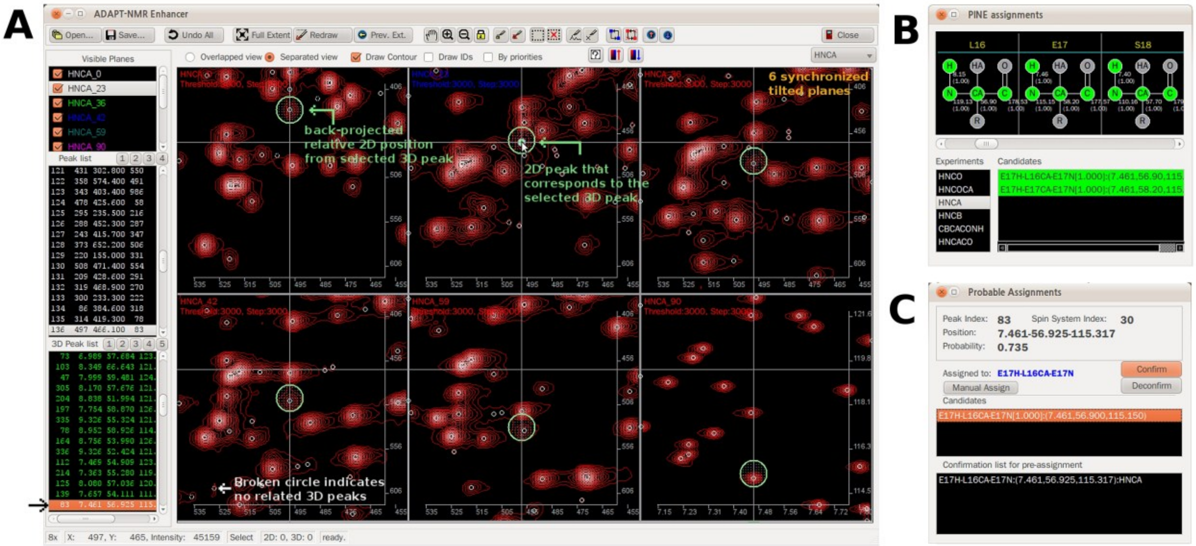

ADAPT-NMR
Sample Preparation
Fast Structure Determination from NMRFAM
Data Collection
Spectral Processing
Resonance Assignment
Peak Picking
PONDEROSA
Torsion Angle Restraint Collection
NOE Peak Picking
Structure Quality Assessment
Structure Calculation/ Refinement
ADAPT-NMR Enhancer
Visualize data
Add/subtract peaks
Assign manually If needed
Re-run ADAPT-NMR
No existing visualization tools can do this with tilted-plane data

Figure 5. Schematic diagram illustrates rationale for fast structure determination from NMRFAM. It shows the rationale for ADAPT-NMR Enhancer by showing how it can be adopted to the ADAPT-NMR and PONDEROSA.


A
ADAPT-NMR Enhancer
Open...
Save...
Undo All
Full Extent
Redraw
Prev. Ext.
Close
○ Overlapped view
● Separated view
☑ Draw Contour
□ Draw IDs
□ By priorities
HNCA
back-projected relative 2D position from selected 3D peak
2D peak that corresponds to the selected 3D peak
6 synchronized tilted planes
Broken circle indicates no related 3D peaks
B
C

Visible Planes

☑ HNCA_0
☑ HNCA_23
☑ HNCA_36
☑ HNCA_42
☑ HNCA_59
☑ HNCA_90

Peak list

```
121  431 302.800 550
122  358 574.400 491
123  343 403.400 986
124  478 425.600  58
125  295 235.500 216
126  288 452.300 287
127  243 415.700 347
128  373 652.200 506
129  220 155.000 331
130  508 471.400 554
131  209 428.600 291
132  319 468.900 270
133  300 233.300 222
134   86 384.600 318
135  314 419.300  78
136  497 466.100  83
```

3D Peak list

```
 73  6.989 57.684 123.
103  8.349 66.643 121.
 47  7.999 59.481 124.
305  8.170 57.676 121.
204  8.838 51.994 121.
197  7.754 58.870 126.
335  9.326 55.324 121.
 78  8.952 58.926 114.
164  8.756 53.990 126.
336  9.326 52.424 121.
112  7.469 54.909 123.
214  7.363 55.280 119.
125  8.080 57.036 120.
139  7.657 54.111 111.
 83  7.461 56.925 115.
```

8x | X: 497, Y: 465, Intensity: 45159 | Select | 2D: 0, 3D: 0 | ready.

PINE assignments

L16	E17	S18

Experiments: HNCO, HNCOCA, HNCA, HNCB, CBCACONH, HNCACO

Candidates:

E17H-L16CA-E17N[1.000]:(7.461,56.90,115.

E17H-E17CA-E17N[1.000]:(7.461,58.20,115.

Probable Assignments

Peak Index: 83 Spin System Index: 30

Position: 7.461-56.925-115.317

Probability: 0.735

Assigned to: E17H-L16CA-E17N

Manual Assign

Confirm

Deconfirm

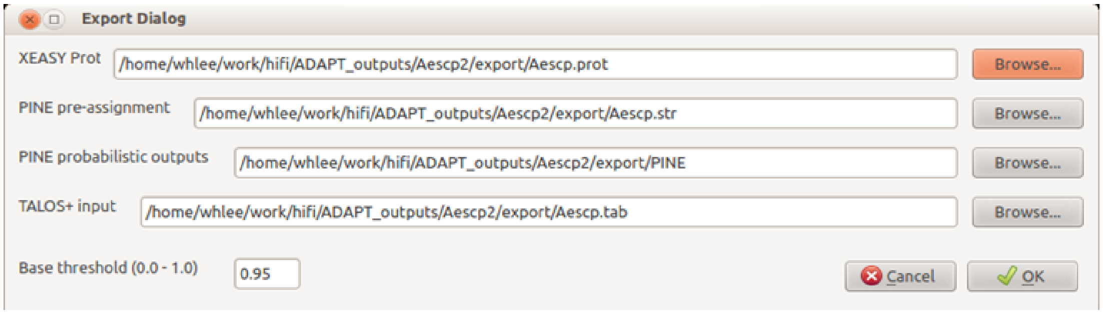
Candidates

E17H-L16CA-E17N[1.000]:(7.461,56.900,115.150)

Confirmation list for pre-assignment

E17H-L16CA-E17N:(7.461,56.925,115.317):HNCA

Figure 6. ADAPT-NMR Enhancer user interface with AeSCP-2. (A) *Main Window Dialog* for tilted plane visualization. (B) *PINE Assignment Dialog* for atom-by-atom assignment. (C) *Probable Assignment Dialog* for peak-by-peak assignment.



XEASY Prot /home/whlee/work/hifi/ADAPT_outputs/Aescp2/export/Aescp.prot Browse...

PINE pre-assignment /home/whlee/work/hifi/ADAPT_outputs/Aescp2/export/Aescp.str Browse...

PINE probabilistic outputs /home/whlee/work/hifi/ADAPT_outputs/Aescp2/export/PINE Browse...

TALOS+ input /home/whlee/work/hifi/ADAPT_outputs/Aescp2/export/Aescp.tab Browse...

Base threshold (0.0 - 1.0) 0.95

ADAPT-NMR Enhancer can be used to make ADAPT-NMR results accessible to various other software programs

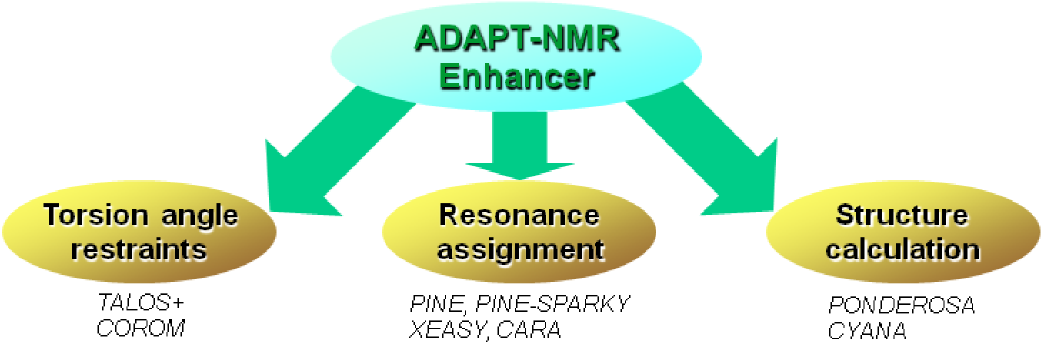
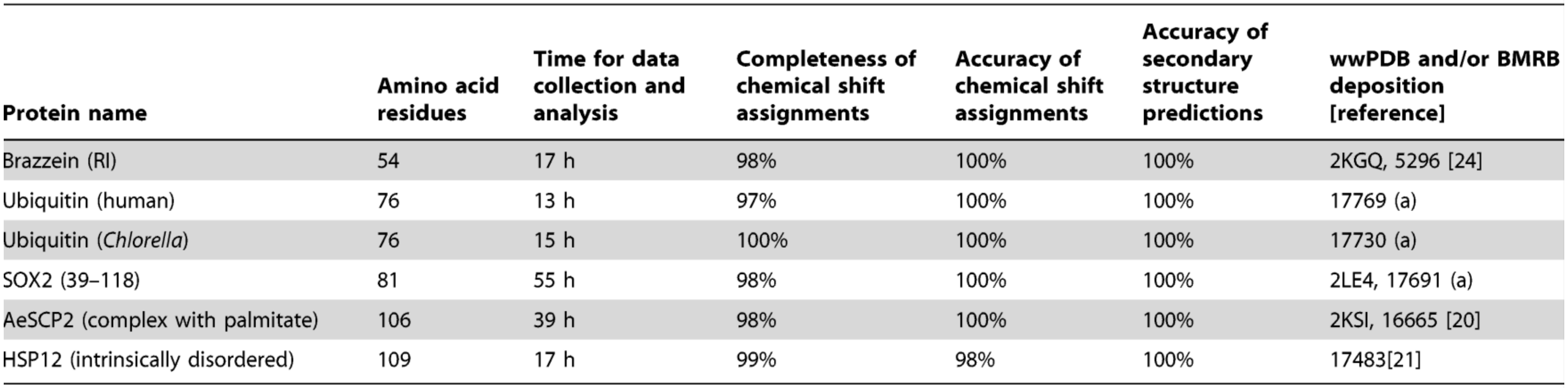


Figure 7. Export capability of ADAPT-NMR Enhancer. Through *Export Dialog*, user can work on structure determination seamlessly from restraint collection to structure calculation.

Table 1. Results from ADAPT-NMR data collection and backbone analysis of six proteins on Varian (Agilent) spectrometers.

Adapted from Bahrami et al., 2012.



Protein name	Amino acid residues	Time for data collection and analysis	Completeness of chemical shift assignments	Accuracy of chemical shift assignments	Accuracy of secondary structure predictions	wwPDB and/or BMRB deposition [reference]
Brazzein (RI)	54	17 h	98%	100%	100%	2KGQ, 5296 [24]
Ubiquitin (human)	76	13 h	97%	100%	100%	17769 (a)
Ubiquitin (*Chlorella*)	76	15 h	100%	100%	100%	17730 (a)
SOX2 (39–118)	81	55 h	98%	100%	100%	2LE4, 17691 (a)
AeSCP2 (complex with palmitate)	106	39 h	98%	100%	100%	2KSI, 16665 [20]
HSP12 (intrinsically disordered)	109	17 h	99%	98%	100%	17483[21]

Table 2. Input table for orthogonal plane collection. The same values of number of scan and increment, carrier position and spectral width were also used for tilt plane collection.

(A) Chlorella ubiquitin

Experiment	Keyword	# of scans	# of increments	Name of the plane	Carrier position(ppm)	Spectral width(ppm)
$^1H,^{15}N$-HSQC	ubiq	2	128	ubiq_NHSQC	118.0	36.0
HNCO	ubiq	4	64	ubiq_HNCO_0	176.0	20.0
HN(CO)CA	ubiq	4	64	ubiq_HNCOCA_0	56.0	32.0
HNCA	ubiq	8	64	ubiq_HNCA_0	56.0	32.0
CBCA(CO)NH	ubiq	8	50	ubiq_CBCACONH_0	45.0	70.0
HNCB	ubiq	8	64	ubiq_HNCB_0	45.0	70.0
HN(CA)CO	ubiq	8	128	ubiq_HNCACO_0	176.0	20.0

(B) BRPF1 bromodomain

Experiment	Keyword	# of scans	# of increments	Name of the plane	Carrier position(ppm)	Spectral width(ppm)
$^1H,^{15}N$-HSQC	Hbromo	2	128	Hbromo_NHSQC	118.0	36.0
HNCO	Hbromo	4	64	Hbromo_HNCO_0	176.0	22.0
HN(CO)CA	Hbromo	4	64	Hbromo_HNCOCA_0	56.0	32.0
HNCA	Hbromo	8	64	Hbromo_HNCA_0	56.0	32.0
CBCA(CO)NH	Hbromo	8	58	Hbromo_CBCACONH_0	45.0	70.0
HNCB	Hbromo	32	80	Hbromo_HNCB_0	45.0	70.0
HN(CA)CO	Hbromo	16	64	Hbromo_HNCACO_0	176.0	22.0

Table 3. Orthogonal planes and tilted planes recorded by ADAPT-NMR on Bruker DRX 500 MHz spectrometer.

(A) Chlorella ubiquitin

Experiment	Orthogonal planes	Tilt planes
^{1}H,^{15}N-HSQC	90°	-
HNCO	0°	73°, 81°,
HNCOCA	0°	76°, 50°, (30°)[1]
HNCA	0°	35°, 58°, 122°, 45°, (17°)[1]
CBCACONH	0°	46°, 39°, 32°, (54°, 28°, 62)[1]
HNCB	0°	37°, 50°, 71°, 22°
HNCACO	0°	49°, 72°, 37°

(B) BRPF1 bromodomain

Experiment	Orthogonal planes	Tilt planes
^{1}H,^{15}N-HSQC	90°	-
HNCO	0°	16°, 54°,
HNCOCA	0°	29°, 20°, 56°
HNCA	0°	24°, 52°, 42°, 34°
CBCACONH	0°	29°, 70°, 47°, 40°, (18°, 53°, 43°, 34)[1]
HNCB	0°	18°, 63°, 54°, 43°, 30°, (23°, 70)[1]
HNCACO	0°	16°, 58°, 48°, 32°, 38°, (40°, 66°, 20) [1]

[1] Tilt angles in the parentheses were recorded after the first continuous recording of tilted planes in the experiment list queue. The angles and experiment types were suggested by ADAPT-NMR engine to the ADAPT_NMR_run AU program.

Table 4. Gradual improvement on assignment quality of Sterol Carrier Protein-2 from Aedes aegypti (AeSCP-2, 110 residues) before and after ADAPT-NMR Enhancer.

PINE probability	# of assignments before peak modification	# of assignments after peak modification	# of assignments after PINE-SPARKY-like tools
>99%	510 (95.5%)	527 (98.7%)	534 (100.0%)
85%-99%	1 (0.2%)	2 (0.4%)	0 (0.0%)
50%-85%	12 (2.3%)	1 (0.2%)	0 (0.0%)
<50%	6 (1.1%)	2 (0.4%)	0 (0.0%)
No assignment (0%)	5 (0.9%)	2 (0.4%)	0 (0.0%)
Total	534 (100.0%)	534 (100.0%)	534 (100.0%)

CHAPTER 5

Structure Determination and Analysis of RV-C02 2A Proteinase from Strain W12 by NMR spectroscopy

Adapted in part from Lee W, Frederick R, Tonelli M, Troupis AT, Reinin N, Suchy FP, Moyer K, Watters K, Aceti D, Palmenberg AC, Markley JL. (2013) Structure determination and analysis of RV-C02 2A proteinase from strain W12 by NMR spectroscopy. *J Virol*. in preparation

5.1 Introduction

The human rhinovirus is known as a disease agent for the common cold (Price, 1956; Gwaltney et al., 1966). It is a nonenveloped and spherical RNA virus of the picornavirus family covered by viral capsid proteins (VP1, VP2, VP3, and VP4) of which more than 100 serotypes have been discovered (Bochkov et al., 2011). HRV-A (77 serotypes) and HRV-B (25 serotypes) are categorized depending on their viral capsid-coding regions, noncoding regions, and some complete genomes (Andries et al., 1990; Lau et al., 2011). However, some unknown HRV-like sequences, later categorized as HRV-C, have been found in patients suffering influenza-illnesses with severe respiratory compromises (Dominguez et al., 2008). Except for three HRV-A serotypes discovered later by Dr. Lau and her colleagues (Lau et al., 2011), Dr. Palmenberg and her colleagues sequenced the genomes for every undetermined HRV in the repository of known serotypes, including HRV-C species that had previously been unknown (Palmenberg et al., 2009). HRV-C serotype genomes, similar to other serotypes code for 2A proteinase proteins, which are known to cleave in the middle of VP1 (Toyoda et al., 1986), to be related to RNA replication (Molla et al., 1993), and also to cleave eIF4G homologues in a manner related to the inhibition of host-cell cap-dependent protein synthesis (Lloyd et al., 1988; Willcocks et al., 1994). One 2A proteinase from strain W12 (HRV-C serotype), RV-C02 $2A^{pro}$, was originally purified by the Palmenberg group. Currently there are no 3D structures available for HRV-C type proteins. NMRFAM (National Magnetic Resonance Resonance Facility at Madison) and the CESG (Center for Eukaryotic Structural Genomics) collaborated with the Palmenberg group to determine the 3D structure of this protein using biomolecular NMR techniques. Dr. Ronnie Frederick and his crew from CESG carried out NMR sample preparation, while I worked on

NMR data collection with Dr. Marco Tonelli and carried out spectral analysis and structure determination. In addition to gaining biological insight into this protein, this project allowed me to test and apply my earlier development of high-throughput NMR structure determination methods such as PINE-SPARKY, ADAPT-NMR on Bruker spectrometers, ADAPT-NMR Enhancer, and PONDEROSA. The 3D structure of RV-C02 2A^{pro} were solved by use of these high-throughput strategies as discussed in this chapter.

