Development of Computer-Assisted Methods for the Resonance Assignment of Heteronuclear 3D NMR Spectra of Proteins

Kuo-Bin Li

Department of Chemistry
McGill University
Montréal, Québec, Canada

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

September 1996

ISBN 0-612-19741-7

Canada

Abstract

An automated sequential assignment protocol for proteins is presented using heteronuclear 3D NMR. For the observed amino acid spin systems, the protocol includes an algorithm to determine their amino acid types. For the detected polypeptides, the protocol includes another algorithm to sequentially map them to the primary sequence. The former algorithm measures the similarity between the detected spin systems and the 20 standard amino acid patterns. Both chemical shift and topological likeness are considered. Knowing the amino acid types, the mapping algorithm assigns the detected polypeptides to proper positions within the protein primary sequence. The assignment protocol can be applied to spin systems generated from many different approaches. To demonstrate the assignment protocol, a few computer algorithms were designed to deduce the backbone and side-chain spin systems of proteins using heteronuclear 3D NMR. Magnetization transfer through peptide bonds can be observed in triple resonance 3D NMR. To automate the backbone assignment using the through-bond correlations, a generic algorithm is proposed. This algorithm searches and merges cross peaks among all available NMR spectra. Individual spin systems can be extracted and linked to create polypeptide chains based on the observed interresidue correlations. The algorithm is not restricted to any particular type of experiment. It is shown to be applicable to two sets of NMR spectra: the five-experiment set of 3D HNCO, HNCA, HN(CO)CA, HCACO, ^{15}N TOCSY-HMQC and the one-experiment set of 3D CBCANH. For the side chain assignment, an automated approach using a constrained partitioning algorithm has been developed to extract side chain spin systems of proteins by analyzing the 3D HCCH-COSY/TOCSY spectra. The extracted amino acid spin systems show the chemical shifts of the component nuclear spins as well as the connectivities between these spins. A 90-residue protein, the N-domain of chicken skeletal troponin-C (1-90), was used to test the implementation of the above algorithms with both simulated and experimental data. Limitations of the algorithms are discussed.

Résumé

Un protocol automatisé pour l'attribution séquentielle des protéines est presentée en utilisant RMN 3D. Le protocole utilise une algorithme pour déterminer les types d'acides aminés en observant les systèmes de spin. Pour les polypeptides détectés le protocole utilise une algorithme pour déterminer la séquence primaire. La première algorithme mesures les similarités entre les systèmes de spin détecter et les vingts acides aminés standards. Le déplacement chimique et similarités topologiques sont aussi pris en considération. Connaisant les types d'acides aminés, l'algorithme peut attribuer aux polypeptides détectés leur propres positions dans la séquence primaire du protéine. Cet protocol peut être appliqué aux systèmes de spin générés par méthodes différentes. Pour démontré le protocole d'attribution, quelques algorithmes sont crées pour déduire la chaîne principale et les systèmes de spin des chaînes latérales en utilisant RMN 3D. Un transfert de magnétisation à travers les liaisons peptides peut être observer en triple résonance RMN 3D. Pour automatiser l'attribution de la chaîne principale en utilisant les liasons à travers les corrélations, une algorithme générale est proposée. Cette algorithme, cherche et unit les pics croisés de tous spectres RMN disponibles. Les systèmes de spin individuel peut être extracté et liés pour créer les chaînes polypeptides basées sur les corrélations observés des interrésidus. Cette algorithme n'est pas en particulier limité a une seule expérience. L'algorithme est démontrée d'être applicable aux deux séries de spectres RMN: la série de cinqs expériences de HNCO, HNCA, HN(CO)CA, HCACO ^{15}N TOCSY-HMQC 3D et à l'expérience de la série CBCANH 3D. Pour l'attribution des chaînes latérales on à développer une approche automatisée en utilisant une algorithme de partition contrainte. Cette algorithme extraire les systèmes de spin des chaînes latérales des protéines en analysant les HCCH-COSY/TOCSY 3D spectres. Les systèmes de spin des acides aminée extracté montre les déplacement chimiques des spins nucléaire composés et les rapports connectifs entre ses spins. Une protéine de 90 ramifications de domaine-N du Troponin C(1-90) squellette de poulet était utilisé pour essayer l'algorithme avec les résultats simulés et expérimentaux. Les limitations de l'algorithme sont aussi discuter.

Acknowledgments

I am deeply grateful to my supervisor, Professor B. C. Sanctuary, for his continuous support and advice he has given me throughout my research and the preparation of this thesis.

I also wish to express my sincere gratitude to Dr. Jun Xu, a former postdoctoral staff in our laboratory. His enthusiasm, advice and assistance are essential to the completion of this work.

The Protein Engineering Network of Center of Excellence at University of Alberta and Prof. B. D. Sykes are acknowledged for providing me with an excellent scientific environment during my one-month visit in 1994. Special thanks go to Dr. Stéphan Gagné for many useful discussions and providing me with experimental data.

I would like to thank Dr. Feng Ni at Biotechnology Research Institute, NRC Canada for reading drafts of my papers and giving me useful opinions.

Thanks also go to my labmates, particularly, Mr. Wing Yiu Choy for providing me with the chemical shift database of protein amide nitrogen and Mr. Alvin Loh for translating the abstract of this thesis into French.

I am grateful to the financial support, the Max Binz fellowship, provided by McGill University.

Finally, I wish to extend my gratitude to my parents, Mrs. and Mr. S. J. Li and my wife Meng-Fang. This thesis could not have been done without their constant encouragement and endless love.

Contents

List of Figures

List of Tables

List of Abbreviations

AAPR	Amino Acid Pattern Recognition
ASPA	Aliphatic Side-chain Partitioning Algorithm
CAPP	Contour Approach Peak Picking
CAPRI	Computer Assisted Peak Resonance Identification
CBCANH	NMR experiment observing correlations between peptide C_β, C_α, NH and N
COSY	COrrelated SpectroscopY
CPA	Constrained Partitioning Algorithm
DBPA	Dipeptide Backbone Partitioning Algorithm
DQF-COSY	Double quantum filtered COSY
GUI	Graphical User Interface
HBA	Heuristic Backtracking Algorithm
HCACO	NMR experiment observing correlations between peptide H_α, C_α and CO
HCCH-COSY	COSY experiment using H-C-C-H magnetization transfer pathway
HCCH-TOCSY	TOCSY experiment using H-C-C-H magnetization transfer pathway
HMQC	Heteronuclear Multiple Quantum Coherence
HN(CO)CA	NMR experiment observing peptide NH, H and previous C_α
HNCA	NMR experiment observing peptide NH, H and C_α
HNCO	NMR experiment observing peptide NH, H and previous CO
HOHAHA	HOmonuclear HArtman HAhn spectroscopy
NCPA	Nitrogen Constrained Partitioning Algorithm
NMR	Nuclean Magnetic Resonance
NOE	Nuclear Overhauser Enhancement
NOESY	NOE SpectroscopY
PGA	Polypeptide Generating Algorithm
PMA	Polypeptide Mapping Algorithm
TOCSY	TOtal Correlated SpectroscopY
TSA	Tree Search Algorithm

Chapter 1

Introduction

This thesis presents automated software for protein resonance assignment from heteronuclear three-dimensional nuclear magnetic resonance (NMR) spectra. The assignment strategy is divided into three steps: (1) the extraction of amino acid spin systems, (2) the determination of amino acid types for the extracted spin systems, (3) the sequence-specific resonance assignments. A generic sequential assignment protocol was proposed under which algorithms were developed to automate the above three steps. The algorithms were implemented into computer programs and validated with simulated and real spectral data.

Using the proposed sequential assignment protocol, this thesis demonstrates that a complete automation of protein resonance assignment is possible, although in practice many aspects, such as the lack of sufficiently accurate automated peak picking software and the uncertainties of the amino acid type determination, have to be overcome before this ultimate goal can be achieved.

1.1 Motivation

Resonance assignment has direct implications on the structure determination of biomolecules from NMR data. In particular, the sequence-specific resonance assignment, as described in this thesis, is the essential analysis step needed before the structure determination and refinement can be conducted.

NMR based protein structure determination techniques have been widely used since early

1980s. The established procedure consists of several major steps [1–4]. First, the spin systems of all of the amino acid residues in the protein are identified, then a sequential assignment procedure attempts to map the extracted spin systems to the target protein's primary sequence. The results of the resonance assignments are then used to interpret the through-space NOE cross peaks, from which a number of distance constraints can be derived from analysis of the NOE data. Finally, these constraints are used to calculate the protein's 3D structure.

Since spectral overlap is proportional to the size of the molecule being studied, spectral analysis of larger molecules using 2D NMR becomes difficult if not impossible. With the recent development of 3D and 4D heteronuclear NMR [5, 6], the techniques of cloning and expressing $^{15}N/^{13}C$ labeled proteins, it is now possible to resolve the severe spectral overlap, resulting in complete structure determination for larger proteins. Many of these multidimensional heteronuclear NMR spectra take advantage of the scalar magnetization transfer through peptide bonds and thus a uniformly $^{15}N/^{13}C$ labeled protein is needed. Although 3D and 4D NMR simplify the overlapped spectra, the analysis of spectra remains difficult as more data is acquired and must be analyzed.

It is generally accepted that the resonance assignment of NMR spectra is tedious and time-consuming work, hence, there have been many attempts [7–19] to automate the resonance assignment part of the structure determination analysis.

Computer-assisted resonance assignment plays an important role for multidimensional NMR data analysis. Although 3D and 4D heteronuclear NMR greatly reduce the spectral overlap, it is at the expense of increased amount of data. Therefore, computer programs are needed and allow a more unambiguous spectral analysis, making it possible to automate the resonance assignment procedure. Similar results are difficult to attain using 2D NMR only.

Numerous approaches [20] have been applied to the automated assignment problem using multidimensional NMR. Vuister *et al.* [21] proposed an assignment strategy for homonuclear 3D NOE-HOHAHA spectrum, Kleywegt *et al.* [9] implemented and extended the strategy for homonuclear 3D [J,NOE]- and [NOE-J]-type NMR spectra of proteins. Oschkinat *et al.* [16] presented an automated strategy making use of homonuclear 3D TOCSY-TOCSY and TOCSY-NOESY. Among the attempts using heteronuclear 3D NMR, Zimmerman *et al.* [19] developed an approach for determining the sequential order of amino acid spin systems using 3D HCC(CO)NH-

TOCSY and constraint propagation methods. Bernstein *et al.* [18] applied the technique of combinatorial minimization to achieve the sequence-specific assignment of proteins using 3D ^{15}N-HMQC-TOCSY and ^{15}N-HMQC-NOESY. Two complete protein automated resonance assignment protocols were proposed, one was done by Meadows *et al.* [17] the other by Morelle *et al.* [22] The first makes use of 4D HNCAHA, HN(CO)CAHA, HC(CO)NH-TOCSY, 3D HNCA and HN(CO)CA while the second protocol uses a set of 2D triple resonance NMR spectra to assign the protein's backbone resonances. Some of these computer programs, for example, Zimmerman's and Bernstein's, automate sequential assignment only. Consequently, the amino acid spin systems must be created and identified using other approaches. Meadow's and Morelle's protocols are able to extract amino acid spin systems but an automated amino acid type recognition routine is lacking. In addition, many of these programs put emphasis on particular kinds of NMR experiments.

1.2 Issues

A self-contained automated assignment strategy should consist of three steps. (1) Extracting the spin coupling segments (amino acids) that make up the biomolecule. (2) Mapping of the spin coupling segments to amino acid residues. (3) Searching for a most probable spin system sequence which matches the protein's primary sequence. These steps can be treated by a series of algorithms: Constrained Partitioning (CPA) [23, 24], fuzzy pattern recognition [25] and tree searching [25, 26], respectively. CPA can automatically extract and identify spin coupling segments from a combination of the 2D COSY and TOCSY spectra where the latter is used as partitioning constraints. The fuzzy pattern recognition algorithm determines the amino acid types for those observed spin coupling segments. Once the amino acid types are determined, each residue's position within the protein sequence can be obtained from the tree searching algorithm.

1.3 Scope of the Thesis

In this thesis, extensions [27] are made to the CPA algorithm so that the aliphatic side chain spin systems can be deduced from heteronuclear 3D NMR data. A generic sequential assignment protocol is proposed. Three algorithms, a protein backbone extraction algorithm [28], an extended amino acid pattern recognition algorithm [29] and a sequential mapping algorithm [28], are applied to the sequential assignment protocol. The methods developed in this work are applicable to a wide variety of heteronuclear 3D NMR experiments. The applications are not restricted to certain special designed NMR experiments. This approach provides a basis for further development of a fully generic, i.e., completely independent on the types of input NMR experiments, automated assignment software.

1.4 Organization of the Thesis

The thesis is organized as follows:

Chapter 2 highlights the NMR based protein structure determination procedures, introduces the NMR experiments used in the thesis and reviews the previous work at the automation of 2D NMR spectrum assignment. It attempts to give an overview of the earlier work this thesis is based upon as well as emphasis of the direct relationship between resonance assignment and the structure determination. The subject of the research, protein resonance assignment, is defined in a formal manner in the same chapter.

Chapter 3 describes two approaches for the determination of protein backbone resonances. The first one makes use of the triple resonance heteronuclear 3D NMR experiments. This approach is able to extract individual spin systems as well as establish the sequential connectivities. Another approach is a direct extension of the two-dimensional CPA algorithm, making CPA possible to process three-dimensional NMR experiments such as ^{15}N TOCSY-HMQC.

Chapter 4 presents an algorithm for the assignment of protein aliphatic side chain resonances. The deduced spin systems can be merged with the backbone spin systems to provide possible candidates for the amino acid type determination.

Chapter 5 deals with the amino acid type determination and the sequence-specific assignment. An algorithm is introduced to merge the previously determined backbone and side chain spin systems. A mathematical graph-theory-based spin pattern recognition algorithm is described. Finally, a sequential mapping algorithm places the recognized spin systems at positions within the primary sequence. The interresidue connectivities can be created by through-bond (from triple resonance 3D NMR) or through-space (from 2D NOESY or 3D ^{15}N NOESY-HMQC) correlations. A sequential assignment protocol is discussed to summarize the above algorithms.

Chapter 6 concludes the thesis by highlighting the significant contributions of the current work, discussing various possibilities of applying the research to real world cases, and pointing out the directions for future investigation.

Chapter 2

Related Issues and Previous Work

2.1 Introduction to 2D NMR spectroscopy

Essentially all contemporary NMR work on biopolymers is done with two-dimensional (2D), or three-dimensional (3D) NMR. In this section, the most commonly seen 2D NMR experiments that are applied to protein resonance assignments are introduced and their information content is described. Discussion emphasizes the experiments that are used in later chapters.

2.1.1 COSY

The basic 2D NMR experiment based on through-bond scalar coupling is COSY(COrrelated SpectroscopY) [30, 31]. The COSY experiment has the simplest pulse sequence of all 2D NMR experiments. The pulse sequence is shown in Figure 2.1. In this experiment the spins undergo precession about one another, in addition to the usual precession about the applied magnetic field. During the mixing period, i.e., the second pulse, of this experiment, J-coupled spins exchange coherence and communicate the information about their precession frequencies. The result, for the COSY experiment, is that a cross peak between two spins, i, and j, will occur at position (δ_i,δ_j) and (δ_j,δ_i) in the spectrum if spin i and j are directly coupled to one another.

The cross peaks in COSY spectra are antiphase in character, that is, half of the multiplet is "up" and the other half is "down" as shown in a simulated 2D COSY spectrum in Figure 2.2.

