DEVELOPMENT OF COMPUTATIONAL METHODS FOR THE RAPID DETERMINATION OF NMR RESONANCE ASSIGNMENT OF LARGE PROTEINS

LI KAI

(B.Sc., Beijing Institute of Technology)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF BIOCHEMISTRY

NATIONAL UNIVERSITY OF SINGAPORE

2003

Acknowledgements

I would like to express my appreciation and gratitude to my supervisors Associate Professor Tan Tin Wee and Assistant Professor Yang Daiwen for their patience, encouragement and guidance during the course of the project.

My special thanks to Dr. Ke Youan for his support and guidance. I also thank him for introducing me to this project.

I would like to thank Professor Lewis E. Kay for providing the NMR experimental data of p53 and MSG.

I am also grateful to my colleagues Ms. Anita Suresh, Mr. Bernett Lee, Dr. Edula Goutham, Mr. Gopalan Vivek, Mr. Kong LeSheng, Ms. Manisha Brahmachary, Mr. Paul Tan Thiam Joo in Bioinformatics Centre, NUS, and Mr. Lin Zhi, Ms. Ran Xiaoyuan, Mr. Sui Xiaogang, Mr. Xu Xingfu, Ms. Xu Ying, Mr. Zheng Yu in Structure Biological Laboratories, Department of Biological Sciences, NUS for many discussions on the subject of this thesis.

My sincere gratitude to administrative staff in Bioinformatics Centre, NUS Mr. Mark De Silva, Mr. Lim Kuan Siong, Ms. Hazel Tan and Ms. Madeline Koh.

I would like to acknowledge National University of Singapore for providing me with a scholarship to carry out my research work.

Finally, I gratefully acknowledge the support and encouragement of my family throughout this endeavor.

Table of Contents

List of Figures

Figure Title

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ASAP	Automated Sequential Assignment of NMR Resonances in Large Proteins
2D	two-dimensional
3D	three-dimensional
4D	four-dimensional
COSY	Correlated Spectroscopy
CPN	Constraint Propagation Network
CSA	Chemical Shift Anisotropy
CSP	Constraint Satisfaction Problems
DD	dipole-dipole
HSQC	Heteronuclear Single Quantum Coherence
MSG	Malate Synthase G
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
NOESY	Nuclear Overhauser Enhancement SpectroscopY
p53	a 67-kDa dimeric construct of p53 (residue 82-360)
PDB	Protein Data Bank
SR	Spectral Resolution
TOCSY	Total Correlation Spectroscopy
TROSY	Transverse Relaxation-Optimized SpectroscopY

Summary

In structural genome projects, structure determination on a large scale is required, which would not be practicable without a high degree of automation. Since protein NMR has become an indispensable tool in protein structure determination, the automation of structure determination process by NMR has become a matter of great urgency. It was also widely accepted that one of the most time-consuming steps towards structure determination is the spectral assignment procedure. This involves sequence-specific resonance assignment of NMR signals and the assignment of NOESY spectra. Resonance assignment forms the basis for characterizing secondary structure, dynamics, intermolecular interactions and 3D structure computation of proteins (Moseley and Montelione, 1999); hence, the first task of automating structure determination is to study how to automate resonance assignment. Almost all currently available programs for automated resonance assignment using 2D and/or 3D NMR experiments are limited by protein size (usually <20kDa). Although some programs utilize 4D experiments successfully for large proteins, they rely on user intervention instead of a completely automatic process. In order to facilitate fully automated resonance assignment for large proteins, algorithm and software for protein resonance assignment based on 4D-TROSY triple resonance NMR spectroscopy are proposed in this thesis.

We have designed a protein resonance assignment strategy consisting of four steps: (1)

the combination of amino acid spin-systems, (2) the determination of amino acid types for combined spin-systems, (3) the identification of sequential connections between these spin-systems, and (4) sequence-specific resonance assignments.

To overcome the severe chemical shift degeneracy and missing peaks for large proteins, we choose 4D TROSY NMR instead of conventional 3D experiments. The increased dimensionality increases the number of correlations obtained in a single data set, which also causes the combination of various experiments to become straightforward and enables the resonance assignment accomplished using a minimal number of spectra.

The determination of amino acid type for a given spin-system relies on analyzing the chemical shifts of $^{13}C^{\alpha}$, $^{13}C^{\beta}$, and ^{13}CO. In order to provide a more reliable and specific estimation of amino acid type, we took into account the information of amino acid type and protein secondary structure for the analysis of chemical shift.

We also applied constraint propagation algorithms to reduce solution space for the identification of sequential relationships. Due to chemical shift degeneracy, it is still not practicable to conduct ‘exhaustive searching’ automatically, which is supposed to provide correct solution from all assignment possibilities. In this case, an approach that combines ‘best-first’ deterministic and ‘exhaustive search’ methods was developped in this thesis to rapidly and accurately assign spin-systems to the protein sequence.

The algorithms developed to automate the above four steps, were implemented through computer programs and validated with real spectral data of large proteins. 234 resultant sequence-specific resonance assignments of p53 agree with 241 previously obtained manual assignments, and 640 automated resonance assignments of MSG agree with 651 manual assignments. Using the proposed resonance assignment program, this thesis demonstrates that an automated resonance assignment work is possible for very large molecules.

Chapter 1

Related Background and Previous Work

1.1 Introduction to protein NMR in structural biology

The dream of having genomes completely sequenced is now a reality. However, an even greater challenge, proteomics — the study of all the proteins coded by the genes under different conditions, awaits biologists to further unravel biological processes. In many cases it will be necessary to know the three-dimensional (3D) structure of a protein to understand its function. The feasibility of such a structural proteomics project was recently demonstrated (Yee *et al*., 2003) and it was shown that two techniques would play a dominant part: X-ray crystallography and Nuclear Magnetic Resonance (NMR). These two main techniques can provide the structures of macromolecules at atomic resolution.

Although X-ray crystallography is still the dominant technique in this field, NMR complements it in many ways. For example, it does not require the growth of crystals as X-ray crystallography does — a task that (if successful) requires months or even years. In addition, NMR can provide the 3D structure of a protein in solution under nearly physiological conditions along with the dynamics information associated with the protein’s function. The important role that NMR plays in structural biology is illustrated by far more than 1000 NMR solution structures deposited in the Protein

Data Bank (PDB) (Berman *et al.*, 2000). With the advent of recent innovations such as heteronuclear NMR and cryoprobes (Ferentz and Wagner, 2000), NMR will play a more significant role in structural biology, particularly in the high-throughput structure production of the Structural Genomics Initiative (Montelione *et al.*, 2000).

NMR does not directly create an image of a protein. Rather, it is able to yield a wealth of indirect structural information from which the 3D structure can only be revealed by extensive data analysis and computer calculation. The typical strategy of a NMR structure determination follows a suite of steps, as described below.

1.2 Protein structure determination from multidimensional NMR spectroscopy

1.2.1 Basic strategies

Figure 1.1 Depicts the basic steps toward determining solution structures from NMR data set.

Protein production in solution

Protein production in *E. coli.* has an established record of being the most successful approach to provide protein targets for structure study. When successful, bacterial expression provides a cost-effective, flexible, reliable, and scalable way to support structural characterizations. Metabolic labeling of biomolecules with stable isotopes (^{15}N, ^{13}C and/or ^{2}H) for NMR spectroscopy was pioneered with *E. coli.* expression systems and has been extended successfully to only a few other systems (Markley and Kainosho, 1993). In the case where the production of proteins are not expressed well in *E. coli.*, some eukaryotic options are available to express these proteins, including yeast, insect, or human cells.

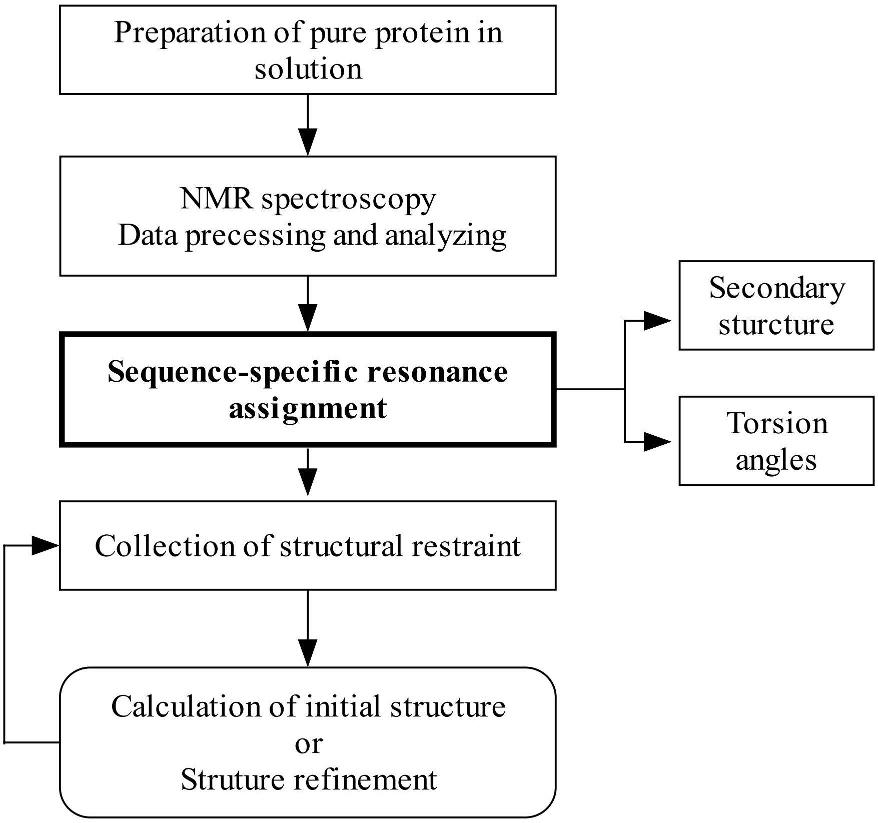

Preparation of pure protein in solution
NMR spectroscopy
Data precessing and analyzing
Sequence-specific resonance assignment
Secondary sturcture
Torsion angles
Collection of structural restraint
Calculation of initial structure
or
Struture refinement

Figure 1.1 The flowchart of protein structure determination by NMR. The sequence-specific resonance assignment that is emphasized by bold plays a key role in protein structure determination. ASAP program proposed in this thesis facilitates automated backbone resonance assignment of large proteins, as described in chapter 2.

The higher the protein concentration, the faster the NMR data can be collected, provided that the protein does not aggregate. Practically, the lower limit concentrations are about 200 µM with ordinary probes and about 60 µM with cryogenic probes. Depending on the length of the detection coil in the probe, a sample volume of 300 to 500 µL is usually required. Some samples may not stable over data collection period.

Cryogenic probes together with higher field can shorten the time of each experiment, which makes it possible to investigate proteins that are less stable over time.

Processing and analyzing multidimensional NMR data

NMR spectrometers produce resonance signals in 1D, 2D, 3D, and 4D spaces, which could reflect both the signature information of amino acid type and the adjacency information between amino acids. The general approach in a biomolecular NMR study is to first convert time-domain data to frequency-domain spectra by Fourier transform. Then peaks are picked out from each spectrum. This identifies real resonance peaks that are generated from protein residues rather than noises. Current protocols for processing NMR data set and peak picking use the programs NMRPipe (Fourier transformation) (Delaglio *et al.*, 1995), XEASY (peak picking and semi-automated assignment) (Bartels *et al.*, 1995), and NMRView (peak picking and spectrum data analysis as well as semi-automated assignment) (Johnson and Blevins, 1994).

Sequence-specific NMR resonance assignment

Once NMR spetra are acquired, individual cross peaks in the experiments have to be assigned to sequence-specific positions in the primary sequence of protein before other structural restraints (e.g., the distance information between residues in the NOESY spectrum) can be fully interpreted. Sequence-specific NMR resonance assignment plays a key role in the whole process of structure determination. The objective of our study is to automate the resonance assignment procedures. Detailed manual and automated assignment approaches are depicted in later sections of this chapter.

Structural restraint extraction

Structural restraints are obtained from the interpretation of data from one or more different classes of NMR experiments. (1) Once all ^{1}H, ^{15}N, and ^{13}C resonances have been assigned, fully analysis of the NOESY spectrum, ‘NOE assignment’, provides the most important restraint, ^{1}H-^{1}H distance constraints (<5Ǻ). (2) Three-bond spin-spin coupling experiments provide torsion angle constraints, two dihedral angles associated with each peptide bond: angle Φ, is the torsion angle between bond ^{15}N-^{1}H^N and C$^\alpha$-H$^\alpha$ while angle Ψ is another torsion angle between bond C$^\alpha$-H$^\alpha$ and C-O. Besides, these torsion angles can also be predicted from the assigned chemical shifts of ^{15}N, C$^\alpha$, CO, and C$^\beta$, as described in program TALOS (Cornilescu *et al.*, 1999). (3) Additional hydrogen bond constraints are determined from hydrogen exchange experiments, chemical shifts, and/or trans-hydrogen-bond couplings (Cordier *et al.*, 1999).

Structure calculation and refinement

NMR structures are obtained from constrained molecular dynamics simulations and energy minimization calculations, with the NOE-derived inter-proton distances being the primary experimental constraints as well as other available constraints. As a consequence of chemical shift degeneracy, many NOE cross peaks may have multiple assignment possibilities, and the results of preliminary structure calculations are used to eliminate unlikely candidates on the basis of inter-proton distances. Refinement continues in an iterative manner until a self-consistent set of experimental constraints produces an ensemble of structures that also satisfies standard covalent geometry and steric overlap considerations.

1.2.2 Important role of sequence-specific resonance assignment

As mentioned above, NMR spectra contain information about the structure of a molecule through the chemical shift which are sensitive to local physicochemical environment, through spin-spin coupling constaints which is sentitive to dihedral angles, and through relaxation (NOE) which is sensitive to the positions of nearby spins. However, before any of this information can be put to use in determining the structure of the moleclue, it must first be determined which resonances come from which spins. The process of associating specific spins in the molecule with specific resonances is called ***sequence-specific assignment of resonances***, on which this thesis will focus.

Sequence-specific resonance assignment is essential in: (1) the structure determination of proteins, (2) intermolecular interactions, and (3) protein dynamics.

Firstly, consider the determination of protein structure from NMR data. Protein chemical shift assignment may be used in at least four different ways in structural analysis including: (i) secondary structure mapping, (ii) generating structural constraints, (iii) three-dimensional structure generation, and (iv) three-dimensional structure refinement. Perhaps the most well-known application of chemical shift in biomolecular NMR is in the area of secondary structure identification and quantification (Dalgarno *et al.*, 1983; Wishart *et al.*, 1991; Wishart *et al.*, 1992; Metzler *et al.*, 1993; Gronenborn and Clore, 1994; Wishart and Sykes, 1994; Wishart and Nip, 1998). The assigned chemcial shifts (H^{α}, ^{15}N, and ^{13}C) provide more reliable information about the secondary structure of the protein than any other computational prediction methods based on sequence similarity. Chemical shifts can also play a useful role in delineating three-dimensional structure of protiens. The structural information mainly derives from NOE cross peaks. A NOE peak correlating two hydrogen atoms is observed if these hydrogens are located at a shorter distance than from each other. Combined with resonance assignment these distance constraints can be attributed to specific sites along the protein chain and therefore the three-dimensional structure can be initialized. In addition, calculated with other constraints derived from chemical shift assginement (e.g., dihedral angles) (Cornilescu *et al.*, 1999) along with the contraints from NOE correlations, the protein's tertiary structure can be formed and furtherly refined.

The second application of sequence-specific resonance assignment is to study protein-protein interactions. Analysis of intermolecular interactions by solving the structures of protein-protein complexes using conventional NMR methodology presents a considerable technical challenge and is highly time-consuming. If the structures of the free proteins are already known at high resolution, and conformational changes upon complexation are either minimal or localized, it is possible to use conjoined rigid body/torsion angle dynamics (Clore and Bewley, 2002) to solve the structure of the complex based solely on intermolecular inter-proton distance restraints, derived from isotope-edited NOE measurements. Nevertheless, unambiguous assignment of intermolecular NOEs is still difficult and time-consuming, particularly for large complexes. In contrast, the mapping of interaction surfaces by $^{1}H^{N}/^{15}N$ chemical shift perturbation (Zuiderweg, 2002) is a simple and rapid procedure and a most widely used NMR method to study protein interactions. In a nutshell, the ^{15}N-^{1}H HSQC spectrum of one protein is monitored when an unlabeled interaction partner is titrated in, and the perturbations of chemical shifts are recorded. The interaction causes environmental changes on the protein interfaces and, hence, affects the chemical shifts of the nuclei in this area. It is easy and straightforward to correlate these value-changed chemical shifts with specific residues according to sequence-specific resonance assignment and therefore, the interaction regions derived from the perturbation of chemical shifts can be discovered.