5.2 Materials and Methods

5.2.1 Protein sample preparation

Dr. Ronnie Frederick from the CESG purified RV-C02 2A^{pro} by collaborating with Dr. Kelly Watters from the Palmenberg group. Kelly provided an initial protein purification protocol to Ronnie for mass sample production, and Ronnie attempted multiple methods to produce a large yield suitable for NMR spectroscopy. Dr. Marco Tonelli and I collaborated with Ronnie extensively to optimize protein sample conditions. Firstly, Ronnie encountered low yields when he followed Kelly's purification protocol of the active form of wildtype 2A^{pro}. The yield of of wildtype RV-C02 2A^{pro} was just sufficient to record a 2D ^{1}H,^{15}N-HSQC spectrum; however, we discovered that the active protein was not stable. The NMR signals differed a week after purification, which suggested that the proteinase was undergoing self cleavage. Ronnie mutated C105, a residue in the catalytic triad, to alanine in order to make an inactive form of 2Apro, which he was able to purify in higher yield (Fig. 1). Finally, Marco and I found sample conditions where the protein was stable for over a month after purification by trying several pH values, temperatures, salt and metal concentrations, and antibiotics. The conditions that yielded a well-dispersed NMR spectrum with the sharpest and the most stable peaks from a 2D ^{1}H,^{15}N-HSQC experiment were 10 mM MES, 20 mM NaCl, 10 mM DTT in 10% ^2H_2O / 90% H_2O at pH 6.5. The protein sample concentration was kept lower than 0.5 mM because aggregation was observed at higher concentrations.

5.2.2 NMR spectroscopy

The sample used for NMR spectroscopy contained 3.4 mg [U-^{13}C,U-^{15}N]-RV-C02 2A^{pro} dissolved in 0.4 ml buffer consisting of 10 mM MES, 20 mM NaCl, 10 mM DTT in 10% ^2H_2O / 90% H_2O at pH 6.5. The approximate protein concentration was 0.5 mM. The solution was placed in a 5 mm Shigemi NMR tube (Allison Park, PA). NMR data were collected at the National Magnetic Resonance Facility at Madison (NMRFAM) on Varian VNMRS spectrometers operating at 600 MHz, 800 MHz and 900 MHz (^{1}H frequency). The sample temperature was regulated at 313 K. ADAPT-NMR was attempted on a 500 MHz Bruker, but it was not successful because of poor peak dispersion and bad signal-to-noise. 3D HNCO, HN(CA)CO, HNCA, HN(CO)CA, CBCA(CO)NH, HBHA(CO)NH, C(CO)NH, H(CCO)NH, H(C)CH-TOCSY, and ^{15}N-edited NOESY data sets were collected on the 600 MHz spectrometer equipped with a triple-resonance cryogenic probe. 2D ^{1}H,^{15}N-HSQC and 3D ^{15}N-edited TOCSY, (H)CCH-TOCSY, and ^{13}C-edited NOESY data sets were collected on the 800 MHz spectrometer equipped with a conventional triple-resonance probe. 2D ^{1}H,^{13}C-HSQC and 3D HNCACB spectra were collected on the 900 MHz spectrometer with a triple-resonance cryogenic probe. Collected time-domain data were processed with NMRPipe (Delaglio et al., 1995) to generate frequency-domain data which were converted to the SPARKY (ucsf) file format for further analysis (T. D. Goddard and D.G. Kneller, SPARKY 3, University of California, San Francisco).

5.2.3 Structure determination of RV-C02 2A^{pro}

Initially, we used the APES program (Shin et al., 2008) to identify backbone resonances in ^{1}H,^{15}N-HSQC, HNCACB, and CBCA(CO)NH spectra. These resonances were used with the restricted peak picking feature in SPARKY to identify signals in other backbone and side chain spectra and these automatically picked peaks were carefully validated by visual inspection in SPARKY. Peak lists generated for each spectrum were then exported to the PINE-NMR server for automated resonance assignments (Bahrami et al., 2009) and the results verified using the PINE-SPARKY package (Lee et al., 2009). Validated chemical shift assignments were then used with the PONDEROSA package (Lee et al., 2011) for the automated assignment of NOE cross-peaks in ^{15}N-edited NOESY and ^{13}C-edited NOESY data sets. We then used SPARKY for manual validation and refinement of NOE peak picking and assignments. The curated peak lists, NOE assignments, distance restraints, and torsion angle restraints were used in further structure refinement steps by manual operation of CYANA version 3.0 followed by fine-tuned structure calculation (Güntert, 2004). Hydrogen bond restraints for regions found to contain regular secondary structure ($d_{N\text{-}O}$ = 2.7 to 3.5 Å; $d_{H^{N}\text{-}O}$ = 1.8 to 2.5 Å) were then added. The torsion angle constraints generated by the TALOS+ program (Cornilescu et al., 1999) module executed by PONDEROSA were validated one-by-one by reference to SPARKY and PyMOL visualizations to remove too tight constraints. Once an acceptable structure, as validated by the PSVS suite server (Bhattacharya et al., 2007), was obtained, we identified C51, C53, C111, and

H113 as the metal coordinating side chains and added a zinc ion to the model. Subsequent CYANA calculations were then run using covalent distance restraints for the zinc coordination side (Cys $S^{\gamma}-Zn = 2.4$ Å and His $N^{\varepsilon 2}$-Zn = 2.20Å.) The 15 best models from a total of 200 models annealed from random structures were chosen on the basis of lowest energy with fewest violations to represent the structure of the proteinase. We used the MOLMOL program (Koradi et al., 1996) to calculate the root mean square deviation (rmsd) and the PyMOL program (Version 1.2r3pre, Schrödinger, LLC) for graphical analysis. In order to generate electrostatic potential surfaces, the APBS plug-in for PyMOL (Baker et al., 2001) was used with PQR files generated from Poisson-Boltzmann electrostatics calculated by the PDB2PQR package (Dolinsky et al., 2004). We used the STRIDE program (Frishman and Argos, 1995) to determine the secondary structural features in the model with the lowest energy. We used the PSVS suite server (Bhattacharya et al., 2007) to validate the quality of the final structure ensemble.

5.2.4 Hydrophobic scale measurement of the aromatic residues

We used the STRIDE program to determine the surface accessibility of the aromatic side chains (His, Phe, Trp, Tyr) in the lowest energy structure. The calculated accessible surface areas were divided by the fully exposed residue accessible surface areas in the corresponding (Gly-X-Gly) tripeptides to obtain the percent exposure (0%: fully buried, 100%: fully exposed) (Eisenhaber and Argos, 1993).

5.2.5 Spin relaxation and amide exchange experiment

^{1}H-^{15}N NOE and ^{15}N relaxation (T_1, T_2) data were recorded on the Varian VNMRS 800 MHz spectrometer equipped with a conventional triple-resonance probe. Multi-interleaved NMR spectra were collected with relaxation delays of 0, 50, 100, 200, 300, 400, 600, 1200, and 1600 ms for the ^{15}N T_1 measurements, and with relaxation delays of 10, 30, 50, 70, 90, and 110 ms for the ^{15}N T_2 measurements. The relaxation rate constants were extracted in SPARKY by fitting the decay of peak height as a function of the relaxation delay to a single exponential function. Interleaved 2D ^{1}H,^{15}N-HSQC spectra with and without a 5 sec proton saturation were collected for the {^{1}H}-^{15}N NOE measurements. The ratio of peak heights between two spectra were calculated by using SPARKY and LibreOffice Spreadsheet programs.

5.2.6 Database accession number

The coordinates of the 15 best models have been deposited in the Protein Data Bank as entry 2M5T, and the chemical shifts have been deposited in the Biological Magnetic Resonance Bank as entry 19079.

5.3 Results and Discussion

We determined the three-dimensional structure of [U-^{13}C,U-^{15}N]- RV-C02 2A^{pro} by multidimensional, multinuclear magnetic resonance methodology. The structure was based on a total of 1440 restraints, 1239 distance constraints, 142 angle constraints, 59 hydrogen bond constraints derived with the PONDEROSA, SPARKY, CYANA and TALOS+ programs. The 15 best models (lowest energy and fewest violations) were chosen to represent the solution structure of RV-C02 2A^{pro}. The rmsd of the regions with regular secondary structurel was 0.6 Å for backbone heavy atoms and 0.8 Å for all heavy atoms. From Procheck, 85.0% of the backbone angles were in the most favored regions, 13.2% in the additionally allowed regions, 1.8% in the generously allowed regions, and none in the disallowed regions; from MolProbity, 93.6% were in the most favored regions, 6.4% in the allowed regions, and none in the disallowed regions. The Z-scores for backbone/all dihedral angles were -2.95/-5.62, and the mean score/Z-score values from MolProbity were 24.03/-2.60 (Table 1). The number of NOE restraints per residue used for structure calculation is shown in Fig. 2A. The lack of NOE assignments for the N-terminal, C-terminal, and for residues 82-86 facing the catalytic triad region (H18, D34, C105A) led to relatively higher rmsd values and lower structural compactness of the models in these regions (Fig. 2B).

STRIDE analysis of the structure determined that the protein consists mostly of β-strands as also found for the structure of the ortholog, RV-A02 $2A^{pro}$ (Petersen et al., 1999). The assigned secondary structural elements are depicted at the top of Fig. 2, and the nomenclature of the labels follows that used for RV-A02 $2A^{pro}$ and other chymotrypsin-related proteinases (Petersen et al., 1999). RV-C02 $2A^{pro}$ consists of an N-terminal and C-terminal domain connected by a loop. The N-terminal domain (orange color in Fig. 3B) contains four strands that constitute an antiparallel β-sheet (β-strands V7−T9 [bI2], A12−N16 [cI], L28−A30 [eI2], L35−G39 [fI]). The C-terminal domain (light blue color in Fig. 3B) contains six strands that constitute an antiparallel β-barrel (β-strands S55−S60 [aII], R65−V79 [bII], H88−E97 [cII], G107−L110 [dII], V115−G123 [eII], H126−D131 [fII]). The long connecting loop (green in Fig. 3) consists of C40−T54. The conserved di-tyrosine flap, observed as a β-hairpin loop in both RV-A02 $2A^{pro}$ (Y85, Y86, P87) and EV-CB4 $2A^{pro}$ (Y89, Y90, P91), also can be seen in this protein (Block arrow in Fig. 3B; Y84, Y85, P86). Three short 3^{10}-helices were identified in the structure, each consisting of three residues that come after β-strands (cI, eI2, and aII).

In order to understand the common structural features of picornaviral 2A proteinases, we compared the structure of RV-C02 $2A^{pro}$ with those of RV-A02 $2A^{pro}$ and EV-CB4 $2A^{pro}$. Multiple sequence alignment (Fig. 4) by the STRAP program (Gille and Frömmel, 2001) showed a sequence identity of 57% for RV-C02 $2A^{pro}$ and RV-A02 $2A^{pro}$ and 41% for RV-C02 $2A^{pro}$ and EV-CB4 $2A^{pro}$. Structurally, and functionally important residues in the primary sequences are well conserved throughout the proteins. The di-tyrosine flap (YYP) is marked by ellipsoid (Fig. 4); the conserved zinc binding-site consists of one histidine and three cysteine residues shown by dashed line boxes in Fig. 4 (side chains colored magenta and zinc ion represented as a gray sphere, in Fig. 3B). The conserved catalytic triads are marked by solid line boxes in Fig. 4 (side chains colored cyan in Fig. 3B). The conserved PGDCGG motif is located between two β-strands of the C-terminal domain (cII and dII in RV-C02 $2A^{pro}$). As with EV-CB4 $2A^{pro}$, the cysteine in the middle of this motif was mutated to alanine to obtain a stable inactive protein. The 3D coordinates of RV-C02 $2A^{pro}$ were aligned and superimposed onto the 3D coordinates of both RV-A02 $2A^{pro}$ and EV-CB4 $2A^{pro}$ (Fig. 5A). Even though the sequence identity between RV-A02 $2A^{pro}$ and RV-C02 $2A^{pro}$ (57%) is higher than that between EV-CB4 $2A^{pro}$ and RV-C02 $2A^{pro}$ (41%), the structural similarities based on rmsd values. were equivalent (both 1.809 Å). We used the APBS package in PyMOL to generate electrostatic potential surfaces for the three proteins (Fig. 5B, C, D). The contouring value was set to ±10kT/e for visualization of the surface charges. The three proteins have similar positive surface charge distributions (red). Larger differences are seen in negative surface charge distributions (blue): RV-C02 $2A^{pro}$ lacks the patches of negative surface seen in RV-A02 $2A^{pro}$ and EV-CB4 $2A^{pro}$.

The aromatic residues of RV-C02 2A^{pro} tend to be exposed to the protein surface, as was found for EV-CB4 2A^{pro} (Baxter et al., 2006), rather than forming a more normal hydrophobic core that stabilizes protein structure (Cox et al., 2000). To assess the exposure of aromatic residues in the structures, we calculated hydrophobic scales by dividing residue solvent accessible surface area (SAS) obtained from STRIDE by the SAS of each fully exposed residue from Gly-His-Gly: (1.94 Å^2), Gly-Phe-Gly: (2.18 Å^2), Gly-Trp-Gly (2.59 Å^2), and Gly-Tyr-Gly: (2.29 Å^2). Of the 18 aromatic residues in the sequence of RV-C02 2A^{pro} (9 Tyr, 6 His, 2 Phe, 1 Trp), 13 are exposed to solvent (7 Tyr, 4 His, 1 Phe, 1 Trp) and only 5 are partially buried (2 Tyr, 2 His, 1 Phe). Only two residues are fully (> 90%) buried (Y58, F129). The hydrophobic cores of these proteinases consist mainly of valine, leucine, and isoleucine residues.

We collected longitudinal (T_1) and transverse (T_2) ^{15}N relaxation data and $\{^1H\}$-^{15}N heteronuclear NOE data (Fig. 6) to explore the dynamic dynamic behavior of RV-C02 2A^{pro}. The rotational correlation time (τ_c) for RV-C02 2A^{pro} was estimated to be about 10.5 ns. Except for the 5 C-terminal residues, the structure shows very little internal motion over the whole sequence including for the loop regions. This appears to be a common feature of picornaviral proteinases (Skern et al., 2002). Even though we found little evidence for internal motion, it is interesting that the peaks in ^{1}H,^{15}N-HSQC data sets do not have uniform intensity across the spectra suggesting some structure heterogeneity. This phenomenon is also in agreement with previous studies on EV-CB4 2A^{pro} by Baxter group.

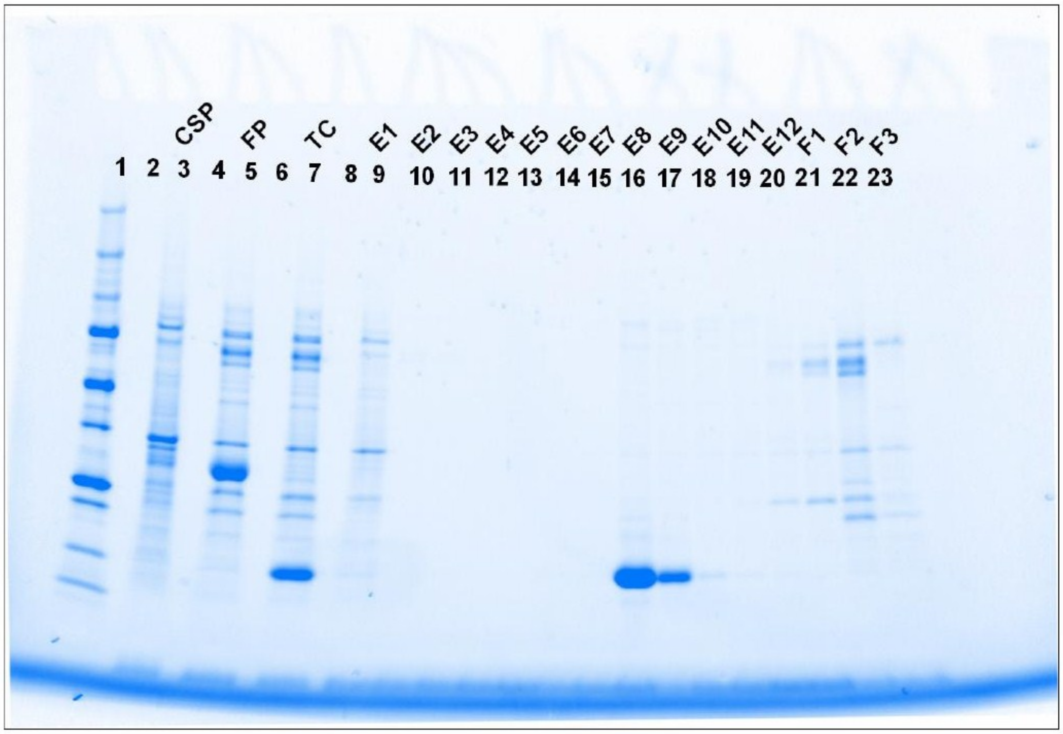


Figure 1. Final purification step for inactive RV-C02 2A^{pro}. Lane 16 and 17 were taken to the sample tube and frozen to the droplet for the storage. Yield was 5.05 mg and it was separated to 3.4 mg and leftover, and 3.4 mg was diluted to 0.5mM for NMR experiments.

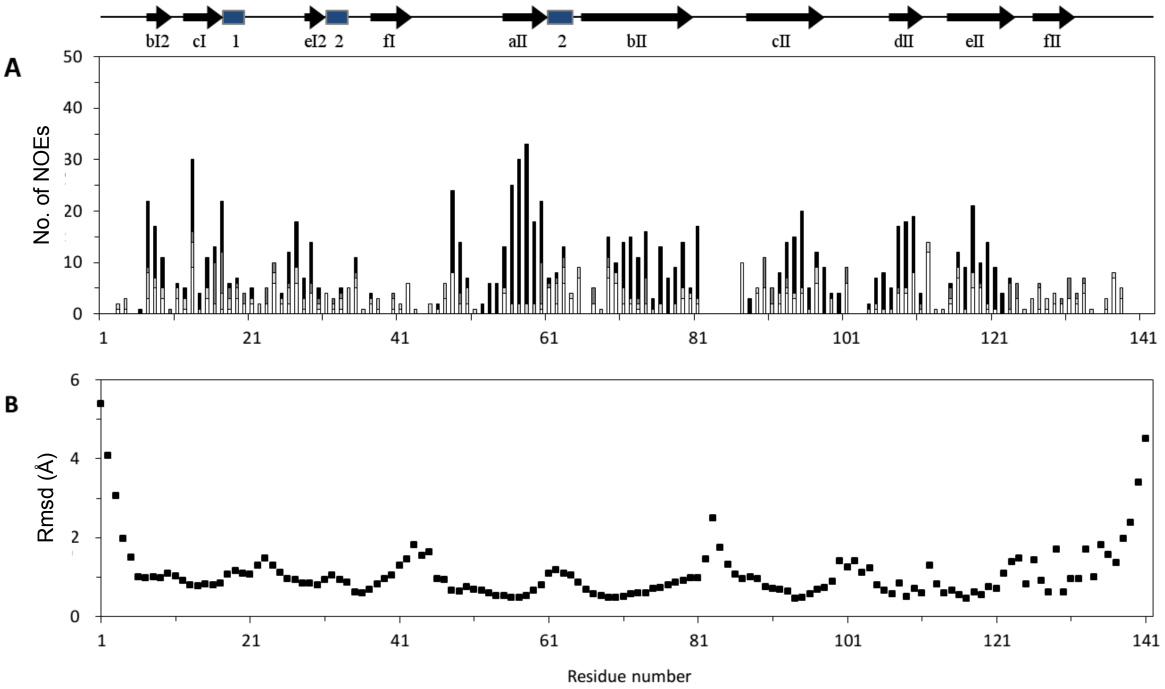


Figure 2. Structural properties of RV-C02 $2A^{pro}$. (A) Bar graph showing the number of NOE constraints used for the structure calculation: (from bottom to top) white bars, intraresidue; light gray bars; sequential; dark gray bars, medium range; and black bars, long range constraints. (B) Rmsd values for backbone atoms (N, Cα, and C′) of the best 15 models from the average structure. Structurally compact regions have rmsd values below 2 Å. Secondary structural features derived from the NMR solution structure are displayed at the top of the figure (arrows represent β-strands, and rectangles represent 3_{10} helices.