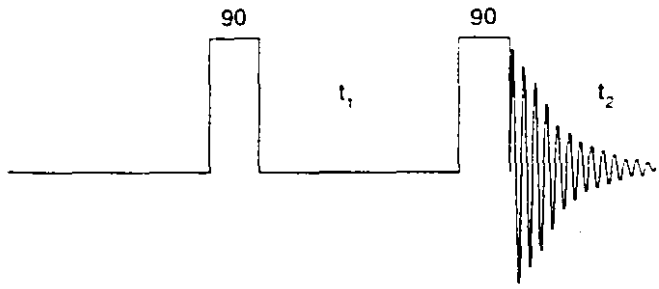


Figure 2.1: The pulse sequence of the 2D correlated spectroscopy (COSY). Two 90-degree pulses are separated by the mixing period t_1.

This feature increases the difficulties of doing automated peak picking. To determine the center of

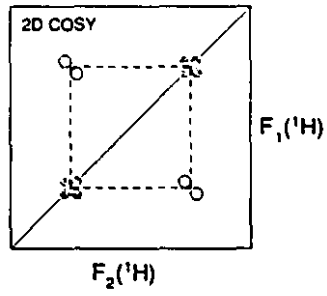


Figure 2.2: Simulated 2D COSY contour plot of two coupling spins. Open circles with solid and dashed lines are cross peaks with positive and negative intensity, respectively. The dispersive diagonal peaks are represented by filled circles.

each cross peak, two local maxima and minima have to be found which is difficult in a severely overlapped spectrum.

In COSY experiment, peaks also occur on the diagonal arising from coherence that remained on the same spin after two pulses of the experiment. In normal COSY the diagonal signals are out of phase (dispersive) relative to the cross peaks and are broader than the absorption signal. This phenomenon interferes with the detection of cross peaks near the diagonal. The cross and diagonal peaks for a weakly coupled systems with two $I = 1/2$ spins are derived in Appendix A. A useful revision to overcome the dispersive diagonal peaks is to apply a double quantum filter, which is discussed in the next section.

2.1.2 *DQF-COSY*

DQF-COSY was introduced by Rance *et al.* [32] All coherence from spins that do not have coupling partners, i.e., singlet in the spectrum and all coherence that remained on the same spin during the evolution period t_1 are added to zero through a phase cycle. The result is the disappearance of the obscure diagonal peaks. DQF-COSY is now the usual 2D COSY experiment for biomolecular applications although sometimes the prefix "DQF" is omitted. Appendix A gives an example of the phase cycling scheme used in DQF-COSY.

2.1.3 *TOCSY*

A more recent 2D NMR experiment for identifying extended couplings is TOCSY(TOtal Correlation SpectroscopY) [33] which is also known as HOHAHA(HOmonuclear HArtman HAhn spectroscopy) [34, 35]. An isotropic mixing is added after the evolution time t_1 by applying a sequence of pulse which effectively averages out chemical shifts. This can be thought of as a sequence of 180° pulses, each of which refocuses the chemical shifts. In effect, all coupled spins will have the same precession frequency, so they will be strongly coupled ($\Delta\delta \ll J$) and their transitions will be thoroughly mixed. In the collected FID, all the coherences return to their original chemical shifts but become labeled with the precession frequencies of all the other spins in the same spin coupling system. For example, four spins $\{i, j, k, l\}$ are within a coupled spin system. In the TOCSY spectrum, cross peaks occur at position (δ_i, δ_j), (δ_i, δ_k), (δ_i, δ_l), (δ_j, δ_i), (δ_j, δ_k), (δ_j, δ_l), . . . , etc. Figure 2.3 shows a simulated TOCSY spectrum for an alanine.

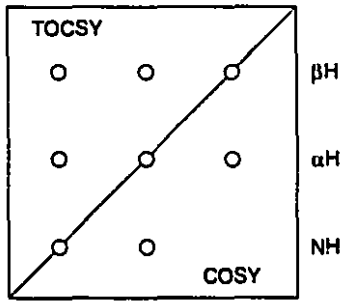


Figure 2.3: Simulated COSY and TOCSY spectra of an alanine spin system with three spins, NH, αH and βH.

In TOCSY the mixing between spins does not occur instantaneously, and the number of spins intermediate between the initial spin and the detected spin can be adjusted by modifying the isotropic mixing period of the experiment. In other words, it is possible to control the number of correlated spins within a spin system. A short mixing time TOCSY may only record cross peaks arising from adjacent protons but not from distant protons. An additional advantage of the TOCSY experiment is that the cross peak multiplets are all in-phase rather than antiphase, so there is no loss of signal intensity for broad lines due to cancelation of overlapping antiphase component.

TOCSY experiment provides redundant information to help resolving chemical shift degeneracy problem (more than one protons having the same frequency). TOCSY data are used as constraints to confirm cross peak merge in the computer algorithm called Constrained Partitioning Algorithm which is described in section 2.6.

2.1.4 NOESY

In COSY and TOCSY experiments, magnetization transfer between spins is mediated by the through-bond scalar couplings. The NOESY(Nuclear Overhauser Enhancement SpectroscopY) [36–38], on the other hand, takes advantage of the through-space dipolar couplings. To describe the NOESY experiment, consider a pair of spin I and S, which are in close spatial proximity so as to have the dipolar interaction. Figure 2.4 shows the pulse sequence of NOESY. The first 90°

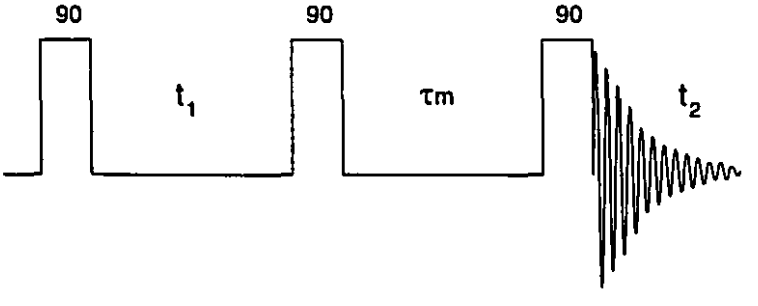


Figure 2.4: The NOESY pulse sequence. The maximum distance to give an observable cross peak depends on the value of τ_m.

pulse brings the magnetization of spin I down to the $x-y$ plane. After the evolving period t_1, the second 90° pulse flips the magnetization of I back to the z axis. During the delay τ_m, cross

relaxation between spin I and S occurs and some of the spin I magnetization is transferred to S. In the detection period t_2, magnetization of spin S is detected but the FID signal (at the frequency of spin S) is amplitude-modulated at the frequency of spin I. The result is the cross peak (δ_I, δ_S) in the NOESY spectrum. By adjusting the mixing time τ_m, the maximum distance between spins for which cross peaks will be seen can be adjusted.

To interpret the intensity of a NOESY cross peak, one must know that NOE is a consequence of modulation of the dipolar coupling between different nuclear spins by the Brownian motion of the molecules in solution. The NOE intensity can be related to the distance r between the pre-irradiated (in the above example, spin I) and the observed (spin S) spins by an equation of the general form [2].

$$\text{NOE} \propto \frac{1}{\langle r_6 \rangle} f(\tau_c) \tag{2.1}$$

The second term is a function of the correlation time τ_c which accounts for the influence of the motional averaging process on the observed NOE. In protein structure determination using NMR spectroscopy, the NOESY experiments provide connectivities, such as $d_{\alpha N}(i, i+1)$, $d_{NN}(i, i+1)$, between sequentially adjacent amino acid residues. Those connectivities are the building blocks for protein sequential assignment.

2.1.5 Heteronuclear 2D NMR

In all of the above NMR experiments, the magnetization transfer is going from proton to proton, resulting in ^{1}H-^{1}H spectra. When two different kinds of nuclear spins are considered, the magnetization transfer can be from ^{1}H to X or from X to ^{1}H, where X stands for ^{13}C, ^{15}N or ^{31}P in biomolecules. Since the chemical shifts of heteronuclei, ^{13}C and ^{15}N, are usually well dispersed while the protons tend to be closely overlapped, it is generally desired to place the crowded ^{1}H spectrum in F_2 dimension where it can be finely digitized and leave the better dispersed ^{13}C or ^{15}N spectrum in F_1 dimension. Another concern of heteronuclear NMR is that the natural abundance of ^{13}C and ^{15}N is low (1.1% of ^{13}C, 0.37% for ^{15}N in comparison with 99.9% of ^{1}H). For example, only one percent of the protons will be attached to a ^{13}C nucleus, the rest will be attached to inactive ^{12}C. Most recently presented heteronuclear 2D and 3D NMR experiments

require uniformly isotope labeled ^{13}C or ^{15}N to overcome the sensitivity problem. The ^{1}H–^{13}C and ^{1}H–^{15}N couplings are large (125-160 Hz for $^{1}J_{CH}$ and ~ 92 Hz for $^{1}J_{NH}$) [5], and the efficiency of magnetization transfer is high even when spectral lines are broad for high molecular weight molecules.

A commonly used heteronuclear 2D NMR is HMQC(Heteronuclear Multiple Quantum Coherence spectroscopy) [39]. The ^{13}C or ^{15}N spins are recorded in the F_1 dimension while the protons scalar coupled to the ^{13}C or ^{15}N are recorded in the F_2 dimension.

2.2 Introduction to Heteronuclear 3D NMR Spectroscopy

In three-dimensional NMR spectra, correlations of three different frequencies are generated through two different mixing times of an experiment. The mixing mechanisms are the same as in 2D NMR, that is, COSY, TOCSY, NOESY types mixing can also be used in 3D NMR. 3D NMR experiments are essentially combinations of two 2D experiments (Figure 2.5). 3D NMR

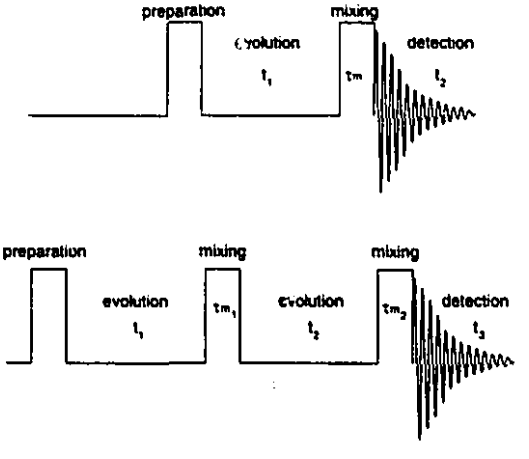


Figure 2.5: 2D and 3D general experiments. In the 2D experiment, the value of t_1 is incremented to obtain the second time-domain information. In the 3D experiment, the value of t_1 and t_2 are incremented to obtain the second and third dimension.

experiments can be classified according to the observed nuclei. Homonuclear 3D NMR observes proton frequencies. Heteronuclear 3D NMR is further classified to double resonance experiments (^{1}H and ^{13}C, ^{1}H and ^{15}N) and triple resonance experiments (^{1}H, ^{13}C and ^{15}N). Since our study

focuses on development of automated assignment tools for heteronuclear 3D NMR, the following description is limited to heteronuclear 3D NMR.

Heteronuclear 3D NMR experiments separate the individual proton resonances according to the chemical shifts of the directly bonded heteronuclei and simultaneously produce important additional information about the chemical shifts of the heteronuclei. Thus problems of proton resonance overlapping that occur for large proteins can be overcome by separating the crowded ^{1}H–^{1}H 2D NMR plane into many planes of a 3D NMR spectrum as shown in Figure 2.6.

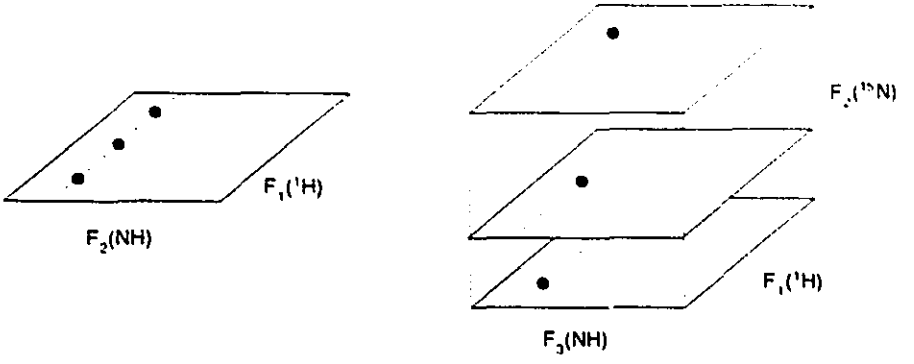


Figure 2.6: Schematic illustration of the relationship between ^{15}N edited 2D and 3D spectra. The closed circles represent three NH-αH cross peaks, which can be separated, in the corresponding 3D spectrum, into three planes depending on the different chemical shifts of the amide nitrogen nuclei.

The spectral line width of NMR spectra is approximately proportional to the inverse of the molecular tumbling rate and therefore increases approximately linearly with the size of the protein [2,5]. For large proteins (> 10 kD) the ^{1}H–^{1}H J couplings are smaller than the spectral line width, making the 2D COSY spectrum ineffective. As mentioned above, the heteronuclear one-bond couplings are much larger than $^3J_{HH}$. As a result, the line broadening problems can be overcome by using the heteronuclear one-bond couplings instead of $^3J_{HH}$ to achieve efficient magnetization transfer of NMR experiments.

The sensitivity, i.e., the signal-to-noise ratio achievable in a unit time interval, of 3D NMR is generally lower than the 2D NMR counterpart [40]. To overcome the problem of loss of sensitivity, more efficient magnetization transfer steps are required because a greater percentage of the nuclear spin magnetization is transferred from one nucleus to another, resulting in stronger signal intensity. Heteronuclear 3D NMR takes advantage of the more efficient magnetization transfer (as much as

50% to 90%) [5] between ^{13}C or ^{15}N and ^{1}H so that a lower concentration protein sample can still produce high sensitivity spectra.

We now briefly introduce some heteronuclear 3D NMR experiments which are used in the following chapters.

2.2.1 *Triple resonance heteronuclear 3D NMR*

For moderate sized proteins ($\sim$ 20 kD), most of the one-bond J couplings are significantly larger than the spectral line width [5]. This means that the magnetization can be transferred efficiently from one nucleus to its directly bonded neighbor. A number of triple resonance NMR experiments have been designed, correlating mainly the backbone resonances. In chapter 3, a computer algorithm is presented to achieve the protein backbone assignment using heteronuclear 3D HNCO, HNCA, HN(CO)CA, HCACO, CBCANH experiments. Schematic representations of Figure 2.7 and listings in Table 2.1 show the nuclei that are correlated in the above 3D experiments. Those experiments are named according to the nuclei they correlated. For example, the HNCO experiment correlates NH(i), N(i) and CO($i-1$).

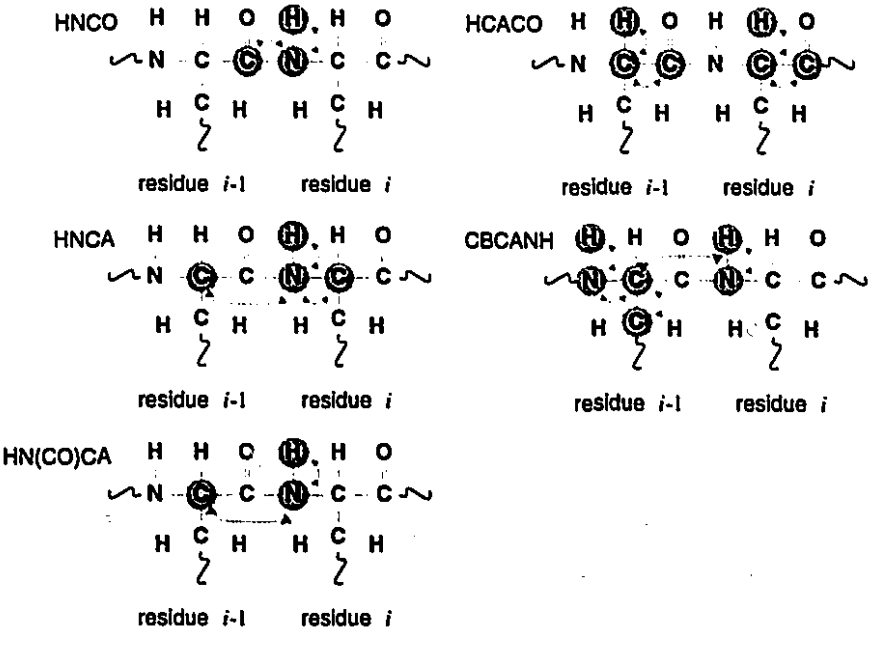


Figure 2.7: Schematic illustration of the correlations shown in Table 2.1.