NMR spectroscopy can also be used to monitor the dynamic behavior of a protein at a multitude of specific sites, which is associated with the specific functions of the protein. Once again, resonance assignment is a prerequisite to determine the residues implicated in the analysis of structural dynamic from nuclear spin relaxation (generally from ^{15}N relaxation).

1.3 Introduction to NMR spectroscopy for large molecules in solution

The foundations of NMR resonance assignment studies are high-quality NMR spectra recorded with good S/N ratio and spectral resolution. With increasing molecular weight, these basic requirements are harder to achieve. Limiting factors are low sensitivity and line broadening due to rapid transverse spin relaxation and extensive signal overlap due to the high complexity of the spectra. Recent advances have been achieved with both novel NMR techniques and new biochemical approaches. In particular, using the NMR technique Transverse Relaxation-Optimized SpectroscopY (TROSY) in combination with suitable isotope labeling schemes (Goto and Kay, 2000), the size limit for the observation of NMR signals in solution has been extended severalfold (Wider and Wüthrich, 1999).

1.3.1 Advantages of TROSY technique in investigating large proteins

During the past decades, the highest polarizing magnetic field for high-resolution NMR has greatly increased, which benefited biomolecular NMR through improved intrinsic sensitivity and spectral resolution in large proteins. However for commonly used heteronuclear experiments, the advantages of using higher magnetic fields were partly offset by field-dependent line broadening, which is a manifestation of increased transverse relaxation rates and will cause loss of the sensitivity and spectral resolution in complex NMR experiments.

In these experiments, the transfer of magnetization along networks of scalar-coupled spins includes long delays during which ^{13}C and ^{15}N magnetization evolve in the transverse plane. Therefore, fast transverse relaxation during these delays and during ^{1}H acquisition, limits the application of triple-resonance NMR experiments with large proteins. $^{13}C^{\alpha}$ is efficiently relaxed by dipole-dipole (DD) interactions with H^{α}, but this transverse relaxation rate can be significantly decreased by deuteration. However the transverse ^{15}N relaxation during coherence transfer steps is only slightly affected by deuteration.

The TROSY technique as introduced by Pervushin and his co-workers (Pervushin *et al.*, 1997) (Pervushin *et al.*, 1998) and improved by Yang and his co-workers (Yang and Kay, 1999), yields substantial reduction of the transverse relaxation rates in ^{15}N-$^{1}H^{N}$ moieties, based on the mutual compensation of DD coupling and chemical shift anisotropy (CSA) interactions. Generally, for large proteins (molecular sizes above 20 kDa), TROSY-type spectra show narrower lines and higher sensitivity compared with conventional COSY (correlated spectroscopy) (Figure 1.2).

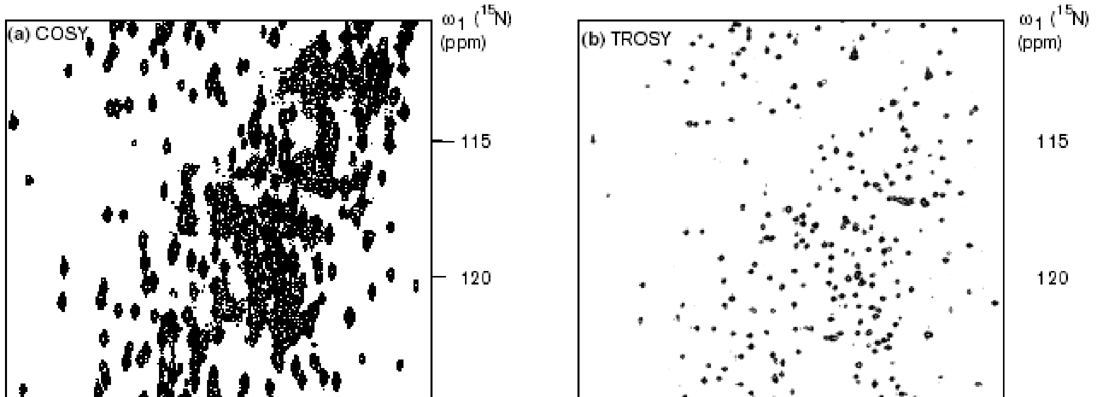


Figure 1.2 A comparison of ^{15}N-^{1}H^{N} correlation spectra of a protein (45 kDa) recording using (a) conventional procedure (COSY) and (b) TROSY. Both spectra were measured at the same condition (Wider and Wüthrich, 1999). Obviously, the cross-peaks in TROSY-type spectrum are distinctly separated, which are overlapped with others in conventional COSY. (Reproduced from the work of Wider and Wüthrich, 1999)

TROSY is, in this way, applicable to studies of proteins with macromolecular structures that have accrued molecular weights of 100 kDa or larger (Riek *et al.*, 2002) and, thus, the introduction of the TROSY technique opens a wide field of new applications for solution NMR.

1.3.2 TROSY triple-resonance experiments for resonance assignments of large proteins

NMR experiments provide a set of unique combinations of neighboring resonance spin system information for resonance assignment. But these approaches require deuterated samples to prevent the fast transverse relaxation of ^{13}C, when applied to proteins in the 20 kDa range or larger (Grzesiek *et al.*, 1993; Yamazaki *et al.*, 1994; Nietlispach *et al.*, 1996; Gardner *et al.*, 1997). Therefore, a generally applicable program for the automated assignment of larger proteins should not rely on side-chain information in the initial sequential assignment process. In addition to avoid increasing signal overlap, more suitable method is to apply a suite of heteronuclear 3D and 4D experiments tracing the protein backbone, including C$^{\beta}$ information.

The implementation of TROSY (Salzmann *et al.*, 1998) (Salzmann *et al.*, 1999) (Yang and Kay, 1999) (Riek *et al.*, 2002) in conducting assignment by triple-resonance experiments with ^{2}H, ^{13}C, ^{15}N-labeled proteins, can additionally prevent fast transverse relaxation of ^{15}N during extended time periods with transverse ^{15}N magnetization, as well as preventing amide proton line broadening. Therefore, the use of the TROSY principle in triple-resonance experiments promises to enable resonance assignments for significantly large proteins such as Malate Synthase G (Mulder *et al.*, 2000), p53 (Tugarinov *et al.*, 2002), etc.), which are much larger than what is achievable today with corresponding conventional NMR experiments. At the same time, the assignment strategies (Konrat *et al.*, 1999; Yang and Kay, 1999; Mulder *et al.*, 2000; Tugarinov *et al.*, 2002) are straightforward and highly amenable to be automated.

Some TROSY-type triple resonance experiments tracing protein backbone and C^{β} information are presented as follows.

1.3.2.1 TROSY-type 3D NMR experiments

A number of three-dimensional (3D) triple-resonance NMR experiments, TROSY-HNCA, TROSY-HNCO (Salzmann *et al.*, 1998), TROSY-HN(CO)CA, TROSY-HN(CA)CO, TROSY-HNCACB, and TROSY-HN(CO)CACB (Salzmann *et al.*, 1999), have been designed for sequential backbone assignments for large proteins.

Listings in Table 1.1 show the nuclei that are correlated in the above 3D experiments. Those experiments are named according to the nuclei they correlates. For example, the TROSY-HNCO experiment correlates H_i, N_i and CO_{i-1}.

TROSY-type experiments	HNCA	HNCO	HN(CO)CA
Correlated nuclei	Ca_i-N_i-H^N_i Ca_{i-1}-N_i-H^N_i	Co_{i-1}-N_i-H^N_i	Ca_{i-1}-N_i-H^N_i
TROSY-type experiments	HN(CA)CO	HNCACB	HN(CO)CACB
Correlated nuclei	Co_i-N_i-H^N_i Co_{i-1}-N_i-H^N_i	Ca_i/Cb_i-N_i-H^N_i Ca_{i-1}/Cb_{i-1}-N_i-H^N_i	Ca_{i-1}/Cb_{i-1}-N_i-H^N_i

Table 1.1. Correlations observed in 3D TROSY-type triple-resonance NMR experiments.

The enhancement of sensitivity compared to the corresponding conventional experiments, is most pronounced for regular secondary structure elements (Salzmann *et al.*, 1999). The gain in sensitivity is of particular interest for TROSY-HNCA, TROSY-HN(CA)CO and TROSY-HNCACB , since these experiments reveal both sequential and intra-residual correlation peaks and thus allows the determination of sequential connectivities in a single experiment. This characteristic of TROSY-HNCACB has been employed in this thesis as described in chapter 2.

1.3.2.2 TROSY-type 4D NMR experiments and other experiments used in this thesis

Although 3D NMR spectra efficiently resolve the proton resonance overlapping that occurs for moderate size proteins, as shown in Figure 1.3 where the crowded ^1H-^1H 2D NMR planes (Figure 1.3A) are separated into many planes of a 3D NMR spectrum (Figure 1.3B), the degeneracy of ^{15}N-$^1H^N$ moiety is still severe for very large proteins. It is difficult to combine cross-peaks from different 3D experiments according to the ^{15}N-$^1H^N$ spin pairs of specific residues along the sequence of a target protein, which is usually the groundwork for resonance assignment.

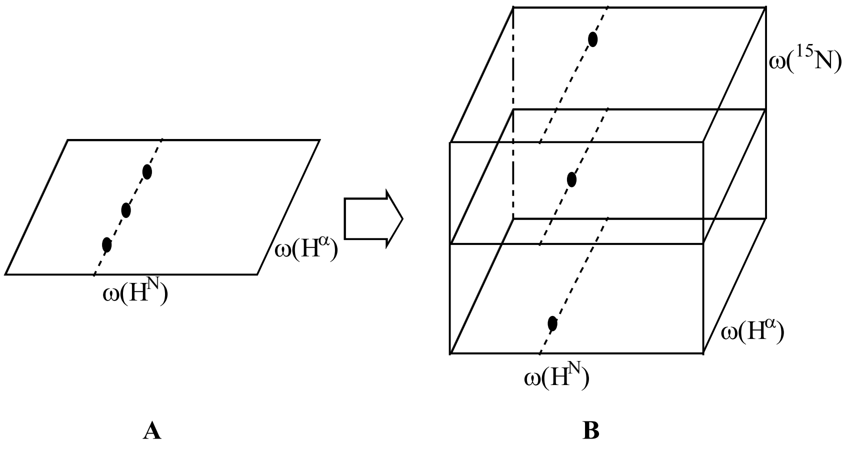


Figure 1.3 Schematic illustration of the relationship between 2D spectra and ^{15}N-edited 3D spectra. The closed circles represent three H^N-H$^\alpha$ cross-peaks, which overlap with each other in H^N coordinate of 2D spectrum but can be separated, in the corresponding 3D spectrum, into three planes depending on the different chemical shifts of the amide nitrogen nuclei, ^{15}N.

A suite of 4D TROSY NMR experiments have been designed to resolve the ambiguity of ^{15}N-^{1}H^N chemical shifts for large proteins. Similar to previous 3D spectra, these 4D experiments separate overlapping cross-peaks in ^{15}N-^{1}H^N coordinates into different planes depending on the different chemical shifts of alpha carbon or carbonyl carbon (details will be discussed in chapter 2). Schematic representations of Figure 1.4 and listings in Table 1.2 show the correlation of 4D TROSY-HNCACO, 4D TROSY-HNCOCA (Yang and Kay, 1999), and 4D TROSY-HNCOCASIM (Konrat *et al.*, 1999), as well as 3D TROSY-HNCACB, 3D TROSY-HN(CO)CACB (Salzmann *et al.*, 1999) and 4D ^{15}N-edited ^{15}N,^{15}N-NOESY (Grzesiek *et al.*, 1995; Venters *et al.*, 1995) experiments, all of which have been used for manual backbone sequential assignments (Mulder *et al.*, 2000; Tugarinov *et al.*, 2002). This thesis employs most of these experiments except the HN(CO)CACB experiment to develop automated resonance assignment, as discussed in the next chapter.

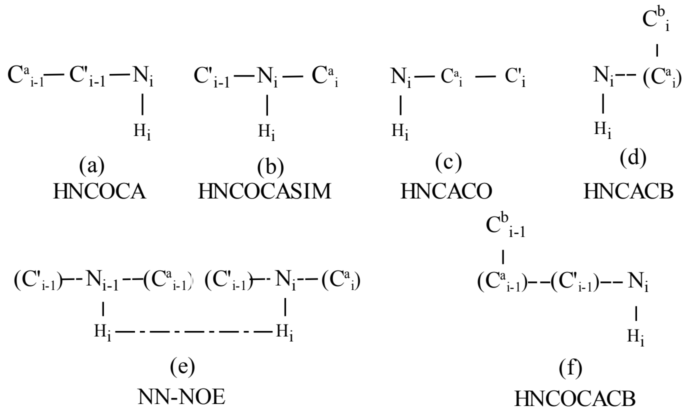


Figure 1.4 Schematic representation of the correlation forms of experimental NMR data used for sequential resonance assignment. Each NMR experiment is depicted as a non-directed graph whose edges reflect the transfer of magnetization through the participation nuclei. Those nuclei that are not detected in a given experiment are shown in parentheses. Experiments (a) and (f) correlate the C^{α}, C^{β} and CO frequencies of residue i-1 with the $^{15}N/^{1}H^{N}$ frequencies of residue i. Experiments (c) and (d) correlate the C^{α}, C^{β} and CO frequencies of residue i with its own $^{1}H^{N}$-^{15}N frequencies. Experiment (b) represents the correlation of C^{α}_{i}/CO_{i-1} spin pair with the $^{15}N_{i}/^{1}H^{N}_{i}$ spin pair. Except for the correlations listed above, experiment (b), (c) and (d) also correlate between $C^{\alpha}_{i-1}/C^{\beta}_{i-1}/CO_{i-1}$ spin pair and the $^{15}N_{i}/^{1}H^{N}_{i}$ spin pair to a lesser extent (which do not display in this figure). Sequential backbone $^{1}H^{N}$-^{15}N correlations are recorded in experiment (e).

TROSY-type experiments	HNCOCA	HNCOCASIM	HNCACO
Correlated nuclei	Ca_{i-1}-Co_{i-1}-N_i-H^N_i	Co_{i-1}-N_i-H^N_i-Ca_i	N_i-H^N_i-Ca_i-Co_i
& in less extent, if any		Ca_{i-1}-Co_{i-1}-N_i-H^N_i	Ca_{i-1}-Co_{i-1}-N_i-H^N_i
TROSY-type experiments	HNCACB	HN(CO)CACB	NN-NOESY
Correlated nuclei	Ca_i/Cb_i-N_i-H^N_i	Ca_{i-1}/Cb_{i-1}-N_i-H^N_i	N_j-H^N_j--N_i-H^N_i
& in less extent, if any	Ca_{i-1}/Cb_{i-1}-N_i-H^N_i		

Table 1.2 Correlations observed in 3D and 4D TROSY-type triple-resonance NMR experiments as well as an NN-NOESY experiment used for very large proteins.

Four-dimensional NMR spectra excel three-dimensional spectra in developing automated resonance assignment software at many aspects. The first computational advantage of using 4D NMR is that a single cross peak in a 4D NMR spectrum represents the magnetic interactions among four nuclei and provides the relationships among four chemical shifts, and therefore resonance assignment could be obtained only using a minimal number of spectra. For example, a cross peak (at ω_H=8.42, ω_N=121.62, $\omega_{C\alpha}$=61.15, ω_{CO}=174.16) in 4D TROSY-HNCACO spectrum represents the adjacency relationship among chemical shifts of all backbone nuclei within the same spin system. To obtain the same information from two 3D spectra, one has to find a pair of 3D cross peaks, in the above example, a HN(CA)CO peak (at ω_H=8.42, ω_N=121.62, ω_{CO}=174.16) and a HNCA peak (at ω_H=8.42, ω_N=121.62, $\omega_{C\alpha}$=61.15), having two chemical shifts, ω_H=8.42 and ω_N=121.62, in common. Finding such pairs is not as straightforward as it is in the case of using 4D NMR. Degenerate chemical shifts, e.g., (ω_H=8.42, ω_N=121.62, ω_{CO}=172.38 from a HN(CA)CO peak), may cause ambiguity when determining which chemical shift, ω_{CO}=172.38 or 174.16, is in the same spin system with the resonances (ω_H=8.35 and ω_N=121.62 ppm).

The second advantage of using 4D NMR is the redundant coverage between different spectra. For example, to group two 4D NMR cross peaks, (H_i,N_i,C^{α}_i,CO_i) and (H_i,N_i,C^{α}_i,CO_{i-1}), one can do so by verifying whether the amide shifts (H_i and N_i) are the same, which is similar to emerge two 3D NMR peaks. For 4D NMR peaks, however, the carbonyl carbon shift must also be consistent between two peaks, which can efficiently resolve most of amide shift degeneracy.

4D NMR experiments provide one more dimension compared to 3D experiments, and this tends to separate peaks away from each other and makes peak shapes more accurate. Peaks with better shapes are more suitable to be picked by automated peak picking softwares, since noise peaks can be more easily separated from real signals.