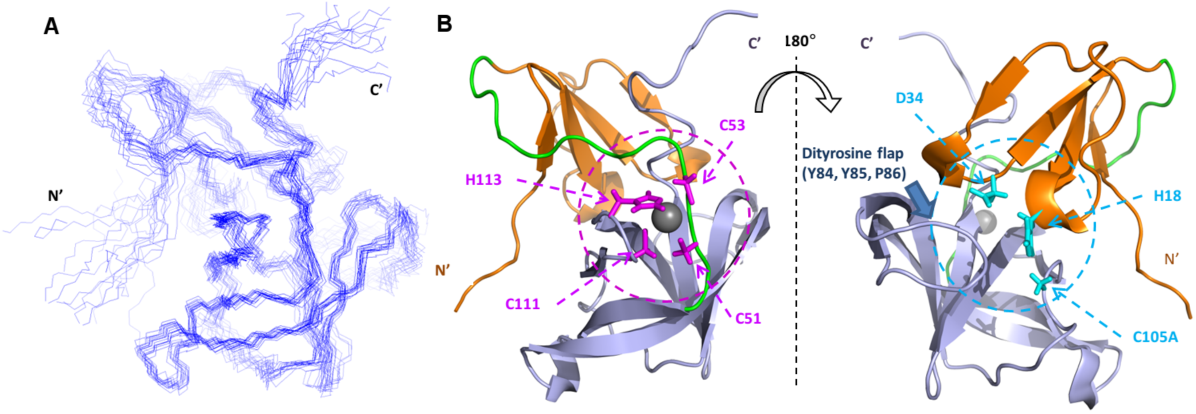


Figure 3. Three-dimensional solution structure of the viral proteinase RV-C02 2A^{pro}. (A) Bundle of the 15 best models with the backbone atoms (N, Cα, C') of the regions of regular secondary structure superimposed by the MOLMOL program (107). (B) Ribbon diagram of the lowest energy model. Shown in magenta are stick representations of the side chains of the residues (C51, C53, C111, H113) ligating the zinc ion (gray sphere) . Side chains of the residues forming the catalytic triad (H18, D34, C105A) are represented in cyan.

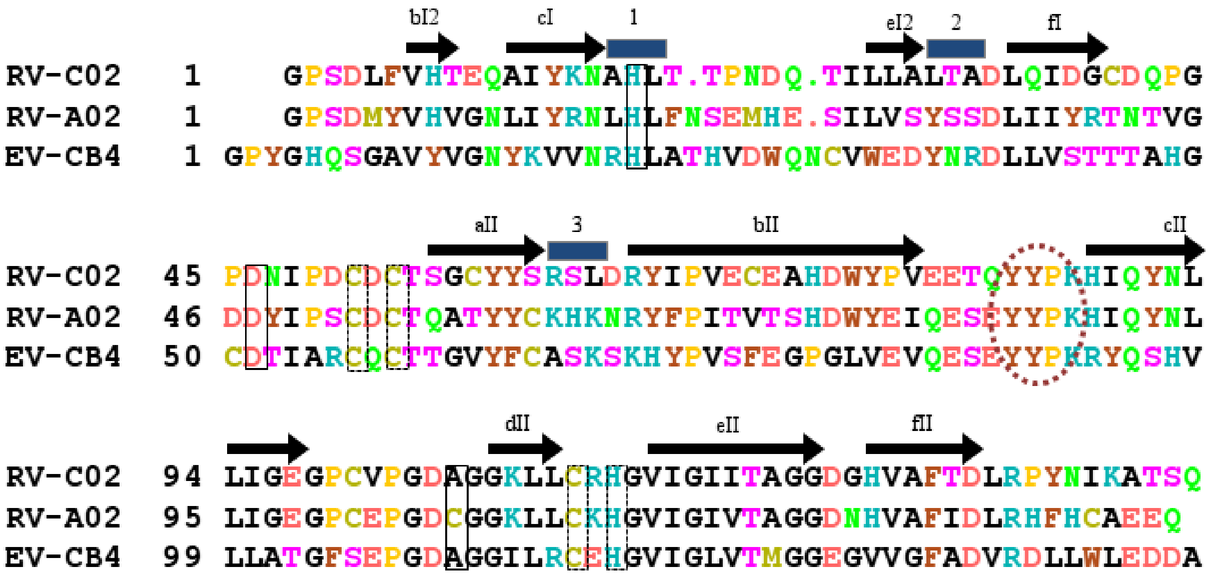

bI2
cI
1
eI2
2
fI
RV-C02 1 GPSDLFVHTEQAIYKNAHLT.TPNDQ.TILLALTADLQIDGCDQPG
RV-A02 1 GPSDMYVHVGNLIYRNLHLFNSEMHE.SILVSYSSDLIIYRTNTVG
EV-CB4 1 GPYGHQSGAVYVGNYKVVNRHLATHVDWQNCVWEDYNRDLLVSTTTAHG
aII
3
bII
cII
RV-C02 45 PDNIPDCDCTSGCYYSRSLDRYIPVECEAHDWYPVEETQYYPKHIQYNL
RV-A02 46 DDYIPSCDCTQATYYCKHKNRYFPITVTSHDWYEIQESEYYPKHIQYNL
EV-CB4 50 CDTIARCQCTTGVYFCASKSKHYPVSFEGPGLVEVQESEYYPKRYQSHV
dII
eII
fII
RV-C02 94 LIGEGPCVPGDAGGKLLCRHGVIGIITAGGDGHVAFTDLRPYNIKATSQ
RV-A02 95 LIGEGPCEPGDCGGKLLCKHGVIGIVTAGGDNHVAFIDLRHFHCAEEQ
EV-CB4 99 LLATGFSEPGDAGGILRCEHGVIGLVTMGGEGVVGFADVRDLLWLEDDA

Figure 4. Multiple sequence alignment of of the picornaviral 2A proteinases RV-C02, RV-A02, and EV-CB4 2A^{pro} by the STRAP program (113). Residues that make up the catalytic triad (H13, D46, and C105A in the RV-C02 and RV-A02 numbering system) are boxed by solid lines. Residues whose side chains ligate the zinc ion (C51, C53, C111, H113) are boxed by dashed lines. The dashed ellipse indicates the conserved YYP sequence that forms the di-tyrosine flap. The symbols above the sequences indicate secondary structural features (bars, helices; arrows, β-strands) coded by the nomenclature used with the RV-A02 2A^{pro} structures.

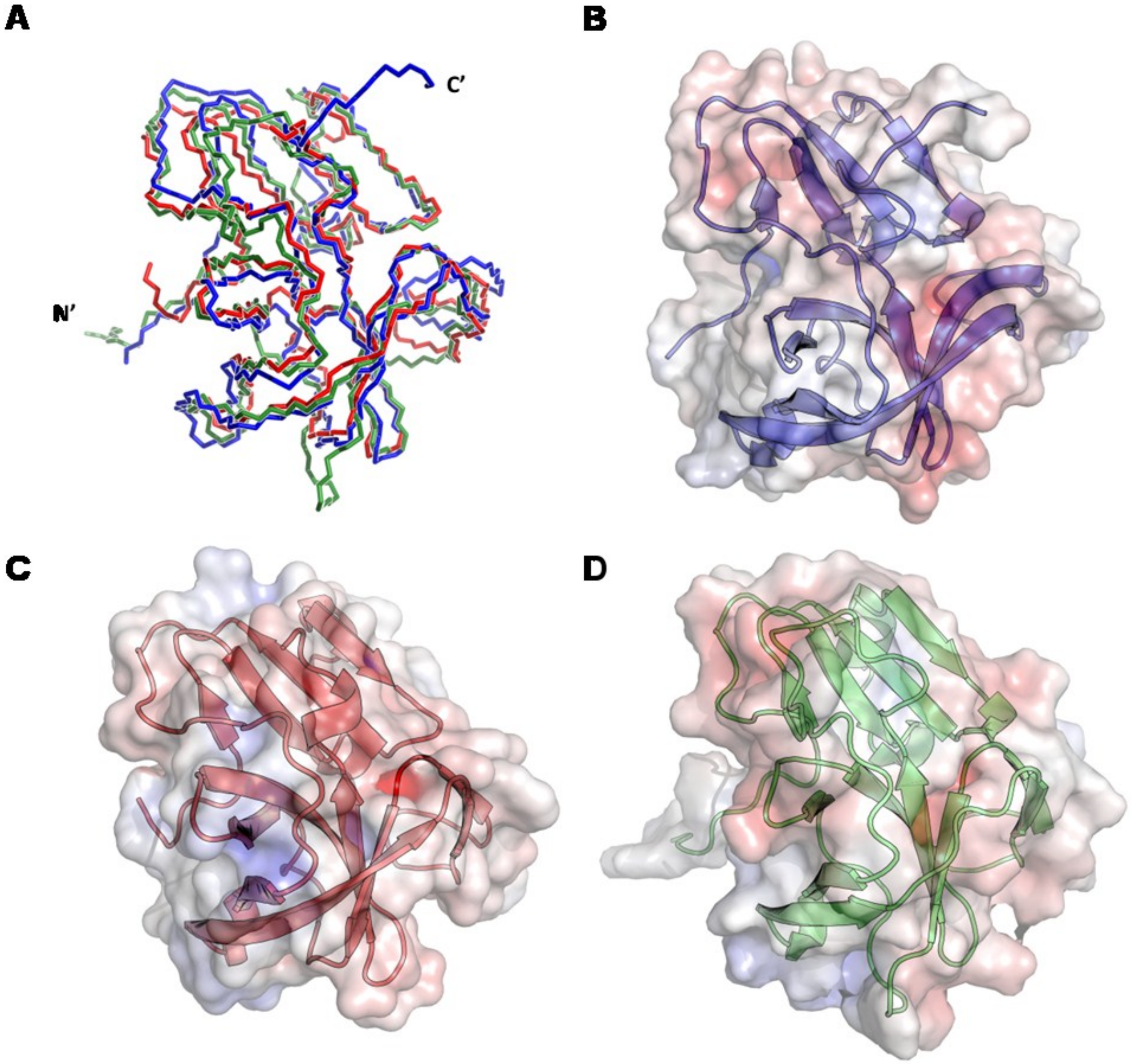


Figure 5. Comparison of the three-dimensional structures of RV-C02 2A^{pro} (blue), HRV-C 2A^{pro} (red), and EV-CB4 2A^{pro} (green). The spatial orientations of the three structures have been kept constant in A, B, C, and D. (A) Superimposition of backbones of the three proteinases showing their structural similarity, pairwise rmsd values are 1.809 Å for both [RV-C02 2A^{pro} and HRV-C 2A^{pro}] and [RV-C02 2A^{pro} and EV-CB4 2A^{pro}]. Poisson-Boltzmann electrostatic potential surfaces calculated by APBS (108) and PDB2PQR (109) with the limiting factor ±10.0 as

illustrated by the PyMOL program for (B) RV-C02 $2A^{pro}$, (C) RV-A02 $2A^{pro}$ and (D) EV-CB4 $2A^{pro}$.

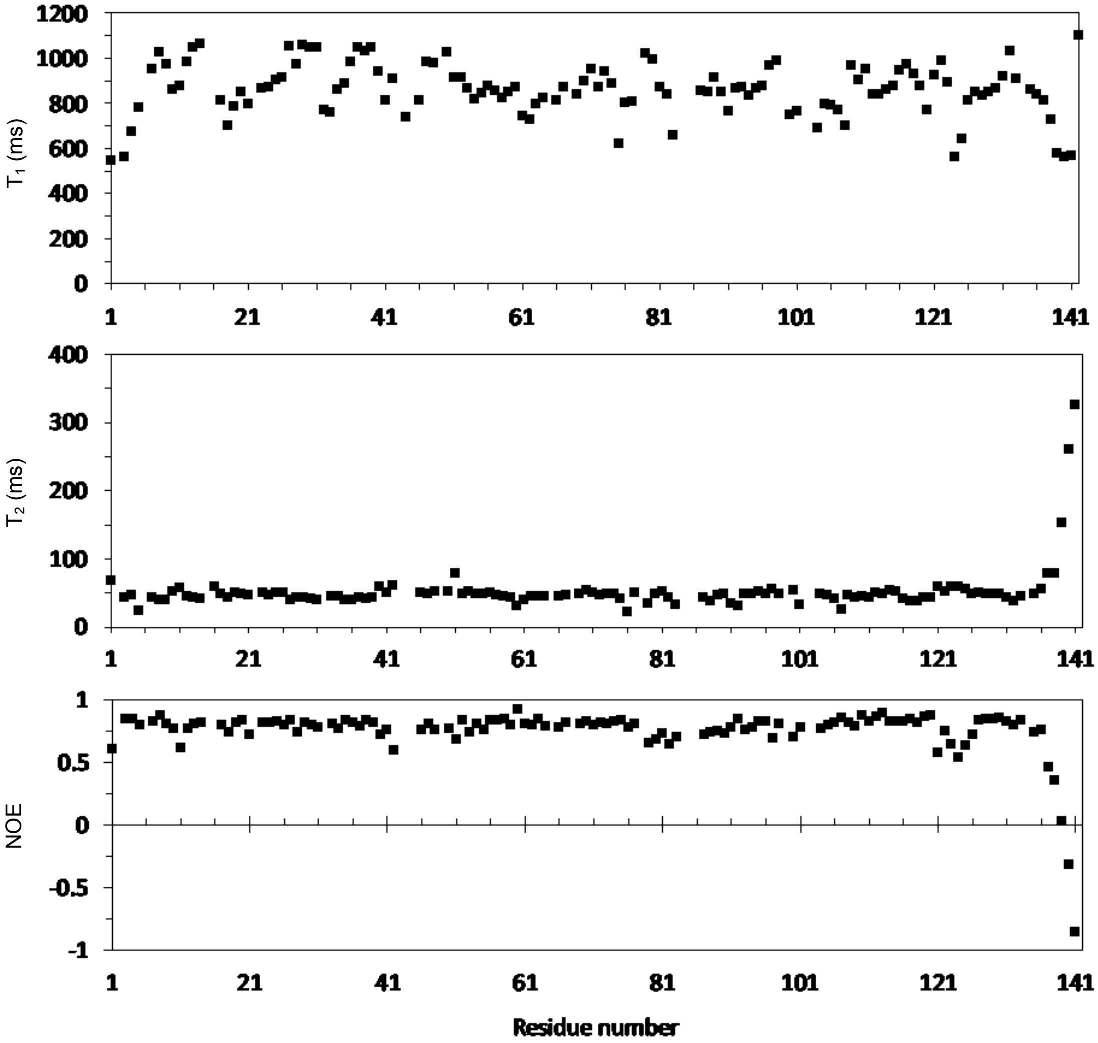


Figure 6. Longitudinal (T_1), and transverse (T_2) relaxation times and {^{1}H}, ^{15}N heteronuclear NOE data for the nitrogen atoms of RV-C02 2A^{pro}.

Table 1. Statistics for2 the NMR structure of RV-C02 $2A^{pro}$

Conformationally restricting distance constraints	
Intraresidue [i = j]	274
Sequential [(i– j) = 1]	181
Medium Range [1 < (i – j) ≤ 5]	148
Long Range [(i – j) > 5]	636
Total	1239
Dihedral angle constraints	
φ	70
ψ	72
Hydrogen-bond constraints	59
CYANA target function [Å]	3.49
Average rmsd to the mean CYANA coordinates [Å]	
Regular secondary structure elements, backbone heavy[1]	0.6
Regular secondary structure elements, all heavy atoms[1]	0.8
Backbone heavy atoms N, Cα, C′ (1–142)	1.5
All heavy atoms (1–142)	1.7
PROCHECK Z-scores (φ and Ψ/all dihedral angles)	-2.95/-5.62
MolProbity Mean score/Z-score	24.03/-2.60
Ramachandran plot summary for selected residue ranges from PROCHECK [%][1]	
Most favored regions	85.0
Additionally allowed regions	13.2
Generously allowed regions	1.8
Disallowed regions	0.0
Ramachandran plot summary for selected residue ranges from MolProbity [%][1]	
Most favored regions	93.6
Allowed regions	6.4
Disallowed regions	0.0
Average number of distance constraint violations per CYANA conformer	
0.2 – 0.5 Å	11
> 0.5 Å	0
Average number of angle constraint violations per CYANA conformer	
> 10°	0

[1] Stretches of regular secondary structure: 7-9, 12-16, 28-30, 35-39, 55-60, 65-74, 78-79, 88-96, 108-110, 115-122, 127-131

CHAPTER 6

PACSY, a Relational Database Management System for Protein Structure and Chemical Shift Analysis.