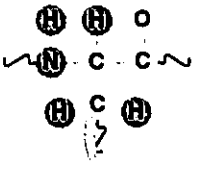
Table 2.1: Correlations observed in the five triple resonance NMR experiments.

HNCO			HNCA			HN(CO)CA		
$CO(i-1)$	$N(i)$	$NH(i)$	$C_\alpha(i-1)$	$N(i)$	$NH(i)$	$C_\alpha(i-1)$	$N(i)$	$NH(i)$
			$C_\alpha(i)$	$N(i)$	$NH(i)$			
HCACO			**CBCANH**					
$C_\alpha(i)$	$CO(i)$	$H\alpha(i)$	$C_\alpha(i-1)$	$N(i)$	$NH(i)$			
$C_\alpha(i-1)$	$CO(i-1)$	$H\alpha(i-1)$	$C_\beta(i-1)$	$N(i)$	$NH(i)$			
			$C_\alpha(i-1)$	$N(i-1)$	$NH(i-1)$			
			$C_\beta(i-1)$	$N(i-1)$	$NH(i-1)$			

2.2.2 *Double resonance heteronuclear 3D NMR*

3D ^{1}H-^{15}N TOCSY-HMQC experiment

This experiment provides intraresidue correlations between aliphatic and NH protons, information which is important for identifying amino acid spin systems of proteins. 3D ^{15}N TOCSY-HMQC is a combination of 2D TOCSY and HMQC experiments. In the first step, magnetization originating on aliphatic protons is transferred to intraresidue NH protons via TOCSY type isotropic mixing pulse sequence. At the end of the t_2 evolution period, ^{1}H magnetization is amplitude modulated by the chemical shift of the directly bonded intraresidue ^{15}N nucleus. The NH protons are finally detected during the t_3 detecting period. For each of the amino acids of a protein, the F_1 dimension records the chemical shifts of the aliphatic αH, βH, ... , etc., the ^{15}N is recorded in the F_2 dimension while F_3 records the NH chemical shifts. Figure 2.8 shows the correlated nuclei by the 3D ^{15}N TOCSY-HMQC experiment.



^{15}N TOCSY-HMQC		
Hα(i)	N(i)	NH(i)
Hβ1(i)	N(i)	NH(i)
Hβ2(i)	N(i)	NH(i)
Hγ1(i)	N(i)	NH(i)
etc.	etc.	etc.

Figure 2.8: Correlations observed in the 3D ^{15}N TOCSY-HMQC experiments.

2.2.3 *Choosing between Three- and Two-dimensional NMR*

As we have seen in earlier discussion, 3D NMR experiments overcome the peak overlap problem by introducing the third dimension and separating overlapped peaks into a number of 2D planes. In the case of heteronuclear 3D NMR, use of larger one-bond couplings reduces the risk of peak overlapping arising from line broadening effect.

Three-dimensional NMR spectra provide some other advantages over 2D spectra as far as the design of an automated software for resonance assignment is concerned. The first computational advantage of using 3D NMR is that a single cross peak in a 3D NMR spectrum represents the magnetic interactions between three nuclei and provides the relationships between three chemical shifts . For example, a cross peak (4.29, 119.50, 8.35) in 3D ^{15}N TOCSY-HMQC spectrum represents the adjacency relationship between the chemical shifts of 119.50 and 8.35. In addition, the chemical shifts of 8.35 and 4.29 must be in the same spin coupling system. To obtain the same information from a 2D spectrum, one has to find a pair of 2D cross peak, in the above example, a COSY peak (4.29, 8.35) and a HMQC peak (119.50, 8.35), having one chemical shift, 8.35, in common. Finding such pairs is not as straightforward as it is in the case of using 3D NMR. Degenerate chemical shifts, e.g., (3.47, 8.35), may cause ambiguity when determining which chemical shifts, 3.47 or 4.29, is in the same spin system with the resonance of 8.35 ppm.

The second advantage of using 3D NMR is that there are two ways of confirming the merging of two 3D NMR cross peaks while there is only one way when merging two 2D peaks. For example, to merge 3D peak (δ_i,δ_j,δ_k) and (δ_j,δ_k,δ_l), one can do so by verifying that the second coordinate of peak 1 and the first coordinate of peak 2 are the same chemical shift. Additionally, peak 1's third coordinate must also be consistent with peak 2's second coordinate.

3D NMR experiments tend to separate peaks away from each other, making peak shapes more predictable. Peaks with better shapes are more suitable to be picked by automated peak picking softwares, since noise peaks can be more readily separated from real signals.

There are, however, several disadvantages of using 3D NMR. The time required to acquire a spectrum increases with the increase of dimensionality. For example, a typical 3D HNCO experiment may take 3 days to acquire [5]. The sensitivity, i.e., the S/N ratios, drops by $\sqrt{2}$ with

increasing of one dimension [40].

Despite the loss of sensitivity and increase of acquisition time, in many cases, especially with large proteins, 3D NMR experiments are the only choice to conduct successful resonance assignments. Moreover, computerized analysis become more desirable in 3D and 4D NMR because of the large amount of data present and the difficulty of visualizing 3D and 4D data spaces.

2.3 Protein structure determination from NMR data

Remarkable progress has been made in applying NMR spectroscopy to the study of protein [3,41] in the past 15 years. NMR method provides complementary information about protein structures to that from X-ray crystallography. For example, in NMR method, the solution conditions can be varied over some ranges, the internal dynamics and chemical exchange phenomena can be characterized and the effects of temperatures can be studied.

The NMR method can also be applied to other biomolecules, such as nucleic acids and polysaccharides as well as small molecules.

In this section a short survey is devoted to the NMR methodology for protein structure determination. In the next section, the resonance assignment, our research subject, is described in detail.

2.3.1 Basic approach

Figure 2.9 depicts the steps for determining solution structures from NMR data. Multidimensional NMR data are acquired as a series of 1D spectra. The time delays required for frequency labeling in the evolution period result in loss of signal intensity, i.e., low sensitivity. In addition to applying certain data manipulation techniques, higher concentration of protein sample generally produces higher sensitivity. The typical concentration of protein sample required for 2D COSY, TOCSY or NOESY experiment is about 2 mM. The required volume of sample solution is about 400 μl [42]. Higher concentration is desired provided that the protein is soluble and does not aggregate, since this not only provides higher sensitivity but also permits shorter experiment

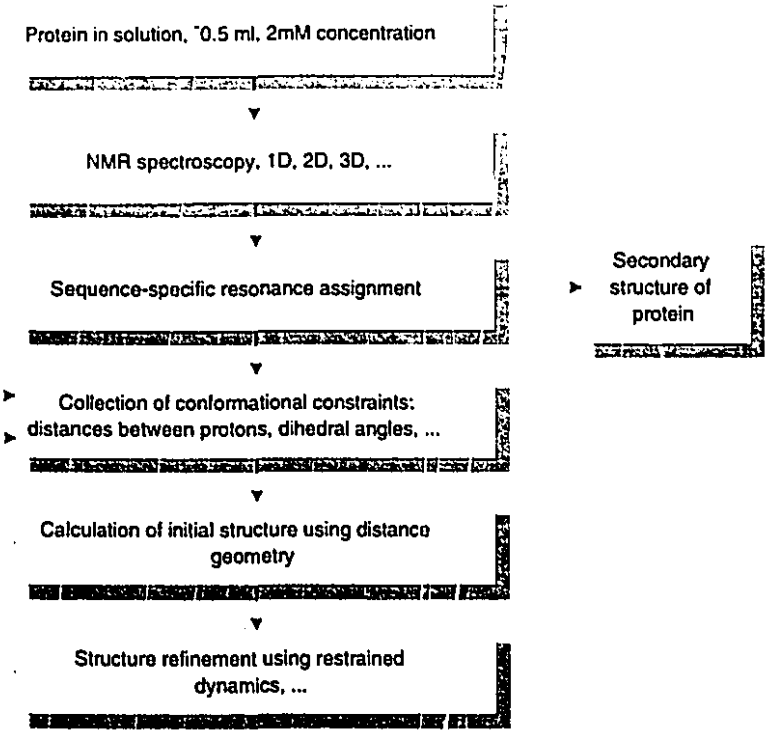


Figure 2.9: The flowchart of the protein structure determination from NMR data.

time.

Once the NMR experiments are acquired, individual peaks in the spectra have to be assigned to sequence-specific locations in the chemical structure of protein before the distance information in the NOESY spectrum can be fully interpreted. Sequence-specific NMR resonance assignment plays a pivotal role in the structure determination process. The objective of our study is to automate the resonance assignment procedures using computers. The detailed manual assignment strategies is described in the next section.

Fully analysis of the NOESY spectrum, the "NOE assignment", provides many distance constraints between the hydrogen atoms of a protein. As described in equation 2.1 , the inter-proton distance can be calculated from the intensity of the NOE cross peaks provided a fixed distance can be found to calibrate equation 2.1. Generally speaking, an NOE peak with strong intensity may indicate that two protons are within 2.5 Å of each other while a weak NOE peak corresponds to an upper limit of 5 Å.

Many other geometrical constraints can be inferred using various methods. One of the constraints available from NMR data is dihedral angles. Two dihedral angles are associated with each peptide bond. Angle ϕ is the torsion angle between bond N – NH and $C_\alpha - \alpha H$ while angle ψ is another torsion angle between bond $C_\alpha - \alpha H$ and C – O.(Figure 2.10)

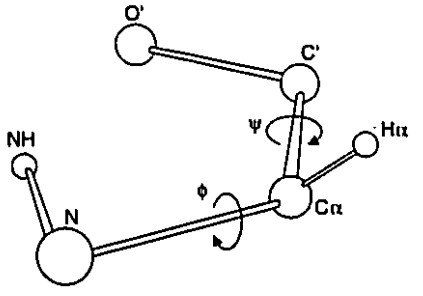


Figure 2.10: The torsion angles of an amino acid residue.

The dihedral angle ϕ can be calculated from the vicinal spin-spin couplings $^3J_{\alpha H-NH}$ using Karplus equation [43,44].

$$^3J_{\alpha H-NH} = 6.4\cos^2\theta - 1.4\cos\theta + 1.9 \tag{2.2}$$

where $\theta = |\phi - 60\,^\circ|$ and ^3J is given in Hertz. With the use of the above equation, measurement of $^3J_{\alpha H-NH}$ present a complementary information to NOE distance constraints for calculating the initial structure of a protein.

The next step is to determine an initial protein structure which is consistent with the thousands of NOE constraints and, frequently, with some other conformational constraints. Distance geometry is the most commonly used mathematical procedure by which distance constraints are converted into three dimensional structures [45]. Distance geometry procedure is essentially a projection from a high-dimensional space (in which thousands of distance relations can be accommodated) into ordinary three-dimensional space. The initial structures calculated from distance geometry almost always violate many of the experimental constraints. Subsequently structure refinement is required to obtain a high resolution protein structure.

2.3.2 *The important role of sequence-specific resonance assignment*

As above described, NMR spectra contain information to determine biomolecular structures in solution. However, none of the embedded information can be used without having the resonances of the biomolecules assigned. In other words, it must first be determined which resonance come from which nuclear spins. This is a common problem or process in all spectroscopies. The process of associating specific spins in the molecule with specific resonances is called *sequence-specific resonance assignment*.

Sequence-specific resonance assignment is essential in three areas of the biomolecular NMR applications: (1) biomolecular structural analysis (2) intermolecular interaction with biopolymers (3) studies of molecular dynamics. The importance of resonance assignment in those three areas is discussed below.

As a first discussion consider the determination of protein structures from NMR data. The structural information mainly comes from NMR cross peaks. An NOE peak between two hydrogen atoms (or groups of hydrogen atoms) is observed if these hydrogens are located at a shorter distance than approximately 5.0 Å from each other. Without sequence-specific resonance assignment it is impossible to determine to which the two hydrogen atoms a specific distance constraint refers. On the other hand, 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 formed.

The second application where resonance assignment is pivotal is the studies of intermolecular interaction. For example, in the study of the protein-DNA binding interaction, the binding sites are the first thing we want to know. The intermolecular NOE peaks can manifest short distances between nuclear spins located in different interacting molecules. Without sequence-specific assignment, such NOE data merely indicate that the intermolecular interaction has occurred. When combined with assigned resonances, the NOE data identify the binding sites of the intermolecular contacts.

The study of protein dynamics has made significant progress during the past several years. These studies rely on the observation of certain spectral properties in distinct NMR lines (peaks)

that can be correlated with intramolecular motions. Once the NMR lines responsible for the study region (such as a methyl group) have been assigned, it is then possible to investigate the desired spectral properties in the corresponding spectra.

2.3.3 *The difference between resonance assignment and NOE assignment*

Before describing the strategy of protein resonance assignment , the sometimes confusing terms "NOE assignment" is clarified first.

The sequence-specific assignment of protein resonances is a process of associating specific nuclear spins in the protein with specific resonances, i.e., chemical shifts. The process may or may not involve NOE data. In traditional resonance assignment strategy using homonuclear 2D NMR, the interresidue connectivities are established from NOESY data. Recently, heteronuclear 3D NMR provides interresidue connectivities through a series of triple resonance experiments, there is hence no need of using NOE data.

NOE assignment is the analysis of the NOESY peak set to locate as many proton-proton distance constraints as possible. The sequence-specific resonance assignment usually assign only a few backbone NOE correlations, such as $d_{\alpha N}(i, i+1)$, $d_{NN}(i, i+1)$, $d_{\alpha N}(i, i+3)$, . . . , etc. The backbone NOE correlations provide the required sequential connectivities for placing amino acid residues to their corresponding locations along the primary sequence. The majority of the NOE peaks, however, remain unassigned in the resonance assignment stage. The NOE assignment process is responsible for determining all the short- and long-range interresidue NOE correlations.

The chemical shift degeneracy sometimes makes the complete NOE assignment difficult in the protein side chain region. For example, consider 10 protons resonating at 1.88 ppm. Now there is an NOE peak (1.88, 2.43) to be assigned. It is difficult to determine which one of the 10 protons gives the above NOE peak.

2.4 Introduction to manual assignment strategy

Resonance assignment has been a major hurdle for protein structural analysis from NMR data. Significant progress has been made through the introduction of 2D, 3D even 4D NMR experiments. Combined with systematic approaches for spectral analysis, although it is still tedious, time-consuming work, the resonance assignment of protein spectra is no longer an unmanageable task.

Except for resonance assignment, most other parts of the protein structure determination rely heavily on computers. Therefore it is natural to ask ourself the question: is it possible to develop a fully automated resonance assignment software? The ultimate goal of this thesis is to accomplish this by developing as fully an automated assignment tool as possible. Before discussing aspects regarding automated resonance assignment, we will describe the traditional but efficient manual assignment strategy.

2.4.1 Manual assignment from homonuclear 2D NMR spectra

After the 2D COSY and NOESY experiments were first applied to proteins, it was realized that the intra- and interresidue covalent linkage can be readily achieved provided that the NMR data are of high quality. The idea for systematic assignment of proton resonances in protein was first proposed by Wüthrich *et al.* [4] in 1982. Another approach, proposed by Englander and Wand [46], uses the same COSY and NOESY information but in different order. This approach is referred to as the Main-Chain-Directed (MCD) assignment.

Wüthrich's assignment strategy includes the following steps:

1. The spin systems of the protons in individual amino acid residues are identified using as many as possible through-bond ^{1}H-^{1}H connectivities, which are mainly provided by 2D COSY experiments.