Separated peaks provide more precise position information (chemical shift), which enables automated assignment program to use a strict tolerance to combine different cross peaks together.

There are, however, several disadvantages of using 4D NMR. The time required to acquire a spectrum increases with the increase of dimensionality. For example, it takes 7 days to acquire a 4D TROSY-type HNCACO data set (Yang and Kay, 1999). Sensitivity, the S/N ratios, drops by $\sqrt{2}$ with the increase of one dimension.

Despite the loss of sensitivity and increase of acquisition time, in many cases, especially with large proteins, 4D NMR experiments are superior to 2D and 3D experiments in the conduct of successful resonance assignments. The abundant information existing in a single 4D NMR makes it straightforward to group different cross peaks and use only a few number (3 or 4) of 4D spectra together with HNCACB and/or HN(CO)CACB experiments to achieve almost complete backbone resonance assignment.

1.4 Introduction to manual assignment strategy

Resonance assignment has been a major role for protein structural analysis by NMR. Significant progress has been made through the introduction of 2D, 3D and recently 4D NMR experiments. Combined with systematic approaches for spectral analysis, however it is still tedious, time-consuming work. The ultimate goal of this thesis is to accomplish this work by developing an automated assignment tool as fully as possible.

Before discussing aspects regarding automated resonance assignment, we will describe traditional but efficient manual assignment strategy.

1.4.1 Manual assignment from homonuclear NMR

The problem of resolving proton resonances and assigning them to specific neclei in proteins, remained an overwhelming challenge untill the work of Wüthrich and his co-workers during the early 1980s (Wüthrich, 1986). They designed, implemented, and refined a logical approach using homonuclear NMR experiments in the sequence specific ^{1}H resonance assignment of proteins as follows:

1. First, J-correlated spectra are used to identify proton resonances belonging to each amino acid sidechain spin system. These spin systems are then classified as to a given type of amino acid and the characterization depends on the ability to discern specific spectral features of unique spin system or classes of spin systems.

2. The next and critical step in the sequential assignment procedure is to link the identified amino acid spin systems within the primary sequence of protein by use of observed nuclear Overhauser effects between main chain amide NH, $C_\alpha H$ and $C_\beta H$ protons.

3. Based on the above information, it is possible to establish chains of amino acid spin systems corresponding to polypeptide segments that are sufficiently long to be unique when compared to the primary sequence of protein. Sequence

specific assignment can then be obtained by matching the identified spin system chains with the corresponding segment in the independently determined protein primary sequence.

While suitable for smaller proteins (<10 kDa), this approach usually fails for larger proteins due to increasing signal overlap (crowded NOESY spectra).

1.4.2 Manual assignment from heteronuclear NMR

The second approach is to edit and resolve ^{1}H-^{1}H interactions not on the basis of an additional interaction with another ^{1}H spin but, on the basis of an interaction of one proton with a bonded heteronucleus. This makes it possible to identify sequential relationships between spin systems without using crowded NOESY spectra. Additionally, recent heteronuclear NMR experiments applying TROSY technique employ cancellation interactions of DD coupling and CSA interactions, to prevent fast transverse relaxation and to prevent amide proton line broadening when using higher magnetic fields. This advantage makes it possible to study the resonance assignment of large proteins. As described in 1.1.3, several 3D- and 4D-TROSY triple resonance NMR experiments have been designed to conduct the sequence-specific resonance assignments.

The inter-residue correlations are traditionally provided by NOE-type experiments where through-space dipolar couplings contribute to the observed cross-peaks. Certain triple resonance NMR experiments, such as 3D HNCA, HN(CA)CO, HNCO, and 4D

HNCACO, HNCOCASIM, HNCOCA also provide inter-residue correlations where through-bond scalar couplings contribute to the observed cross-peaks. Properly combining several triple resonance NMR experiments, it is possible to establish a sequential walk from one residue to the next without using NOE information. Figure 1.5 shows two examples where assignments are carried out by overlapping previously assigned frequencies in each subsequent spectrum.

In the first step of Figure 1.5A (HNCA and HN(CA)CO), the NH and ^{15}N frequencies of residue (i) are used to obtain the assignments of the C_α and CO of the same residue. Then, the CO frequency is used to obtain assignments for the HN and ^{15}N of residue ($i+1$) with the HNCO experiment. Finally, the NH and ^{15}N frequencies are used to find the C_α frequency of residue ($i+1$) with the HNCA spectrum, thus completing one cycle of the assignment. Due to the severe chemical shift degeneracy of large proteins, a set of 4D experiments have been designed for a similar assignment scheme (Figure 1.5B), where the number of overlapping frequences increases up to 2 or 3 while not more than 2 in Figure 1.5A. In chapter 2, a similar but more rigorous algorithm is described to assign the protein backbone and C_β resonances using heteronuclear 4D-TROSY triple resonance experiments.

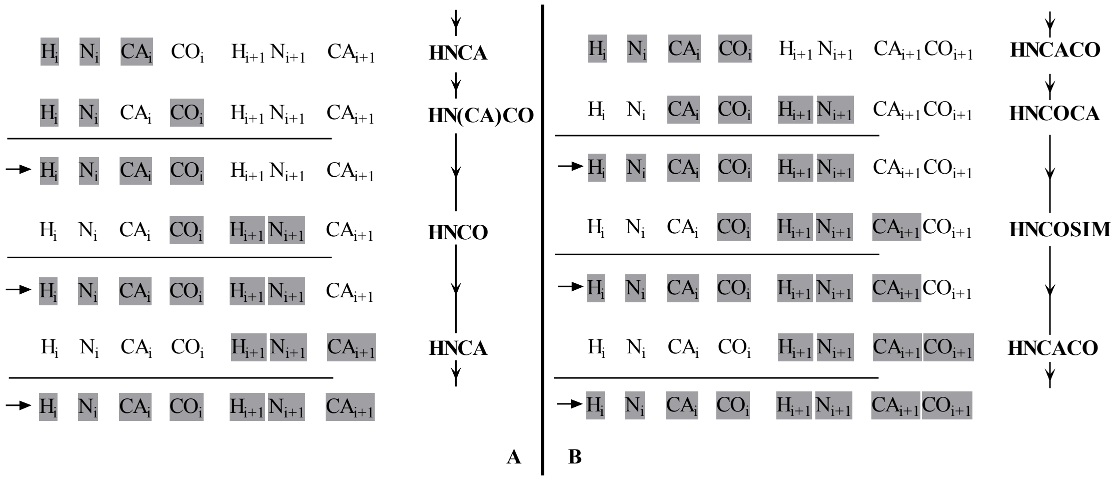

H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1
HNCA
H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1
HN(CA)CO
H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1
H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1
HNCO
H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1
H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1
HNCA
H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1
A
H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1 CO_i+1
HNCACO
H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1 CO_i+1
HNCOCA
H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1 CO_i+1
H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1 CO_i+1
HNCOSIM
H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1 CO_i+1
H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1 CO_i+1
HNCACO
H_i N_i CA_i CO_i H_i+1 N_i+1 CA_i+1 CO_i+1
B

Figure 1.5 The assignment scheme using heteronuclear NMR based on the through-bond correlations. The assignment is conducted by overlapping previously assigned (shadowed) frequencies in each subsequent spectrum. (A) using 3D experiments and (B) using 4D ones.

1.5 Literature review of the automated analysis of NMR resonance assigment

We have discussed the important role of resonance assignment in protein structure determination by NMR, as well as the actual strategy used to carry out manual assignments from NMR spectra. In this section, several systems are described, which have been developed to provide the automated analysis of triple-resonance spectra.

All of the programs follow a common process, though details regarding the kinds of data used as input and specific issues of implementation differ from one program to another. As a usual starting point, a high-resolution *root* spectrum (e.g., 2D HSQC, 3D HNCO, etc.) is used to identify the backbone ^{15}N-$^{1}H^{N}$ resonances of most residues. Each cross peak in that spcetrum is initially interpreted as the *root* of an individual spin system, and the remaining spectra are then examined to identify additional intra-residue and inter-residue cross peaks whose amide resonances fall within some specified tolerances of that root. Once these spin systems have been identified and collated, most implementations attemp to give each spin system a classification or measure of relative merit regarding possible amino acid types to which it can be assigned. With this information in hand, the search for logically consistent and 'optimal' sequential assignments proceeds by establishing matches between the intra-residue resonances of spin system i and the sequential resonances of spin system j.

In general, these systems can be divided into four classes according to their final mapping methods: (1) those that use genetic algorithms (Bartels *et al.*, 1997; Gronwald *et al.*, 1998), (2) those that use simulated-annealing methods (Lukin *et al.*, 1997; Leutner *et al.*, 1998; Hitchens *et al.*, 2003), (3) those that use constraint-based deterministic algorithms (Zimmerman *et al.*, 1994; Zimmerman *et al.*, 1997; Moseley and Montelione, 1999; Moseley *et al.*, 2001), and (4) those that employ an exhaustive search for resonance assignment (Coggins and Zhou, 2003). They are described as follows.

1.5.1 Genetic algorithm

GARANT (Bartels *et al.*, 1997) represents resonance assignment as an optimal match between expected cross peaks and experimentally observed peaks. The expected peaks (of COSY- or TOCSY- or NOESY-type spectra) are derived based on the primary structure of proteins and the knowledge about magnetization transfer pathways (through-bond or through-space). If available, the structure or chemical shifts of a homologous protein can be used for its scoring scheme, which give rise to a more restrictive scoring rule. Since the optimal solution of matching between expected and observed peaks requires excessive computing time, the program uses a general evolutionary algorithm combined with a specific local optimization routine, which avoids such excessive calculations and finds nearly optimal solutions.

CAMRA (Gronwald *et al.*, 1998) works in a similar way to GARANT, which achieves resonance assignment by matching between predicted signals and observed signals. It consists of three units: ORB, CAPTURE, and PROCESS. ORB predicts chemical shifts for unassigned proteins using the chemical shifts of previously assigned proteins together with the statistical analysis of individual chemical shift with respect to its residue type, atom, and secondary structure type. CAPTUTRE, on the other hand, groups peaks from different NMR spectra into distinct spin systems. Finally, PROCESS combines the chemical shifts predicted by ORB with the spin systems identified by CAPTURE, to obtain sequence-specific resonance assignment.

1.5.2 Simulated-annealing methods

Jonathan Lukin and his colleagues used Bayesian statistical method to combine resonance peaks from several 3D NMR experiments to form intra-residual segments and then tried to find the maximum likelihood assignment by using simulated annealing (Lukin *et al.*, 1997). They performed a large number of (about 10) 3D NMR experiments to overcome the problems of chemical shift degeneracy and missing peaks. A deterministic, ‘best-first’ procedure is used to combine the cross peaks into segments of six chemical shifts (N, H^N, H^α, C^α, C^β, CO), which then become the units of the program. The program then evaluates the probability of linking overlapping segments and assigning a segment to a given position along protein backbone. By arranging segments using Monte Carlo simulation so as to maximize this overall probability, the optimal resonance assignment can be obtained.

PASTA (Leutner *et al.*, 1998), is a combinatorial minimization strategy as used in the program of Lukin *et al.* However the slow convergence of the simulated annealing procedure is resolved by a threshold-accepting algorithm.

MONTE (Hitchens *et al.*, 2003), also, uses Monte Carlo methods as a basis for automated assignment programs. Its distinct advantage over previous assignment programs using simulating annealing methods, such as (Lukin *et al.*, 1997), is that it provides a general software package for chemical shift assignments of proteins, independent of any particularly ‘required’ experimental data. In addition, a wealth of source data, such as inter-residue scalar connectivity, inter-residue dipolar (NOE) connectivity and residue specific information, can be utilized in this program.

1.5.3 Constraint-based deterministic algorithms

AutoAssign (Zimmerman *et al.*, 1994; Zimmerman *et al.*, 1997; Moseley and Montelione, 1999; Moseley *et al.*, 2001), characterizes the assignment problem as a constraint satisfaction problem. It utilizes seven to eight specific types of NMR experiments. Cross peaks from these spectra are combined into the self-defined union of AutoAssign, GS (generic spin system), which is classified into two classes: unambiguous GS and degenerate GS (with very similar ^{15}N-$^{1}H^{N}$ shifts to others). “Constrained-based matching” generates resonance assignment by progressively relaxing the criteria used to establish sequential connectivity between GSs, and degenerate GSs can be involved after unambiguous GSs have been connected using restrictive criteria and have been reliably assigned to specific positions of protein. Using methods of Artificial Intelligence (AI), AutoAssign performs a CPN (constraint propagation network) along the whole procedure. CPN enables the constraints used in resonance assignment to be propagated from previously finished assignment to continuously undone assignment, and enables those to be ‘ruled out’ on the basis of inconsistencies or contradictions with the finished assignment. In some sense, CPN makes the computer accomplish assignment work straightforwardly and progressively just like what a human being does, but much faster.

1.5.4 Exhaustive searching

Although PACES (Coggins and Zhou, 2003) uses an algorithm that conducts an exhaustive search of all spin systems both for establishing sequential connection and for fulfilling sequence-specific assignment, it is actually a semi-automated program. Its iterative run with user intervention based on the similar constraints as what AutoAssign utilizes after each cycle efficiently reduces the ambiguities in the assignments. Although the possible residue type information for each spin system is determined simply by the statistical chemical shift distribution without weighing the probabilities, the chemical shift ranges used by PACES are progressively enlarged along the assignment procedure and hence this algorithm works efficiently for determining amino acid types. Similar to MONTE, it can also directly utilize a variety of sources as additional constraints of resonance assignment, such as residue-type information and side-chain assignment data.

Since most of these programs conduct resonance assignment from 2D or 3D NMR experiments, the test protein size for these programs is usually below 200 residues and only a few touch the size above 200 amino acids. Although PACES can provide mostly complete resonance assignment for a 723-residue protein, it forms spin-systems using simulative data set from manual resonance assignment instead of real experiments. In order to develop a program to automated resonance assignment for large molecules as fully as possible, this thesis proposes a program using 4D NMR spectra, as discussed in the next chapter.

Chapter 2

Automated Backbone Resonance Assignment Using 4D NMR

2.1 Introduction

This chapter reports a suite of computer algorithms that conduct the backbone resonance assignment of large proteins using 4D NMR experiments. These experiments are the same as those used for manual assignment (as described in Figure 1.4 and Table 1.3) except 3D-TROSY HN(CO)CACB experiment, which may be not practicable for large proteins. Since the modern trend of performing NMR experiments for large proteins at high fields runs into the problem of efficient transverse relaxation of carbonyl by the chemical shift anisotropy mechanism, the sensitivity advantage of HN(CO)CACB experiment is lost in some cases.

A nearly fully automated backbone resonance assignment program package is presented in this thesis. This package is able to (1) extract backbone spin-systems; (2) identify amino acid types; (3) obtain adjacency relationship between spin systems; and (4) map spin-systems to dipeptide sites in protein sequence. The resultant spin-system assignment provides resonance assignment as well as backbone NOE assignment which is responsible for secondary structure confirmation.

Two target protein samples were used: a 67-kDa dimeric construct of p53 (residue 82-360) and an 81 kDa monomeric enzyme, Malate Synthase G (MSG, residue 1-723). The resultant assignments agree well with previously studied data (234 out of 241 residues were excellent for p53, and 640 out of 651 residues for MSG).

2.2 Theory and Methods

The program can be run interactively or in batch mode. Figure 2.1 shows a schematic overview of default execution sequence when running in batch mode. There are three main parts of the program: initialization, constraint-based match cycle, and processing stages. Inputs of the program include a list of cross peaks from the required experiments (Table 1.2), the primary sequence of the target protein, chemical shift deviation at each coordinate (e.g., $^{1}H^{N}$, ^{15}N, $^{13}C^{\alpha}$, and ^{13}CO) of each experiment compared with corresponding coordinate of the reference experiment (4D TROSY-HNCOCA), chemical shift references derived from previous statistical analysis and secondary structure information.

Firstly, ‘Initialization’ manipulates all these input data and preserves them as global variables of the program. In addition, ‘initialization’ combines cross peaks into groups where peaks share common characteristics, and propagates these combinations (clusters and spin-systems) to later procedures as the minimal structural elements contributing to the resonance assignment.

Secondly, the ‘Constraint-based match cycle’ identifies amino acid type for each spin-system, and then maps each spin system to its expected positions within the primary sequence. The remaining task is to identify adjacency relationships between spin-systems, where not only the overlap chemical shifts between two spin-systems need to be consistent but their possible predicted dipeptide sites along target sequence also need to be consecutive. In addition, this cycle performs a set of tightly coupled routines that are triggered each time a sequential link or assignment of spin-system is established.