Adapted in part from Lee W, Yu W, Kim S, Chang I, Lee W, Markley JL (2012) PACSY, a relational database management system for protein structure and chemical shift analysis. *J Biomol NMR* 54:169

6.1 Introduction

The importance of three-dimensional structures of proteins derives from their relevance to biological function. It was recognized early on, when only a handful of X-ray structures of proteins had been solved, that it would be valuable to make the information available from a publicly accessible data bank, and this led to the establishment of Protein Data Bank (PDB) (Bernstein et al., 1977). The data format of the PDB has been extended, and the current Worldwide Protein Data Bank (wwPDB) now encompasses structural data from NMR spectroscopy as well as X-ray crystallography (Berman et al., 2007). Comparisons of three-dimensional structures provide information on evolutionary relationships, and analyses of this kind are available from the SCOP database (Structural Classification of Proteins, Murzin et al., 1995) and the CATH database (a hierarchic classification of protein domain structures, Orengo et al., 1997). Currently most of the structures in the PDB have been solved by either X-ray crystallography (87.7 %) or NMR spectroscopy (11.8 %); but a growing number of structures are being determined by electron microscopy (0.5 %). Although structure determination by NMR spectroscopy has limitations in that it is not as highly automated as X-ray crystallography and not as successful with large proteins or protein complexes, it offers several interesting features. NMR structures can be solved in solution under molecular conditions similar to those in vivo. NMR can be used to determine dynamic properties of proteins (both local and global). In addition, NMR as a spectroscopic approach can be used to determine thermodynamic and kinetic properties of proteins and their interactions with other molecules. The Biological Magnetic Resonance Bank (BMRB) (Ulrich et al., 2008) provides an archive for the full range of biomolecular NMR data, in addition to its role as the repository of chemical shifts and restraints associated with three-dimensional NMR structures as a partner in the wwPDB (Markley et al., 2008).

Structure calculations from NMR data typically depend on determining a variety of constraints, including distance constraints from NOE measurements, dihedral angle restraints from chemical shifts or spin-spin couplings, and/or projection angles between bond vectors from residual dipolar coupling measurements. It has long been recognized that NMR chemical shifts contain information on local structure, and this is the basis for the approaches used to determine secondary structure from NMR chemical shifts (Wishart and Sykes 1994; Wishart et al., 1992; Eghbalnia et al., 2005). Currently, TALOS (Torsion Angle Likelihood Obtained from Shifts and sequence similarity) (Cornilescu et al., 1999) and its successor TALOS+ (Shen et al., 2009) are the most popular software packages used to predict dihedral angles from NMR chemical shifts for use as angle constraints in structure calculations. The accuracy of such predictions can be improved by making use of homology modeling (Berhanskii et al., 2006). Several software packages have been developed that provide robust determinations of 3D structures from the available constraints: these include CYANA (Güntert 2004), ARIA (Bardiaux et al., 2012), CNS (Brunger et al., 1998), and Xplor-NIH (Schwieters et al., 2003).

Newer approaches to protein NMR data collection and analysis are streamlining and automating the steps in protein structure determination. Reduced dimensionality and sparse sampling approaches (Kim and Szyperski 2003; Schulte-Herbrüggen et al., 1999; Hiller et al., 2005; Eghbalnia et al., 2005; Gledhill and Wand 2011; Stanek and Kozminski 2010; Hyberts et al., 2010; Bahrami et al., 2012) are speeding up NMR data collection. Furthermore, the use of protein modeling approaches along with a chemical shifts as constraints appears very promising, particularly for small proteins (Sgourakis et al., 2011; Shen et al., 2009; Shen et al., 2008).

Although clear relationships have been found between 3D structure and NMR parameters (e.g., chemical shifts, J-coupling constants, RDC values), tools are lacking that enable the combined analysis of data from the PDB, BMRB, and SCOP databases. One of the reasons for this is that PDB and BMRB data are stored in flat-file formats, versions of the Self-defining Text Archive and Retrieval (STAR) file format (Hall and Spadaccini 1994). As an aid to easier and faster handling of the huge information content of these databases, we have developed the PACSY (Protein structure And Chemical Shift spectroscopY) database, which utilizes a relational database management system (RDBMS), to manage information derived from the PDB, BMRB, and SCOP databases. We describe how information from each database is extracted and processed to make them cross-related one another to enable queries

6.2 Materials and Methods

6.2.1 Database design

The PACSY database was designed to store and distribute information from protein structures and NMR experiments. PACSY makes use of an RDBMS (Relational Database Management System) to implement its data submission and request features. The data are stored and maintained by the RDBMS server, and the SQL language is used for data management. An RDBMS offers advantages over a file-based database server. First, it is possible to avoid database anomalies by separating tables through database normalization (Codd 1970). In addition, data consistency can be maintained by synchronous management and parallel control and data can be standardized by organizing methods for data expression. Data integrity and recovery are additional benefits of an RDBMS database server. We developed a tool called “PACSY Maker” to create and maintain the database, and because the SQL language is not easy to learn to operate and manage, we have developed a second tool called “PACSY Analyzer” to facilitate queries.

Fig. 1 illustrates how PACSY is organized. Data from the BMRB ftp archive are acquired as a dbmatch.csv file. Structural information from the PDB and chemical shift information from BMRB are extracted. The PACSY Maker software then processes these data with STRIDE (Frishman and Argos 1995), combines them with SCOP data, and parses the resulting data into a set of tables and fields in the prepared RDBMS server. The data stored in the RDBMS server can be accessed by various database client application interfaces (APIs): open database connectivity (ODBC) software, Oracle’s Database Express, MySQL Connector/PHP, or Microsoft's ActiveX Data Objects (ADO). The PACSY Maker program automates the building and updating of the database. It generates SQL dump files and an insertion script file.

PACSY consists of six different types of tables (Table 1). When the database is being built, PACSY Maker extracts and processes necessary information to fill these tables from PDB, BMRB, and SCOP. STRIDE is used to calculate secondary structure and solvent accessible surface area (SAS) (Lee and Richards 1971), and hydrophobicity scales are calculated from the SAS. The SAS values from residues are divided by those calculated from Gly-X-Gly by numerical integration to yield the relative solvent exposure. The separation of table types avoids storage of repetitive information (known as data anomalies). The “X” in front of a table type, stands for one of the 20 standard amino acids. Thus tables, *X_DB*, *X_STRC_DB*, *X_CS_DB* and

X_COORD_DB are each actually 20 tables. Each type of table has a *KEY_ID* field. Thus, if chemical shift information about a certain residue is requested, it can be performed by querying both the *X_CS_DB* and *X_DB* with same *KEY_ID.* Whereas other table types each consist of 20 amino-acid-specific tables, *SEQ_DB* and *SCOP_DB* are single tables. They also have a *KEY_ID* field, whose value matches that of the *X_CS_DB* for the first residue of protein sequence.

6.2.2 Software design

The PACSY Maker software was developed in C++ with the Qt Developer Library (http://qt.nokia.com) for automated database generation. It builds the PACSY database by automating the flowchart shown in Fig. 2. It has the simple graphical user interface (GUI) shown in Fig. 3A, which is used to set up a working directory to store downloaded files from the PDB, BMRB, and SCOP databases along with processed files, such as SQL dump files and an insertion script file. Once a root of the working directory is set up, other directories for storage and processes are created automatically as relative directories. The user can modify those directories for more detailed setup. PACSY Maker downloads *dbmatch.csv* from BMRB ftp archive when it is executed (Fig. 2). The file, *dbmatch.csv*, contains information on how BMRB entries are related to entries in other databases such as PDB, Swiss-Prot, and EMBL (Gattiker et al., 2003; Guenter et al., 2002). PACSY Maker processes the file to contain only information from PDB and BMRB submitted by a common author, and checks for needed updates by comparing the results to a recently processed *dbmatch.csv* file. Next, PACSY Maker downloads the SCOP database, and parses it to add structural classification information to each PDB entry. Finally, PACSY Maker downloads PDB and BMRB files from the respective web archive that match the update list made by comparing the new and old processed *dbmatch.csv* files. Because BMRB has not converted fully from the old NMR-STAR v2.1 to the new NMR-STAR v3.1 file format, PACSY Maker has a parser for both file formats. PACSY Maker downloads the v3.1 file if it exists or, if not, downloads the 2.1 file. Of all the processes, this step takes the longest time, and the duration depends on the Internet bandwidth of the computer building the database. The initial run of PACSY Maker typically takes two hours, but after the initial database creation, updates take only a few minutes. Because the PDB entry for a protein structure typically contains coordinates for multiple conformers, PACSY separates these prior to analysis by STRIDE. The model splitter module in PACSY Maker splits the downloaded PDB entry into files containing single structural models. PACSY Maker then creates output files with residues classified into six secondary structure types (H; α-helix, E; β-strand, T; turn, G; 3_{10} helix, C; coil, B; isolated β-bridge), solvent accessible surface area (SAS), and dihedral angles (PHI, PSI). PACSY Maker reads the outputs, and calculates the hydrophobicity scale of each residue from its SAS by dividing by pre-defined values of SAS of Gly-X-Gly.

The next step is to build the PACSY database. PACSY Maker generates SQL dump files and an insertion script files for RDBMS servers. First, an SQL dump file, *initdb.dmp*, is generated for initialization. It cleans existing tables and creates new tables. To (re)generate a completely new PACSY DB, the user reactivates commented-out lines in *initdb.dmp* to erase all pre-existing data. Otherwise, this file is left unedited. Second, SQL dump files containing actual data, *X_DB_#.dmp*, *COORD_DB_#.dmp*, and *CS_DB_#.dmp*, are generated. The “#” characters indicate incremented indices that start from zero. For compatibility with 32-bit operating systems that handle files only smaller than 2 GB, PACSY Maker utilizes a strategy to limit file sizes.

Finally, PACSY Maker generates an *insertSQL.sh* file for executing other SQL dump files. The *insertSQL.sh* file, which is specific for the relational database, has the following structure forMySQL :

mysql -u USERNAME -pPASSWORD DBNAME < SQL DUMP FILE

If PostgreSQL, another popular open source database server, is used, the script file would be:

psql -U USERNAME -d DBNAME -f SQL DUMP FILE

Because PACSY Maker generates SQL dump files with general SQL sentences, only minimal changes are needed for field types in the *initdb.dmp*. However, field types are not always compatible between database servers. For example, MySQL requires a specific length of characters for the *TEXT* field type, whereas PostgreSQL supports variable length of *TEXT* field type. Another difference is in the nomenclature of the 8-bit floating variable: DOUBLE is used by MySQL, whereas FLOAT8 is used by for PostgreSQL. These minor changes can be easily carried out by use of any text editor.

After the *insertSQL.sh* file and *initdb.dmp* have been modified as needed, the *insertSQL.sh* can be executed for database creation. These database creation steps took one day for an initial run on a 2.4 GHz quad-core machine running CentOS 5.5 64 bit.

Through its interface to the PACSY RDBMS server the client software PACSY Analyzer provides an easy graphical user interface (GUI) to the PACSY database (Fig. 3B). Although the SQL language supports a powerful and standardized way to query a database, its complexity can be a barrier to non-specialists. PACSY Analyzer provides graphical user interface that allows the user to select for search tables and fields in a dialog window. Once the selections are made, PACSY Analyzer generates an SQL sentence to be executed with the PACSY database. PACSY Analyzer is written in PASCAL language (FPC version 2.6) using Lazarus IDE (Integrated Development Environment, http://www.lazarus.freepascal.org) version 0.9.30. PACSY Analyzer supports any database server that has ODBC (Open Database Connectivity).

The dialog window has two tab controls, *Input Filters* and *Output Filters*, that are used to specify the input and output. The *Input Filters* tab sets the conditions for a search. For example, if the user wants to browse proteins whose data were collected between pH 6.0 and pH 8.0, the *pH* field in the *SEQ_DB* table is set to 6.0 and 8.0, and the "Add button" is clicked. If the user wants to search chemical shifts of only CA atoms, a filter is set in the *ATOM_NAME* field of the *X_CS_DB* table by typing CA in the text box. Filters can be set to select for any conditions supported by the PACSY database.

The *Output Filter* tab is the place where the user describes the desired output of the information to be grabbed by the *Input Filter*. From the example above of a search for proteins whose data were collected between pH 6.0 and pH 8.0, if the user wants a list of the PDB entries that satisfy this condition, the *PDB_ID* field in the *SEQ_DB* table is chosen as an output filter. From the example of a search for CA chemical shifts, if the user wants to see the mean value of chemical shifts satisfying the condition, AVG in the Statistics and C_SHIFT field of the X_CS_DB table is selected. After adding all input and output filters, the *Make* button is clicked to create the SQL sentence that will run the user's request. The generated sentence appears in a large text box. Users can verify or edit the SQL sentence as needed to refine the search. To commit the sentence, the *Query!* button is clicked.

Depending on the SQL query, the search can take seconds or hours. Simple queries, such as browsing chemical shifts under certain conditions, are very fast (usually less than a second). The example shown Table 4 requesting the statistics on alanine alpha-carbon chemical shifts from proteins with 80-100 residues at low pH (pH 3-5) took only one second. However, if the search is complex or if multiple searches are requested, more time will be required to complete the query. When multiple queries are entered, PACSY Analyzer generates a new SQL sentence after the previous one has been executed. The queried results are shown in a grid. PACSY Analyzer has a function that allows results to be exported in tab-delimited text format for use in a spreadsheet program such as Microsoft Excel or OpenOffice Spreadsheet.

6.3 Results

6.3.1 Database build

The PACSY database was built and installed for testing at the National Magnetic Resonance Facility at Madison (NMRFAM). PACSY Maker ran on a 64-bit CentOS 5.5 developmental server for an entire day to build and upload SQL dump script files for the initial database. The number of downloaded PDB and BMRB files were both 3745, and a data file was downloaded for SCOP. 473 Mb were consumed by BMRB files, whereas 18 Gb were consumed by PDB files. The size of SCOP database was only 5.8 Mb.

A MySQL 5 server was installed with default parameters. The uploading process was carried out by executing the *insertSQL.sh* file after editing the user account in the *insertSQL.sh* file to change the preset *USER* and *PASSWORD* values. Execution of the *insertSQL.sh* shell script made the stored PACSY database ready for use by the MySQL server. 8460 files were generated: 648 of *X_DB_*.dmp*, 204 of *CS_DB_*.dmp,* 7590 of *COORD_DB_*.dmp*, 16 of *SEQ_DB_*.dmp*, *initdb.dmp*, *insertSQL.sh* and *update.log* files. The total size of the files was 6943 Mb (mostly SQL dump files). It took approximately 4 h to upload the PACSY data into the prepared MySQL server.

6.3.2 Database composition

*A*fter the files were uploaded to the server, the overall volume of PACSY storage was estimated at 5639 Mb. Because PACSY contains only data, its size is smaller than the SQL files, which contain commands, brackets, quotes, and other SQL-related information. Of the six table types in PACSY (Table 1), the *X_COORD_DB* tables are the most space-consuming, because they contain the atom coordinates from each of the multiple conformers that represent the NMR structure of the protein as deposited in the PDB. The file *X_STRC_DB* also contains all of the structural models in the PDB entry, however, *X_STRC_DB* is smaller because it contains only one record per residue whereas *X_COORD_DB* contains all of the atom coordinates. *X_STRC_DB* has a field named *MODEL_NO*, which indicates the model number in the PDB. This makes it possible to select a particular structural model, such as the one with the lowest energy. As in SCOP and CATH, the SEQ_DB refers to chains rather than to structures; currently, 7395 chains are represented.

6.3.3 Nomenclature

PACSY is consistent with the IUPAC recommendations (Markley et al., 1998), which are followed by PDB and BMRB. PACSY Maker adopts the atom names from PDB and BMRB. PACSY does not use pseudo atom nomenclature; however, ambiguously assigned atoms are represented by the same chemical shift values. A field named *AMBIGUITY* in the X_CS_DB tables carries the information; as in the BMRB, a value of 1 in the field indicates that the assignment is unambiguous, whereas a value of 2 indicates ambiguity.

6.3.4 PACSY statistics

Statistics were collected from PACSY to confirm both the availability and feasibility of database queries. Because the PACSY database employs a client-server concept, it supports many different options, including remote operation (Fig. 1). Because PACSY Analyzer utilizes an ODBC connection to the database server, in our case MySQL 5.0, we first installed and set up ODBC Connector. Next, we used PACSY Analyzer to determine the structural classification of PACSY entries as defined by the SCOP database (Table 2). SCOP does not cover all PDB entries because full classification is not automated. Csaba's study in 2009 revealed that the SCOP database version 1.73 covered 35.5% of all PDB entries whereas CATH database version 3.1.0 covered 32.0% (Csaba et al., 2009). Furthermore, Jefferson and co-workers found that for single domain classifications of the type commonly found in NMR structures, coverage of CATH by SCOP was greater than that of SCOP by CATH (Jefferson et al., 2008). We found that the SCOP 1.73 database provided 43% coverage. Because PACSY contains structural classification information, it is possible to investigate proteins by fold class. Apart from unclassified entries, the largest class of PDB and BMRB entries were for all-alpha proteins (745 entries, Table 2). Other major classes are well represented, except for multi-domain proteins (no entries, Table 2).

We also determined the mean and standard deviation values of the chemical shifts of the backbone atoms ($^{13}C^{\alpha}$, $^{13}C'$, ^{15}N, ^{1}H, $^{1}H^{\alpha}$) as a function of 6 secondary structure types. The values were calculated by a short Python script for the 20 amino acids and 6 secondary structure types. Strong relationships between local structure and chemical shifts are known to exist (Iwadate et al., 1999; Moon and Case 2007; Vila et al., 2009; Meiler 2003; Kohlhoff et al., 2009; Han et al., 2011). Thus, we expected to see distinct chemical shift differences between secondary structures, particularly three major secondary structure types, α-helix, β-strand, and random coil residues. Fig. 4A and Table 3A shows results for the alpha carbon ($^{13}C^{\alpha}$) chemical shifts. To visualize the distinctions between amino acid and structure types, we calculated differences for each of the 6 structure types from the average over all 6 (Fig. 4B). Statistics for the chemical shifts of the four other backbone atom types are also in Table 3. The results show that mean chemical shifts differ by amino acid type, atom type, and secondary structure class, α-helix (H), β-strand (E), or coil (C). For example, the mean ^{15}N chemical shifts of Ala in α-helical (121.72 ppm) and coil (124.48 ppm) environments differ by 2.76 ppm, whereas those for β-strand (124.85 ppm) and coil (124.48 ppm) differ by only 0.37 ppm. By contrast, the mean ^{15}N chemical shifts of Thr in α-helical (114.86 ppm) and coil (114.99 ppm) environments differ by only 0.13 ppm, whereas those for β-strand (117.70 ppm) and coil (114.99 ppm) differ by 2.71 ppm. This kind of analysis can be refined by any of the conditions available in the PACSY database, e.g., pH, temperature, hydrophobicity scale or solvent accessible surface area (SAS).