2. Sequentially neighboring amino acid ^{1}H spin systems are identified from observation of the sequential NOE connectivities $d_{\alpha N}(i, i+1)$, $d_{NN}(i, i+1)$, or possibly $d_{\beta N}(i, i+1)$.

3. Combining the information in the above, it is possible to establish chains of amino acid spin

systems corresponding to peptide 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.

2.4.2 *Identification of amino acid proton-proton spin systems*

The identification of proton-proton the spin systems of individual amino acid residues is usually achieved by analysis of ^{1}H COSY spectrum in D_2O solution after replacement of all labile protons with deuterium. One tries to collect all J-coupled resonances arising from the same amino acid residue. The 20 common amino acid residues produce 10 different COSY connectivity patterns for the aliphatic protons and four pattern for the aromatic rings. Figure 2.11 shows all of the 14 patterns on COSY spectrum. In principle, it is impossible to distinguish a spin system with one αH and two βH's to be Ser, Cys, Asp, Asn, Phe, Tyr, His or Trp. All have the same connectivity pattern on a COSY spectrum (Figure 2.11). However, different amino acids have different chemical shift ranges, making it possible to reduce the candidate number by inspecting the chemical shifts of the deduced spin systems. For example, serines have relatively downfield chemical shifts for their two βH's ($\sim$ 3.8 ppm), making serine an easily identified spin system.

In crowded COSY spectrum, spectral overlap and chemical shift degeneracy make the identification of unique patterns difficult. A RELAYED-COSY or TOCSY spectrum, which provides redundant information about the amino acid patterns, often allows the ambiguous assignments to be solved. An example is given in section 2.6.

2.4.3 *Sequential assignment via proton-proton NOE*

Using 2D COSY and possibly TOCSY spectra the ^{1}H amino acid spin systems can be identified. As show in Figure 2.11, certain amino acids have unique connectivity patterns, such as Val, Ile, Ala, Gly, Leu and Thr. It is possible to assign the deduced spin systems to those unique amino acids directly. However, for AMX-type spin systems (one αH and two βH's), unique assignments

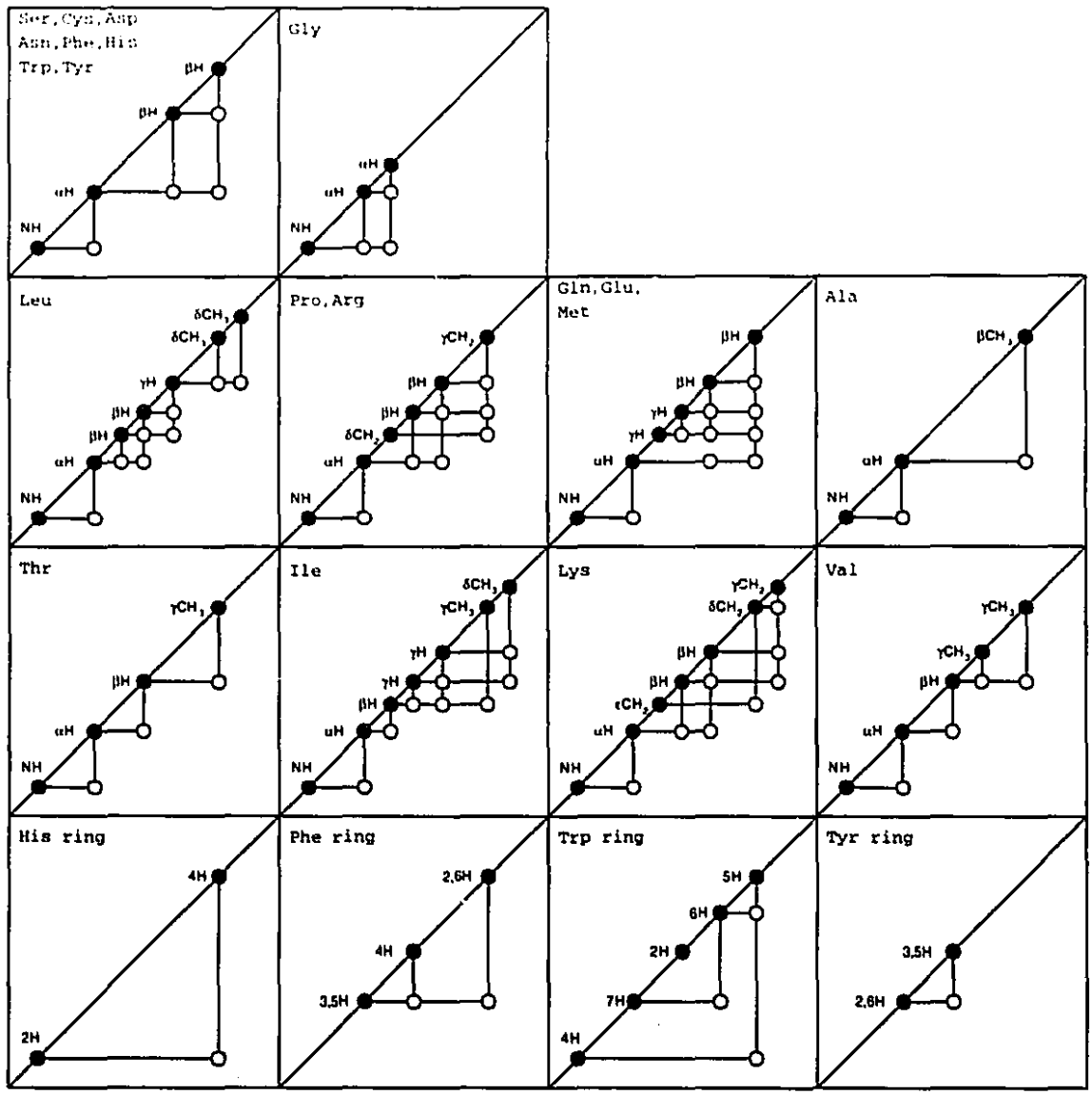


Figure 2.11: The simulated COSY coupling patterns of the 20 common amino acids.

are generally unachievable. Wüthrich [3] proposed four different methods to classify the amino acid types, they are summarized in Table 2.2.

Before the NOE information can be used to create sequential connectivities, the deduced spin systems must be classified according to one the the above amino acid types. This task is achieved by inspecting the chemical shifts and the spin coupling patterns. In chapter 5 an automated approach is described where the determination of amino acid types can be accomplished by

Table 2.2: The four different methods to classify the 20 amino acids.

category	number of amino acid types in this category	Descriptions
1	8	Gly, Ala, Val, Leu, Ile, Thr, (all $\alpha CH - \beta CH_2$), (all others)
2	13	Gly, Ala, Val, Leu, Ile, Thr, Phe, Tyr, Trp, His Ser, (Cys,Asp,Asn), (all others)
3	15	Gly, Ala, Val, Leu, Ile, Thr, Phe, Tyr, Trp, His Ser, (Cys,Asp,Asn), Pro, (Lys,Arg), (Met,Glu,Gln)
4	18	Gly, Ala, Val, Leu, Ile, Thr, Phe, Tyr, Trp, His Ser, Cys, (Asp,Asn), Pro, Lys, Arg, Met, (Glu,Gln)

computers.

Wüthrich and his coworkers [3] also found that there is a very high possibility that at least one proton among the NH, αH, or βH from one residue will be near (less than 3.5 Å, i.e., within the allowed NOE range) to the NH of the following residue. Thus by searching appropriate $d_{\alpha N}$, d_{NN} or $d_{\beta N}(i, i+1)$ NOE correlations in the NOESY spectrum, it should be straightforward to step from one residue to the next along the primary sequence of the protein. Once the connections between spin systems are established, the connected spin systems must be matched with the known protein primary sequence. To illustrate the final sequential matching, consider the following example. From the NOESY spectrum, an 8-residue long polypeptide chain was found. The corresponding amino acid types of the 8 residues were determined previously as Ala-Val-Leu-□-Thr-Δ-Gly-□ where □ represents all the $\alpha CH - \beta CH_2$ spin systems in the category 1 of Table 2.2, and Δ represents the amino acid type including Pro, Lys, Arg, Met, Glu and Gln. To find an unambiguous matching of the 8-residue chain on the amino acid sequence, one has to inspect the protein primary sequence to make sure there is only one segment fulfilling the Ala-Val-Leu-□-Thr-Δ-Gly-□ pattern. If such a unique matching is found, the sequence-specific assignment for the 8-residue chain is obtained. If not, the length of the connected polypeptide may need to be increased in order to obtain a unique matching.

2.4.4 *Manual assignment from heteronuclear 3D NMR*

Heteronuclear 3D NMR experiments make use of larger one-bond couplings, $^1J_{H-X}$, where X=^{13}C or ^{15}N, to overcome the spectral line broadening problem. As described in section 2.2.1, several triple resonance NMR experiments have been designed to conduct the sequence-specific resonance assignments without using crowded NOESY spectra.

The interresidue 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, HNCO, HCA(CO)N, also provide interresidue correlations where one-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 2.12 is an example where assignment is carried out by overlapping two previously assigned frequencies in each subsequent spectrum. In the first two steps (HNCA and TOCSY-HMQC), the NH and ^{15}N frequencies of residue (i)

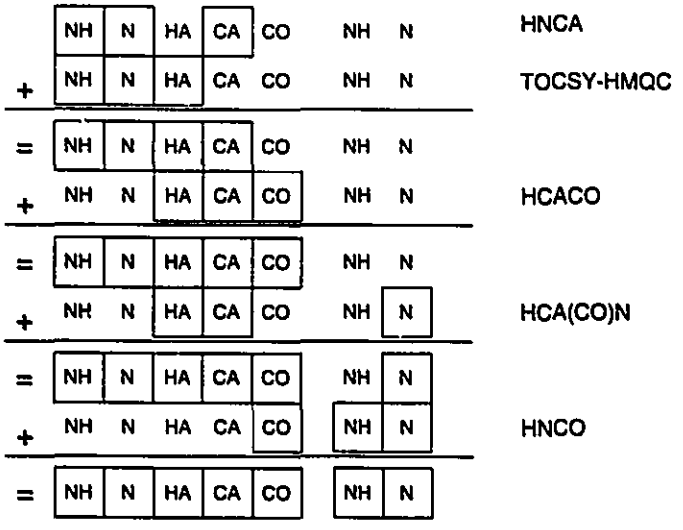


Figure 2.12: The assignment scheme using heteronuclear 3D NMR based on the through-bond correlations. The assignment is conducted by overlapping two previously assigned frequencies in each subsequent spectrum.

are used to obtain the assignment of the C_α and αH of the same residue. Then, the C_α and αH frequencies are used to obtain assignments for the CO of residue (i) and ^{15}N of residue ($i+1$) with the HCACO and HCA(CO)N experiments. Finally, the CO and ^{15}N frequencies are used

to find the NH proton frequency of residue (i) with the HNCO spectrum, thus completing one cycle of the assignment. In chapter 3, a similar but more rigorous algorithm is described to assign the protein backbone resonances using heteronuclear 3D triple resonance experiments. Subsequent assignment of protein side chain can be conducted using 2D DQF-COSY, TOCSY or 3D HCCH-COSY/TOCSY. The corresponding automated approaches are described in chapter 4.

2.5 General description of the automated resonance assignment

We have discussed the importance of resonance assignment in the protein structure determination from NMR data. The actual strategy to carry out a manual assignment is also described. In this section, the characteristics of automated resonance assignment tools are discussed, some important problems and the limitations of automated assignments are also addressed.

The strategy of automated resonance assignment essentially parallels the manual assignment strategy. The assignment is divided into two parts: the spin system identification and the establishment of sequential connectivities. Although integration of resonance assignment and structure calculation [20] have been proposed, almost all of the published attempts are designed for spin system identification, sequential assignment or both. In other words, structure calculations are usually separated from resonance assignments.

In terms of a complete automated assignment software, an automated tool must be provided to extract spin systems from available spectral data. Furthermore, an automated amino acid type determination tool should also be provided. As for the sequence-specific assignment, both common approaches, i.e., use of interresidue NOE and use of triple resonance heteronuclear 3D NMR, should be taken into consideration. The design must allow the sequential connectivities to be created in a reasonable amount of time, for example, in several hours. A variety of algorithms has been applied to implement the above requirements, including the ones using systematic approaches [20] as well as artificial intelligence such as expert systems [12, 13], neural network [47,48], constraint propagation [19] and genetic algorithm [49].

An important characteristic of a good automated assignment software is that it should have the flexibility to accept many different types of NMR data from various experiments. NMR spec-

troscopists are continuously creating novel experiments. The advance of NMR hardware and biotechnology also enable them to design specific experiments for a specific protein sample. A well-designed automated assignment software should not restrict itself to certain types of experiments. However, algorithms designed for specific types of experiments sometimes outperform general-purpose algorithms, because general-purpose algorithms might be unable to take full advantage of all the information embedded in a spectrum.

Although the ultimate goal of resonance assignment is complete automation, human intervention is inevitable in today's automated assignment tools simply due to the complexity of the spectral data which make complete automation difficult to achieve. An automated assignment software should not become a black box which prevents users from understanding the internal actions and process. It is better to allow the software to have the capability of interacting with users at various stages during the assignment process while keep the integrity of the software from becoming merely a bookkeeping tool. As an example, our spin system extraction algorithm, which is described in chapter 3 and 4, generates output files containing all the deduced spin systems. Sometimes degenerate chemical shifts result in strange spin systems. Such a case is a spin system with one αH and 4 βH's that can be generated due to degenerate αH chemical shifts. Although it is easy for computer algorithms to determine which spin systems are incompatible with the 20 common amino acids using the spin coupling patterns, human inspection might still be necessary to separate the degenerate chemical shifts.

To obtain accurate assignment, a program should ideally be able to use as much information as is available. Knowledge about the structural information, such as a helix, coil or β-strand, may make it possible to predict the chemical shift range of certain protons. In subsequent assignment, the known chemical shift ranges can be treated as an additional evidence to confirm or deconfirm the assignment. The experimental conditions under which the spectra are acquired may help users to predict which peaks are present in the spectra, and which aren't. A mutant or homologous protein may be assigned rapidly as long as the original protein has been sequentially assigned [50, 51]. A 2D ^{13}C HMQC spectrum may help to unfold the ^{13}C chemical shifts of a 3D spectrum. Such miscellaneous information sometimes is indispensable for a successful resonance assignment.

In terms of the quality of the NMR data, a good automated assignment tool should be able to

overcome problems caused by false and missing peaks. The software should tolerate peak missing to a considerable extent just as it should also be to reject false data. Again our own programs are chosen to illustrate these points. A leucine is shown in Figure 2.13(a). It comprises 11 hydrogen

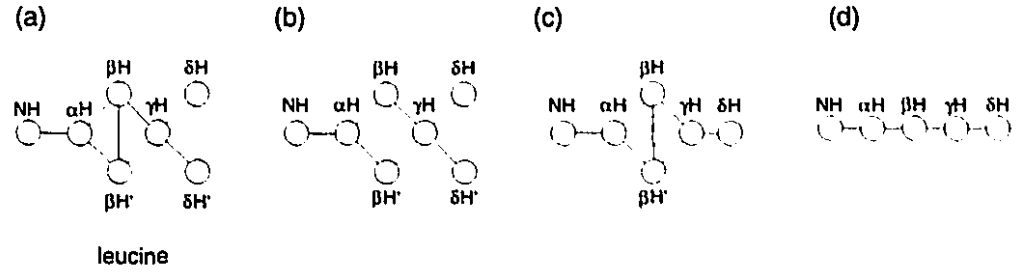


Figure 2.13: A leucine and its three possible candidate spin systems. (a) The normal leucine spin system. (b) A leucine without the βH-βH' connection. (c) A leucine without one δH. (d) A leucine without one βH and one δH.

atoms and 8 $^3J_{H-H}$ couplings. Suppose the $^3J_{\beta H-\beta H'}$ cross peak is missing due to the broad diagonal in the COSY spectrum. The extracted spin system will probably look like the one shown in Figure 2.13(b). Furthermore, after missing another peak of the methyl group, the deduced spin system is shown in Figure 2.13(c). Finally, another missed βH reduces the spin systems to the one shown in Figure 2.13(d). According to the spin system pattern recognition algorithm we designed, all the spin systems in Figure 2.13(b), (c), (d) can be matched with Figure 2.13(a). In other words, they all have chances to be assigned to a leucine. Certainly Figure 2.13(b) has the greatest probability to be assigned because its spin coupling topology has the highest similarity with an ordinary leucine.