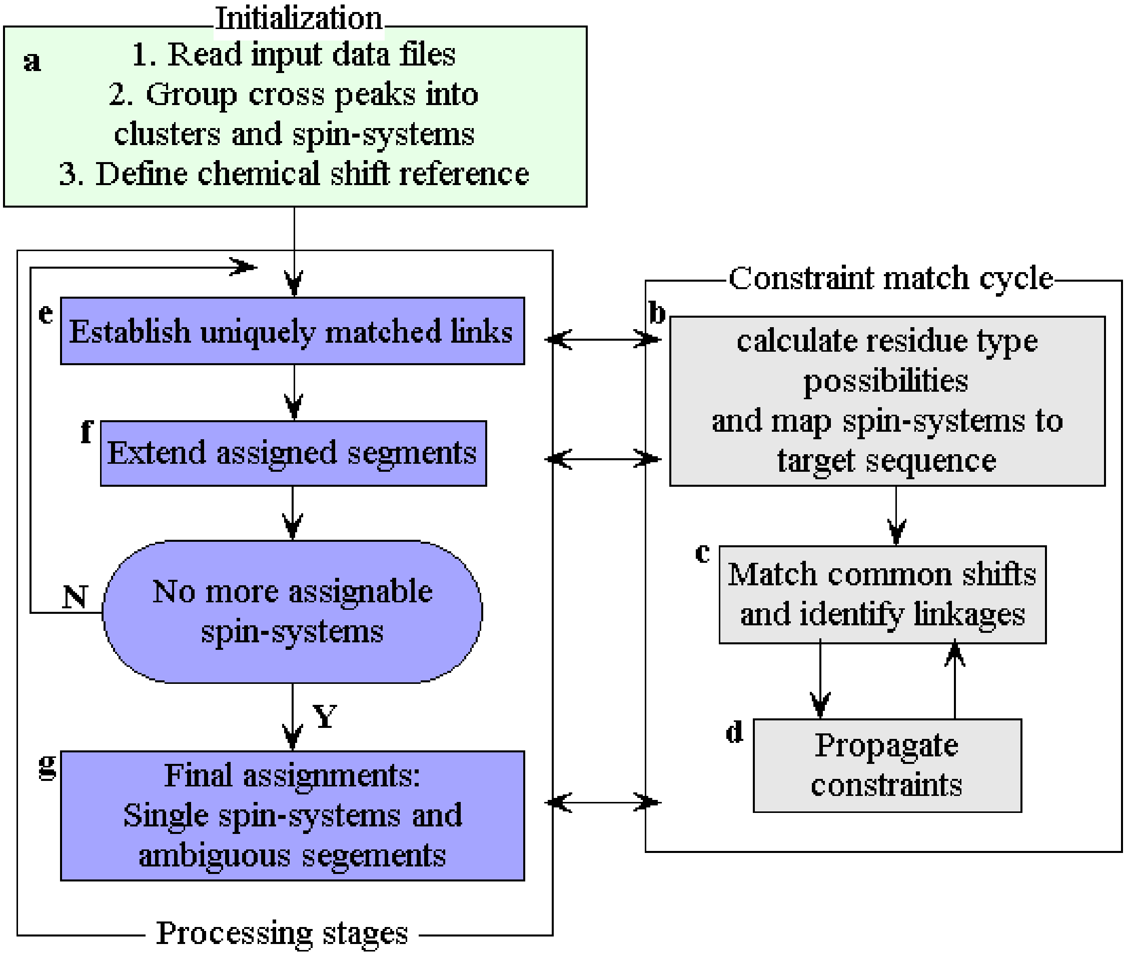

Initialization
a
1. Read input data files
2. Group cross peaks into clusters and spin-systems
3. Define chemical shift reference
Constraint match cycle
b
calculate residue type possibilities and map spin-systems to target sequence
c
Match common shifts and identify linkages
d
Propagate constraints
e
Establish uniquely matched links
f
Extend assigned segments
N
No more assignable spin-systems
Y
g
Final assignments: Single spin-systems and ambiguous segements
Processing stages

Figure 2.1 Schematic overview of default execution sequence. This program includes three main parts for batch mode: Initialization, Constraint-based match cycle, and Processing stages. These three processing stages (e, f, g, as discussed in the text) follow the initialization routines that process the input data. Depending on the execution stage, different criteria are used to establish sequential links between spin-systems based on the adjacency connectivity identified

by the constraint-based match cycle and to assign spin-systems to target sequence.

‘Processing stages’ manipulate the information from previous parts and conduct resonance assignment. Sequential connectivities between spin-systems are established based on identified adjacency relationship, where the used criteria are loosened gradually when more spin-systems are assigned to the target sequence. Spin-system assignments are obtained by iteratively performed two stages: establishing uniquely matched link (Figure 2.1e) and extending assigned segments (Figure 2.1f). Parts of remaining unassigned spin-systems, isolated spin-systems (which do not exist adjacency relationship with others) and those forming ambiguous segments (which can not be mapped to unique position within primary sequence), are assigned based on their characteristic chemical shifts at the ‘final assignments’ stage (Figure 2.1g).

The following will describe more details about each part of the program.

2.2.1 Input data

This program uses five different types of lists of peak positions (peak list). The initial peak lists are generated from the following five experiments using common NMR software (e.g., NMRView (Johnson and Blevins, 1994)): 4D TROSY-HNCACO (correlation form, [$\omega_{C\alpha}$(i), ω_{CO}(i), ω_{N}(i), ω_{HN}(i)]), 4D TROSY-HNCOCA ([$\omega_{C\alpha}$(i-1), ω_{CO}(i-1), ω_{N}(i), ω_{HN}(i)]), 4D TROSY-HNCO$_{i-1}$CA$_i$ or HNCOCASIM ([$\omega_{C\alpha}$(i), ω_{CO}(i-1), ω_{N}(i), ω_{HN}(i)]), 4D ^{15}N,^{15}N-NOESY ([ω_{N}(j), ω_{HN}(j), ω_{N}(i), ω_{HN}(i)]), and 3-D TROSY-HNCACB ([$\omega_{C\alpha}$(i)/$\omega_{C\beta}$(i), ω_{N}(i), ω_{HN}(i)]) (as described in Figure 1.4 and Table 1.2). The chemical shifts (coordinates) of the spins that are correlated in each cross-peak are indicated in square brackets with the symbol $\omega_X(i)$ denoting the chemical shift of spin X in residue i. In some cases, inter-residue correlations of the form [$\omega_{C\alpha}(i-1)$, $\omega_{CO}(i-1)$, $\omega_N(i)$, $\omega_{HN}(i)$] are observed in the 4D TROSY-HNCACO and 4D TROSY-HNCOCASIM experiments along with the correlations shown above. The existence of this redundant information may complicate data analysis because discrimination of these correlations from the desired ones is necessary. On the other hand, when these correlations cannot be observed in the 4D HNCOCA experiment, they become extremely valuable. Similarly, except for the intra-residue information, the 3D HNCACB can provide inter-residue correlations [$\omega_{C\alpha}(i-1)/\omega_{C\beta}(i-1)$, $\omega_N(i)$, $\omega_{HN}(i)$] in the cases of residues located at non-helical regions. Such inter-residue information is difficult to be obtained from the traditional 3D HN(CO)CACB or CBCA(CO)NH experiment for large proteins. This is caused by the rapid decay of the magnetization involving ^{13}CO spin, especially at high field. In this program, the correlations [$\omega_{C\alpha}(i-1)/\omega_{C\beta}(i-1)$, $\omega_N(i)$, $\omega_{HN}(i)$] are mainly utilized to reduce ambiguities in the establishment of fragments instead of being a major connectivity factor.

Initial peak lists are filtered against a ^{1}H-^{15}N HSQC spectrum to remove most of artificial peaks. Using interactive graphics, the peak lists are then edited manually to identify and eliminate most of the extraneous peaks mainly resulting from sinc-wiggles for very intense peaks as well as side-chain cross peaks in HNCOCA and HNCOCASIM experiments. The side-chain peaks feature largely in their small ^{1}H chemical shifts (around 7 ppm) and appear in pair (identical $^{13}C^{\alpha}$, ^{13}CO, ^{15}N chemcial shifts, only differ in $^1H^N$ chemical shift). The final peak lists are output in ASCII text format.

2.2.2 Spin-system combination from 4D NMR experiments

Resonance assignment results from a suite of specific NMR experiments. Concerning automated work, it will be very efficient to combine cross peaks of each experiment with corresponding peaks of other experiments and then to interpret each of these combinations to represent specific residue(s) along the target protein sequence, since the information derived from such combination is obviously abundant compared with that from any single peak of certain experiment.

Most of current programs group peaks from various experiments into a series of **Spin Systems,** where all peaks are associated with common amide proton and nitrogen. As a result of chemical shift degeneracy in both $^{1}H^{N}$ and ^{15}N dimensions, however, the ambiguity to establish spin-systems associated with the amide proton and nitrogen of specific residues is severe for large proteins. This program utilizes the redundant coverage (3-atom coverage) between specific 4D NMR experiments to efficiently solve ($^{1}H^{N}$-^{15}N) shift degeneracy in three steps: (1) clustering, (2) identifying spin-systems, and (3) modifying established spin-systems.

2.2.2.1 Clustering

Each of the cross-peaks used in this program has the chemical shifts of amide $^{1}H^{N}$ and ^{15}N. When a set of cross-peaks that have the same $^{1}H^{N}$ and ^{15}N chemical shifts within a certain tolerance, they are grouped together to form a cluster. This procedure is called clustering in this thesis.

The aim of clustering is to accelerate the performance of the program. In subsequent steps of constructing spin-systems and modifying established spin-systems as well as assigning sequential NOEs along with backbone resonance assignment, the program will only examine the peaks available within a certain cluster instead of searching peaks from the whole peak lists at the expense of excessive computing time.

Before initializing clusters, the program detects all inter-residue cross-peaks through filtering HNCOCASIM and HNCACO peak lists against HNCOCA peak list, since these redundant cross-peaks have the same ($^1H^N$, ^{15}N, $^{13}C^{\alpha}$, ^{13}CO) chemical shifts as those in the HNCOCA list. In this program, clusters are firstly generated by comparing the cross-peaks from HNCOCA with the cross-peaks from the rest of the peak lists

$$|\omega_H(i, \text{HNCOCA}) - \omega_H(j)| \leq \delta_H \tag{2.1}$$

$$|\omega_N(i, \text{HNCOCA}) - \omega_N(j)| \leq \delta_N \tag{2.2}$$

where $\omega_H(i, \text{HNCOCA})$ and $\omega_H(j)$ are the $^1H^N$ chemical shifts of cross-peak i in the HNCOCA peak list and cross-peak j in the rest of the peak lists; $\omega_N(i, \text{HNCOCA})$ and $\omega_N(j)$ are the ^{15}N chemical shifts of cross-peak i in the HNCOCA peak list and cross-peak j in the rest of the peak lists; δ_H and δ_N represent user-defined chemical shift tolerances of $^1H^N$ and ^{15}N, respectively. Secondly, the cross-peaks which do not meet the above conditions are compared with the HNCOCASIM peak lists to form other clusters, provided that the cross-peaks have the same chemical shifts for both $^1H^N$ and ^{15}N nuclei within the tolerances of δ_H and δ_N. The cross-peaks that cannot form clusters through these two steps will be deleted. In this way, most of spurious artificial peaks that are present in HNCACO, HNCACB and NN-NOESY data are effectively filtered out.

In order to reduce the number of artificial NOE cross-peaks, cross-peaks having no symmetrical partners are excluded since NOE is a kind of mutual effect and a pair of cross-peaks ([$\omega_N(j)$, $\omega_{HN}(j)$, $\omega_N(i)$, $\omega_{HN}(i)$]), ([$\omega_N(i)$, $\omega_{HN}(i)$, $\omega_N(j)$, $\omega_{HN}(j)$]) are anticipated in the 4D NN-NOESY spectrum.

2.2.2.2 Identifying Spin-systems

Each complete spin-system associated with a specific pair of amide shifts, contains three parts (Figure 2.2): $^1H^N_i$ and $^{15}N_i$ pair (root); $^{13}C^\alpha_i$, $^{13}CO_i$, and $^{13}C^\beta_i$ (intra-part); $^{13}C^\alpha_{i-1}$, $^{13}CO_{i-1}$, and $^{13}C^\beta_{i-1}$ (inter-part), where the subscript i denotes the residue number. An incomplete system may miss chemical shifts of one or several spins at intra-, and/or inter-part.

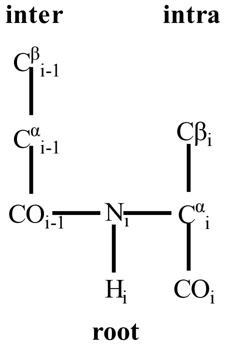


Figure 2.2 Construction of a complete spin-system. It encloses three parts: $^1H^N_i$ and $^{15}N_i$ pair (root); $^{13}C^\alpha_i$, $^{13}CO_i$, and $^{13}C^\beta_i$ (intra-part); $^{13}C^\alpha_{i-1}$, $^{13}CO_{i-1}$, and $^{13}C^\beta_{i-1}$ (inter-part).

In the case where ($^1H^N$, ^{15}N) spin pairs are unique, one cluster corresponds to one spin-system. Unfortunately, many clusters contain two or more spin systems in the application to large proteins. In this case, we do not initially enumerate the full set of combinations of the peaks in one cluster, instead we use HNCOCASIM data ([$\omega_{C\alpha}$(i), ω_{CO}(i-1), ω_N(i), ω_{HN}(i)]) to separate the peaks into different spin systems. Considering one possible combination of the peaks in cluster i: cross-peak j ([$\omega_{C\alpha}$(j), ω_{CO}(j), ω_N(i), ω_{HN}(i)]) from HNCOCA, peak k ([$\omega_{C\alpha}$(k), ω_{CO}(k), ω_N(i), ω_{HN}(i)]) from HNCACO and peak m ([$\omega_{C\alpha}$(m), ω_{CO}(m), ω_N(i), ω_{HN}(i)]) from HNCOCASIM, these peaks will be designated to the same spin system, provided that the following conditions are met:

$$|\omega_{Co}(j) - \omega_{Co}(m)| \leq \delta_{Co} \quad (2.3)$$

$$|\omega_{C\alpha}(k) - \omega_{C\alpha}(m)| \leq \delta_{C\alpha} \quad (2.4)$$

where δ_{Co} and $\delta_{C\alpha}$ are the chemical shift tolerances of ^{13}CO and $^{13}C^{\alpha}$, respectively. In the case where the chemical shifts of three spins ($^1H^N$, ^{15}N, ^{13}CO) or ($^1H^N$, ^{15}N, $^{13}C^{\alpha}$) for two or more cross-peaks in the same experiment are identical, the solution of assigning peaks to spin systems using HNCOCASIM is not unique and thus all the possible combinations are considered.

Figure 2.3 schematically shows how one proper combination (spin-system) is established from the cross-peaks of cluster i, based on matching identical chemical shifts.

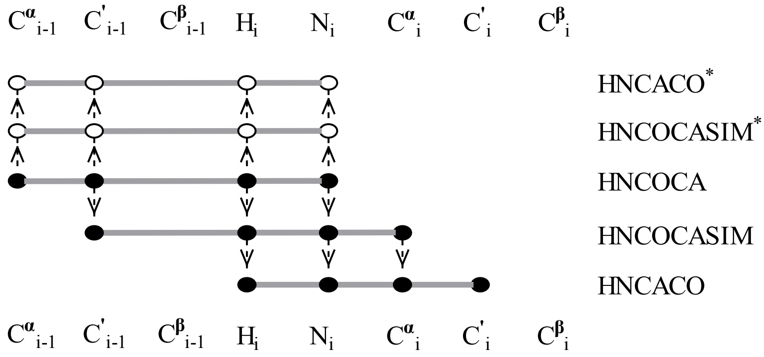


Figure 2.3 Schematic diagram of establishing spin-system based on matching identical chemical shifts. Closed circles indicate the resonances detected by each experiment. In particular, inter-residue cross peaks of HNCACO and HNCOCASIM experiments are filtered out from intra-residue peak by identical $^{1}H^{N}$, ^{15}N, $^{13}C^{\alpha}$, and ^{13}CO chemical shift to HNCOCA peaks. Their resonances are depicted as filled circles, but other resonances from HNCOCA peaks and intra-residue peaks of HNCOCASIM and HNCACO are depicted as hollow circles. Intra-residue peaks of HNCOCASIM are detected with identical $^{1}H^{N}$, ^{15}N, and ^{13}CO shifts to HNCOCA (eq. 2.3), and intra-residue peaks of HNCACO are detected with identical $^{1}H^{N}$, ^{15}N, and $^{13}C^{\alpha}$ shifts to intra-residue HNCOCASIM peaks (eq. 2.4).