6.3.5 Practical example

We provide an example of how PACSY can be used in practice (Table 4). We assume that a novel protein of interest contains 90-residues and has been shown to be all α-helical and stable at low pH. As an aid to assignment, we are interested in knowing the range of $^{13}C^{\alpha}$ chemical shifts for an alanine residue under these conditions and how the chemical shift may depend on backbone torsion angles (PHI, and PSI), solvent accessible surface (SAS), and hydrophobicity scale (HDO_PBT).

The conditions to be searched were inserted into the Input Filter tab of PACSY Analyzer (Table 4A). To limit the size of proteins to those near 90 residues, we set the residue number (SEQ_COUNT) to "80" (minimum) and "100" (maximum); to limit the output to alanine residues, we set CLASS to "A"; to include a range of low pH values we set PH to "3"(minimum) and "5" (maximum); because we were not interested in comparing multiple conformers representing solution structures, we set MODEL_NO to "1"; since we were interested in all helical proteins, we set SND_STRC to "H" and EDGE to "N" (no mixed secondary structure); to limit the atom queried to Cα, we set ATOM_NAME to "CA".

We specified the desired output data in the Output Filter tab of PACSY Analyzer (Table 4B). Items requested from the sequence database (SEQ_DB) were: the PDB structure identifier (PDB_ID); the chain designator (for proteins containing more than one peptide) (CHAIN_ID); the BiomagResBank, accession number (BMRB_ID), the total number of residues in the chain (SEQ_COUNT), the pH (PH) and temperature (TEMP) at which the NMR data were acquired. Requested from the chain database (X_DB) was the residue number (SEQ_ID) for the particular chain (CHAIN_ID). Items requested from the structure database (STRC_DB) were the ϕ (PHI) and ψ (PSI) torsion angles, the hydrophobicity scale (HDO_PBT), and the solvent accessible surface (SAS). The only request from the chemical shift database was the ^{13}C chemical shift (C_SHIFT).

PACSY Analyzer automatically generated the SQL sentence to be submitted to the PACSY database (Table 4C). The advanced search performed by the PACSY database took less than 10 s. The *Export* feature of PACSY Analyzer was used to save the result in comma-separated value (.csv) format for input into a spreadsheet or text editing program (Table 4D).

The search of the PDB, BMRB and SCOP databases identified 50 alanine residues in six proteins (PDB ID: 1HUE, 1AAB, 1QPU, 1XSX, 2JN6, 2JS1). Whereas the ^{13}C$^{\alpha}$ chemical shifts in the full BMRB have a mean value of 53.2 ppm and a standard deviation of 2.4 ppm range, this restricted search yielded a mean value of 53.8 with a standard deviation of 1.30. The results could be filtered further, for example on the basis of ϕ,ψ angles.

We wrote a short Python script to show another practical application of PACSY (available from the PACSY website). We used PACSY to determine the correlation between chemical shifts and hydrophobicity scale values. Fig 5 shows how chemical shift and hydrophobicity are related in for alanine residues. In addition, the trend is also depended on secondary structure type. If the secondary structure was α-helix, the chemical shift tended to increase when the residue was more exposed to the solvent. On the other hand, if the secondary structure was β-strand, the chemical shift tended to decrease when the residue was more exposed. The exposure rate predicted from the chemical shift could be used to add another target function to structure calculations.

6.4 Conclusions

PACSY introduces a way of both storing and categorizing structural and chemical shift data. It supports easy data queries based on information from the PDB, BMRB and SCOP databases. To create this environment, we first defined the database structure and table descriptions; then we created the PACSY Maker program that automatically downloads, parses, processes, and stores data from PDB, BMRB, and SCOP. PACSY Analyzer was designed to make the PACSY database accessible to users without experience in creating SQL queries. PACSY Analyzer has a graphical user interface and automatic SQL generating function. As an initial test of the PACSY database, we carried out a query that returned the dependence of protein backbone chemical shifts on amino acid residue type and classification of their secondary structure (Table 3). The script used in that query is available from the PACSY website (http://pacsy.nmrfam.wisc.edu). We also show an example of how PACSY Analyzer can be used to generate a complex SQL query.

PACSY will enable research focused on the relationship between local structure and chemical shifts. Studies can employ as variables, temperature, pH, SCOP class, or sequence length. PACSY is easily extensible because it makes use of an RDBMS server. Users can make use of powerful SQL queries to edit the PACSY database for specific purposes. If a new feature needs to be added, the JOIN or ALTER commands can be used to modify table structures or to add another field. If an added feature is quite distinct from pre-existing tables, the table can be included in PACSY by specifying a KEY_ID field that refers to the PACSY database. We envision that PACSY will be found useful as a tool for assisting NMR peak assignments as illustrated by the practical example. Researchers interested in protein structure prediction from chemical shifts can filter the PACSY database to test hypotheses. These can involve coordinates from the X_COORD_DB, structure types from the X_STRC_DB, sequence information from SEQ_DB, and chemical shift information from X_CS_DB. PACSY also can be used as a NOESY simulator for known structures. Coordinates of hydrogens closer than 5 Å can be searched from the PACSY database, and the matching chemical shifts from the X_CS_DB table can be assembled to simulate NOESY.

The PACSY website (http://pacsy.nmrfam.wisc.edu) accepts SQL command line requests from users. Users unfamiliar with SQL can use PACSY Analyzer to generate SQL commands. For those who wish to build their own PACSY database, executable files for PACSY Maker and PACSY Analyzer are available from the website.

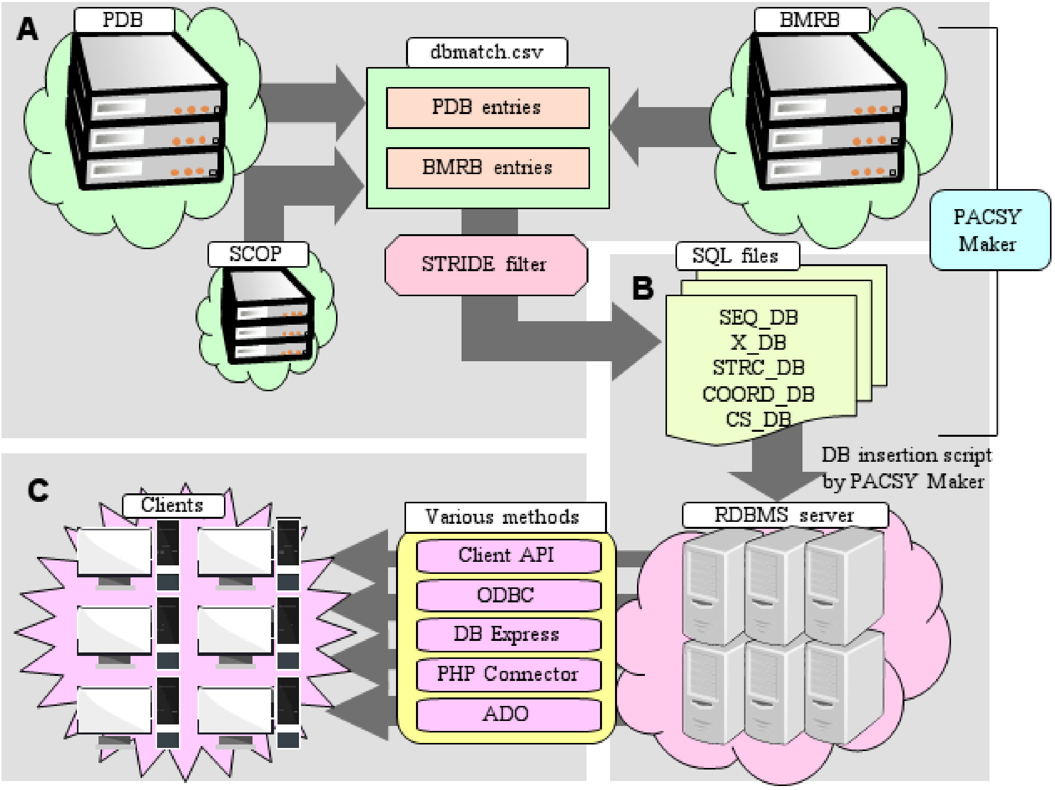


Figure 1. Schematic diagram of PACSY database. Building and maintaining the PACSY database consists of three steps. (A) First, PACSY Maker is used to generate SQL dump files and insertion scripts. Next dbmatch.csv file from the BMRB FTP site is analyzed to determine which entries should be incorporated in the database. (B) An RDBMS server should be set up before inserting the SQL files. The insertion script is written for MySQL; however, it can be modified for other RDBMS servers as long as SQL dump files are in general SQL grammars. The settings can be optimized to the particular server environment to improve performance. (C) The database can be served by the various connection methods supported by the RDBMS server.

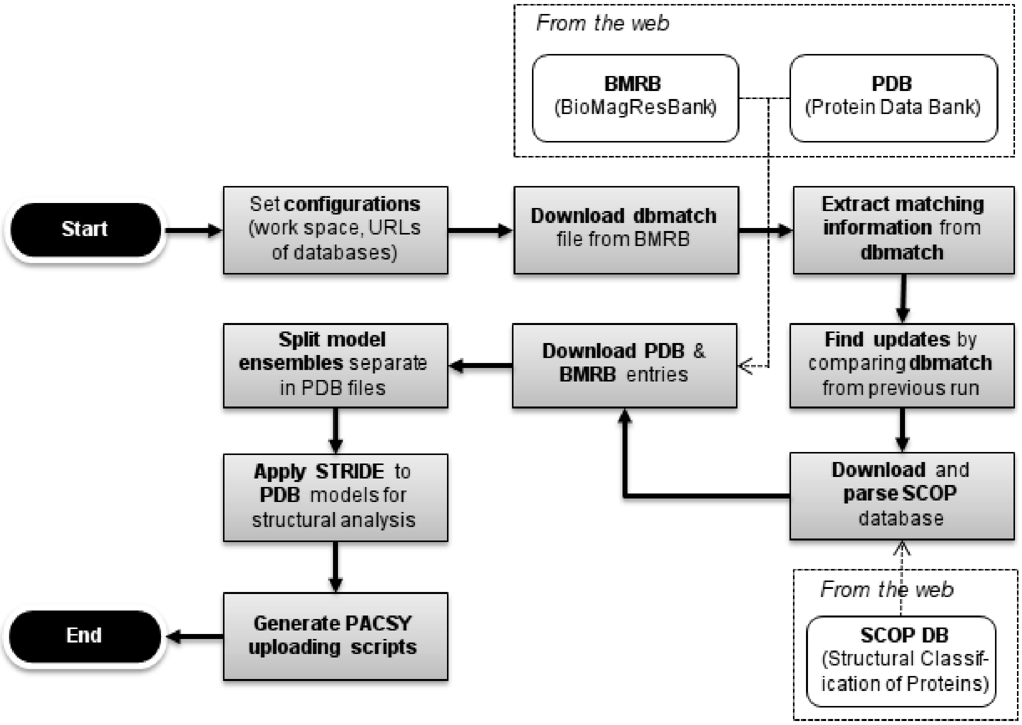
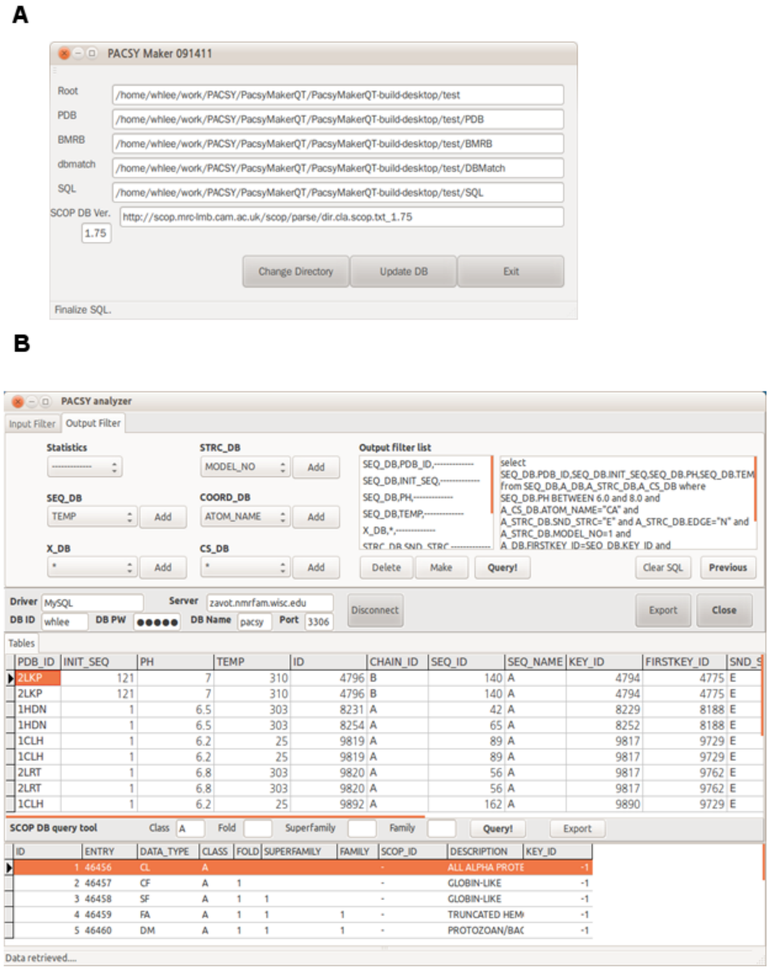

From the web
BMRB (BioMagResBank)
PDB (Protein Data Bank)
Start
Set configurations (work space, URLs of databases)
Download dbmatch file from BMRB
Extract matching information from dbmatch
Split model ensembles separate in PDB files
Download PDB & BMRB entries
Find updates by comparing dbmatch from previous run
Apply STRIDE to PDB models for structural analysis
Download and parse SCOP database
From the web
SCOP DB (Structural Classif-ication of Proteins)
End
Generate PACSY uploading scripts

Figure 2. Flowchart of PACSY Maker. The PACSY Maker program identifies entries to be updated and downloads them for processing. The STRIDE program provides structural information from PDB coordinates, such as dihedral angles, secondary structure and solvent accessible surface area. The hydrophobicity scale is calculated from the solvent accessible surface area from STRIDE. After the downloading, processing, and updating steps, PACSY Maker generates SQL dump files and an insertion script file for MySQL server.



Root /home/whlee/work/PACSY/PacsyMakerQT/PacsyMakerQT-build-desktop/test

PDB /home/whlee/work/PACSY/PacsyMakerQT/PacsyMakerQT-build-desktop/test/PDB

BMRB /home/whlee/work/PACSY/PacsyMakerQT/PacsyMakerQT-build-desktop/test/BMRB

dbmatch /home/whlee/work/PACSY/PacsyMakerQT/PacsyMakerQT-build-desktop/test/DBMatch

SQL /home/whlee/work/PACSY/PacsyMakerQT/PacsyMakerQT-build-desktop/test/SQL

SCOP DB Ver. 1.75 http://scop.mrc-lmb.cam.ac.uk/scop/parse/dir.cla.scop.txt_1.75

Finalize SQL.

Statistics -------------

SEQ_DB TEMP

X_DB *

STRC_DB MODEL_NO

COORD_DB ATOM_NAME

CS_DB *

Output filter list

```
SEQ_DB,PDB_ID,-------------
SEQ_DB,INIT_SEQ,-------------
SEQ_DB,PH,-------------
SEQ_DB,TEMP,-------------
X_DB,*,-------------
STRC_DB,SND_STRC,-------------
```

```
select
SEQ_DB.PDB_ID,SEQ_DB.INIT_SEQ,SEQ_DB.PH,SEQ_DB.TEM
from SEQ_DB,A_DB,A_STRC_DB,A_CS_DB where
SEQ_DB.PH BETWEEN 6.0 and 8.0 and
A_CS_DB.ATOM_NAME="CA" and
A_STRC_DB.SND_STRC="E" and A_STRC_DB.EDGE="N" and
A_STRC_DB.MODEL_NO=1 and
A_DB.FIRSTKEY_ID=SEQ_DB.KEY_ID and
```

Driver MySQL Server zavot.nmrfam.wisc.edu

DB ID whlee DB PW ●●●●● DB Name pacsy Port 3306

PDB_ID	INIT_SEQ	PH	TEMP	ID	CHAIN_ID	SEQ_ID	SEQ_NAME	KEY_ID	FIRSTKEY_ID	SND_S
2LKP	121	7	310	4796	B	140	A	4794	4775	E
2LKP	121	7	310	4796	B	140	A	4794	4775	E
1HDN	1	6.5	303	8231	A	42	A	8229	8188	E
1HDN	1	6.5	303	8254	A	65	A	8252	8188	E
1CLH	1	6.2	25	9819	A	89	A	9817	9729	E
1CLH	1	6.2	25	9819	A	89	A	9817	9729	E
2LRT	1	6.8	303	9820	A	56	A	9817	9762	E
2LRT	1	6.8	303	9820	A	56	A	9817	9762	E
1CLH	1	6.2	25	9892	A	162	A	9890	9729	E

SCOP DB query tool Class A Fold Superfamily Family

ID	ENTRY	DATA_TYPE	CLASS	FOLD	SUPERFAMILY	FAMILY	SCOP_ID	DESCRIPTION	KEY_ID
1	46456	CL	A				-	ALL ALPHA PROTE	-1
2	46457	CF	A	1			-	GLOBIN-LIKE	-1
3	46458	SF	A	1	1		-	GLOBIN-LIKE	-1
4	46459	FA	A	1	1	1	-	TRUNCATED HEM	-1
5	46460	DM	A	1	1	1	-	PROTOZOAN/BAC	-1

Data retrieved....

Figure 3. Screen shots of PACSY Maker and PACSY Analyzer. (A) PACSY Maker is a program with a simple user interface for setting up working directories. It is fully automated and does not require any user management. It takes a full day to download and process the PDB, BMRB, and SCOP databases. (B) We developed the PACSY Analyzer program to provide a user interface for users not fluent in the SQL language. *Input Filter* and *Output Filter* tabs allow the used to specify the input and output PACSY queries. The output from both SQL and PACSY Analyzer queries can be exported in comma-separated file format for external use.