To reject false peaks, automated assignment algorithms should inspect all logical relationships that exist between the suspicious peak and its surroundings. A genuine peak must have several coupled neighboring peaks whereas a false peak may have one connected neighbor but less likely to have two or three neighbors.

Data processing prior to assignment also plays a significant role in the design of automated assignment softwares. Spectral artifacts which might be confusing in automated assignment procedures should be removed prior to the start of the actual assignment processing. Before performing a Fourier transform on the time-domain data, zero filling, linear prediction [52] and Karhunen-

Loéve transformation [53] may be applied. After Fourier transform, ridges of t_1 noise can be removed manually [54, 55].

The most critical pre-assignment processing is the peak picking procedure. The simplest approach is either to pick all points above a given threshold or to use a maxima detecting procedure to find local maxima. These simple approaches seem incapable, to date, of providing reliable peak lists, a great volume of peaks, many of them noise, can be generated. A more advanced approach is to implement user-defined peak shapes (for example, ellipsoid) and search for peaks having those shapes in the spectrum. Garrett *et al.* [56] have designed a software called CAPP based on this approach. Artificial neural networks [57], after training with examples, also have the capability to distinguish real from bad peaks.

Spectral alignment is another pre-assignment problem. Almost all assignment strategies use several different types of spectra. The same hydrogen atom may appear at slightly different positions in those spectra. This chemical shift inconsistency can cause problems when comparing chemical shifts or peaks from two or more different spectra. If the inconsistency is systematic, i.e., all nuclear spins shift toward the same direction with roughly the same distance, the correction is straightforward. Otherwise a usual approach is to introduce tolerance values in the actual assignment stage. Every comparison between two chemical shifts from different spectra must pass the tolerance. Of course, some incorrect matches are inevitable.

Some people argue that automated assignment tools don't have much use simply because computers can do no more than human beings can. Although the argument is true, this doesn't imply that the computer-assisted assignments are valueless. Complete automation of resonance assignment still remains a goal due to the complexity of the task. However, properly designed automated assignment softwares do reduce the effort and the time required to assign a spectrum.

Another common argument is that automated assignment tools should be able to get the results with fewer data than human need. Many of the present automated assignment programs simply emulate manual assignment strategies. It is apparent that to achieve the goal of "use fewer NMR experiments" one must implement different assignment strategies exclusively for computers. We would like to emphasize, however, that computer programs cannot achieve what people can't. If a person cannot get the assignment using a limited data set in an unlimited amount of time, there

is no reason to ask computers to succeed.

Manual assignment is not 100% deterministic. That is, independently obtained assignments from two persons might differ because of the human bias and intuition participated during the assignment period. On the contrary, every step is deterministic in computer assignment. Intuition and bias are not involved. If a person is able to assign a protein NMR data without using any personal bias or intuition, i.e., every step must have a clear logical basis, computer-assisted assignment tools should be able to produce identical assignment in much shorter time. This is probably the main advantage of using automated resonance assignment tools.

2.6 Spin System Identification Using Constrained Partitioning Algorithm (CPA)

2.6.1 Introduction

Parallel to the manual assignment strategy, automated assignment begins with the identification of spin systems. Here the meaning of "identification" is two-fold. First of all the spin systems must be extracted from NMR data. Secondly, the amino acid types of those spin systems must be determined. Traditionally, 2D DQF-COSY and TOCSY provide sufficient NMR data for extracting spin systems, at least for moderate sized proteins. The amino acid types are determined mainly by human experience along with possibly other available chemical information. A computer algorithm to extract spin systems from 2D ^{1}H DQF-COSY and TOCSY spectra is introduced in this section. The remaining amino acid type determination task is discussed in the next section where a spin pattern recognition algorithm determines the amino acid types of spin systems automatically.

2.6.2 The Constrained Partitioning Algorithm

The most commonly used 2D NMR experiments for assigning protein resonances are DQF-COSY and TOCSY. Both experiments observe proton-proton couplings and represent them as cross peaks in the NMR spectra. The COSY experiment observes couplings between adjacent

protons (within three bonds) while TOCSY experiment observes long range correlations between all protons within a spin coupling system.

The algorithm responsible for the spin system extraction is called the Constrained Partitioning Algorithm (CPA). It partitions NMR data into amino acid spin systems based on fulfilling certain constraints. CPA takes the peak list of DQF-COSY spectrum as the major data input. The TOCSY peak list is treated as a database where constraint peaks can be found. The basic operation CPA performs is the cross peak merge. CPA attempts to find all cross peaks belonging to a spin system, merges these peaks together and constructs the spin system. For example, a serine spin system is composed of four spins and four cross peaks: an NH, an αH ,two βH's, NH-αH, αH-βH_1, αH-βH_2 and βH_1-βH_2 (Figure 2.14). Merging one cross peak at a time, CPA can construct

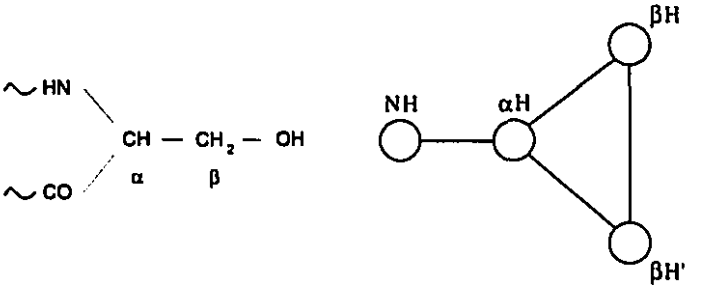


Figure 2.14: A serine and its spin coupling system.

a serine spin system in three steps. Before discussing the details of the spin system constructions, the basic operation of merging of two peaks is first described.

Each 2D NMR cross peak correlates two spins. Therefore, as a result of the merge of two peaks, a three-spin system is created. Figure 2.15 shows such a simple merge. A cross peak (δ_i,

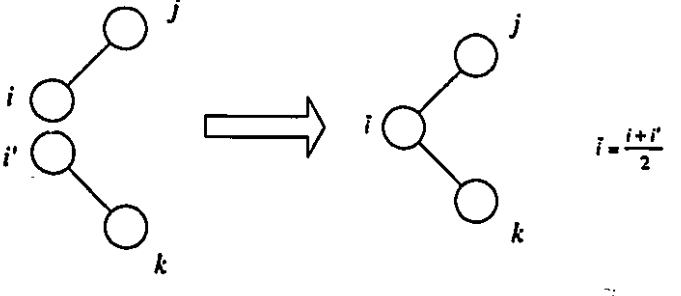


Figure 2.15: Schematic illustration of the merge of two 2D NMR cross peaks.

δ_j) is merged with another peak ($\delta_{i'}$, δ_k). The condition to justify the merge is that $|\delta_i - \delta_{i'}|$ be less than a merge tolerance, which is a value with the unit of chemical shift. A three-spin system $\{i, j, k\}$ can be thus formed. In a crowded NMR spectrum, many cross peaks might fulfilled the above merge condition. That is, other than the peak ($\delta_{i'}$, δ_k), one might also observe ($\delta_{i''}$, δ_l), ($\delta_{i'''}$, δ_m), ... , etc., whose first coordinates are all located within the merge tolerance of δ_i. A way is needed to distinguish the peak that should be merged from those that simply satisfy the tolerance requirement. CPA implements a constraint checking procedure which requires each candidate peak, ($\delta_{i'}$, δ_k), ($\delta_{i''}$, δ_l), ($\delta_{i'''}$, δ_m), ... , to provide additional evidence, i.e., a constraint peak, to support the merge. CPA has a ranking system which selects the most reliable evidence from all the possible candidates. The actual merge takes place between the original peak and the candidate having the most reliable evidence. The evidence peaks usually come from COSY or TOCSY spectrum. Figure 2.16 shows two typical merge CPA conducts. In Figure 2.16(a), a

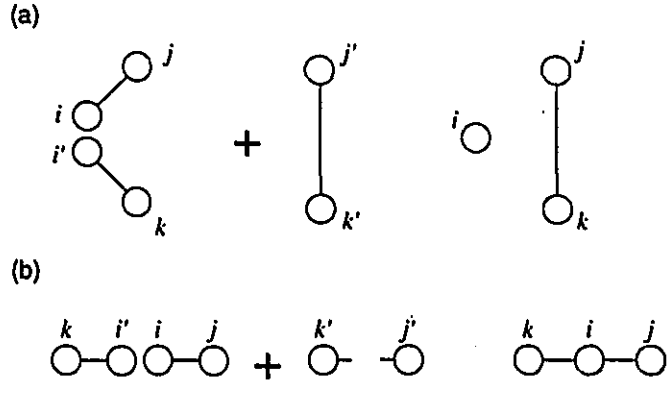


Figure 2.16: Two typical merge conducted by algorithm CPA.

COSY or TOCSY constraint peak ($\delta_{j'}$, $\delta_{k'}$) is required to construct the three-spin system $\{i, j, k\}$. In Figure 2.16(b), only a TOCSY constraint ($\delta_{j'}$, $\delta_{k'}$) can provide the eligibility of the merge. The mechanism shown in Figure 2.16 reduces the chance of incorrect merge and makes it possible for CPA to process overlapped NMR spectra. Later in this section the limitations of CPA are discussed where the ambiguities that CPA is unable to resolve are listed. The ranking system CPA implements calculates a parameter which measures the deviation of the chemical shifts between merging peaks. Suppose two COSY peaks (δ_i, δ_j) and ($\delta_{i'}$, δ_k) are about to be merged. This means

$|\delta_i - \delta_{i'}|$ is less than a merging tolerance T_m, typically 0.02 ppm for proton. Another TOCSY peak $(\delta_{j'}, \delta_{k'})$, the expected constraint, is also observed. Both $|\delta_j - \delta_{j'}|$ and $|\delta_k - \delta_{k'}|$ are less than another chemical shift tolerance T_c. The ranking parameter A is defined as

$$A = 1 - \sqrt{d_1 \times d_2} \tag{2.3}$$

where

$$d_1 = \frac{|\delta_i - \delta_{i'}|}{T_m}$$

$$d_2 = \frac{|\delta_j - \delta_{j'}| + |\delta_k - \delta_{k'}|}{2T_c}$$

$$0 \leq d_1 \leq 1 \text{ since } |\delta_i - \delta_{i'}| \leq T_m$$

$$0 \leq d_2 \leq 1 \text{ since } |\delta_j - \delta_{j'}|, |\delta_k - \delta_{k'}| \leq T_c$$

therefore

$$0 \leq A \leq 1$$

Figure 2.17 is the pictorial representation of the ranking parameter. Depending on how close

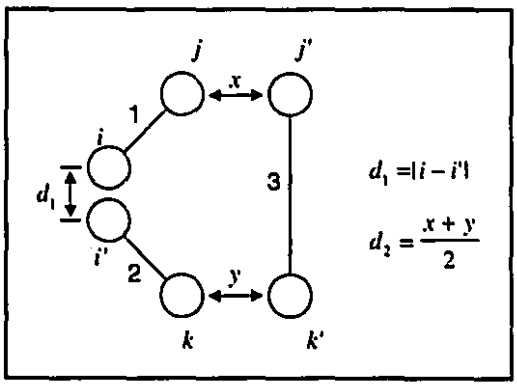


Figure 2.17: Pictorial representation of the variables used in calculating the ranking parameter.

δ_i and $\delta_{i'}$, δ_j and $\delta_{j'}$, δ_k and $\delta_{k'}$ are, the ranking parameter A bears a value from 0 to 1. A higher value of A corresponds to a better match between the three peaks, hence, a more reliable merge is expected.

We now proceed to describe the construction of amino acid spin systems . CPA's main goal is to extract spin systems from NMR data. The extracted spin systems are processed as mathematical

graphs and represented as adjacency lists [58]. Each graph represents an individual spin system. The nodes of a graph correspond to the spins while the edges of a graph correspond to the cross peaks. Figure 2.18 illustrates a graph and its corresponding spin system. The following pseudo

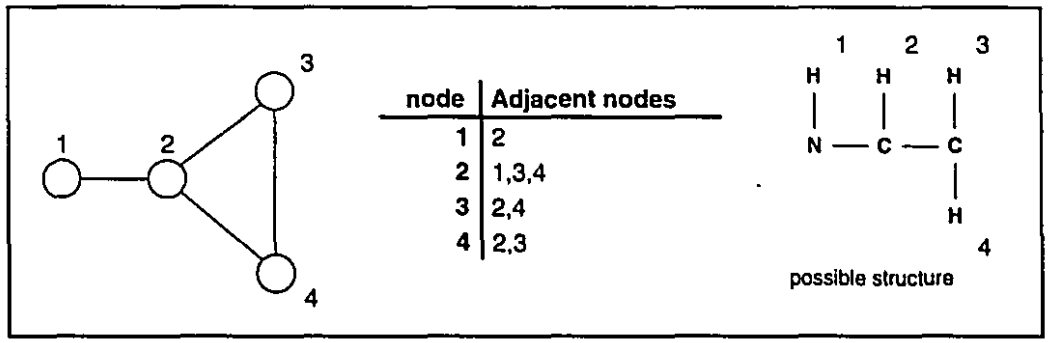


Figure 2.18: A spin coupling graph, its mathematical representation and the corresponding chemical structure.

codes are responsible for constructing spin systems from NMR data.

```
void CreateSpinSystem(Peaklist_type 2D DQF-COSY, 2D TOCSY)
{
// Input : 2D DQF-COSY and TOCSY peak lists
// Output: spin systems represented as graphs

  for each input COSY peak i {
    add peak i into an empty spin system Si ;
    for each input COSY peak j {
       in the COSY peak list, find a peak n which is the most likely peak
       to be merged with peak j ;

       if peak j is a member of the spin system Si
          add peak n into Si;
       else if peak n is a member of the spin system Si
          add peak j into Si;
    }
  }
  output all Si;
}
```

The above segment of computer codes produces N spin systems for a COSY data set containing N peaks. There are, however, usually many redundant spin systems being formed. For example, starting from cross peak 1, CPA might construct a spin system containing four peaks $\{1, 2, 3, 4\}$. Furthermore, the same spin system can also be created starting at peak 2, 3 or 4 independently. In this case, four identical spin systems can be created starting from four different

peaks. A post-partitioning subroutine should be conducted over the extracted spin systems to remove such redundancies.

2.6.3 *Discussion and Limitation of CPA*

Some technical problems related to CPA are discussed here. The first one involves symmetrical peaks. Homonuclear 2D NMR experiments produce symmetrical cross peaks on two sides of the diagonal. Since both dimensions record proton frequencies, the two symmetrical peaks have redundant information. Hence the symmetrical cross peaks are removed in CPA. The processed peak data is then sent to the subroutine `CreateSpinSystem( )` to initiate the real partitioning. Another technical problem involves chemical shift tolerances. As seen in equation 2.3, two types of tolerances are introduced. The tolerance T_m is used for merging two peaks. Another tolerance T_c is used for comparing evidence peak with the query peaks. In Figure 2.17, T_m is the tolerance for merging peak 1 and 2 while T_c is applied to judge if peak 3 is qualified as an evidence for the merge. The "to-be-merged" peaks usually come from the same NMR data, in this case, a 2D DQF-COSY spectrum. However, the evidence peak might come from a TOCSY spectrum which could have a small inconsistency in the chemical shift positions. The tolerance T_c might then need to be set to a greater value than T_m to reflect this inconsistency. The default values for T_m and T_c are set to 0.02 ppm. The users are encouraged to set reasonable values for those tolerances based on their knowledge about the NMR data. Using smaller tolerances means that all merge is carefully verified, so that the risks of incorrect merge are low. However, small tolerances might leave a number of peaks unpartitioned, that is, many peaks might be unable to find their coupling partners. On the contrary, large tolerance values risk merging incorrect peaks into a spin system which might have strange (unrecognizable) spin coupling pattern. Applying appropriate tolerances relies on human experience and a trial-and-error approach may be needed for determining appropriate tolerances.