If cross-peak j ([$\omega_{C\alpha}(j)$, $\omega_{CO}(j)$, $\omega_{N}(i)$, $\omega_{HN}(i)$]) from HNCOCA is not available, considering another kind of possible combination of the peaks in cluster i (Figure 2.4): cross-peak j' ([$\omega_{C\alpha}(j')$, $\omega_{CO}(j')$, $\omega_{N}(i)$, $\omega_{HN}(i)$]) from HNCOCASIM, peak k ([$\omega_{C\alpha}(k)$, $\omega_{CO}(k)$, $\omega_{N}(i)$, $\omega_{HN}(i)$]) from HNCACO and peak m ([$\omega_{C\alpha}(m)$, $\omega_{CO}(m)$, $\omega_{N}(i)$, $\omega_{HN}(i)$]) from HNCOCASIM, these peaks will be designated to the same spin system, provided that the following conditions are met:

$$|\omega_{C\alpha}(k) - \omega_{C\alpha}(m)| \leq \delta_{C\alpha} \qquad (2.4)$$

$$|\omega_{Co}(j') - \omega_{Co}(m)| \leq \delta_{Co} \qquad (2.5)$$

and peak j' presents a less intensity compared with peak m.

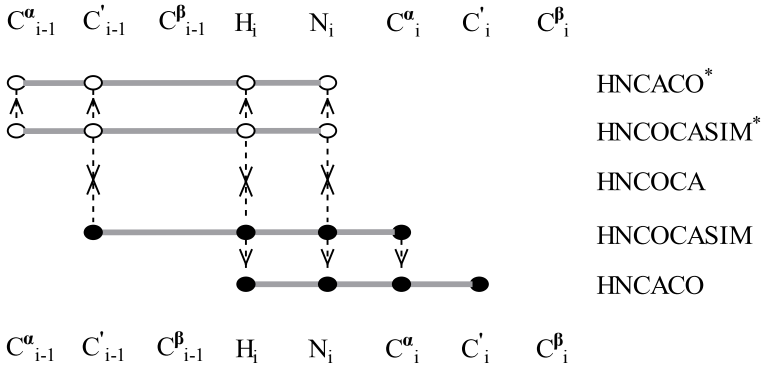


Figure 2.4 Schematic diagram of establishing spin-system based on matching identical chemical shifts without HNCOCA cross peak. Two peaks from HNCOCASIM with identical $^{1}H^{N}$, ^{15}N, ^{13}CO chemical shifts are combined together (eq. 2.5), one of which with more intensity is used to find corresponding intra-residue peak from HNCACO (eq. 2.4).

Although the HNCOCASIM can resolve ambiguities in ^{13}CO and $^{13}C^{\alpha}$ spins, assignment of $^{13}C^{\beta}$ resonance to spin system remains to be ambiguous at this stage for the systems have the same ($^{1}H^{N}$, ^{15}N) chemical shifts within given tolerances. In this case, the spin-system identified is not completed yet and a null for $^{13}C^{\beta}$ resonance is set.

2.2.2.3 Modification of spin-systems

Parts of identified spin-systems may be meaningless derived from the fully enumerated possible combinations for certain clusters when: (1) 4D peaks from HNCOCASIM are not available to resolve ambiguities in ^{13}CO and $^{13}C^{\alpha}$ spins; or (2) HNCOCASIM fails to reduce ambiguities in backbone carbon shifts. In addition, some assignments of $^{13}C^{\beta}$ resonance to spin systems remain to be incomplete due to the ($^{1}H^{N}$, ^{15}N) shift degenerate. This redundant and imperfect information will affect the program to quickly achieve a complete resonance assignment.

In order to resolve above problems, before each stage of this program every unassigned spin-system will be checked whether its construction is correct, and/or the spin-system will be designated more information after some ambiguity has been reduced according to assigned spin-systems. The examinations for each spin-system include: (1) whether its intra-residue cross peak of HNCOCASIM experiment belongs to other spin-systems, and (2) whether its intra-residue peak of HNCACO experiment belongs to others.

For the first role, the program examines all the spin-systems containing the same intra-residue peaks of HNCOCASIM experiment, verifies which one of them consists of incorrect peaks, and removes unreliable peaks if possible. Rules applied to checking each unassigned spin-system are as follows:

a) Check whether this peak belongs to other assigned spin-system (at inter- or intra-part). If so and the chemical shift ($^1H^N$ and ^{15}N) difference between this peak and the root of unassigned spin-system is larger than 1.5 times of spectral resolution (SR), this peak will be removed from its present spin-system;

b) Check whether this peak belongs to other unassigned spin-systems at intra-part. If there exist some other spin-systems containing the same peak at their intra-parts, the program will compare the Euclidean distances (in $^1H^N$, ^{15}N, and ^{13}CO dimensions) among all these spin-systems. The peak will be removed from the currently examined spin-system, on condition that the distance of the spin-system is obviously less than (1/4 times less than) that of one of other spin-systems.

To the second role, rules are the same as the previous one to check intra-residue peaks of HNCACO experiment for each unassigned spin-system. Since some spin-systems have been cut down part of unreliable information (peaks), they may become ‘redundant’ units for the program. Other spin-systems with the same intra-residue peak of HNCOCASIM or HNCACO experiment might contain all peak information as the same as that of these ‘redundant’ spin-systems but also contain more peaks. In this case, the program will automatically scan such ‘redundant’ spin-systems and delete them along with the correction of spin-systems.

On the other hand, once some spin-systems have been assigned to a target sequence, $^{13}C^{\beta}$ chemical shifts of each spin system can also be designated if they are not yet assigned in the step of spin system identification. This is done in three steps: (1) predicting the $^{13}C^{\beta}$ chemical shifts of each spin-system from the amino acid types and secondary structure that can be obtained from $^{13}C^{\alpha}$ and ^{13}CO chemical shifts, (2) comparing the predicted intra-residue (inter-residue) $^{13}C^{\beta}$ shift with each of the observed $^{13}C^{\beta}$ chemical shift in the cluster from which the concerned spin-system originates, and (3) assigning $^{13}C^{\beta}$ if only one observed $^{13}C^{\beta}$ resonance matches the predicted one.

2.2.3 Amino acid type identification using carbon chemical shift

A chemical shift reference is derived from the statistical analysis of ^{13}CO, $^{13}C^{\alpha}$, and $^{13}C^{\beta}$ chemical shifts for each type of 20 amino acids with respect to their secondary structure (Table 2.1) (Lukin *et al.*, 1997).

Amino	alpha helix			beta sheet			coil			Amino	alpha helix			beta sheet			coil		
Acid	$^{13}C'$	$^{13}C^a$	$^{13}C^b$	$^{13}C'$	$^{13}C^a$	$^{13}C^b$	$^{13}C'$	$^{13}C^a$	$^{13}C^b$	Acid	$^{13}C'$	$^{13}C^a$	$^{13}C^b$	$^{13}C'$	$^{13}C^a$	$^{13}C^b$	$^{13}C'$	$^{13}C^a$	$^{13}C^b$
G	175.35	46.71	0	171.87	44.97	0	173.96	45.41	0	D	178.44	57.11	40.25	175.4	53.5	42.38	176.45	54.12	40.83
A	179.37	54.77	18.4	175.9	51.15	21.02	177.3	52.42	19.03	N	176.73	55.62	38.46	174.55	52.47	39.78	174.65	53.22	38.74
S	175.69	61.31	62.98	173.6	57.11	65.08	174.41	58.27	64.14	E	178.73	59.06	29.3	175.28	54.96	31.93	176.27	56.66	30.13
C	176.68	61.62	26.75	174.18	56.08	29.02	174.84	58.01	28.2	Q	178.35	58.87	28.46	174.54	54.77	31.57	175.54	55.78	29.34
M	178.04	58.34	32.87	175.53	54.53	35.13	175.45	55.34	33	R	178.68	59.18	30.02	174.85	54.73	32.54	176.05	56.25	30.56
K	178.51	59.03	32.19	175.07	55.23	34.75	176.39	56.59	32.62	H	176.88	58.71	29.63	174.34	54.46	32.13	174.54	55.78	29.78
V	177.73	66.16	31.41	174.5	61.01	34.19	175.79	62.13	32.65	F	177.04	60.75	38.95	174.43	56.34	41.45	174.79	57.91	39.34
T	176.35	65.85	68.29	173.71	61.18	70.7	174.75	61.62	69.83	Y	177.09	60.82	38.57	174.22	56.57	41.19	175.8	57.77	38.88
I	177.52	64.61	37.75	174.79	60.09	40.47	175.69	60.98	38.87	W	178.14	59.51	29.26	174.78	56.49	30.98	175.85	57.5	29.09
L	178.64	57.52	41.41	175.76	54.01	43.88	177.15	54.82	42.82	P	179.45	65.28	30.9	175.56	62.71	32.03	176.6	63.27	32.09

Table 2.1 Statistical mean chemical shifts

In step b, whenever the $^{13}C^{\alpha}$, ^{13}CO, and $^{13}C^{\beta}$ chemical shifts of the intra- or inter-part of one spin-system are defined or partially defined, the amino acid type probabilities associated with that part are computed as follows.

The variation from average chemical shifts is defined as the difference between each individually observed chemical shift and the statistical mean value for the originating amino acid type, e.g.

$$\delta_{(i,j)}\,\omega_{C\alpha} = \omega_{C\alpha} - < \delta_{(i,j)}\omega_{C\alpha} > \qquad (2.6)$$

where $\omega_{C\alpha}$ is the observed chemical shift and $<\delta_{(i,j)}\omega_{C\alpha}>$ is the statistical mean of $^{13}C^{\alpha}$ chemical shifts for amino acid type i and secondary structure j. The variations of $^{13}C^{\alpha}$, ^{13}CO, and $^{13}C^{\beta}$ are pooled together and fitted by a three-dimensional Guassian distribution:

$$G(\delta_{(i,j)}\omega_{C\alpha}, \delta_{(i,j)}\omega_{Co}, \delta_{(i,j)}\omega_{C\beta}) = \exp\left\{-\frac{1}{2}\left[A1\left(\frac{\delta_{(i,j)}\omega_{C\alpha}}{\delta\omega_{C\alpha}}\right)^2 + A2\left(\frac{\delta_{(i,j)}\omega_{Co}}{\delta\omega_{Co}}\right)^2 + A3\left(\frac{\delta_{(i,j)}\omega_{C\beta}}{\delta\omega_{C\beta}}\right)^2\right]\right\} \qquad (2.7)$$

with the fitted standard deviations $\sigma\omega_{C\alpha}$=1.42 ppm, $\sigma\omega_{Co}$=1.32 ppm, and $\sigma\omega_{C\beta}$=1.31 ppm. The three variables ($\omega_{C\alpha}$, ω_{Co}, $\omega_{C\beta}$) are assumed to be normally distributed and independent, with the means and standard deviations to each amino acid type i and secondary structure j, where two variables (i, j) are also assumed to be independent. A1, A2, and A3 are normally equal to one, but in the case where one of the chemical shifts is not available corresponding Ax will be set to zero and the probability is still estimated from the available data. Using Bayes' theorem, this Guassian distribution can be interpreted as the Bayesian class conditional probability of observing a set of chemical shifts given amino acid type and fractional secondary structure content of the protein (Lukin *et al.*, 1997).

$$P(\omega_{C\alpha},\omega_{Co},\omega_{C\beta}\mid i,j)\propto G(\delta_{(i,j)}\omega_{C\alpha},\delta_{(i,j)}\omega_{Co},\delta_{(i,j)}\omega_{C\beta}) \tag{2.8}$$

The information about secondary structure is sometimes available from previous crystallographic studies of the protein now under NMR investigation, or from homologous proteins. Unfortunately, the secondary structure may not be available at the stage of chemical shift assignment. At this point, two possible approaches can provide this information. One is to use secondary-structure prediction programs, such as Psipred (Jones, 1999) with approximately 80 percent accuracy for assigning a residue to an α-helix, a β-sheet, or a loop. The other way is to estimate the fractional secondary structure content of the protein by circular dichroism (CD) spectroscopy or by counting the number of NMR cross peaks within certain ranges of chemical shifts (Wishart and Sykes, 1994). Suppose that, of all the residues of a protein, a fraction f1 is found in α-helices, f2 in β-sheets, and f3 in coils, where f3=1-f1-f2. According to Total Probability Theorem (Papoulis, 1984), the probability of observing a set of chemical shifts given amino acid of type i is proportional to

$$P(\omega_{C\alpha},\omega_{Co},\omega_{C\beta}\mid i)=\sum_{j=1}^{3}f_jP(\omega_{C\alpha},\omega_{Co},\omega_{C\beta}\mid i,j) \tag{2.9}$$

Finally, the Bayesian class posterior probability (Duda and Hart, 1973) is computed as:

$$P(i\mid\omega_{C\alpha},\omega_{Co},\omega_{C\beta})=\frac{P(i)P(\omega_{C\alpha},\omega_{Co},\omega_{C\beta}\mid i)}{\sum_i P(i)P(\omega_{C\alpha},\omega_{Co},\omega_{C\beta}\mid i)} \tag{2.10}$$

where $P(i|\omega_{C\alpha},\omega_{Co},\omega_{C\beta})$ is the probability that amino acid of type i occurs for an observed set of chemical shift values $(\omega_{C\alpha},\omega_{Co},\omega_{C\beta})$, and P(i) is the occurrence of amino acid of type i in a protein sequence ($\sum P(i)=1$).

With the possible amino acid types of their intra- and inter-parts, each spin-system is restricted to a list of dipeptide sites in the protein sequence $\{X_{i-1}\text{-}Y_i\}$, where X and Y are one of the 20 amino acids and i denotes the residue number. For example (Figure 2.5), spin-systems 612 (i) and 586 (j) are mapped to a list of dipeptide sites for MSG: $\{P_{26}\text{-}G_{27}, P_{72}\text{-}G_{73}, V_{127}\text{-}G_{128}, V_{340}\text{-}G_{341}\}$ and $\{G_{27}\text{-}T_{28}, G_{296}\text{-}T_{297}, A_{541}\text{-}T_{542}, A_{633}\text{-}T_{634}\}$, respectively.

2.2.4 Constraint-based match cycle for identifying adjacency relationship between spin-systems

Constraint propagation algorithm is applied recursively to reduce solution space for the identification of sequential relationships. Before describing details about how to identify the adjacency relationship between spin-systems, the related issues to the constraint propagation algorithm are mentioned first.

2.2.4.1 Introduction to Constraint Satisfaction Problems and Constraint Propagation Theory

Constraint Satisfaction Problems (CSP) (Tsang, 1995) have been a subject of research in Artificial Intelligence for many years. Constraint propagation [Barták, 2001 #4] is a common way of solving CSP.

What is a constraint? A constraint is simply a logical relationship among several variables, each possessing a value in a given domain. A constraint thus restricts the possible values that variables can take; it represents some partial information about the variables of interest. For instance, "the triangle is inside the circle" relates two objects without precisely specifying their positions, i.e., their coordinates. Now, one may move the triangle or the circle and he or she is still able to maintain the relation between these two objects.

The CSP is a problem where one is given:

1. a finite set of variables,
2. a function which maps every variable to a finite domain,
3. a finite set of constraints which restrict the combination values that a set of variables may take simultaneously.

A solution of a CSP is an assignment to each variable a value from its domain satisfies all the constraints.

The constraint propagation algorithm resolves CSP as follows. When a given variable is assigned a value, either directly by the user or automatically by the system, the algorithm recalculates the possible value sets and assigns values for all its dependent variables. This process continues iteratively until there are no more changes in the domain expression. More specifically, when a variable X changes its value, the system evaluates the domain expression of each variable Y dependent on X. This may generate a new set of possible values for Y. If this set changes, the preference constraint is evaluated to select one of the possible values as the new assigned value for Y. If this assigned value is different from the previous one, it causes the system to recalculate the values for further downstream variables. Values that have been assigned are always adopted as long as they are consistent with the defined constraints.