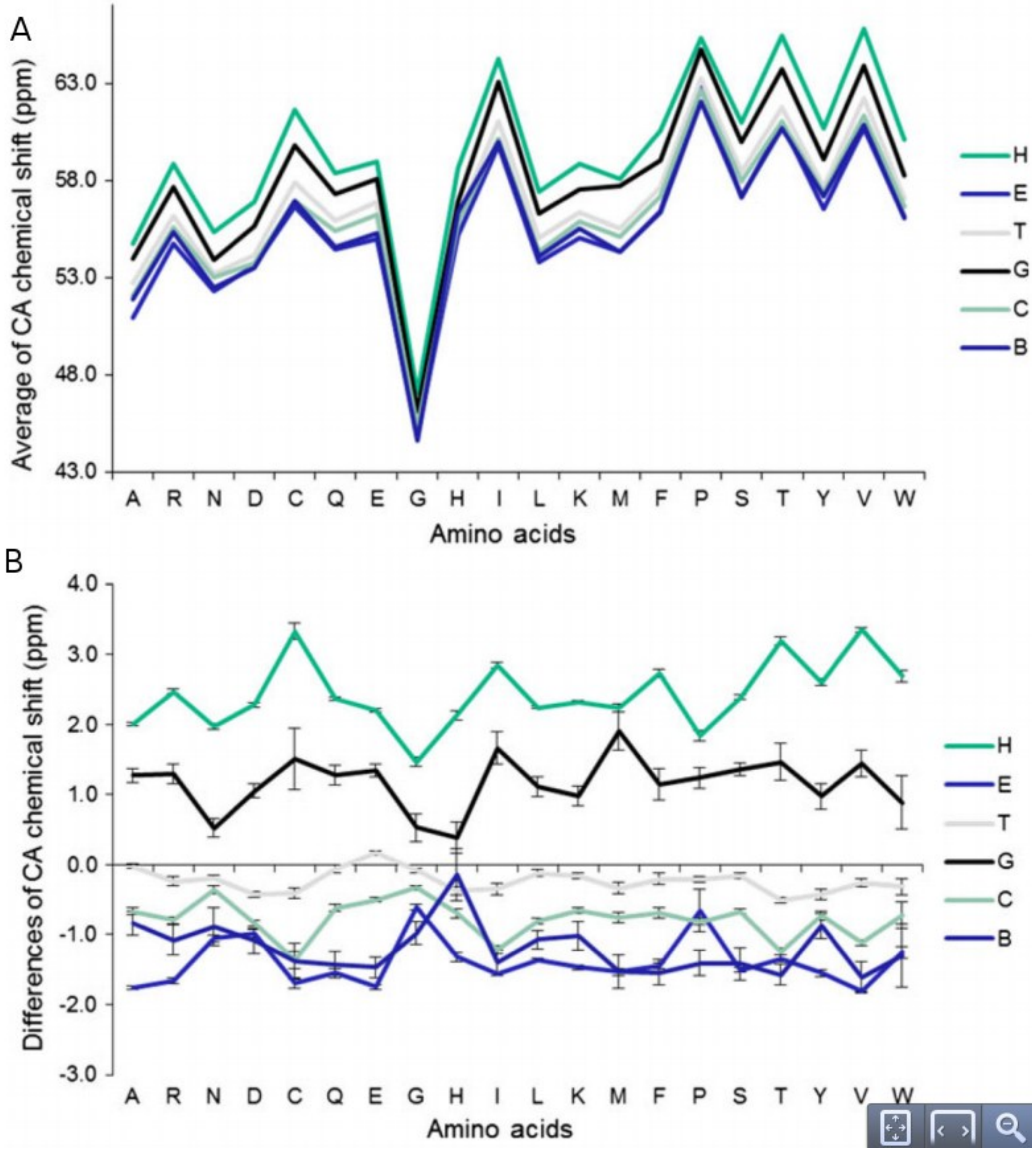


Figure 4. Mean $^{13}C^{\alpha}$ chemical shifts for different types of amino acids in the different classes of secondary structure in the PACSY database. We wrote a short Python script (available from the PACSY website) to collect chemical shift statistics on 5 major backbone atoms in the PACSY database only one of which is plotted here. The abbreviations for secondary structure classes are: α-helix (H), β-strand (E), turn (T), 3_{10} helix (G), coil (C), and isolated β-bridge (B).

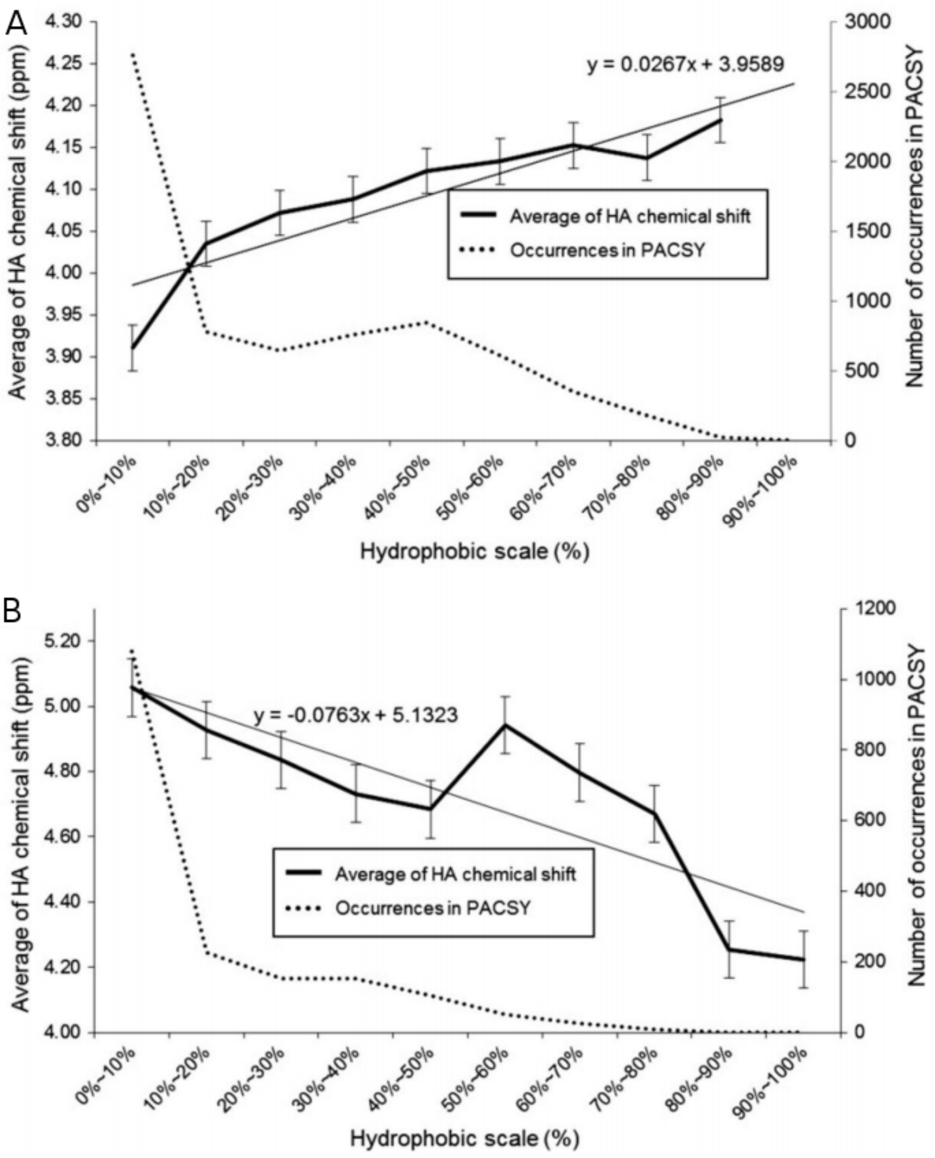


Figure 5. Mean alanine $^{1}H^{\alpha}$ chemical shifts as a function of the hydrophobicity value. The short Python script used to acquire the text data from PACSY and to draw this plot is available from the PACSY website. (A) Double Y-axis plot showing the mean alanine $^{1}H^{\alpha}$ chemical shifts for residues in α-helix and the number of occurrences as a function of hydrophobicity. (B) Double Y-axis plot drawn showing mean alanine $^{1}H^{\alpha}$ chemical shifts for residues in β-strand the and number of occurrences as a function of hydrophobicity.

Table 1. Description of the six types of tables in PACSY.

Table type	# of tables	# of fields	# of records	Contents
SEQ_DB	1	14	7,395	Basic information for entries: sequence, pH, temp, etc.
SCOP_DB	1	10	143,428	Information on the structural classification of proteins (SCOP).
X_DB	20	5	374,631	Residue-related information: e.g., chain ID, sequence ID, amino acid type.
X_STRC_DB	20	9	6,098,716	Structural information for a residue: e.g., secondary structure, dihedral angles, hydrophobicity scale, SAS, # of model.
X_COORD_DB	20	7	75,899,756	Coordinate information for an atom.
X_CS_DB	20	5	2,035,722	Chemical shift information for an atom including assignment ambiguity.

[*] This table represents PDB and BMRB data downloaded on February 7, 2012, and data from SCOP version 1.75.

Table 2. Classes of PACSY entries according to SCOP database.

Class	Number of entries
A- All alpha proteins	745
B- All beta proteins	555
C- Alpha and beta proteins (A/B)	580
D- Alpha and beta proteins (A+B)	443
E- Multi-domain proteins (alpha and beta)	0
F- Membrane and cell surface proteins and peptides	14
G- Small proteins	467
H- Coiled coil proteins	18
I- Low resolution protein structures	1
J- Peptides	166
K- Designed proteins	99
Unassigned	4,307
Total	7,395

[*] This table represents PDB and BMRB data downloaded on February 7, 2012, and data from SCOP version 1.75.

Table 3. Statistics for protein backbone chemical shifts as a function of amino acid type and secondary structure category derived from the PACSY database.

(A) $^{13}C^{\alpha}$ chemical shifts

Secondary structure category:	Helix (H)		β-strand (E)		Turn (T)		3_{10} helix (G)		Coil (C)		Isolated β-bridge (B)	
Amino acid	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.
Ala (A)	54.74	1.80	50.97	1.41	52.71	1.97	53.98	1.69	52.07	2.50	51.89	1.54
Arg (R)	58.89	1.91	54.75	1.97	56.17	2.74	57.70	1.89	55.62	1.75	55.34	2.08
Asn (N)	55.36	1.91	52.33	1.58	53.19	1.84	53.87	1.52	53.03	2.36	52.51	2.05
Asp (D)	56.88	1.70	53.61	2.02	54.19	2.31	55.63	1.72	53.78	1.90	53.52	1.65
Cys (C)	61.66	3.27	56.62	2.24	57.91	2.49	59.86	3.16	56.99	2.83	56.94	2.01
Gln (Q)	58.39	1.73	54.48	1.81	55.97	2.46	57.30	1.75	55.42	1.72	54.58	1.36
Glu (E)	58.95	1.78	55.01	1.47	56.93	1.96	58.11	1.84	56.25	1.86	55.27	1.60
Gly (G)	47.09	2.66	45.01	1.76	45.54	2.28	46.15	2.20	45.29	2.09	44.64	1.35
His (H)	58.62	2.43	55.17	1.88	56.12	2.00	56.88	2.00	55.80	2.02	56.35	2.25
Ile (I)	64.24	2.31	59.83	1.65	61.04	3.02	63.05	2.57	60.17	2.12	59.99	1.88
Leu (L)	57.41	1.71	53.79	1.58	55.05	2.29	56.28	2.41	54.36	1.92	54.08	1.55
Lys (K)	58.87	1.86	55.07	1.42	56.38	2.28	57.53	2.31	55.89	2.03	55.52	2.01
Met (M)	58.09	2.26	54.31	1.43	55.51	2.14	57.74	2.18	55.10	2.17	54.28	1.27
Phe (F)	60.61	2.51	56.43	1.86	57.69	2.58	59.03	2.49	57.19	2.40	56.33	1.75
Pro (P)	65.33	2.08	62.83	7.72	63.27	2.26	64.71	1.95	62.67	2.59	62.07	1.35
Ser (S)	61.01	1.69	57.12	1.52	58.47	2.26	59.97	1.56	57.96	1.70	57.20	2.45
Thr (T)	65.47	2.41	60.92	2.55	61.77	2.19	63.72	2.74	61.05	2.12	60.69	1.96
Tyr (Y)	60.70	2.08	56.54	1.95	57.68	2.35	59.04	2.06	57.37	2.00	57.23	1.59
Val (V)	65.84	2.22	60.67	1.75	62.23	2.46	63.94	2.22	61.37	1.88	60.90	2.25
Trp (W)	60.09	2.34	56.14	2.30	57.05	2.23	58.28	2.54	56.68	3.54	56.09	1.89

(B) $^{13}C'$ chemical shifts

Secondary structure category:	Helix (H)		β-strand (E)		Turn (T)		3_{10} helix (G)		Coil (C)		Isolated β-bridge (B)	
Amino acid	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.
Ala (A)	179.03	3.04	175.71	1.48	177.29	1.57	178.24	1.43	176.68	1.86	176.82	1.39
Arg (R)	178.11	1.47	174.79	1.49	175.73	3.56	177.17	1.58	175.60	1.55	175.46	1.67
Asn (N)	176.79	1.66	174.40	1.62	174.82	1.49	175.34	1.65	174.82	1.54	175.18	1.66
Asp (D)	178.04	1.47	175.35	1.63	175.93	1.80	176.73	1.36	175.76	1.39	175.97	1.62
Cys (C)	176.27	1.52	173.72	2.15	175.39	1.75	175.35	1.88	174.46	2.00	175.31	2.52
Gln (Q)	177.88	1.43	174.64	1.48	175.56	1.47	176.48	1.67	175.37	1.54	175.59	1.58
Glu (E)	178.31	2.91	175.07	1.33	176.25	1.35	177.10	1.53	175.80	3.36	175.39	1.38
Gly (G)	175.46	1.64	172.00	2.31	173.87	3.31	174.76	1.59	173.96	2.08	173.12	2.25
His (H)	176.76	1.59	173.96	2.66	174.87	1.60	175.65	1.95	174.66	1.69	175.10	2.15
Ile (I)	177.44	1.47	174.76	1.50	175.58	1.65	176.33	1.93	175.26	1.60	175.37	1.83
Leu (L)	178.26	1.54	175.43	1.76	176.59	1.66	177.75	1.64	176.24	1.67	176.65	1.70
Lys (K)	178.09	2.94	175.09	1.39	176.09	1.53	176.45	8.49	175.79	1.55	175.55	1.50
Met (M)	177.74	1.52	174.57	1.60	175.76	1.56	176.99	1.90	175.02	1.84	175.08	1.94
Phe (F)	176.85	1.57	174.27	1.81	175.30	1.65	175.84	1.57	174.77	1.82	174.12	1.73
Pro (P)	178.39	1.47	175.72	10.83	176.64	1.81	177.77	1.26	176.27	4.07	176.97	1.37
Ser (S)	176.06	1.47	173.36	1.70	174.42	1.56	175.23	1.58	174.22	1.42	174.02	2.12
Thr (T)	175.91	1.37	173.54	1.69	174.61	1.58	175.41	1.49	174.26	1.50	174.69	1.70
Tyr (Y)	177.15	1.57	174.28	1.63	175.14	1.72	175.36	1.60	174.74	1.57	174.42	2.15
Val (V)	177.32	3.96	174.63	1.53	175.61	1.56	176.88	1.42	175.24	1.44	175.30	1.83
Trp (W)	177.71	1.55	175.06	1.90	175.75	1.53	177.02	1.58	175.34	1.91	175.08	1.29

(C) ^{15}N chemical shifts

Secondary structure category:	Helix (H)		β-strand (E)		Turn (T)		3_{10} helix (G)		Coil (C)		Isolated β-bridge (B)	
Amino acid	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.
Ala (A)	121.72	2.58	124.87	4.27	123.62	3.86	121.96	3.93	124.48	5.29	124.54	3.82
Arg (R)	119.37	2.71	122.57	4.98	120.29	5.10	119.42	3.68	121.70	3.92	121.09	4.33
Asn (N)	117.63	2.63	121.37	5.63	118.77	4.66	117.19	3.30	119.47	3.90	120.30	4.68
Asp (D)	119.65	2.51	123.14	3.90	120.02	5.88	118.50	6.02	121.31	3.74	121.85	3.52
Cys (C)	118.46	3.52	121.93	4.88	119.94	4.83	118.42	3.77	120.79	4.50	121.67	4.49
Gln (Q)	118.57	3.31	122.02	5.27	119.42	4.88	118.09	2.93	120.74	4.89	122.28	3.67
Glu (E)	119.31	3.03	122.66	4.02	120.67	4.82	119.04	3.50	121.66	5.03	121.15	3.63
Gly (G)	107.81	4.80	109.77	4.08	110.25	4.86	108.24	4.34	110.13	4.05	109.54	3.20
His (H)	118.52	2.92	121.84	5.16	119.22	4.24	117.15	3.56	120.28	4.07	121.67	4.08
Ile (I)	119.71	2.97	123.28	4.36	120.58	4.48	118.94	5.65	121.50	4.90	122.40	4.55
Leu (L)	119.90	2.76	124.88	4.21	121.42	5.54	120.26	4.14	122.47	3.76	122.93	3.82
Lys (K)	119.47	2.73	123.18	4.09	120.90	4.07	119.01	3.46	122.11	3.76	121.95	3.99
Met (M)	118.56	2.71	122.37	3.91	120.26	3.65	119.76	4.28	120.93	4.59	121.09	3.63
Phe (F)	119.59	3.33	121.50	5.09	119.57	4.41	118.51	4.08	120.90	4.33	119.89	4.09
Pro (P)	122.53	9.07	106.32	36.34	123.54	12.32	123.45	8.52	125.82	12.08	122.45	4.33
Ser (S)	115.08	2.85	117.54	4.27	115.92	4.40	114.52	3.73	116.95	3.91	116.56	3.71
Thr (T)	114.87	4.24	117.68	5.93	113.80	5.67	112.88	5.67	114.99	4.18	113.82	4.90
Tyr (Y)	119.58	3.23	121.60	5.20	119.63	4.58	118.64	3.85	120.72	3.94	121.23	3.26
Val (V)	119.39	3.26	122.66	4.71	120.18	4.60	118.94	4.82	120.91	4.59	120.68	4.82
Trp (W)	120.43	4.94	122.94	4.30	119.53	7.86	121.33	3.75	121.96	7.41	122.78	2.60