CPA is designed to overcome spectral overlap. The adoption of additional constraints during the merging stage helps to resolve many spectral overlap problems. Moreover, the number of constraints used in the algorithm is not fixed. If a single TOCSY peak does not resolve the spectral

ambiguity, one can add other constraints such as an additional 2D spectrum or the third coordinate of a 3D NMR spectrum. In practice, if only 2D COSY and TOCSY data are provided, CPA fails to separate certain spin systems under conditions of severe chemical shift degeneracy. In Figure 2.19 two amino acid residues having degenerate NH and αH resonances are shown. Spectroscopists

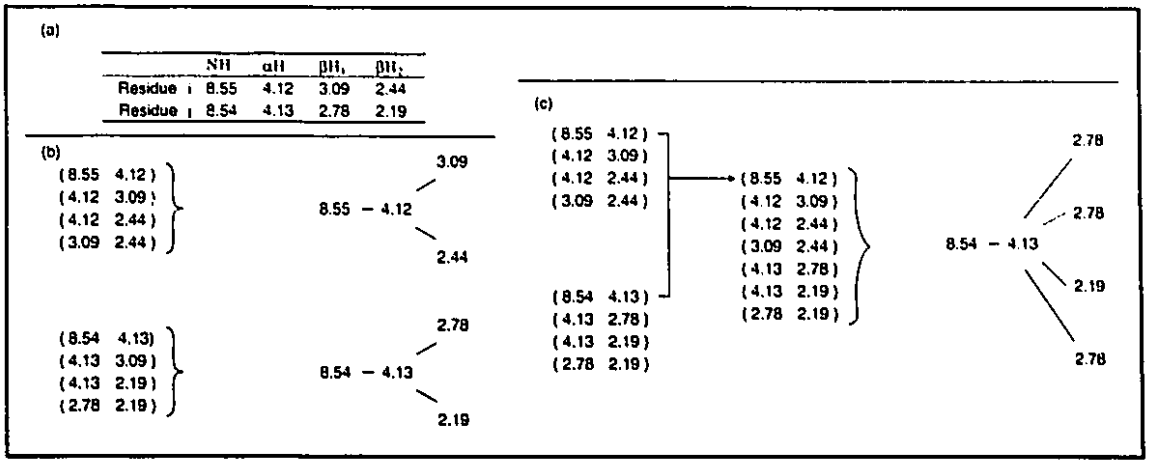


Figure 2.19: Schematic illustration of the chemical shift degeneracy problem. (a) Two spin systems having degenerate NH and αH chemical shifts. (b) If a small chemical shift tolerance is chosen in CPA, it is possible to resolve the degenerate NH and αH. (c) if a larger tolerance is used, an overlapped spin system will be created due to the degenerate NH and αH resonances.

might be able to distinguish the cross peak (8.55, 4.12) from (8.54, 4.13). However, it is difficult for computer programs to separate such nearly overlapped peaks. The cross peak data for the two hypothetical amino acid residues are listed in Figure 2.19. In Figure 2.19(b), a small tolerance, e.g., 0.005 ppm, is chosen. This tolerance is able to resolve the overlap which occurs at the two degenerate NH and αH peaks. In Figure 2.19(c), an ordinary tolerance of 0.02 ppm is chosen. In this case, CPA considers peak (8.54, 4.13) a redundant peak of (8.55, 4.12) and discards the former. A large spin system containing 6 resonances is constructed as a result. As mentioned above, although a small tolerance solves the problem of chemical shift degeneracy, using small tolerances may leave a lot of peaks unpartitioned. In practice, a moderate tolerance (0.01 to 0.03 ppm) is preferred. On the one hand, there won't be too many unpartitioned peak. On the other hand, unreasonably large spin systems generated by a small tolerance can still be manually examined and resolved.

In general, CPA is unable to resolve the spectral overlap caused by two or more degenerate

resonances within a spin system. An extension of the 2D CPA algorithm is described in Chapter 4 where heteronuclear 3D NMR data are used to enhance the capability of overcoming spectral overlap.

2.7 Determination of amino acid types

2.7.1 Introduction

An algorithm called CPA (Constrained Partitioning Algorithm) is described in the previous section. CPA traces and extracts spin coupling systems from homonuclear 2D NMR spectra. The observed spin systems have to be sequentially assigned to the proper positions within the primary sequence of the protein. Before the sequential assignment can be done, the identities, the amino acid types, of those spin systems must be determined. Although it is difficult to determine exactly to which specific amino acid an observed spin system corresponds, it is, however, possible to find a number of amino acid candidates to which a spin system may be assigned. Traditionally this task is done manually. Knowing the number of protons and their chemical shifts, experienced spectroscopists are able to identify the amino acid types of observed spin systems. A simple example using the traditional strategy to determine a glycine spin system is that almost all spin systems having one proton with chemical shift around 8 ppm and the other two protons around 4 ppm can be identified as glycines.

In this chapter, attempts are made to automate the amino acid type determination. Algorithms are proposed to allow computers to "visualize" the spin system patterns, i.e., to recognize distinct spin patterns. The recognition is based on chemical shift as well as topological matches. In the example of glycine, the spin pattern recognition algorithm not only makes sure the observed chemical shifts are indeed in the expected ranges but also examines the topology of the pattern, i.e., there are fewer than 3 spins and they should be connected to each other through scalar couplings. To achieve this goal, the proposed algorithms use the mathematics of graph theory and the simple fuzzy subset theory. Background of those topics is introduced in first followed by detailed description of the pattern recognition algorithm.

The pattern recognition algorithm was originally proposed and applied to homonuclear 2D NMR data by Xu *et al.* [25] in 1993. An extended version is described in chapter 5 where the application is extended to heteronuclear 3D NMR.

2.7.2 *Background*

2.7.2.1 *Introduction to graph theory*

The mathematical graphs are graphical representations of nodes and lines. The nodes are called *vertices* and the linking lines are called *edges*. When the linking lines are directed, they are referred to as *arcs*. Mathematically speaking, a graph consists of a vertex and an edge sets. The exact definition of graphs can be given as the follows [59]: *a graph $\mathcal{G}$ consists of a vertex set $\mathcal{V}$ on which a pair relation $\mathcal{E}$ is defined.*

$$\mathcal{G} = \{\mathcal{V}, \mathcal{E}\} \tag{2.4}$$

A set of vertex pairs can be defined by means of the pair relationship. The vertex pairs can be *ordered* or *unordered*. (see Figure 2.20) Two vertices are *adjacent* if they are connected by

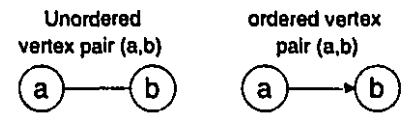


Figure 2.20: Ordered and unordered vertex pairs.

an edge. The linking edge and the two vertices are said to be *incident* to one another. Graphs composed only of arcs are called *directed graphs*. The number of edges incident with a given vertex is called the *degree*, g, of that vertex. Two degrees are assigned to each vertex in a directed graph. The *indegree*, g^-, counts the number of arcs ending on this vertex. The *outdegree*, g^+, counts the number of arcs originating from this vertex. The concepts of indegree and outdegree are used later in this chapter.

Figure 2.21 shows a pair of ordered graph. A *linearly ordered* graph is the graph with both indegree and outdegree of each vertex equal to 1. A *partially ordered* graph is the graph whose

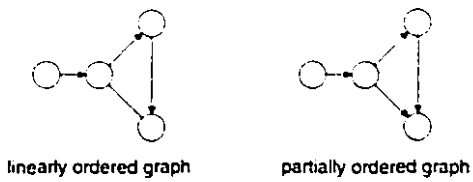


Figure 2.21: Linearly ordered and partially ordered graphs.

indegree and outdegree of each vertex can be greater than 1. A *walk* is a sequential collection of edge pairs, originating from one vertex and ending on another. There is no restriction on how many times a vertex can be traversed through a walk. For directed graphs, of course, the traverse can only be conducted through the direction of the arcs.

2.7.2.2 *Graph Representation*

It is necessary to represent the mathematical structure of a graph using some kind of data structure in order to solve graph related problems by computer programs. Since our pattern recognition algorithm demands random access to the vertices of a graph, in the implementation, graphs are represented as adjacency lists where each vertex keeps an array holding all the adjacent vertices. A typical implementation may look like this:

```
//MAX is the maximum number of vertices in the graph
typedef int AdjacencyList_type[MAX];
typedef struct {
                int n;
                int valence[MAX];
                AdjacencyList_type A[MAX];
               }Graph_type;
```

2.7.2.3 *The Concept of Fuzzy Subsets*

The concept of fuzzy subsets was first introduced in 1965 by Zadeh [60]. It is a novel way of representing fuzziness happened everyday in our life. The fuzzy subset theory is a generalization of conventional mathematical set theory.

There are two kinds of imprecision or vagueness in data or information recorded from our environment. The first one is statistical, like flipping up a coin, the outcome is not certain but

can be predicted statistically. The other imprecision is non-statistical. For example, two persons are much alike. One application of fuzzy subset theory is to quantitatively describe the similarity between two objects. This is also the main feature which is applied to our amino acid recognition algorithm.

For conventional sets, the membership of the elements is determined by precise properties. For example, the set of numbers H from 6 to 8 is crisp; we write $H = \{r \in \mathcal{R} \mid 6 \leq r \leq 8\}$. Equivalently, H is described by its membership function, m_H:

$$m_H(r) = \begin{cases} 1 & : \quad 6 \leq r \leq 8 \\ 0 & : \quad \text{otherwise} \end{cases} \tag{2.5}$$

The above membership function corresponds to a 2-values logic, that is, is an element of the set or isn't.

On the other hand, a fuzzy subset contains elements having imprecise properties which in turn lead to multi-values membership function. The rigorous definition of the fuzzy subset was given by Zadeh [61]: let **E** be a set, denumerable or not, and let x be an element of **E**. Then a fuzzy subset $\tilde{\mathbf{A}}$ of **E** is a set of ordered pairs

$$\{(x \mid \mu_{\tilde{\mathbf{A}}}(x))\}, \forall x \in \mathbf{E} \tag{2.6}$$

where $\mu_{\tilde{\mathbf{A}}}(x)$ is the grade or degree of membership of x in $\tilde{\mathbf{A}}$. Thus, if $\mu_{\tilde{\mathbf{A}}}(x)$ takes its values in a set **M**, called the *membership set*, one may say that x takes its value in **M** through the *membership function* $\mu_{\tilde{\mathbf{A}}}(x)$. Note that $\tilde{\mathbf{A}}$ is called a fuzzy subset and not fuzzy set, since the reference set **E** is not fuzzy.

Consider the following example. A finite set with five elements:

$$\mathbf{E} = \{x_1, x_2, x_3, x_4, x_5\} \tag{2.7}$$

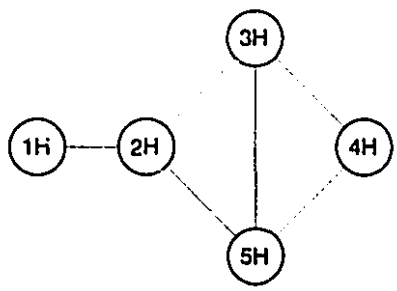
A fuzzy subset $\tilde{\mathbf{A}}$ can be defined by the expression

$$\tilde{\mathbf{A}} = \{(x_1 \mid 0.2), (x_2 \mid 0), (x_3 \mid 0.3), (x_4 \mid 1), (x_5 \mid 0.8)\} \tag{2.8}$$

where x_i is an element of the reference set **E** and where the number placed after the bar is the value of the membership function for the element. Fuzzy subset $\tilde{\mathbf{A}}$ contains a little x_1, does not contain x_2, a little x_3, contains x_4 completely, and a large part of x_5.

2.7.3 *Amino acid type identification*

The through-bond correlations observed in NMR spectra can be used to extract spin systems. The computer algorithm CPA (Constrained Partitioning Algorithm), which is described in section 2.6, automatically extracts amino acid spin systems using through-bond scalar couplings. CPA takes input from correlation spectra, such as 2D DQF-COSY and TOCSY. Figure 2.22 is a sample spin system extracted from CPA.



Node	chemical shift	adjacency list
1H	8.969	2
2H	4.432	1,3,5
3H	2.022	2,4,5
4H	2.035	3,5
5H	2.120	2,3,4

Figure 2.22: A five-spin system. Its spin coupling graph and the corresponding mathematical representation using an adjacency list.

The spin system shown in Figure 2.22 has five spins. An adjacency list is used to represent the connectivity relationships between those five nodes. An important remaining question is "which amino acid does this spin system belong to?". It might be a leucine as there are two βH's and one γH. It might also be a methionine, a glutamine, an arginine or a lysine since they all have two βH's and one γH. On the other hand, it is obvious that this observed spin system must not be a glycine, an alanine, or a serine ... , etc., because these amino acids don't have the γH. This kind of analysis inspired us to design computer programs to automate the determination of amino acid types. A spin pattern recognition algorithm was developed to accomplish this task. The algorithm determines the amino acid types of the extracted spin coupling systems using topological analysis, such as the numbers of βH and γH, as well as chemical shift analysis. Using Figure 2.22 as an example, suppose the two βH's have chemical shifts 2.022 and 2.120 ppm, respectively. It is more likely that the query spin system is a glutamine than it is an arginine, because the former has the expected βH chemical shifts of 1.92 and 2.10 ppm [62] while the latter has βH chemical shifts of 1.63 and 1.79 ppm. (See Table 2.3 for the expected chemical shifts for the 20 amino acids)

The following section describes the basic principles of the spin pattern recognition algorithm. The original version of the algorithm is applied to spin systems obtained from homonuclear 2D NMR.

2.7.3.1 Graph representation of the amino acids

As shown in Figure 2.22, spin coupling patterns can be defined by mathematical structures called graphs. Each spin corresponds to a vertex of the graph and each J coupling connection corresponds to an edge of the graph. Mathematically a graph is represented as a set of vertices and edges.

$$\mathcal{G} = \{\mathcal{V}, \mathcal{E}\} \tag{2.9}$$

Groß and Kalbitzer [62] produced a chemical shift database for the 20 amino acids using published NMR assignment data. The averaged chemical shifts and standard deviations for each proton in the 20 commonly seen amino acid were reported. Using those data, with respect to each of the 20 amino acids, the reference set of an amino acid graph can be constructed as

$$RS(i) = \{\mathcal{V}_{ref}, \mathcal{E}_{ref}\}, i = \text{Ala, Gly, Thr, } \ldots \tag{2.10}$$

where $\mathcal{V}_{ref}$ is the set of chemical shifts of NH, αH, βH , ... , a so-called *cluster*, and $\mathcal{E}_{ref}$ is the set of edges connecting vertices in the cluster $\mathcal{V}_{ref}$. $\mathcal{V}_{ref}$ has a corresponding chemical shift standard deviation set $\Delta\mathcal{V}_{ref}$ where the data is taken from Groß's database. Table 2.3 lists the expected chemical shifts and the standard deviation data from the 20 amino acids. A sample reference set of alanine $RS(\text{alanine}) = \{\{8.15, 4.24, 1.32\}, \{8.15 - 4.24, 4.24 - 1.32\}\}$ and its standard deviation set $\Delta\mathcal{V}_{ref} = \{8.15/0.62, 4.24/0.38, 1.32/0.28\}$ are shown in Figure 2.23.