2.2.4.2 Adjacency relationship identification based on constraint propagation

In step c (Figure 2.1c), the program identifies all adjacency (preceding or succeeding) connectivities (Figure 2.5) between spin-systems, where constraint propagation algorithm is used actively to reduce solution space by filtering connectivities that cannot take part in any solution. Variables of CSP used for this program are spin-systems. Each spin-system (e.g., i) takes a set of values including chemical shifts

$$\{\omega_{C\beta}(i),\omega_{C\alpha}(i),\omega_{Co}(i),\omega_{H}(i),\omega_{N}(i),\omega_{C\beta}(i\text{-}1),\omega_{C\alpha}(i\text{-}1),\omega_{Co}(i\text{-}1),\omega_{H}(i,k),\omega_{N}(i,k)\} \quad (2.11)$$

and possible dipeptide $\{X_{i-1}\text{-}Y_i\}$ lists. The first six chemical shifts come from the three parts of one spin-system, and the latter two are obtained from sequential NOE cross-peaks (correlating $^{15}N/^{1}H^{N}$ shifts of different residue). Two spin-systems i and j, are considered to have sequential relationship (e.g., spin-system i is the preceding one of spin-system j) when the following constraints are satisfied:

$$|\omega_{Co}(i) - \omega_{Co}(j\text{-}1)| \leq \delta_{Co} \quad (2.12)$$

$$|\omega_{C\alpha}(i) - \omega_{C\alpha}(j\text{-}1)| \leq \delta_{C\alpha} \quad (2.13)$$

$$|\omega_{C\beta}(i) - \omega_{C\beta}(j\text{-}1)| \leq \delta_{C\beta} \quad (2.14)$$

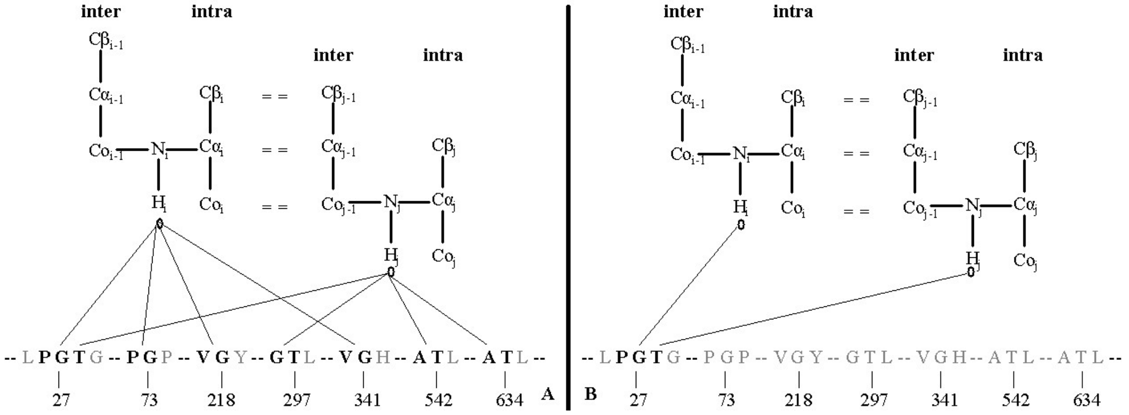
and at least one pair of dipeptides $\{X_{i-1}\text{-}Y_i\}$ and $\{X_{j-1}\text{-}Y_j\}$ satisfies:

$$Y_i = X_{j-1} \quad (2.15)$$

If there exist missing resonances in a spin system, comparison (eqs. 2.12-14) will be performed for the available data alone to examine adjacency relationships. In this case, high ambiguities might occur due to weakness in connectivity. As spin-systems i and j present an adjacency relationship, the ^{1}HN-^{1}HN NOESY cross-peak in the cluster from which spin-system i originates will be considered as a sequential NOE peak of spin-system i, if

$$|\omega_H(i,k) - \omega_H(j)| \leq \delta_H \tag{2.16}$$
$$|\omega_N(i,k) - \omega_N(j)| \leq \delta_N \tag{2.17}$$

Figure 2.5 presents an example to identify sequential connectivities by satisfying previously defined constraints. Spin-system i (spin-system ID: 612 for MSG), missing C^β resonance at intra-part, can represent dipetides, P-G or V-G (For MSG, its possible sites include: P_{26}-G_{27}, P_{72}-G_{73}, V_{217}-G_{218}, V_{340}-G_{341}), while spin-system j (ID: 586), also without C^β at inter-part, represent dipeptides, A-T or G-T (possible sites along MSG sequence: G_{27}-T_{28}, G_{296}-T_{297}, A_{541}-T_{542}, A_{633}-T_{634}). Equations 2.12-15 are all satisfied except Equation 3.14 because of the lost C^β chemical shifts. Therefore, the sequential connectivity between these two spin-systems is identified, corresponding to a tri-peptide, P-G-T (P_{26}-G_{27}-T_{28} along the primary sequence of MSG).



c Match common chemical shifts and identify sequential connectivities. Spin-system i without C^β information at intra-part, represents dipeptides (P-G or V-G) along MSG's primary sequence. Spin-system j without C^β at inter-part, represents dipeptides (A-T or G-T). These two spin-systems can be sequentially connected as a tripeptide (P_{26}-G_{27}-T_{28}) since their overlap C^α

and CO shifts are all consistent with each other.

In step d (Figure 2.1d), a "constraint propagation network" (CPN) (Zimmerman *et al.*, 1994; Zimmerman *et al.*, 1997) is applied to reducing the ambiguities encountered in the assembly of connectivity fragments and in mapping of the fragments to the protein sequence via a progressive procedure. A constraint, ‘unique-link’, is applied to the assembly of connecting fragments, which permits each spin-system to take only one preceding spin-system and only one succeeding spin-system. On account of this constraint, concerning spin-system j that can be linked to spin-system k or m without preference (two possible assignments), whenever spin-system k is linked to spin-system n, spin-system k is removed from the lists of possible neighbors of j, leading to a unique assignment of spin-system m linked to j. Since the protein data under NMR investigation are usually obtained from the single monomer construct in solution, another constraint, ‘unique-assignment’ which assumes that each residue in the protein can only have a unique assignment, is also utilized to map spin-systems (or fragments) onto the protein sequence. Due to this constraint, when a stretch of residues is assigned to a particular segment of linked spin-systems, these sites must be removed from the list of possible assignments that are stored for all other spin-systems, which in turn might lead to the unique sequential assignments for these spin-systems.

Step c and d together with b, form a constraint-based match cycle, which will perform along all the following three processing stages, to calculate the probabilities of amino acid types for intra- and inter-part of each spin-system, to map each spin-system to possible dipeptide sites in protein sequence, and to identify sequential relationship between spin-systems.

2.2.5 Strategies of spin-system assignment

As discussed previously, concerning assembling spin-systems and mapping the resultant segments to the protein sequence, a variety of methods (e.g., deterministic best-first, simulated annealing, genetic algorithm, exhaustive search, etc.) are applied to solve the problems arising from resonance overlap, missing resonances and extra resonances due to artifacts and impurities in the spectra. The best-first method achieves the goal via propagating constraints from the best candidate to good ones towards the end of the assignment process based on tightly matched criteria. However, any error made early can propagate to later assignments, leading to the failure of automation, especially for proteins with significant resonance overlap and/or missing resonances. Although the simulated annealing method is not sensitive to incomplete peak lists and overlap, it costs enormous computation work and can be susceptible to become trapped at local minima that correspond to incorrect assignment configurations. For the genetic algorithm method, experimentally observed peaks are mapped to predicted peaks based on homologous proteins with known assignments. This method is limited to the study of one of the members in a protein family. Exhaustive search attempts to enumerate all of the assignment solutions and eliminate improbable ones gradually through constraints established from experiments. In the cases where there are not enough connectivity constraints, for example, when many $^{13}C^{\beta}$ resonances are missing, the huge number of the possible solutions due to degeneracy of resonances makes it practically impossible to fully automatically obtain a unique solution for medium or large size proteins.

In respect with all the advantages and disadvantages of different methods mentioned above, this thesis proposes a set of routines for the backbone assignment based on the combination of best-first and exhaustive search methods. Instead of performing impractical exhaustive searches for the whole solution space at the beginning, the program starts from assigning some short but reliable segments to a target sequence using best-first method, which provides a footstone for the latter processes. Subsequently, extending assigned segments is achieved by exhaustive searches. In order to avoid local energy minimization and to be insensitive for incomplete peak lists, a new algorithm is also developed to select the correct segments from all possibilities resulting from the extension of reliable assigned short segments.

2.2.5.1 Establishing uniquely matched links

After defining the sequential relationships, the program first selects spin-systems that have unique sequential relationships with other spin-systems through the simultaneous matches of both $^{13}C_{\alpha}$ and ^{13}Co chemical shifts (eqs. 2.12-13). Secondly, spin-systems having more than one proceeding or succeeding spin-systems will be considered, provided that these systems contain information of $^{13}C_{\beta}$ and/or sequential NOE cross-peaks. If the ambiguities in sequential connectivity can be completely resolved by using C_{β} spin according to eq. 2.14 and/or using NOE according to eqs. 2.16-17, these spin systems will also be selected. Using all of the spin systems selected above, unique linkages among these spin-systems can be established and then some short reliable segments will be constructed and will be mapped to the sequence. The longer the segment, the higher the reliability to map the segment onto a specific site in the protein sequence based on CPN will be. These short but reliable segments provide a footstone for the next procedure, extending assigned segments.

2.2.5.2 Extending assigned segments

An exhaustive search algorithm is applied to find the remaining solutions based on the reliable segments assigned using uniquely matched spin-systems. The program extends each assigned short segment towards two directions (N-terminal or C-terminal) from both ends of the segment. Extension is not done by selecting one spin system over another according to the goodness of the matching (eqs. 2.12-14 and 2.16-17) between the spin-systems at the end of the segment and the one considered, which might be sensitive to incomplete peak lists. Instead, starting from one of the ends, all possible downstream paths (segments) are traced out. Selection of a correct path from all possibilities is based on an overall score

$$score = len + amsc * 1.5 + alink * 3 + fnoe \times 2 + fC_{\beta} * 2 - npenalty * 2 \tag{2.18}$$

$$amsc = \frac{1}{len - 1} \sum_{\substack{i=1 \\ j=i+1}}^{len-1} msc_{ij} \tag{2.19}$$

$$msc_{ij} = a_1 e^{-\Delta C\alpha_{ij} / \delta_{c\alpha}} + a_2 e^{-\Delta Co_{ij} / \delta_{co}} + a_3 e^{-\Delta C\beta_{ij} / \delta_{c\beta}} \tag{2.20}$$

$$alink = \frac{\sum_{i=1, j=i+1}^{len-1} link_{ij}}{len - 1} \tag{2.21}$$

where len is the number of the spin systems in the path (segment); amsc is an average matching score over the segment; msc_{ij} is the matching score between spin-system i and j; $\Delta C\alpha_{ij}$ (ΔCo_{ij} , $\Delta C\beta_{ij}$) is the $^{13}C^{\alpha}$ (^{13}CO and $^{13}C^{\beta}$) chemical shift difference between two spin-systems i and j; a_i equals to 1 if the resonances corresponding to the spin in the concerned term are available for both i and j spin systems, otherwise a_i is set to 0 in the absence of resonances for the concerned spin; alink is an average link-up factor; $link_{ij}$ is the number of link-up factor between spin systems i and j (one link-up factor corresponds to one match in either $^{13}C^{\alpha}$, ^{13}CO, $^{13}C^{\beta}$ or NOE); fnoe and fC_{β} are the fractions of spin-systems confirmed by sequential NOEs and matches of $^{13}C^{\beta}$ chemical shifts in the segment, respectively; npenalty is the fraction of weak links that are established using only one of the three spins ($^{13}C^{\alpha}$, ^{13}CO and $^{13}C^{\beta}$) due to missing resonances. All coefficients used in equation 2.18 are empirical values, which work well in the program. At the beginning of extension, only the segments with five or more spin systems will be considered. After assigning all of these long segments, shorter segments will be considered. With the increase of assignments, the ambiguities of the unassigned spin systems (segments) decrease based on the CPN procedure. Then, the short segments can be picked out more reliably. The program will repeat the procedure of constructing uniquely matched links and extending the established segments until no more spin-systems can be assigned.

2.2.5.3 Final assignments

At this point, 90% or more spin-systems can be assigned by recursively performing the previous two stages. For the rest of spin-systems, some have no linkage with others while others form short segments but cannot be uniquely mapped to the protein sequence. In this step, the probability of placing one segment in each possible set of positions will be calculated according to the overall amino acid type probability of the spin-systems which form the segments. The amino acid type probability of each spin-system is obtained from eq. 2.10. The one with higher overall probability will be chosen as the final assignment. After this step, isolated spin systems will be similarly placed according to amino acid type probability.

2.3 Results

2.3.1 Implementation

This program was written in the Tcl language. It can run independently in a batch mode, as well as run interactively with NMRView (Johnson and Blevins, 1994), which is also implemented in Tcl/Tk. The interface for displaying results and for checking potential errors is developed using Tk complementary package. All tests were conducted on a PC Intel Pentium III system (1 GHz) running Linux with Tcl/Tk 8.3.

Currently, online access of this program is also available at http://nmr5.dbs.nus.edu.sg. It consists of two parts: ASAP (Automated Sequential Assignment of NMR Resonances in Large Proteins), which runs on a server; and ASAPView, which runs independently on users' computers. ASAP web interface uses perl codes as CGI to call the ASAP program. These codes automatically process the input files of users and then send resonance assignment results back to users. ASAPView used in conjunction with NMRView can display the statistical analysis of resonance assignment results and check potential errors. ASAPView is downloadable from the same website.

2.3.2 MSG and p53 backbone resonance assignment

This program was tested using triple resonance data sets obtained from two distinctly different proteins: a 67 kDa dimeric protein p53 consisting of both DNA-binding and tetramerization domains (residues 82-360) (Mulder *et al.*, 2000), which mainly contains β-sheet secondary structures (residues 95-289, PDB entry ID 1KZY), and Malate Synthase G (residues 1-723, PDB ID 1D8C) (Tugarinov *et al.*, 2002), a largely α-helical protein. These data sets were gifts of Professor Lewis E. Kay.

According to eqs. 2.1-2, clusters were generated using tolerance values of 0.027/0.023 (for MSG and p53 respectively) and 0.38/0.34 ppm for $^{1}H^{N}$ and ^{15}N, corresponding to 1.75 times the respective spectral resolutions (1.75*SR). For the clusters with more than one spin-systems (e.g., more than one HNCOCA or intra-HNCOCASIM cross-peaks), we reduced the tolerances to 1.5 times the spectral resolutions. At the same time, we employed HNCOCASIM peak list to further designate cross-peaks into their corresponding spin-systems using tolerance values of 0.21/0.42 and 0.23/0.24 ppm (1.5*SR) for $^{13}C^{\alpha}$ and ^{13}CO according to eqs. 2.3-4. After this step, the tolerance values for $^{1}H^{N}$ and ^{15}N were increased to 2 times the digital resolutions to handle the peaks not assigned to any spin-systems. Finally, 26 and 12 clusters still could not be separated into unambiguous spin-systems for MSG and p53, respectively. Overall, 694 and 240 spin-systems were formed for MSG and p53.

Since most of NMR spectra are obtained from a unique monomer species in solution, this program is proposed to conduct one unique assignment for each set of NMR data. However, there are two configurations for p53. One corresponds to a major monomer species that is folded. The other is a minor species in solution that is unfolded (Mulder *et al.*, 2000). In this thesis, cross peaks corresponding to the unfolded domain were manually removed at this step, but the additional test of automated assignment with those data were also specially designed as discussed later.

Due to the missing peaks in NMR experiments, the information presented within each spin-system may not be complete. For MSG, we obtained 625 spin-systems with both intra and inter $^{13}C^{\alpha}$ (^{13}CO) chemical shifts and 71 spin-systems without either inter or intra $^{13}C^{\alpha}$ (^{13}CO) chemical shift. There were 228 spin-systems with both intra and inter $^{13}C^{\beta}$ chemical shifts, 126 spin-systems with intra $^{13}C^{\beta}$ chemical shifts, and 23 spin-systems with inter $^{13}C^{\beta}$ chemical shifts whose intra part corresponds to Gly residues since only Gly displays no $^{13}C^{\beta}$ cross-peak in the HNCACB spectrum. For p53, we obtained 211 spin-systems with both intra and inter $^{13}C^{\alpha}$ (^{13}CO) chemical shifts and 30 spin-systems without either inter or intra $^{13}C^{\alpha}$ (^{13}CO) chemical shift. There were 103 spin-systems with both intra and inter $^{13}C^{\beta}$ chemical shifts, 50 spin-systems with intra $^{13}C^{\beta}$ chemical shifts, and 12 spin-systems with inter $^{13}C^{\beta}$ chemical shifts.

All of the spin-systems were used to form segments based on sequential connectivity (eqs. 2.12-17) using tolerance values of 0.24/0.49 (1.75*SR), 0.27/0.28 (1.75*SR), and 0.45/0.25 (1.5*SR) ppm for $^{13}C^{\alpha}$, ^{13}CO, and $^{13}C^{\beta}$. Relying on the uniquely matched links, 284 and 99 spin-systems formed 40 and 22 segments and were mapped onto the protein sequences for MSG and p53, respectively. Using these reliable segments, assignments were extended according to eqs. 2.12-14 using tolerance values of 0.28/0.56 (2*SR), 0.31/0.31 (2*SR) and 0.45/0.25 (1.5*SR) ppm for $^{13}C^{\alpha}$, ^{13}CO, and $^{13}C^{\beta}$. At the end of the extension process, 637 and 223 spin-systems were assigned with 1 and 0 errors for MSG and p53, respectively. For MSG, residue V155 (Figure 2.6B) was wrongly assigned because its corresponding spin-system and one unassigned spin-system shared nearly identical ($^{13}C^{\alpha}$, ^{13}CO, $^{13}C^{\beta}$) chemical shifts for both intra- and inter-residue parts.