(D) ^{1}H chemical shifts

Secondary structure category:	Helix (H)		β-strand (E)		Turn (T)		3_{10} helix (G)		Coil (C)		Isolated β-bridge (B)	
Amino acid	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg	Std. Dev.	Avg.	Std. Dev.
Ala (A)	8.10	0.54	8.63	0.65	8.17	0.60	8.11	0.63	8.21	0.56	8.32	0.66
Arg (R)	8.07	0.52	8.63	0.69	8.21	0.67	8.08	0.63	8.23	0.64	8.26	0.79
Asn (N)	8.19	0.55	8.61	0.64	8.41	0.72	8.19	0.56	8.29	0.60	8.43	0.81
Asp (D)	8.21	0.54	8.54	0.60	8.35	1.82	8.26	0.67	8.29	0.54	8.41	0.70
Cys (C)	8.16	0.68	8.76	0.63	8.24	0.71	8.06	0.55	8.37	0.67	8.42	0.64
Gln (Q)	8.08	0.94	8.59	0.64	8.22	0.66	8.19	0.63	8.27	0.58	8.54	0.64
Glu (E)	8.21	0.59	8.58	0.59	8.41	0.62	8.46	0.83	8.32	0.53	8.16	0.62
Gly (G)	8.27	0.55	8.33	0.83	8.42	1.55	8.35	0.51	8.23	0.58	8.24	0.71
His (H)	8.00	0.70	8.71	0.65	8.26	0.83	7.97	0.81	8.20	0.67	8.50	0.82
Ile (I)	8.03	0.51	8.74	0.59	8.01	0.71	8.04	0.82	8.07	0.74	8.43	0.81
Leu (L)	8.07	0.54	8.74	0.60	8.10	0.67	8.05	0.65	8.11	0.63	8.50	0.71
Lys (K)	7.99	0.55	8.55	0.61	8.22	0.65	8.01	0.56	8.22	0.54	8.27	0.63
Met (M)	8.11	0.49	8.72	0.63	8.19	0.59	8.14	0.68	8.27	0.51	8.45	0.64
Phe (F)	8.21	0.59	8.79	0.64	8.09	0.68	7.97	0.64	8.18	0.74	8.58	0.71
Pro (P)	-	-	-	-	-	-	-	-	-	-	-	-
Ser (S)	8.15	0.49	8.55	0.64	8.26	0.64	8.19	0.63	8.32	0.55	8.38	0.67
Thr (T)	8.05	0.50	8.59	0.62	8.13	0.66	8.03	0.72	8.25	0.62	8.35	0.65
Tyr (Y)	8.12	0.60	8.78	0.65	8.02	0.70	7.92	0.76	8.14	0.70	8.48	0.68
Val (V)	8.02	0.58	8.70	0.59	8.01	0.72	8.01	0.67	8.12	0.60	8.29	0.74
Trp (W)	8.25	1.86	8.70	0.65	7.84	0.83	8.01	0.62	8.21	0.80	8.32	0.70

(E) $^{1}H^{\alpha}$ chemical shifts

Secondary structure	Helix (H)		β-strand (E)		Turn (T)		3_{10} helix (G)		Coil (C)		Isolated β-bridge (B)	
Amino acid	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.
Ala (A)	4.03	0.31	4.89	0.48	4.24	0.34	4.11	0.34	4.33	0.35	4.54	0.37
Arg (R)	3.98	0.31	4.83	0.50	4.26	0.41	4.07	0.41	4.37	0.39	4.56	0.41
Asn (N)	4.47	0.27	5.06	0.47	4.63	0.36	4.63	0.30	4.70	0.32	5.02	0.51
Asp (D)	4.40	0.22	4.94	0.43	4.58	0.30	4.50	0.28	4.64	0.32	4.87	0.45
Cys (C)	4.16	0.56	5.02	0.74	4.62	0.56	4.43	0.55	4.69	0.50	4.78	0.50
Gln (Q)	4.00	0.28	4.82	0.45	4.27	0.36	4.15	0.33	4.37	0.34	4.61	0.34
Glu (E)	4.02	0.26	4.84	0.47	4.21	0.34	4.08	0.32	4.33	0.31	4.73	0.48
Gly (G)	3.80	0.37	4.05	0.63	3.93	0.37	3.89	0.42	3.96	0.31	4.00	0.52
His (H)	4.34	0.54	5.03	0.96	4.58	0.70	4.50	0.58	4.64	0.48	4.78	0.85
Ile (I)	3.68	0.34	4.66	0.45	4.12	0.38	3.95	0.44	4.26	0.43	4.49	0.44
Leu (L)	3.99	0.30	4.88	0.43	4.30	0.34	4.11	0.31	4.39	0.33	4.68	0.42
Lys (K)	3.99	0.30	4.79	0.46	4.25	0.40	4.10	0.39	4.32	0.36	4.58	0.46
Met (M)	4.10	0.37	4.97	0.47	4.42	0.37	4.21	0.37	4.45	0.34	4.83	0.45
Phe (F)	4.16	0.42	5.10	0.47	4.55	0.42	4.38	0.43	4.63	0.40	4.98	0.35
Pro (P)	4.20	0.31	4.56	0.48	4.37	0.39	4.20	0.43	4.43	0.32	4.65	0.40
Ser (S)	4.18	0.29	4.99	0.46	4.43	0.34	4.24	0.38	4.51	0.88	4.76	0.46
Thr (T)	3.98	0.30	4.91	0.45	4.37	0.36	4.13	0.37	4.45	0.34	4.79	0.41
Tyr (Y)	4.14	0.40	5.08	0.51	4.50	0.44	4.35	0.48	4.62	0.46	4.82	0.37
Val (V)	3.61	0.36	4.63	0.46	4.10	1.61	3.83	0.54	4.19	0.43	4.47	0.54
Trp (W)	4.27	0.40	5.15	0.46	4.57	0.44	4.39	0.45	4.68	0.45	4.95	0.48

Table 4. Example of the use of PACSY analyzer in carrying out an advanced search of the PACSY database.

(A) Input filter list for PACSY Analyzer

Table name	Field name	Value1	Value2
SEQ_DB	SEQ_COUNT	80	100
	CLASS	A	-
	PH	3	5
STRC_DB	MODEL_NO	1	-
	SND_STRC	H	-
	EDGE	N	-
	ATOM_NAME	CA	-

Abbreviations: SEQ_DB, sequence database; SEQ_Count, number of residues; CLASS, amino acid type (set at A for alanine); PH, pH value; STRC_DB, structure database; MODEL_NO, structural model designator (set as 1 for the first model), SND_STRC, secondary structure type (set at H for helix); EDGE, setting to allow for multiple secondary structure types (set at N to indicate no—only helix; ATOM_NAME, name of the atom queried (set at CA for α-carbon.

(B) Output filter list for PACSY Analyzer

Table name	Field name	Statistics
SEQ_DB	PDB_ID	-
	CHAIN_ID	-
	BMRB_ID	-
	SEQ_COUNT	-
	PH	-
	TEMP	-
X_DB	CHAIN_ID	-
	SEQ_ID	-
STRC_DB	PHI	-
	PSI	-
	HDO_PBT	-
	SAS	-
CS_DB	C_SHIFT	-

Abbreviations not given above: PDB_ID, Protein Data Bank, structure designator; BMRB_ID, BioMagResBank, accession number; TEMP, temperature at which NMR data were collected; X_DB, peptide chain database; CHAIN_ID, peptide chain designator; SEQ_ID, residue number; HDO_PBT, hydrophobicity; SAS, solvent accessible surface area (Å^2); CS_DB, chemical shift database; C_SHIFT, chemical shift.

(C) SQL sentence generated by PACSY Analyzer to be submitted to the PACSY database by PACSY Analyzer.

```
select SEQ_DB.PDB_ID, SEQ_DB.CHAIN_ID, SEQ_DB.BMRB_ID, SEQ_DB.SEQ_COUNT,
SEQ_DB.PH, SEQ_DB.TEMP, A_DB.SEQ_ID, A_STRC_DB.PHI, A_STRC_DB.PSI,
A_STRC_DB.HDO_PBT, A_STRC_DB.SAS, A_CS_DB.C_SHIFT from SEQ_DB,A_DB,
A_STRC_DB, A_CS_DB where SEQ_DB.CLASS="A" and SEQ_DB.PH BETWEEN 3 and 5 and
A_STRC_DB.MODEL_NO=1 and A_STRC_DB.SND_STRC="H" and A_STRC_DB.EDGE="N"
and A_CS_DB.ATOM_NAME="CA" and SEQ_DB.SEQ_COUNT BETWEEN 80 and 120 and
A_DB.KEY_ID BETWEEN (SEQ_DB.KEY_ID) and
(SEQ_DB.KEY_ID+SEQ_DB.SEQ_COUNT) and A_DB.KEY_ID=A_STRC_DB.KEY_ID and
A_DB.KEY_ID=A_CS_DB.KEY_ID;
```

(D) Query result of the SQL sentence submitted to the PACSY database

PDB_ID	CHAIN_ID	BMRB_ID	SEQ_COUNT	PH	TEMP	SEQ_ID	PHI	PSI	HDO_PBT	SAS	C_SHIFT
1HUE	A	4047	90	4.6	311	9	-65.60	-44.82	12.74	14.40	53.30
1HUE	A	4047	90	4.6	311	11	-65.63	-25.19	10.97	12.40	54.30
1HUE	A	4047	90	4.6	311	21	-70.36	-37.91	0.00	0.00	53.30
1HUE	A	4047	90	4.6	311	24	-58.27	-26.85	2.57	2.90	53.30
1HUE	A	4047	90	4.6	311	27	-121.04	-56.33	0.00	0.00	52.80
1HUE	A	4047	90	4.6	311	35	-60.78	-54.71	6.81	7.70	53.40
1HUE	A	4047	90	4.6	311	88	-47.91	-46.24	34.07	38.50	52.40
1HUE	B	4047	90	4.6	311	9	-66.16	-45.36	7.79	8.80	53.30
1HUE	B	4047	90	4.6	311	11	-65.79	-25.11	10.53	11.90	54.30
1HUE	B	4047	90	4.6	311	21	-71.54	-36.45	0.00	0.00	53.30
1HUE	B	4047	90	4.6	311	24	-60.39	-34.54	2.30	2.60	53.30
1HUE	B	4047	90	4.6	311	27	-120.10	-48.25	0.00	0.00	52.80
1HUE	B	4047	90	4.6	311	35	-66.46	-61.59	7.70	8.70	53.40
1HUE	B	4047	90	4.6	311	88	-51.72	-36.65	51.15	57.80	52.40
1AAB	A	4079	83	5	293	16	-53.21	-28.30	28.94	32.70	54.85
1AAB	A	4079	83	5	293	63	-59.60	-28.52	7.88	8.90	55.04
1AAB	A	4079	83	5	293	65	-69.24	-33.52	47.61	53.80	54.44
1AAB	A	4079	83	5	293	68	-73.81	-34.79	34.87	39.40	54.57
1QPU	A	4759	106	4.8	298	24	-55.91	-63.82	57.96	65.50	53.20
1QPU	A	4759	106	4.8	298	29	-62.99	-50.99	8.32	9.40	52.60
1QPU	A	4759	106	4.8	298	35	-57.01	-45.46	53.63	60.60	52.20
1QPU	A	4759	106	4.8	298	36	-68.23	-31.78	4.34	4.90	51.70
1QPU	A	4759	106	4.8	298	37	-72.24	-34.66	2.65	3.00	52.00
1QPU	A	4759	106	4.8	298	40	-60.75	-33.13	1.24	1.40	51.40
1QPU	A	4759	106	4.8	298	75	-62.42	-38.01	2.21	2.50	52.80

PDB_ID	CHAIN_ID	BMRB_ID	SEQ_COUNT	PH	TEMP	SEQ_ID	PHI	PSI	HDO_PBT	SAS	C_SHIFT
1QPU	A	4759	106	4.8	298	79	-60.99	-35.59	1.77	2.00	53.30
1QPU	A	4759	106	4.8	298	87	-62.33	-48.44	0.00	0.00	53.20
1QPU	A	4759	106	4.8	298	89	-62.56	-52.52	48.41	54.70	52.70
1QPU	A	4759	106	4.8	298	90	-67.20	-14.28	22.12	25.00	52.30
1QPU	A	4759	106	4.8	298	91	-61.59	-31.55	3.89	4.40	51.80
1QPU	A	4759	106	4.8	298	100	-57.60	-58.87	43.45	49.10	52.30
1XSX	A	5891	95	5	308	12	-64.59	-34.13	17.17	19.40	55.13
1XSX	A	5891	95	5	308	16	-64.33	-34.37	18.41	20.80	54.50
1XSX	A	5891	95	5	308	35	-54.51	-48.33	76.19	86.10	55.13
1XSX	B	5891	95	5	308	12	-64.63	-33.07	15.93	18.00	55.13
1XSX	B	5891	95	5	308	16	-63.85	-40.06	19.03	21.50	54.50
1XSX	B	5891	95	5	308	35	-55.52	-49.18	73.36	82.90	55.13
2JN6	A	15086	97	4.5	298	14	-60.40	-45.57	0.18	0.20	55.82
2JN6	A	15086	97	4.5	298	16	-64.40	-33.47	26.55	30.00	55.82
2JN6	A	15086	97	4.5	298	30	-64.91	-37.35	0.18	0.20	56.41
2JN6	A	15086	97	4.5	298	60	-60.55	-31.08	38.85	43.90	53.37
2JN6	A	15086	97	4.5	298	61	-93.40	1.20	53.01	59.90	53.42
2JN6	A	15086	97	4.5	298	65	-49.69	-32.62	34.69	39.20	55.03
2JS1	A	15350	80	4.5	298	29	-61.20	-45.12	0.00	0.00	55.33
2JS1	A	15350	80	4.5	298	50	-71.93	-49.37	18.94	21.40	55.24
2JS1	A	15350	80	4.5	298	62	-70.41	-58.10	2.12	2.40	55.54
2JS1	B	15350	80	4.5	298	29	-75.48	-45.25	0.00	0.00	55.33
2JS1	B	15350	80	4.5	298	50	-67.90	-50.22	17.96	20.30	55.24
2JS1	B	15350	80	4.5	298	62	-72.04	-49.39	3.72	4.20	55.54

Concluding remarks

All the research in this thesis has been done during my doctoral program. They are all newly developed methods and focusing on the high-throughput NMR-based protein 3D structure determination. They are ready to be applied to the practical structure determination field, and they also can be improved by further study as the following.

1) PINE-SPARKY

PINE-SPARKY survey and NMRFAM workshop revealed the usability. However, it will be a better contribution if PINE-SPARKY can support various outputs as current version only supports PINE outputs and ADAPT-NMR Enhancer outputs.

2) PONDEROSA

PONDEROSA shows its ability by participating CASD-NMR. It automates the use of aliphatic ^{13}C-NOESY, ^{15}N-NOESY and torsion angles. It will be better to automate the use of aromatic NOESY and residual dipolar couplings as well.

3) ADAPT-NMR and ADAPT-NMR Enhancer

ADAPT-NMR and ADAPT-NMR Enhancer need to be tested with many practical examples. Especially for ADAPT-NMR Enhancer, it will come up with some good ideas if someone new to the ADAPT-NMR has an opportunity to solve 3D structure of a real protein.

4) Rhinovirus 2A Proteinase

As the 3D structure of 2A proteinase is determined, the next step may be the biological study to show how this protein takes a role in common cold. Furthermore, if the role of this protein is discovered to be critical in the cold disease, this can be further studied for drug discovery as well.

5) PACSY

PACSY is a relational database with plenty of possible applications. As it can generate a refined set of chemical shift statistics with various conditions, one can make a new table, which can substitute a table from BMRB. In addition, a subset of queries to the PACSY can be used for structure predictions.

Bibliography

Andries, K., Dewindt, B., Snoeks, J., Wouters, L., Moereels, H., Lewi, P.J., Janssen, P.A., 1990. Two groups of rhinoviruses revealed by a panel of antiviral compounds present sequence divergence and differential pathogenicity. *J. Virol.* 64, 1117-1123.

Alipanahi, B., Gao, X., Karakoc, E., Donaldson, L., Li, M., 2009. PICKY: a novel SVD-based NMR spectra peak picking method. *Bioinformatics* 25, i268–275.

Bahrami, A., Assadi, A.H., Markley, J.L., Eghbalnia, H.R., 2009. Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. *PLoS Comput. Biol.* 5, e1000307.

Bahrami, A., Tonelli, M., Sahu, S.C., Singarapu, K.K., Eghbalnia, H.R., Markley, J.L., 2012. Robust, Integrated Computational Control of NMR Experiments to Achieve Optimal Assignment by ADAPT-NMR. *PLoS ONE* 7, e33173.

Baker, N.A., Sept, D., Joseph, S., Holst, M.J., McCammon, J.A., 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A*. 98, 10037–10041.

Bardiaux, B., Malliavin, T., Nilges, M., 2012. ARIA for solution and solid-state NMR. *Methods Mol. Biol.* 831, 453–483.

Bartels, C., Billeter, M., Güntert, P., Wüthrich, K., 1996. Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. *J. Biomol. NMR* 7, 207–213.

Bartels, C., Xia, T.H., Billeter, M., Güntert, P., Wüthrich, K., 1995. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol.*

NMR 6, 1–10.

Bax, A., 1994. Multidimensional nuclear magnetic resonance methods for protein studies. *Curr. Opin. Struct. Biol.* 4, 738–744.

Baxter, N.J., Roetzer, A., Liebig, H.-D., Sedelnikova, S.E., Hounslow, A.M., Skern, T., Waltho, J.P., 2006. Structure and dynamics of coxsackievirus B4 2A proteinase, an enyzme involved in the etiology of heart disease. *J. Virol.* 80, 1451–1462.

Benson, D., Lipman, D.J., Ostell, J., 1993. GenBank. Nucleic Acids Res. 21, 2963–2965.

Berjanskii, M.V., Neal, S., Wishart, D.S., 2006a. PREDITOR: a web server for predicting protein torsion angle restraints. *Nucleic Acids Res.* 34, W63–69.

Berman, H., Henrick, K., Nakamura, H., Markley, J.L., 2007. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* 35, D301–303.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr, Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M., 1977. The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur. *J. Biochem.* 80, 319–324.

Bhattacharya, A., Tejero, R., Montelione, G.T., 2007. Evaluating protein structures determined by structural genomics consortia. *Proteins* 66, 778–795.

Bochkov, Y.A., Palmenberg, A.C., Lee, W.-M., Rathe, J.A., Amineva, S.P., Sun, X., Pasic, T.R., Jarjour, N.N., Liggett, S.B., Gern, J.E., 2011. Molecular modeling, organ culture and reverse genetics for a newly identified human rhinovirus C. *Nat. Med.* 17, 627-632.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T., Warren, G.L., 1998. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* 54, 905–921.

Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., Richardson, D.C., 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* 66, 12–21.

Chignola, F., Mari, S., Stevens, T.J., Fogh, R.H., Mannella, V., Boucher, W., Musco, G., 2011. The CCPN Metabolomics Project: a fast protocol for metabolite identification by 2D-NMR. *Bioinformatics* 27, 885–886.

Chothia, C., 1976. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 105, 1–12.

Chylla, R.A., Markley, J.L., 1995. Theory and application of the maximum likelihood principle to NMR parameter estimation of multidimensional NMR data. *J. Biomol. NMR* 5, 245–258.

Codd, E.F., 1970. A relational model of data for large shared data banks. *Communications of the ACM* 13, 377–387.

Cornilescu, G., Delaglio, F., Bax, A., 1999. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* 13, 289–302.

Cornilescu, G., Marquardt, J.L., Ottiger, M., Bax, A., 1998. Validation of Protein Structure from

Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *J. Am. Chem. Soc.* 120, 6836–6837.

Cox, J.D., Hunt, J.A., Compher, K.M., Fierke, C.A., Christianson, D.W., 2000. Structural influence of hydrophobic core residues on metal binding and specificity in carbonic anhydrase II. *Biochemistry* 39, 13687–13694.

Crick, F., 1970. Central dogma of molecular biology. Nature 227, 561

Csaba, G., Birzele, F., Zimmer, R., 2009. Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. *BMC Struct. Biol.* 9, 23.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., Bax, A., 1995. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* 6, 277–293.

Dolinsky, T.J., Nielsen, J.E., McCammon, J.A., Baker, N.A., 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* 32, W665–667.

Dominguez, S.R., Briese, T., Palacios, G., Hui, J., Villari, J., Kapoor, V., Tokarz, R., Glod? M.P., Anderson, M.S., Robinson, C.C., Holmes, K.V., Lipkin, W.I., 2008. Multiplex MassTag-PCR for respiratory pathogens in pediatric nasopharyngeal washes negative by conventional diagnostic testing shows a high prevalence of viruses belonging to a newly recognized rhinovirus clade. *J. Clin. Virol.* 43, 219-222.

Doreleijers, J.F., Sousa da Silva, A.W., Krieger, E., Nabuurs, S.B., Spronk, C.A.E.M., Stevens, T.J., Vranken, W.F., Vriend, G., Vuister, G.W., 2012. CING: an integrated residue-based structure validation program suite. *J. Biomol. NMR* 54, 267–283.

Duggleby, R.G., Kaplan, H., 1975. A competitive labeling method for the determination of the chemical properties of solitary functional groups in proteins. *Biochemistry* 14, 5168–5175.

Eghbalnia, H.R., Bahrami, A., Tonelli, M., Hallenga, K., Markley, J.L., 2005a. High-resolution iterative frequency identification for NMR as a general strategy for multidimensional data collection. *J. Am. Chem. Soc.* 127, 12528–12536.

Eghbalnia, H.R., Wang, L., Bahrami, A., Assadi, A., Markley, J.L., 2005b. Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements. *J. Biomol. NMR* 32, 71–81.

Eisenhaber, F., Argos, P., 1993. Improved strategy in analytic surface calculation for molecular systems: Handling of singularities and computational efficiency. *Journal of Computational Chemistry* 14, 1272–1280.

Frishman, D., Argos, P., 1995. Knowledge-based protein secondary structure assignment. *Proteins* 23, 566–579.

Guerry, P., Herrmann, T., 2011. Advances in automated NMR protein structure determination. *Quarterly Reviews of Biophysics* 44, 257-309.

Güntert, P., 2004. Automated NMR structure calculation with CYANA. *Methods Mol. Biol.* 278, 353–378.

Gallegos, A.M., Atshaves, B.P., Storey, S.M., Starodub, O., Petrescu, A.D., Huang, H., McIntosh, A.L., Martin, G.G., Chao, H., Kier, A.B., Schroeder, F., 2001. Gene structure, intracellular localization, and functional roles of sterol carrier protein-2. *Prog. Lipid Res.* 40, 498–563.

Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A.H., Coudert, E., Lima, T., Kersey, P., Pagni, M., Sigrist, C.J.A., Lachaize, C., Veuthey, A.L., Gasteiger, E., Bairoch, A., 2003. Automated annotation of microbial proteomes in SWISS-PROT. *Comput Biol Chem* 27, 49–58.

Gille, C., Frömmel, C., 2001. STRAP: editor for STRuctural Alignments of Proteins. *Bioinformatics* 17, 377–378.

Gledhill, J.M., Jr, Wand, A.J., 2012. Al NMR: a novel NMR data processing program optimized for sparse sampling. *J. Biomol. NMR* 52, 79–89.

Goddard, T.D., Kneller, D.G., n.d. SPARKY 3. *University of California, San Francisco.*

Gwaltney, J.M., Jr, Hendley, J.O., Simon, G., Jordan, W.S., Jr, 1966. Rhinovirus infections in an industrial population. I. The occurrence of illness. *N. Engl. J. Med.* 275, 1261-1268.

Hall, S.R., Spadaccini, N., 1994. The STAR File: detailed specifications. *Journal of Chemical Information and Modeling* 34, 505–508.

Han, B., Liu, Y., Ginzinger, S.W., Wishart, D.S., 2011. SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR* 50, 43–57.

Herrmann, T., Güntert, P., Wüthrich, K., 2002a. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* 319, 209–227.

Herrmann, T., Güntert, P., Wüthrich, K., 2002b. Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J. Biomol. NMR* 24, 171–189.

Hiller, S., Fiorito, F., Wüthrich, K., Wider, G., 2005. Automated projection spectroscopy (APSY). *Proc. Natl. Acad. Sci. U.S.A.* 102, 10876–10881.

Holder, A.A., Wootton, J.C., Baron, A.J., Chambers, G.K., Fincham, J.R., 1975. The amino acid sequence of Neurospora NADP-specific glutamate dehydrogenase. Peptic and chymotryptic peptides and the complete sequence. *Biochem. J.* 149, 757–773.

Hyberts, S.G., Takeuchi, K., Wagner, G., 2010. Poisson-gap sampling and forward maximum entropy reconstruction for enhancing the resolution and sensitivity of protein NMR data. *J. Am. Chem. Soc.* 132, 2145–2147.

Iwadate, M., Asakura, T., Williamson, M.P., 1999. C alpha and C beta carbon-13 chemical shifts in proteins from an empirical database. *J. Biomol. NMR* 13, 199–211.

Jefferson, E.R., Walsh, T.P., Barton, G.J., 2008. A comparison of SCOP and CATH with respect to domain-domain interactions. *Proteins* 70, 54–62.

Johnson, B.A., Blevins, R.A., 1994. NMR View: A computer program for the visualization and analysis of NMR data. *J. Biomol. NMR* 4, 603–614.

Jung, Y.-S., Zweckstetter, M., 2004. Mars -- robust automatic backbone assignment of proteins. *J. Biomol. NMR* 30, 11–23.

Kim, J.H., Füzéry, A.K., Tonelli, M., Ta, D.T., Westler, W.M., Vickery, L.E., Markley, J.L., 2009. Structure and dynamics of the iron-sulfur cluster assembly scaffold protein IscU and its interaction with the cochaperone HscB. *Biochemistry* 48, 6062–6071.

Kim, S., Szyperski, T., 2003. GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. *J. Am. Chem. Soc.* 125, 1385–1393.

Kohlhoff, K.J., Robustelli, P., Cavalli, A., Salvatella, X., Vendruscolo, M., 2009. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J. Am. Chem. Soc.* 131, 13894–13895.

Koradi, R., Billeter, M., Engeli, M., Güntert, P., Wüthrich, K., 1998. Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J. Magn. Reson.* 135, 288–297.

Koradi, R., Billeter, M., Wüthrich, K., 1996. MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* 14, 51–55, 29–32.

López-Méndez, B., Güntert, P., 2006. Automated protein structure determination from NMR spectra. *J. Am. Chem. Soc.* 128, 13112 –13122.

Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R., Thornton, J.M., 1996. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* 8, 477–486.

Lau, S.K.P., Woo, P.C.Y., Lai, K.K.Y., Huang, Y., Yip, C.C.Y., Shek, C.-T., Lee, P., Lam, C.S.F., Chan, K.-H., Yuen, K.-Y., 2011. Complete Genome Analysis of Three Novel Picornaviruses from Diverse Bat Species. Journal of Virology 85, 8819-8828.

Lee, B., Richards, F.M., 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55, 379–400.

Lee, W., 2008. Computation development for NMR-based protein structure determination. *Graduate School, Yonsei University.*

Lee, W., Bahrami, A., Markley, J.L., 2013. ADAPT-NMR Enhancer: complete package for reduced dimensionality in protein NMR spectroscopy. Bioinformatics 29, 515-517.

Lee, W., Kim, J.H., Westler, W.M., Markley, J.L., 2011. PONDEROSA, an automated 3D-NOESY peak picking program, enables automated protein structure determination. *Bioinformatics* 27, 1727–1728.

Lee, W., Westler, W.M., Bahrami, A., Eghbalnia, H.R., Markley, J.L., 2009. PINE-SPARKY: graphical interface for evaluating automated probabilistic peak assignments in protein NMR spectroscopy. *Bioinformatics* 25, 2085–2087.

Lee, W., Yu, W., Kim, S., Chang, I., Lee, W., Markley, J.L., 2012. PACSY, a relational database management system for protein structure and chemical shift analysis. *J. Biomol. NMR* 54, 169–179.

Linge, J.P., Habeck, M., Rieping, W., Nilges, M., 2003. ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* 19, 315–316.

Lloyd, R.E., Grubman, M.J., Ehrenfeld, E., 1988. Relationship of p220 cleavage during picornavirus infection to 2A proteinase sequencing. *J. Virol.* 62, 4216-4223.

Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E., Wüthrich, K., 1998. Recommendations for the presentation of NMR structures of proteins and nucleic acids. IUPAC-IUBMB-IUPAB Inter-Union Task Group on the Standardization of Data Bases of Protein and Nucleic Acid Structures Determined by NMR Spectroscopy. *J. Biomol. NMR* 12, 1–23.

Markley, J.L., Ulrich, E.L., Berman, H.M., Henrick, K., Nakamura, H., Akutsu, H., 2008. BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J. Biomol. NMR* 40, 153–155.

McErlean, P., Shackelton, L.A., Andrews, E., Webster, D.R., Lambert, S.B., Nissen, M.D., Sloots, T.P., Mackay, I.M., 2008. Distinguishing Molecular Features and Clinical Characteristics of a Putative New Rhinovirus Species, Human Rhinovirus C (HRV C). *PLoS ONE* 3, e1847.

Meiler, J., 2003. PROSHIFT: protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR* 26, 25–37.

Molla, A., Paul, A.V., Schmid, M., Jang, S.K., Wimmer, E., 1993. Studies on dicistronic polioviruses implicate viral proteinase 2Apro in RNA replication. *Virology* 196, 739-747.

Moon, S., Case, D.A., 2007. A new model for chemical shifts of amide hydrogens in proteins. *J. Biomol. NMR* 38, 139–150.

Moseley, H.N.B., Sahota, G., Montelione, G.T., 2004. Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. *J. Biomol. NMR* 28, 341–355.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540.

Orekhov, V.Y., Ibraghimov, I.V., Billeter, M., 2001. MUNIN: a new approach to multi-dimensional NMR spectra interpretation. *J. Biomol. NMR* 20, 49–60.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M., 1997. CATH--a hierarchic classification of protein domain structures. *Structure* 5, 1093–1108.

Palmenberg, A.C., Spiro, D., Kuzmickas, R., Wang, S., Djikeng, A., Rathe, J.A., Fraser-Liggett, C.M., Liggett, S.B., 2009. Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution. *Science* 324, 55-59.

Pervushin, K., Riek, R., Wider, G., Wüthrich, K., 1997. Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl. Acad.*

Sci. U.S.A. 94, 12366–12371.

Petersen, J.F., Cherney, M.M., Liebig, H.D., Skern, T., Kuechler, E., James, M.N., 1999. The structure of the 2A proteinase from a common cold virus: a proteinase responsible for the shut-off of host-cell protein synthesis. *EMBO J.* 18, 5463–5475.

Price, W.H., 1956. THE ISOLATION OF A NEW VIRUS ASSOCIATED WITH RESPIRATORY CLINICAL DISEASE IN HUMANS. *Proc. Natl. Acad. Sci. U.S.A.* 42, 892-896.

Rosato, A., Aramini, J.M., Arrowsmith, C., Bagaria, A., Baker, D., Cavalli, A., Doreleijers, J.F., Eletsky, A., Giachetti, A., Guerry, P., Gutmanas, A., Güntert, P., He, Y., Herrmann, T., Huang, Y.J., Jaravine, V., Jonker, H.R.A., Kennedy, M.A., Lange, O.F., Liu, G., Malliavin, T.E., Mani, R., Mao, B., Montelione, G.T., Nilges, M., Rossi, P., Van der Schot, G., Schwalbe, H., Szyperski, T.A., Vendruscolo, M., Vernon, R., Vranken, W.F., Vries, S. de, Vuister, G.W., Wu, B., Yang, Y., Bonvin, A.M.J.J., 2012. Blind testing of routine, fully automated determination of protein structures from NMR data. *Structure* 20, 227–236.

Rosato, A., Bagaria, A., Baker, D., Bardiaux, B., Cavalli, A., Doreleijers, J.F., Giachetti, A., Guerry, P., Güntert, P., Herrmann, T., Huang, Y.J., Jonker, H.R.A., Mao, B., Malliavin, T.E., Montelione, G.T., Nilges, M., Raman, S., Van der Schot, G., Vranken, W.F., Vuister, G.W., Bonvin, A.M.J.J., 2009. CASD-NMR: critical assessment of automated structure determination by NMR. *Nat. Methods* 6, 625–626.

Schmoldt, A., Benthe, H.F., Haberland, G., 1975. Digitoxin metabolism by rat liver microsomes. *Biochem. Pharmacol.* 24, 1639–1641.

Schulte-Herbrüggen, T., Briand, J., Meissner, A., Sørensen, O.W., 1999. Spin-state-selective TPPI: a new method for suppression of heteronuclear coupling constants in multidimensional NMR experiments. *J. Magn. Reson.* 139, 443–446.

Schwieters, C.D., Kuszewski, J.J., Tjandra, N., Clore, G.M., 2003. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* 160, 65–73.

Sgourakis, N.G., Lange, O.F., DiMaio, F., André, I., Fitzkee, N.C., Rossi, P., Montelione, G.T., Bax, A., Baker, D., 2011. Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. *J. Am. Chem. Soc.* 133, 6288–6298.

Shen, Y., Bax, A., 2010. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* 48, 13–22.

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

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K.K., Lemak, A., Ignatchenko, A., Arrowsmith, C.H., Szyperski, T., Montelione, G.T., Baker, D., Bax, A., 2008. Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4685–4690.

Shen, Y., Vernon, R., Baker, D., Bax, A., 2009b. De novo protein structure generation from incomplete chemical shift assignments. *J. Biomol. NMR* 43, 63–78.

Shin, J., Lee, W., Lee, W., 2008. Structural proteomics by NMR spectroscopy. *Expert Rev Proteomics* 5, 589–601.

Singarapu, K.K., Radek, J.T., Tonelli, M., Markley, J.L., Lan, Q., 2010. Differences in the structure and dynamics of the apo- and palmitate-ligated forms of Aedes aegypti sterol carrier protein 2 (AeSCP-2). *J. Biol. Chem.* 285, 17046–17053.

Skern, T., Hampoelz, B., Bergmann, E., Guarné, A., Petersen, J., Fita, I., James, M., n.d. Structure and function of picornaviral proteinases, in: Molecular Biology of Picornaviruses. *ASM Press, Washington, D.C.*, pp. 199–212.

Stanek, J., Koźmiński, W., 2010. Iterative algorithm of discrete Fourier transform for processing randomly sampled NMR data sets. *J. Biomol. NMR* 47, 65–77.

Stoesser, G., Baker, W., Van den Broek, A., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., Nardone, F., Stoehr, P., Tuli, M.A., Tzouvara, K., Vaughan, R., 2003. The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res*. 31, 17–22.

Toyoda, H., Nicklin, M.J., Murray, M.G., Anderson, C.W., Dunn, J.J., Studier, F.W., Wimmer, E., 1986. A second virus-encoded proteinase involved in proteolytic processing of poliovirus polyprotein. *Cell* 45, 761-770.

Ulrich, E.L., Akutsu, H., Doreleijers, J.F., Harano, Y., Ioannidis, Y.E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C.F., Tolmie, D.E., Kent Wenger, R., Yao, H., Markley, J.L., 2008. BioMagResBank. *Nucleic Acids Res.* 36, D402 –408.

Vila, J.A., Arnautova, Y.A., Martin, O.A., Scheraga, H.A., 2009. Quantum-mechanics-derived 13Calpha chemical shift server (CheShift) for protein structure validation. *Proc. Natl. Acad. Sci. U.S.A.* 106, 16972–16977.

Volk, J., Herrmann, T., Wüthrich, K., 2008. Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. *J. Biomol. NMR* 41, 127–138.

Wüthrich, K., 1990. Protein structure determination in solution by NMR spectroscopy. *J. Biol. Chem*. 265, 22059–22062.

Wagner, G., Hyberts, S.G., Havel, T.F., 1992. NMR structure determination in solution: a critique and comparison with X-ray crystallography. *Annu Rev Biophys Biomol Struct* 21, 167–198.

Willcocks, M.M., Carter, M.J., Roberts, L.O., 2004. Cleavage of eukaryotic initiation factor eIF4G and inhibition of host-cell protein synthesis during feline calicivirus infection. *J. Gen. Virol.* 85, 1125-1130.

Wishart, D.S., Sykes, B.D., 1994. The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. *J. Biomol. NMR* 4, 171–180.

Wishart, D.S., Sykes, B.D., Richards, F.M., 1992. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* 31, 1647–1651.

Yang, D., Kay, L.E., 1999. TROSY Triple-Resonance Four-Dimensional NMR Spectroscopy of a 46 ns Tumbling Protein. *J. Am. Chem. Soc.* 121, 2571–2575.

Zhang, L., Yang, D., 2006. SCAssign: a sparky extension for the NMR resonance assignment of aliphatic side-chains of uniformly 13C,15N-labeled large proteins. *Bioinformatics* 22, 2833–2834.