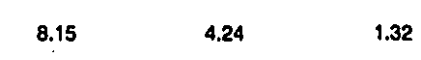


Figure 2.23: A simple alanine spin system.

The chemical shifts of deduced spin systems usually have a certain deviation from the expected values. Protein secondary structures and local chemical environments are factors to effect

Table 2.3: The expected proton chemical shifts for the 20 amino acids. The standard deviations are also given. Data are taken from Groß's paper. All numbers are in ppm.

amino acid	NH	αH	βH	Others
Ala	8.15/0.62	4.24/0.38	1.32/0.28	
Arg	8.20/0.83	4.28/0.35	1.63/0.43, 1.79/0.34	γH 1.52/0.34, 1.56/0.34 δH 3.11/0.19, 3.14/0.19 εH 7.21/0.16
Asn	8.29/0.62	4.73/0.30	2.69/0.32, 2.95/0.27	δH 7.18/0.55, 7.78/0.32
Asp	8.31/0.51	4.65/0.28	2.63/0.31, 2.93/0.33	
Cys	8.25/0.70	4.64/0.75	2.86/0.38, 3.19/0.38	
Gly	8.31/0.62	3.74/4.17 4.17/0.28		
Gln	8.28/0.61	4.43/0.45	1.92/0.27, 2.10/0.20	γH 2.29/0.25, 2.35/0.20 εH 6.85/0.38, 7.61/0.29
Glu	8.22/0.60	4.34/0.42	1.97/0.20, 2.04/0.18	γH 2.27/0.20, 2.34/0.21
His	8.28/0.57	4.54/0.19	2.94/0.39, 3.26/0.29	δ_2H 6.99/0.33, ε_1H 8.10/0.36
Ile	8.26/0.72	4.13/0.52	1.74/0.37	γH 1.01/0.26, 1.30/0.32 0.78/0.24, δH 0.69/0.25
Leu	8.19/0.60	4.25/0.49	1.60/0.37, 1.71/0.31	γH 1.51/0.30 δH 0.68/0.40, 0.83/0.25
Lys	8.28/0.65	4.23/0.42	1.74/0.38, 1.84/0.34	γH 1.30/0.39, 1.36/0.37 δH 1.54/0.24, 1.57/0.23 εH 2.91/0.13, 2.97/0.10 ζH 7.53/0.50
Met	8.10/0.44	4.41/0.51	1.89/0.19, 2.03/0.21	γH 2.55/0.17, 2.60/0.13 εH 1.98/0.21
Phe	8.49/0.80	4.69/0.48	2.85/0.28, 3.16/0.28	δH 7.12/0.27, εH 7.17/0.30 ζH 7.08/0.29
Pro		4.48/0.31	1.88/0.35, 2.18/0.40	γH 1.92/0.50, 2.02/0.45 δH 3.62/0.28, 3.77/0.29
Ser	8.48/0.58	4.50/0.47	3.72/0.44, 3.89/0.43	
Thr	8.30/0.75	4.53/0.43	4.17/0.31	γH 1.15/0.16
Trp	8.43/0.37	4.29/0.80,	3.06/0.23, 3.42/0.22	ε_1H 10.15/0.30, δ_1H 7.18/0.30 ε_3H 7.39/0.24, ζ_3H 7.00/0.30 η_2H 7.17/0.17, ζ_2H 7.41/0.32
Tyr	8.57/0.89	4.64/0.49	2.81/0.19, 3.04/0.28	δH 7.00/0.20, εH 6.70/0.20
Val	8.20/0.61	4.16/0.55	2.02/0.25	γH 0.76/0.22, 0.88/0.18

the exact position of chemical shifts. Therefore, it is difficult to determine that a vertex of an observed spin system is exactly a certain specific spin. A more appropriate representation is that the vertex of a spin system is more likely to be one proton, e.g., an αH than another one, e.g., a βH. The fuzziness of the mappings that appear in this case indicates that fuzzy subsets are proper representations for the experimentally observed spin systems.

A deduced spin system, as the one shown in Figure 2.22, can be represented as a fuzzy subset

$$FS = \{\mathcal{V}, \mathcal{E}, \mu\} \tag{2.11}$$

where $\mathcal{V}$ is a chemical shift subset, in the case of Figure 2.22, $\mathcal{V}$ = { 8.969, 4.432, 2.022, 2.385, 2.120}, $\mathcal{E}$ is the subset of all the connections between elements in $\mathcal{V}$, μ is the membership subset. Suppose there exists a *homomorphic mapping* between FS and RS(leucine). In other words, FS is a subgraph of RS, or FS can be assigned to a leucine. Assume the deviation between the experimental and expected chemical shifts follows the normal distribution, the membership function $\mu(j)$ can be defined as

$$\mu(j) = \exp\left\{-\left[\frac{\mathcal{V}_j - (\mathcal{V}_{ref}(leucine))_j}{(\Delta\mathcal{V}_{ref}(leucine))_j}\right]^2 / 2\right\} \tag{2.12}$$

where $\mathcal{V}_j$ is the jth chemical shift of the observed spin system of FS, $(\mathcal{V}_{ref}(leucine))_j$ is the corresponding chemical shift of a leucine in the amino acid database, $(\Delta\mathcal{V}_{ref}(leucine))_j$ is the standard deviation of $(\mathcal{V}_{ref}(leucine))_j$. $\mu(j)$ represents the degree of membership of mapping jth spin of FS to the corresponding position of RS.

Table 2.4 lists the membership values calculated using equation 2.12. It is obvious that chemical shift 4.432 ppm is the most likely resonance to be mapped to the αH of a leucine whereas 2.385 ppm has a low membership in terms of mapping to the γH of a leucine.

Table 2.4: The comparison of chemical shifts between a fuzzy subset FS, i.e., the observed spin system, and a reference set RS, the expected amino acid spin system. The calculated membership values are also shown.

observed chemical shifts of FS (in ppm)	expected chemical shifts of RS(leucine) /standard deviation	degree of membership
8.969	NH 8.19/0.60	0.43
4.432	αH 4.25/0.49	0.93
2.022	βH 1.60/0.37	0.52
2.120	βH' 1.71/0.31	0.42
2.385	γH 1.51/0.30	0.14

2.7.3.2 *Pattern Recognition Algorithm*

The actual pattern recognition algorithm involves two major stages. In the first stage, a homomorphic graph mapping algorithm is used to topologically determine if an observed spin system is a subgraphs of an amino acid. In the second stage, a similarity value is calculated between the observed spin system and that amino acid based on the membership functions. A standard leucine spin system and an observed spin system are shown in Figure 2.24. The first stage of the pattern

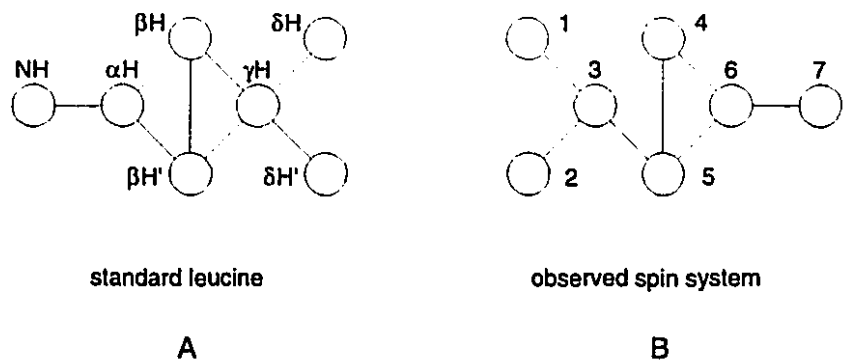


Figure 2.24: A standard leucine spin coupling graph and the observed spin system which might be assigned to the leucine.

recognition algorithm determines if graph B can be mapped to graph A topologically. Once the mapping if confirmed, the subsequent task involves the determination of similarity between A and B numerically.

The homomorphic graph mapping algorithm was implemented through a Heuristic Backtracking Algorithm (HBA) [63]. If HBA finds at least one mapping between a query graph QG and a supergraph SG, QG is said to be a subgraph of SG, namely $QG \subseteq SG$. HBA is composed of two procedures. In the first procedure, a "walking" procedure travels through a QG to find all of the possible routes connecting every nodes of QG. The following codes explains the principle of the walking procedure.

```
void walking(QueryGraph_type, ... )
{
//This function generates partially ordered graphs on each of the
//input query graph
//
```

```
//Input:  Query graphs, observed from NMR spectral data
//Output: All of the possible partially ordered graphs, also known
//        as routes.
//        These routes are stored in a data structure called ROUTE
//        which will
//        be used in a subsequent algorithm of HBA().

  arbitrarily choose a node from QG as the entrance node;
  save this node as the first element of a new route;
  push this entrance node into BranchStack;

  while BranchStack is not empty {
      pop a node from BranchStack;
      append current node into route;
      while there are still branches to walk {
           choose any branch to keep on walking while save the rest
             in BranchStack;
      }
  }
  store route into ROUTE;
}
```

The procedure first arbitrarily chooses an entrance node on QG, then all the untravelled nodes at each branch are saved into the data structure of a stack. Once the walking comes across an ending node, a node in the stack is popped out and the walking is resumed starting at that node. Using graph B of Figure 2.24 as an example of QG, the possible routes include $7-6-4-5-3-2-1, 7-6-5-4-3-2-1, 7-6-4-5-3-1-2, \ldots$, etc. All those routes are saved in a large data structure called ROUTE.

The second part of HBA performs the actual mapping actions. Once all the routes, also known as the partially ordered graphs, are created and saved in ROUTE, HBA walks on SG following each of the routes in the data structure of ROUTE. If the complete walk for a given route on the supergraph SG is accomplished, a mapping between QG and SG is determined and that route has all the information about this mapping. The entire procedure is explained using the following codes.

```
void HBA(ROUTE_type ROUTE, SuperGraph_type SG)
{
//HBA (Heuristic Backtracking Algorithm) walks on the supergraph SG.
//Information saved in ROUTE controls the walking. If the complete
//walk for a given route is accomplished, a mapping between
//the route and SG is determined.

//Input: 1. The data structure ROUTE, generated by the function walking().
```

```
//          ROUTE contains all the partially ordered routes of the
//          query graph QG.
//       2. The supergraph SG.

  for each of the route in ROUTE {
      while there are still untouched nodes left in SG {
           choose a node in SG as the entrance node;
           while ( not arrive at the end of the route) &&
                 ( there are still branches to walk on SG) {
                 look for a branch on SG matching current node of
                  route, essentially we examine adjacent degrees and
                  the chemical shift differences;
                 if a matching branch is found {
                      if not arrive at the end of the route
                          walk to the next node of SG;
                 else
                      go back one node on SG, choose another branch;
           }
           if arrives at the end of the route
               a mapping between QG and SG is determined, the
                actual mapping is the one saved in the route;
      }
  }
}
```

It is emphasized that there might exist more than one mapping between a QG and an SG. For instance, in Figure 2.24 the query graph B is a subgraph of supergraph A, but there are four different ways of mappings between B and A. The mappings are listed in Table 2.5.

Table 2.5: The four different mappings between the observed spin system and the standard leucine. See Figure 2.24.

mapping	$QG(B) \rightarrow SG(A)$						
1	$7 \rightarrow$ NH	$6 \rightarrow \alpha$H	$4 \rightarrow \beta$H	$5 \rightarrow \beta$H′	$3 \rightarrow \gamma$H	$1 \rightarrow \delta$H	$2 \rightarrow \delta$H′
2	$7 \rightarrow$ NH	$6 \rightarrow \alpha$H	$4 \rightarrow \beta$H	$5 \rightarrow \beta$H′	$3 \rightarrow \gamma$H	$2 \rightarrow \delta$H	$1 \rightarrow \delta$H′
3	$7 \rightarrow$ NH	$6 \rightarrow \alpha$H	$5 \rightarrow \beta$H	$4 \rightarrow \beta$H′	$3 \rightarrow \gamma$H	$1 \rightarrow \delta$H	$2 \rightarrow \delta$H′
4	$7 \rightarrow$ NH	$6 \rightarrow \alpha$H	$5 \rightarrow \beta$H	$4 \rightarrow \beta$H′	$3 \rightarrow \gamma$H	$2 \rightarrow \delta$H	$1 \rightarrow \delta$H′

In order to select the best mapping, an evaluation scheme must be introduced to discriminate all the mappings. This problem is solved by implementing a similarity evaluation system, which is discussed in the following.

In terms of fuzzy mathematics, the query graphs, i.e., the observed spin systems, are fuzzy subsets (FS) with respect to the 20 amino acid reference sets (RS). Our goal is to determine an

overall similarity value between a fuzzy subset FS and its reference set RS. Consider a query spin system which can be represented as the following fuzzy subset FS

$$FS = \{\{v_1, v_2, v_3, \ldots, v_n\}, \{v_1 \leftrightarrow v_3, v_1 \leftrightarrow v_4, v_3 \leftrightarrow v_4, \ldots\}\}. \tag{2.13}$$

This spin system has n spins and a number of couplings between v_1 and v_3 , v_1 and v_4, ... , etc.

Suppose this spin system can be mapped to the amino acid $RS(k)$

$$RS(k) = \{\{u_1, u_2, u_3, \ldots\}, \{u_1 \leftrightarrow u_2, u_2 \leftrightarrow u_3, \ldots\}, \{\sigma_{u_1}, \sigma_{u_2}, \sigma_{u_3}, \ldots\}\} \tag{2.14}$$

where u_i = NH, αH, βH, ..., k =Ala or Gly or Thr, and $\sigma'_{u_i}s$ are the sets of standard deviations of u_i's. $RS(k)$ has a total of N spins. Suppose there are M different mappings between FS and $RS(k)$

$$FS \subseteq_m RS(k), m = 1 \text{ to } M \tag{2.15}$$

The similarity for the mth mapping between FS and $RS(k)$ is defined as

$$\text{Similarity } S(m) = \sqrt{\frac{\sum_{l=1}^{n} \left[\mu(v_j \rightarrow u_l)\right]^2}{n}} \tag{2.16}$$

where $\mu(v_j \rightarrow u_l)$ is the degree of membership of mapping the jth spin of FS onto the lth node of $RS(k)$. Apparently the best mapping is the one having the maximum $S(m)$ therefore the overall similarity between FS and $RS(k)$ is given by

$$S(FS \rightarrow RS(k)) = \max(S(m)), m = 1 \text{ to } M. \tag{2.17}$$

As the final example, Figure 2.25 shows an observed spin system. Using HBA, the spin system can be mapped to valine, leucine, glutamine and arginine. There are two different ways of mapping the spin system to valine while there are 16, 24, and 116 ways of mapping it to leucine, glutamine and arginine, respectively. The mappings are summarized in Table 2.6. As an example, the first proton (7.754 ppm) of the observed spin system can be assigned to the NH of valine, which has an expected chemical shift of 8.20 ppm. Using equation 2.12, the degree of membership for

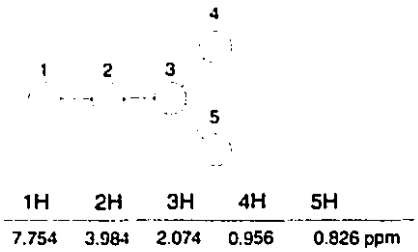


Figure 2.25: A deduced 5-spin system which might be assigned to Val, Leu, Glu or Arg. See Table 2.6.