Protein	Sequential resonance assignment				
	Residue number	Available data	Assigned residues	Incorrect number	Accuracy
Malate Synthase G	723	652	642	1	99.80%
p53[1]	279	241	234	0	100%
p53[2]	279	241+37	234	0	100%

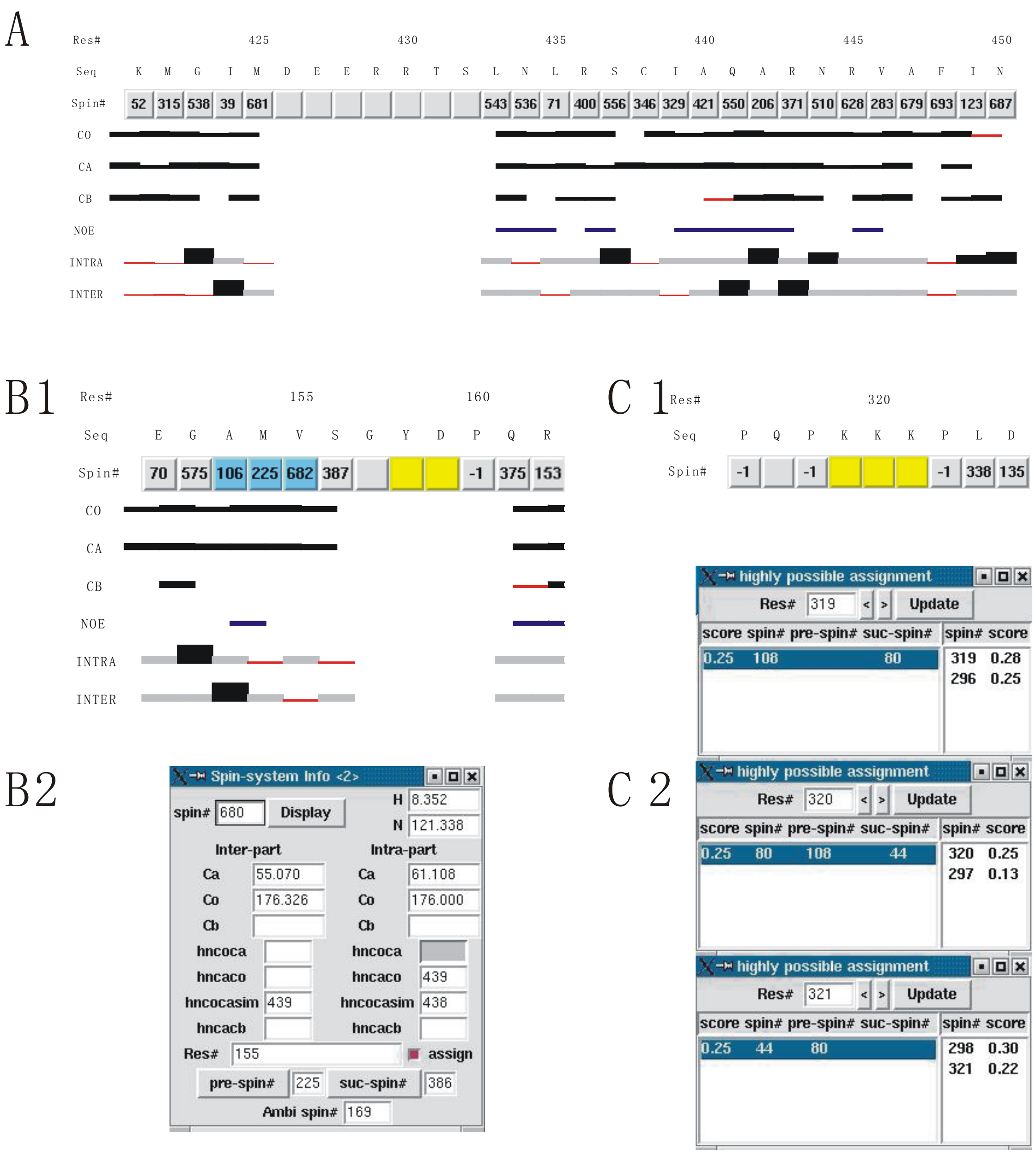
Table 2.2 Sequential resonance assignment of MSG and p53. p53[1] represents the test where unfolded signal in all experiments were all manually removed before automatically conducting assignment algorithms. p53[2] represents the test where all signals (additional ones for 37 residue assignments corresponding to local minor of folded tet and unfolded tet domain) observed in experiments were utilized.

The remaining systems were assigned according to the probability of a segment or spin-system located at a specific site in the sequence. At this step, 3 and 11 spin-systems were correctly assigned for MSG and p53 respectively. Table 2.2

summarizes the final assignment results. These results were obtained in 120 and 21 minutes for MSG and p53, respectively, using the standard desktop computer indicated above.

2.3.3 Comparison with manual assignment

In comparison with manual assignment results, several spin-systems were not uniquely assigned because the amino acid types of the segments could not be properly predicted. For MSG, a segment with three spin-systems, which corresponds to Q_{116}-L_{117}-V_{118}-V_{119}, has three possible assignments because of incorrect prediction of amino acid type for V_{119}. The $^{13}C^{\alpha}$ (55.5 ppm) and ^{13}CO (171.3 ppm) chemical shifts of V_{119} are much smaller than their respective statistical values (62.8 and 176.0 ppm). Several isolated spin-systems have two or more possible sites since they lack characteristic information in amino acid types. Similarly, two short segments of p53 (H_{297}-E_{298}-L_{299}, P_{318}-K_{319}-K_{320}-K_{321}) could not be placed onto unique sites in the protein sequence. If all of the residues excluding prolines give signals, in principle, the segment with 2 or 3 spin-systems should be uniquely assigned. In practice, a number of regions in both p53 and MSG do not display NMR resonances. Since some of the assignments are less reliable, a graphic interface was designed to manually check the assignments with weak linkages and ambiguities. Figure 2.6A shows one part of the assignments, where weak linkages in $^{13}C^{\alpha}$, ^{13}CO, $^{13}C^{\beta}$ are indicated by red lines. The residues displaying unusual chemical shifts in intra-residue (inter-residue) are also highlighted in red on the INTRA (INTER) row in the figure. Figure 2.6C shows additional suggestions for an ambiguous assignment colored in yellow. When the residue button marked in yellow (C1) is clicked, a sub-window (C2) will pop up to show all of the possible assignments for this residue. For example, three sequential spin-systems (108, 80 and 44) can be placed at polypeptides P_{318}-K_{321} or P_{295}-E_{298}, with scores for individual spin-systems (e.g., spin-system can be placed at P_{318}-K_{319} and P_{295}-H_{296} with scores of 0.28 and 0.25, respectively). Such ambiguities might be resolved through manual inspection of the spectra.


A
Res# 425 430 435 440 445 450
Seq K M G I M D E E R R T S L N L R S C I A Q A R N R V A F I N
Spin# 52 315 538 39 681 543 536 71 400 556 346 329 421 550 206 371 510 628 283 679 693 123 687
CO
CA
CB
NOE
INTRA
INTER
B1
Res# 155 160
Seq E G A M V S G Y D P Q R
Spin# 70 575 106 225 682 387 -1 375 153
CO
CA
CB
NOE
INTRA
INTER
B2
C 1
Res# 320
Seq P Q P K K K P L D
Spin# -1 -1 -1 338 135
C 2

Spin-system Info <2>

spin# 680 Display

H 8.352

N 121.338

Inter-part		Intra-part	
Ca	55.070	Ca	61.108
Co	176.326	Co	176.000
Cb		Cb	
hncoca		hncoca	
hncaco		hncaco	439
hncocasim	439	hncocasim	438
hncacb		hncacb	

Res# 155 assign

pre-spin# 225 suc-spin# 386

Ambi spin# 169

highly possible assignment

Res# 319 < > Update

score	spin#	pre-spin#	suc-spin#
0.25	108		80

spin#	score
319	0.28
296	0.25

highly possible assignment

Res# 320 < > Update

score	spin#	pre-spin#	suc-spin#
0.25	80	108	44

spin#	score
320	0.25
297	0.13

highly possible assignment

Res# 321 < > Update

score	spin#	pre-spin#	suc-spin#
0.25	44	80	

spin#	score
298	0.30
321	0.22

Figure 2.6 Graphic output of resonance assignment. (A) Graphical interface showing sequential connectivities. The thickness of the bars represents the relative goodness of the overlaps between inter- and intra-parts of the spin-systems. Weak linkages in ^{13}CO, ^{13}C$^{\alpha}$ and ^{13}C$^{\beta}$ are indicated by red bars. Blue bars show spin-systems with sequential NOEs. The consistency between the observed and predicted (^{13}CO, ^{13}C$^{\alpha}$ and ^{13}C$^{\beta}$) chemical shifts is shown by the bars on the INTRA (INTER) line for intra-part (inter-part) of the spin-system. The thicker the bar, the smaller the difference between observed and predicted chemical shifts. Residues displaying unusual (^{13}CO, ^{13}C$^{\alpha}$ and ^{13}C$^{\beta}$) chemical shifts are also highlighted in red. (B) Graphical interface showing degenerate spin-systems. In (B1), two spin-systems are considered to be degenerate when they have the same chemical shifts in both inter- and intra-residue (^{13}CO, ^{13}C$^{\alpha}$) spin pairs.

The assignment for degenerate spin-systems is indicated by the blue button. When the button is clicked, a sub-window (B2) will pop up to show the details of this spin-system, including chemical shifts of all spins, peak numbers of the available correlations observed in four spectra (HNCOCA, HNCOCASIM, HNCACO, and HNCACB), the spin-system numbers of the preceding and succeeding residues, and the number of the spin-system that is degenerate with the present one. (C) Graphical interface for examining ambiguous assignment. In (C1), the button for ambiguous assignment is colored yellow. When the button is clicked, a sub-window will pop up to show all possible assignments for the spin-system filled at that site. In the first sub-window corresponding to residue 319, an average score of amino acid type for the segment (0.25) is displayed in the left panel of this window. The numbers of the spin-systems corresponding to the current, preceding and succeeding systems follow this score. In the right-panel of the first sub-window, possible assignments for the present spin-system are listed. For each possibility, residue number and score of residue type are given, e.g., the score to assign spin-system 108 to residue 319 is 0.28 while the score is 0.25 for the assignment of spin-system 108 to residue 296. The second and third sub-windows correspond to residues 320 and 321.

2.3.4 Backbone NOE assignment

NN-NOESY experiment provides correlations of the form ([$\omega_{N}(i)$, $\omega_{HN}(i)$, $\omega_{N}(j)$, $\omega_{HN}(j)$]). The program utilizes NOE cross-peaks to resolve carbon shift degenerate and also assigns backbone NOE cross-peaks. The correlation is referred to as sequential NOE when $|i-j|=1$, medium-range NOE when $|i-j|<4$, and long-range NOE when $|i-j|>4$. In total, we obtained 588 (143) sequential NOEs, 9 (11) medium-range NOEs and 20 (33) long-range NOEs for MSG and p53, respectively. To obtain more medium- and long-range NOEs, a model structure is required.

2.3.5 p53 assignment with data responsible for both major and minor monomer species

The program was also tested for p53, with total spectra data corresponding to both folded species in solution and unfolded one that had been manually removed for the previous test. The proposed assignment algorithm in this thesis assumes that each site in a target sequence can only be occupied by a unique spin-system. Therefore as long as one spin-system is assigned to a site (e.g., residue j) in sequence, CPN will remove the possibility of assigning any other spin-systems to residue j ('unique-assignment' constraint), which shrinks the solution space and progressively conducts resonance assignment towards final complete assignment. However, if there exists more than one spin-system corresponding to the same sites in the protein sequence like the additional set of assignment for the minor monomer of p53, the program might fail to receive a correct outcome for the assignment of either the major or minor monomer of the protein. For example, if spin-systems k (corresponding to major assignment) and g (from the minor assignment) both responsible for residue m were both predicted as the candidates of residues m and n, once spin-system k is assigned to residue m, the CPN will find that spin-system g can only be assigned to residue n so that the program incorrectly assigns spin-system g to residue n. In the case where spin-systems from unfolded part with characteristic chemical shifts are recognized as specific residue types (e.g., Ala, Gly, Thr, Ser, etc.) and therefore can be mapped to few but specific sites in sequence, the problem of assigning unfolded spin-systems to incorrect positions in sequence will be more severe.

On the other hand, it will be helpful to avoid the risk of assigning additional assignment data at incorrect positions in sequence if we can know which spin-systems belong to the additional assignment set before assigning them. For this purpose, we started from two 'marker' spin-systems that indicated the presence of additional assignment. The first one was predicted as dipeptide A-G and could only be mapped to residues A_{355}-G_{356} in the sequence of p53. Although another spin-system could also be uniquely mapped to residues A_{355}-G_{356}, its highest intensity compared with the 'marker' spin-system implied that it belonged to the major species (folded species) while the 'marker' spin-system belonged to unfolded species. The other one represented dipeptides: D-G, L-G, N-G, Y-G or F-G, however, only if assigned to residues D_{324}-G_{325} this spin-system could be detected presenting strong sequential connectivity with others to form a reliable segment which could be uniquely mapped to the sequence. Although another spin-system existed, which could associate with others and give birth to a credible segment with respect to the same set of residues, the difference in intensity again assisted to distinguish between the spin-systems arising from folded and unfolded species.

Parts of spin-systems belonging to additional assignments were obtained by extending the 'marker' spin-system into two segments. One corresponds to polypeptide (residues P_{322}-G_{334}), and the other corresponds to residues R_{342}-E_{358}. The connectivities between these spin-systems were established simply by 'best-first' method (e.g., choose the unassigned spin-systems with the highest scores of msc_{ij} in eq. 2.20) from the reliable identified sequential connectivities (e.g., msc_{ij}>0.45). However, other undiscovered spin-systems corresponding to the unfolded part might share similar carbon shifts with unassigned spin-systems of the folded part, which give rise to carbon shift degeneracies during further extending these two segments. In addition, some overlapped carbon shifts between identified sequential spin-systems (for unfolded part) were distinctly different (e.g., msc_{ij}<0.45). These two problems prohibit the program from identifying all spin-systems of the unfolded part (residues P_{322}-E_{358}).

The discovered spin-systems belonging to additional assignment, especially for the spin-systems with characteristic chemical shifts, were prohibited in resonance assignment for the major species. As a result, CPN can work correctly in this instance: 234 spin-systems (the same as that when artificially removing dada from additional assignment) were correctly assigned to the sequence of p53, corresponding to major species in solution (Table 2.2).

2.4 Discussion

Using the automated program developed here, we have achieved assignment for about 96% of the resonances observed in large proteins, MSG and p53. The assignment completeness and accuracy yielded by the automated program are slightly lower than those obtained by manual methods. However, the time needed for the assignment of large proteins is much shorter using the automated method than using manual methods. For example, automatic assignment of MSG (652 residues with data) only takes two hours, while manual assignment of the same protein takes at least two weeks even for an experienced researcher. Assignments with potential errors are often associated with weak linkages or degeneracies in chemical shifts of three or more spins. Hence such potential errors can be easily identified from a graphical interface where sites with weak linkages and degeneracy are marked by different colors. The weak sites may be reviewed manually to confirm the assignment.

The performance of the program can be affected by the quality of the spectra and tolerance values used. The tolerances used for generating spin-systems through combination of different peak lists and for establishing sequential connectivity between spin-systems are critical for automated assignment procedures. Suitable tolerances can reduce computational time and increase the reliability of resonance assignment. Too large tolerances will produce serious ambiguities and increase computational complexity. Too small tolerances may result in the formation of incorrect spin-systems and segments due to uncertainty of peak positions. In most cases, a set of fixed tolerances will either fall into the first situation or have to encounter the second problem.

Our program selected different tolerances at different stages. All of them were evaluated with respect to the spectral resolutions of the NMR spectra. To reduce ambiguities in the construction of spin systems, the tolerance values of δ_H and δ_N (1.5*SR) should be smaller than those used for building clusters (1.75*SR). After most of spin-systems are identified, the remaining resonances were regrouped and then designated to spin-systems using slightly larger tolerance values for $^1H^N$ and ^{15}N spins (2*SR). In this way, wrong assignments arising from uncertainty of peak positions can be greatly avoided. At the same time, the computational time will not be increased significantly. Similarly, to establish segments creditably, the chemical shift difference between the inter-part of one spin-system and the intra-part of another should not be too large and thus relatively smaller tolerance values (1.75*SR) could be used. To extend assigned segments, however, the tolerance values need to be large enough

(2*SR) to take all possibilities into account.

Severe chemical shift degeneracy commonly exists in ($^{1}H^{N}$, ^{15}N) or/and ($^{13}C^{\alpha}$, ^{13}CO) pairs for large proteins. Taking the advantage of the 4D-HNCOCASIM spectrum, the problem arising from degeneracy in ($^{1}H^{N}$, ^{15}N) pairs of chemical shifts can be resolved and thus most of spin-systems excluding $^{13}C^{\beta}$ resonances can be identified. The presence of a few clusters with more than one spin-system allows us to list all of the possible combinations. The problem arising from degeneracy in ($^{13}C^{\alpha}$, ^{13}CO) pairs of chemical shifts can be, in principle, overcome by using sequential NOEs and/or both intra- and inter-residue $^{13}C^{\beta}$ chemical shifts.

The inter-residue correlation [$\omega_{C\beta}$(i-1), ω_{N}(i), ω_{HN}(i)] is difficult to be established from the traditional HN(CO)CACB experiment for high molecular weight molecules. This results from the rapid decay of the transverse magnetization of ^{13}CO spin at high magnetic field through the chemical shift anisotropy relaxation mechanism. On the other hand, the intra-residue correlation [$\omega_{C\beta}$(i), ω_{N}(i), ω_{HN}(i)] can be easily obtained from a sensitive experiment HNCACB. Except for the intra-residue information, there exists the inter-residue information in the same HNCACB spectrum, especially for residues located in β-strands and non-structural regions. These inter-residue correlations normally complicate spectral analysis since they make the spectrum more crowded. Methods have been developed to suppress this so-called redundant information. In this program, we did not include the inter-residue data from the insensitive HN(CO)CACB experiment. Instead, we fully utilized the sequential correlations ([$\omega_{C\beta}$(i-1), ω_N(i), ω_{HN}(i)]) observed in the 3-D HNCACB experiment. Many residues in MSG and p53 displayed both intra- and inter-residue correlations. However, only a part of them can be designated into their corresponding spin-systems in the case where one cross-peak in the ^{1}H-^{15}N HSQC corresponds to one residue. For MSG, both intra- and inter-residue $^{13}C^{\beta}$ correlations of about 35% residues can be designated in the spin-system construction process. For p53, the intra- and inter-residue $^{13}C^{\beta}$ information is higher (42%) due to its high content in β-sheet secondary structure. Sequential NOEs are often observed for residues located in α-helices and turns, which compensate the lack of inter-residue $^{13}C^{\beta}$ correlations of these residues. The combination of sequential NOEs and $^{13}C^{\beta}$ correlations allows us to assign the resonances automatically.

2.5 Summary

In summary, we have developed a computer program for the assignment of backbone and $^{13}C^{\beta}$ resonances of large proteins based on the combination of best-first and exhaustive search methods. Many NOEs can also be assigned using this program. The excellent performance on two test proteins (p53 and MSG) demonstrates that the insensitive experiment HN(CO)CACB is not necessary for large proteins. Then this program will accelerate sequential assignment and facilitate the study of large proteins by NMR.

Chapter 3

Conclusion and Future work

3.1 Conclusion

This thesis aimed at designing automated approaches for backbone resonance assignment from heteronuclear 4D NMR spectra of large proteins. The study consisted of the extraction of backbone spin-systems, the identification of amino acid types, the identification of adjacency relationship between spin-systems, and finally the mapping of spin-systems to dipeptides.

There is normally more severe chemical shift degeneracy in $^{1}H^{N}/^{15}N$ pairs for large proteins than that for small or medium size proteins. 4D-HNCOCASIM experiment excels in resolving the problems arising from degeneracy in $^{1}H^{N}/^{15}N$ chemical shifts, and hence most of spin-systems excluding $^{13}C^{\beta}$ resonances have been identified.

Statistical values of chemical shifts with respect to certain secondary structure type were applied to studying amino acid types. The computed Bayesian class posterior probability provided correct amino acid type information for most of the combined spin-systems.

Adjacency relationships were identified by overlapped carbon chemical shifts (CO, C^{α} and/or C^{β}) between intra-part of one spin-system and inter-part of another. When using these identified relationships to establish sequential linkage, the use of sequential NOEs and/or $^{13}C^{\beta}$ assisted to resolve the ambiguity arising from $^{13}C^{\alpha}/^{13}CO$ chemcial shift degeneracy.

An algorithm described as the combination of ‘best-first’ and ‘exhaustive search’ methods was also developed to rapidly and accurately assign spin-systems to the protein sequence.

Finally, the proposed algorithm was validated with experimental data based on implemented computer programs. Except for several small segments without characteristic chemical shifts, most of spin-systems and resultant segments were correctly assigned to the protein sequence, which were consistent with previous manual assignment.

This thesis demonstrates that an automated resonance assignment work is possible for very large proteins. Since resonance assignment is the basis of protein structure determination by NMR, the study of the structure determination of large proteins will be practicable in the future. Biologists, who currently apply X-ray crystallography technique in researching large proteins, will be able to take the special advantages of NMR to investigate biomolecules in solution under nearly physiological conditions along with the dynamics information associated with protein functions. Even without the information of protein structure, the resonance assignment of large proteins resulting from the program proposed in this thesis, will still be able provide protein-protein interaction and protein dynamics information. These will also enable the biologists to accelerate proteomics projects in unraveling biological functions.

3.2 Future work

There are still several related aspects from the current work that can be further studied. Some of them are described as follows.

Automation of the peak picking

All of the algorithms described in this thesis require peak lists for input. Therefore, a reliable automated peak picking method becomes pivotal. A fully automated resonance assignment program can be realized only with a robust peak picking procedure. Current peak picking algorithms can distinguish real peaks from false ones by analyzing peak shapes. However, they may fail to provide precise positions (chemical shifts) for overlapped cross-peaks. A possible extension from our studies is to develop an intelligent peak picking algorithm which analyzes the suspicious peaks and their surroundings to study and provide precise coordinates for overlapped peaks. Along with such investigation of peak picking, it should be possible to employ restrictive tolerance during the identification of spin-systems and adjacency relationships. This will make the the current resonance assignment more reliable.

Utilization of information from various methods

Besides the protein primary sequence, which is necessary for the sequence-specific resonance assignment, other information obtained from physical or chemical methods may be helpful in designing an automated assignment program. For example, the residue specific information provides more detailed amino acid type information than that obtained from the estimation of observed chemical shifts. The chemical shift assignment of previously studied homologous proteins can serve as a reference for the automated program, since the chemical shifts among a protein family change little and only differ in a few regions. These are useful criteria which should be considered when doing the sequential mapping of spin-systems. Currently our program does not include systematic approaches. A well-designed expert system might be necessary to make use of all such types of miscellaneous information.

Employment of more comprehensive statistical chemical shift values in determining amino acid types

The standard deviations of chemical shifts (σ_{CO}, $\sigma_{C\alpha}$ and $\sigma_{C\beta}$ in eq. 2.7) used to predict amino acid type are identical with all kinds of amino acid in this program. When more reliable resonance assignment data are available, further statistical analysis of chemical shifts can be conducted. For instance, the standard deviation of chemical shifts according to certain amino acid and its secondary structure might provide more specific and reliable amino acid type information, which will speed up the performance of the program and make its result more accurate.

References

Bartels, C., Guntert, P., Billeter, M. and Wüthrich, K. (1997). GARANT - a General Algorithm for Resonance Assignment of Multidimensional Nuclear Magnetic Resonance Spectra. J Comput Chem 18: 139-149.

Bartels, C., Xia, T. H., Billeter, M., Guntert, P. and Wüthrich, K. (1995). The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR. 5: 1-10.

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

Clore, G. M. and Bewley, C. A. (2002). Using conjoined rigid body/torsion angle simulated annealing to determine the relative orientation of covalently linked protein domains from dipolar couplings. J Magn Reson. 154: 329-335.

Coggins, B. E. and Zhou, P. (2003). PACES: Protein sequential assignment by computer-assisted exhaustive search. J Biomol NMR. 26: 93-111.

Cordier, F., Rogowski, M., Grzesiek, S. and Bax, A. (1999). Observation of through-hydrogen-bond 2hJHC' in a perdeuterated protein. J Magn Reson. 140: 510-512.

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

Dalgarno, D. C., Levine, B. A. and Williams, R. J. (1983). Structural information from NMR secondary chemical shifts of peptide alpha C-H protons in proteins. Biosci Rep. 3: 443-452.

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

Duda, R. O. and Hart, P. E. (1973). Pattern Classification and Scene Analysis. New York, John Wiley & Sons.

Ferentz, A. E. and Wagner, G. (2000). NMR spectroscopy: a multifaceted approach to macromolecular structure. Q Rev Biophys. 33: 29-65.

Gardner, K. H., Rosen, M. K. and Kay, L. E. (1997). Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. Biochemistry. 36: 1389-1401.

Goto, N. K. and Kay, L. E. (2000). New developments in isotope labeling strategies for protein solution NMR spectroscopy. Curr Opin Struct Biol. 10: 585-592.

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

Gronwald, W., Willard, L., Jellard, T., Boyko, R. F., Rajarathnam, K., Wishart, D. S., Sonnichsen, F. D. and Sykes, B. D. (1998). CAMRA: chemical shift based computer aided protein NMR assignments. J Biomol NMR. 12: 395-405.

Grzesiek, S., Anglister, J., Ren, H. and Bax, A. (1993). Carbon-13 line narrowing by deuterium decoupling in deuterium/carbon-13/nitrogen-15 enriched proteins. Application to triple resonance 4D J connectivity of sequential amides. J Am Chem Soc. 115: 4369-4370.

Grzesiek, S., Wingfield, P., Stahl, S., Kaufman, J. D. and Bax, A. (1995). Four-Dimensional 15N-Separated NOESY of Slowly Tumbling Perdeuterated 15N-Enriched Proteins. Application to HIV-1 Nef. J Am Chem Soc. 117: 9594-9595.

Hitchens, T. K., Lukin, J. A., Zhan, Y., McCallum, S. A. and Rule, G. S. (2003). MONTE: An automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. J Biomol NMR. 25: 1-9.

Johnson, B. A. and Blevins, R. A. (1994). NMRView: A computer program for the visualization and analysis of NMR data. J Biomol NMR. 4: 603-614.

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

Konrat, R., Yang, D. and Kay, L. E. (1999). A 4D TROSY-based pulse scheme for correlating 1HNi,15Ni,13Cαi,13C'i-1 chemical shifts in high molecular weight, 15N,13C, 2H labeled proteins. J Biomol NMR. 15: 309-313.

Leutner, M., Gschwind, R. M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H. (1998). Automated backbone assignment of labeled proteins using the threshold accepting algorithm. J Biomol NMR. 11: 31-43.

Lukin, J. A., Gove, A. P., Talukdar, S. N. and Ho, C. (1997). Automated probabilistic method for assigning backbone resonances of (13C,15N)-labeled proteins. J Biomol NMR. 9: 151-166.

Markley, J. L. and Kainosho, M. (1993). Stable Isotope Labeling and Resonance Assignments in Larger Proteins. **In** NMR of Biological Macromolecules: A Practical Approach. (G. C. K. Roberts, Ed.). Oxford University Press, Oxford, pp 101-152.

Metzler, W. J., Constantine, K. L., Friedrichs, M. S., Bell, A. J., Ernst, E. G., Lavoie, T. B. and Mueller, L. (1993). Characterization of the three-dimensional solution structure of human profilin: 1H, 13C, and 15N NMR assignments and global folding pattern. Biochemistry. 32: 13818-13829.

Montelione, G. T., Zheng, D., Huang, Y. J., Gunsalus, K. C. and Szyperski, T. (2000). Protein NMR spectroscopy in structural genomics. Nat Struct Biol. 7 Suppl: 982-985.

Moseley, H. N., Monleon, D. and Montelione, G. T. (2001). Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. Methods Enzymol. 339: 91-108.

Moseley, H. N. and Montelione, G. T. (1999). Automated analysis of NMR assignments and structures for proteins. Curr Opin Struct Biol. 9: 635-642.

Mulder, F. A., Ayed, A., Yang, D., Arrowsmith, C. H. and Kay, L. E. (2000). Assignment of 1H(N), 15N, 13C(alpha), 13CO and 13C(beta) resonances in a 67 kDa p53 dimer using 4D-TROSY NMR spectroscopy. J Biomol NMR. 18: 173-176.

Nietlispach, D., Clowes, R. T., Broadhurst, W., Ito, Y., Keeler, J., Kelly, M., Ashurst, J., Oschkinat, H., Domaille, P. J. and Laue, E. D. (1996). An Approach to the Structure Determination of Larger Proteins Using Triple Resonance NMR Experiments in Conjunction with Random Fractional Deuteration. J Am Chem Soc. 118: 407-415.

Papoulis, A. (1984). Probability, Random Variables, and Stochastic Processes. New York, McGraw-Hill.

Pervushin, K., Ono, A., Fernandez, C., Szyperski, T., Kainosho, M. and Wüthrich, K. (1998). NMR scalar couplings across Watson-Crick base pair hydrogen bonds in DNA observed by transverse relaxation-optimized spectroscopy. Proc Natl Acad Sci U S A. 95: 14147-14151.

Pervushin, K., Riek, R., Wider, G. and 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.

Riek, R., Fiaux, J., Bertelsen, E. B., Horwich, A. L. and Wüthrich, K. (2002). Solution NMR techniques for large molecular and supramolecular structures. J Am Chem Soc. 124: 12144-12153.

Salzmann, M., Pervushin, K., Wider, G., Senn, H. and Wüthrich, K. (1998). TROSY in triple-resonance experiments: new perspectives for sequential NMR assignment of large proteins. Proc Natl Acad Sci U S A. 95: 13585-13590.

Salzmann, M., Wider, G., Pervushin, K., Senn, H. and Wüthrich, K. (1999). TROSY-type Triple-Resonance Experiments for Sequential NMR Assignment of Large Proteins. J Am Chem Soc. 121: 844-848.

Tsang, E. (1995). Foundations of Constraint Satisfaction. London, Academic Press.

Tugarinov, V., Muhandiram, R., Ayed, A. and Kay, L. E. (2002). Four-dimensional NMR spectroscopy of a 723-residue protein: chemical shift assignments and secondary structure of malate synthase g. J Am Chem Soc. 124: 10025-10035.

Venters, R. A., Metzler, W. J., Spicer, L. D., Mueller, L. and Farmer, B. T. (1995). Use of 1HN-1HN NOEs to Determine Protein Global Folds in Perdeuterated Proteins. J Am Chem Soc. 117: 9592-9593.

Wider, G. and Wüthrich, K. (1999). NMR spectroscopy of large molecules and multimolecular assemblies in solution. Curr Opin Struct Biol. 9: 594-601.

Wishart, D. S. and Nip, A. M. (1998). Protein chemical shift analysis: a practical guide.

Biochem Cell Biol. 76: 153-163.

Wishart, D. S. and 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. and Sykes, B. D. (1994). Chemical shifts as a tool for structure determination. Methods Enzymol. 239: 363-392.

Wishart, D. S., Sykes, B. D. and Richards, F. M. (1991). Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. J Mol Biol. 222: 311-333.

Wishart, D. S., Sykes, B. D. and 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.

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

Yamazaki, T., Lee, W., Arrowsmith, C. H., Muhandiram, D. R. and Kay, L. E. (1994). A Suite of Triple Resonance NMR Experiments for the Backbone Assignment of 15N, 13C, 2H Labeled Proteins with High Sensitivity. J Am Chem Soc. 116: 11655-11666.

Yang, D. and Kay, L. E. (1999). Improved 1HN-detected triple resonance TROSY-based experiments. J Biomol NMR. 13: 3-10.

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

Yee, A., Pardee, K., Christendat, D., Savchenko, A., Edwards, A. M. and Arrowsmith, C. H. (2003). Structural proteomics: toward high-throughput structural biology as a tool in functional genomics. Acc Chem Res. 36: 183-189.

Zimmerman, D., Kulikowski, C., Wang, L., Lyons, B. and Montelione, G. T. (1994). Automated sequencing of amino acid spin systems in proteins using multidimensional HCC(CO)NH-TOCSY spectroscopy and constraint propagation methods from artificial intelligence. J Biomol NMR. 4: 241-256.

Zimmerman, D. E., Kulikowski, C. A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R. and Montelione, G. T. (1997). Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol. 269: 592-610.

Zuiderweg, E. R. (2002). Mapping protein-protein interactions in solution by NMR spectroscopy. Biochemistry. 41: 1-7.

Publications and poster presentations

a. *Poster presentation*

Kai Li, *Tin-Wee Tan, Daiwen Yang*, (2002).* **Automated backbone assignment in large proteins using 4D TROSY NMR experiments.** *2nd International Conference on Structural Biology and Functional Genomic (poster presentation).*

http://surya.bic.nus.edu.sg/~likai/Publication/poster_v1.1.pdf

b. *Paper to be published*

Kai Li, *Tin-Wee Tan, Daiwen Yang*.* **ASAP: Automated Sequential Assignment of NMR Resonances for Large Proteins.** *(To be submitted to J Mol Biol).*

Online access of ASAP program: http://nmr5.dbs.nus.edu.sg