Table 2.6: The mapping between a 5-spin system and various amino acids. For example, the observed 7.754 ppm spin node can be assigned to valine's NH proton, which has an expected chemical shift of 8.20 ppm. There are two different ways of mapping the observed spin system to a valine.

amino acid	actual mappings (all numbers are in ppm)				
	1H(7.754)	2H(3.084)	3H(2.074)	4H(0.956)	5H(0.826)
Val	8.20	4.16	2.02	0.76	0.88
2 mappings	8.20	4.16	2.02	0.88	0.76
Leu	8.19	4.25	1.60	1.71	1.51
16 mappings	8.19	4.25	1.71	1.60	1.51
	0.68	1.51	1.60	1.71	4.25
	...	...	...	...	...
Glu	8.22	4.34	1.97	2.04	2.27
24 mappings	8.22	4.34	2.04	2.27	2.34
	...	...	...	...	...
Arg	8.20	4.28	1.63	1.79	1.52
116 mappings	8.20	4.28	1.79	1.52	1.56
	...	...	...	...	...

all the proton mappings can be obtained. For instance, the membership between mapping 7.754 and 8.200 ppm is 0.77.

$$0.77 = \exp\left\{-\frac{\left[\frac{8.20-7.754}{0.61}\right]^2}{2}\right\} \tag{2.18}$$

where 8.20 is the expected chemical shift for valine's NH while 0.61 is its standard deviation. The similarity between the spin system shown in Figure 2.25 and various amino acids are listed in the last column of Table 2.7. The figures are calculated using equation 2.16. As an example, the overall similarity of mapping the spin system to Val is 0.92, which is the maximum value between 0.92 and 0.87 and is obtained from equation 2.17.

Having accomplished all the procedures, the spin systems derived from CPA now have asso-

Table 2.7: The similarity values between the observed spin system, (Figure 2.25) and various candidate amino acids. The membership values are calculated by equation 2.12. The similarity values are calculated by equation 2.16.

	membership to various protons of the amino acids					similarity
amino acid	1H(7.754)	2H(3.084)	3H(2.074)	4H(0.956)	5H(0.826)	
Val	0.77	0.95	0.98	0.67	0.96	0.87
2 mappings	0.77	0.95	0.98	0.91	0.96	0.92
Leu	0.77	0.86	0.44	0.052	0.074	0.55
16 mappings	0.77	0.86	0.50	0.22	0.074	0.57
	...	...	...	...	...	...
Glu	0.74	0.70	0.87	0.00	0.00	0.60
24 mappings	0.74	0.70	0.98	0.00	0.00	0.00
	...	...	...	...	...	...
Arg	0.87	0.70	0.59	0.050	0.12	0.57
116 mappings	0.87	0.70	0.71	0.25	0.097	0.60
	...	...	...	...	...	...

ciated amino acid type information. It is possible to construct a "deduced-spin-systems to amino-acids" table where the candidate amino acids for each spin systems are listed. As a subsequent processing, the Tree Search Algorithm (TSA) is responsible for achieving the sequential assignment. Figure 2.26 is an sample "deduced-spin-systems to amino-acids" table.

```
S1:  Asp/0.901 Asn/0.829 Phe/0.803 ......
S2:  Ala/0.778 Arg/0.732 Leu/0.715 ......
S3:  Gly/0.738 Thr/0.555 Phe/0.551 ......
S4:  Phe/0.803 Ser/0.705 Ile/0.648 ......
S5:  Leu/0.760 Arg/0.731 Lys/0.720 ......
.....  .....  .....  ..... .....
```

Figure 2.26: A "deduced-spin-system to amino-acids" table. For example, spin system S1 might be assigned to Asp, Asn, or Phe, ... , etc. The overall similarity values between the spin system and the amino acids are also shown. A higher similarity indicates a better match.

2.8 Sequence-specific resonance assignment using Tree Search Algorithm (TSA)

The Tree Search Algorithm (TSA) [25] was designed to obtain the sequential assignment of protein NMR data based on the spin systems extracted by CPA. TSA takes input of the spin systems and the associated amino acid type information determined from CPA and the pattern recognition algorithm, respectively. The output of TSA is the final sequentially assigned amino acid residues.

The entire protein resonance assignment can be divided into three stages. In the first stage, CPA is used to extract the individual amino acid spin systems from 2D DQF-COSY and TOCSY spectral data. The second stage of the assignment involves amino acid type recognition algorithm, which is described in the previous section. The spin pattern recognition algorithm determines all of the possible amino acids to which a spin system might be assigned. The information is listed in a "deduced-spin-system to amino-acids" table as the one shown in Figure 2.26. Once the table is prepared, TSA is responsible for mapping deduced spin systems into corresponding positions within the protein primary sequence in the final stage of the resonance assignment. The inter-residue correlations required to establish sequential connectivities are provided by NOE type of experiments. TSA relies on exhaustive searches over all possible sequential assignments which are satisfied with the protein primary sequence and the "spin-system to amino-acids" table. Several rules are prepared to determine a globally optimized final assignment. Reliable assignment can be obtained provided that the assigned polypeptide segments are sufficiently long.

Before the exhaustive searches can be started, information obtained from the pattern recognition algorithm must be converted to an appropriate format. The original information depicts the amino acid types of each observed spin system. However, TSA needs to know all the candidate spin systems of each residue. A preliminary conversion the original data is necessary for this purpose. The following codes illustrate the conversion:

```
void ConverTable()
{
//Input:  1. Protein primary sequence.
//        2. "spin-system" to "amino-acids" table.
//Output: 1. "residue" to "spin systems" table.
```

```
for (i=1;i<=n;i++)
  for (j=1;j<=f(i);j++)
    for (k=1;i<=N;k++)
       if (A_ij ==R_k)
          put S_i in the candidate list of R_k;
sort the candidates of R_i, (i=1 to N) according to
their overall similarity value;
}
```

$$\left[\begin{array}{c|l} S_1 & A_{11}\ A_{12}\ A_{13}\ \cdots\ A_{1f(1)} \\ S_2 & A_{21}\ A_{22}\ A_{23}\ \cdots\ A_{2f(2)} \\ S_3 & A_{31}\ A_{32}\ A_{33}\ \cdots\ A_{3f(3)} \\ \cdots & \cdots\cdots\cdots\cdots\cdots \\ S_n & A_{n1}\ A_{n2}\ A_{n3}\ \cdots\ A_{nf(n)} \end{array}\right]$$

1. a total of n deduced spin systems
2. each spin system S_i has
   $f(i)$ possible amino acid candidates.
   $A_{ij} \in \{$ Ala, Gly, Thr, $\ldots\}$

$\Downarrow$ converted to

$$\left[\begin{array}{c|l} R_1 & B_{11}\ B_{12}\ B_{13}\ \cdots \\ R_2 & B_{21}\ B_{22}\ B_{23}\ \cdots \\ R_3 & B_{31}\ B_{32}\ B_{33}\ \cdots \\ \cdots & \cdots\cdots\cdots\cdots \\ R_N & B_{N1}\ B_{N2}\ B_{N3}\ \cdots \end{array}\right]$$

3. The protein has N residues,
4. $R_1 - R_2 - \ldots - R_N$ is the primary sequence.
   $B_{mn} \in \{S_1, S_2, \cdots, S_n\}$.

The remaining task is to map each spin system to its expected position within the primary sequence. Recalling that the pattern recognition algorithm determines all the possible amino acid candidates of the observed spin systems, a mathematical similarity is calculated for each pair of the mapping between a spin system and an amino acid. For example, the similarity for the mapping between spin system S10 and the alanine is 0.87 while the similarity for mapping S10 to the threonine is 0.53. This means that S10 has a higher chance of being assigned to the alanine than to the threonine. Suppose another spin system S18 can also be assigned to an alanine with a similarity 0.94. As far as the assignment of the alanine is concerned, S18 is a better candidate than S10 is because of the higher similarity value. The spin system candidates in the "residue to spin-systems" table are sorted in descending order of each candidate spin system's similarity. The assumption made here is that a spin system candidate with higher similarity has greater probability to be assigned to its corresponding amino acid.

Table 2.8: An "amino-acid-residue to spin-systems" table. For example, the spin system No.9, 3, 14, 28, ..., all may be assigned to Ala10. However, only one assignment is actually chosen.

Residue	spin system candidates							
...	...							
Ala10	9	3	14	28	49	51	7	14
Phe11	4	8	11	24	33	17	25	
Asp12	11	4	5	8	33	19	17	25
Tyr13	21	4	8	39	50	30		
Ser14	8	2	4	21	39	33	25	11
Lys15	14	23	54	69	31	38	42	
Arg16	23	14	51	52	54	37		
Ile17	1	6	12	27	15	59		

The actual procedures of the sequential assignment are illustrated using an example shown in Table 2.8. According to the assumptions made, the most probable assignment for Ala10 is spin system S9; the most probable assignment for Phe11 is S4; ... , etc. Therefore the most probably sequential assignment for the polypeptide segment Ala10 – Phe11 – ··· – Ile17 is S9-S4-S11-S21-S8-S14-S23-S1. However, there is no way to guarantee that the spin system with the highest similarity value is always the right one to be assigned. To cope with this problem, TSA searches *all* possible assignment combinations and in a subsequent step discriminates them with certain criteria in order to determine the most probable sequential assignment. In the above example, the possible assignment combinations for the query polypeptide include S9-S4-S11-S21-S8-S14-S23-S1, S9-S4-S11-S21-S8-S14-S23-S6, S9-S4-S11-S21-S8-S14-S23-S12, ... , etc.. There are a total of $8\times7\times8\times6\times8\times7\times6\times6 = 5419008$ paths to be traversed. In practice, not all of the paths are valid. For instance, once the spin system S4 is assigned to Phe11, S4 can't be assigned to another residue in the following assignment, i.e., a spin system can't occur twice in a sequential assignment. Having applied this restriction, there is no need to traverse all the 5419008 combinations. However, the actual amount of searching is still a heavy load in terms of the computing time.

A number of criteria are set to determine the most probable or the best assignment. The most important criterion is the observation of interresidue correlations. TSA counts the number of NOE cross peaks observed between each adjacent spin system pairs. Knowing these numbers, TSA is able to determine the total number of observed NOE peaks within each assignment path. In Table 2.9 the number of NOE peaks observed between spin system S9 and S4 is two, between S4

Table 2.9: The possible assignment of an 8-residue polypeptide. There are two NOE cross peaks between spin system S9 and S4, one NOE cross peak between S4 and S11, . . . , etc.

Residue	Assignment	No. of NOE peaks	NOE evidence
Ala10	S9		
		2	Yes
Phe11	S4		
		1	Yes
Asp12	S11		
		3	Yes
Tyr13	S21		
		4	Yes
Ser14	S8		
		1	Yes
Lys15	S14		
		0	No
Arg16	S23		
		5	Yes
Ile17	S1		
		total 16 NOE peaks	total 6 NOE evidences

and S11 is one, between S11 and S21 is three, . . . , etc. The total observed NOE for the assignment of S9-S4-S11-S21-S8-S14-S23-S1 to Ala10-Phe11-Asp12-Tyr13-Ser14-Lys15-Arg16-Ile17 is 16. TSA was designed to keep the assignment with the greatest number of observed NOE correlations. Two things must be noticed here. First, the original version of TSA [25] does not discriminate NOE peaks. In other words, all the NOE peaks are considered to have the same contribution in terms of interresidue correlations. The fact that backbone NOE peaks such as $d_{\alpha N}(i, i+1)$ and $d_{NN}(i, i+1)$ are more important in establishing sequential connectivity than NOE between side chain protons is not taken into consideration. In the commercial version of TSA [64], which is bundled into a resonance assignment package called CAPRI, sequential NOE peaks do receive higher weights than side chain NOE peaks. The second feature of TSA is that it allows the absence of NOE connections in an assignment. In Table 2.9, for example, no NOE correlation between spin system S14 and S23 is observed. As missing data arising from spectral overlap or incomplete peak picking procedures is not rare in protein NMR, it is dangerous to discard the entire assignment for lacking of one evidence of NOE cross peak. Hence TSA permits the absence of NOE connection in order not to lose any potential assignment.

In the situation that two or more assignments have the same number of total NOE peaks,

other rules are necessary to pick up the best assignment. Information which hasn't been used to this point is the mathematical similarities, obtained from equation 2.17. Suppose an assignment maps spin system S_i to residue R_i, S_{i+1} to residue R_{i+1}, ... , S_{i+N-1} to residue R_{i+N-1}. The similarity between S_i and R_i is $v(i)$. The TSA similarity parameter for the above assignment can be defined as

$$V = \sqrt[N]{v(i) \times v(i+1) \times \cdots \times v(i+N-1)} \tag{2.19}$$

If more than one assignment has the same number of NOE peaks, their TSA similarities are calculated using equation 2.19. The assignment having the greatest TSA similarity remains while the rest are discarded.

If the above two criteria are not sufficient to resolve the best assignments, TSA is able to measure the chemical shift deviation between observed NOE cross peaks and the corresponding spins in the spin systems. For example, an NOE peak is found between spin system S_i and S_j. S_i has five spins: i_1, i_2, i_3, i_4 and i_5. S_j has four spins: j_1, j_2, j_3 and j_4. Suppose the distance between i_1 and j_2 are close enough to produce an NOE cross peak (δ_a, δ_b) where $|\delta_a - \delta_{i_1}|$ and $|\delta_b - \delta_{j_2}|$ are within a proton chemical shift tolerance. Ideally, $|\delta_a - \delta_{i_1}|$ and $|\delta_b - \delta_{j_2}|$ should be zero. TSA defines a parameter to measure the difference between the observed NOE peak (δ_a, δ_b), and their original spins , i_1 and j_2 in this particular case. This parameter is essentially the geometric mean of the two absolute values. The definition of this parameter is described in the following:

Suppose the mth spin in one spin system and the nth spin in another spin system are in close proximity to produce an NOE peak. The observed cross peak in 2D NOESY is (δ_a, δ_b) where δ_a and δ_b are the observed chemical shifts for spin m and n, respectively. Parameter p is defined as

$$p = \sqrt{\frac{|\delta_a - \delta_m|}{T} \times \frac{|\delta_b - \delta_n|}{T}} \tag{2.20}$$

where T is the chemical shift tolerance. For the assignment of an N-residue polypeptide, i.e., the assignment maps spin system S_i to residue R_i, S_{i+1} to residue R_{i+1}, ... , S_{i+N-1} to residue R_{i+N-1}, $(N-1)$ parameters of p can be defined. Smaller parameters indicates better matching between the NOE peaks and their corresponding spins. Therefore, TSA defines an overall NOE

parameter P as

$$P = 1 - \sqrt[N-1]{(p_{i\to i+1})\times(p_{i+1\to i+2})\times\cdots\times(p_{i+N-2\to i+N-1})} \tag{2.21}$$

A larger P corresponds to a better assignment.

To summarize the sequential assignment for an N-residue polypeptide R_i-$\cdots$-R_{i+N-1}, TSA first constructs a "residue to spin-systems" table. In this table residue R_j has C_j candidate spin systems. TSA then exhaustively searches all $\prod_{j=i}^{i+N-1} C_j$ assignment combinations to determine the most probable assignment. The number of the actually traversed assignment combinations is fewer than the estimated one because a single spin system can not appear twice in any assignment. TSA adopts a few criteria to determine the final assignments. First the assignments with the greatest number of total NOE peaks are kept. If more than one assignment has the same number, TSA computes the TSA similarity for each of the assignment using equation 2.19. If this similarity cannot break the tie between the assignments, equation 2.21 is used to further discriminate the assignments.

Having discussed the way TSA selects the most probable assignment, we now investigate the effect of the length of polypeptide chain on the sequential assignment. TSA is designed based on a global optimization assumption. The optimization is conducted on the number of total NOE correlations, the TSA similarity in equation 2.19 and the parameter P in equation 2.21. It is assumed that a better result comes out when a longer protein chain is adopted as the assigning target. In other words, for an N-residue protein, TSA has the highest chance of producing the correct assignment provided that residue 1 to residue N are set to be assigned simultaneously. If the N-residue protein is divided into several segments, for example, residue 1-20, 17-40, 37-N (N > 40, of course) and TSA is conducted over these segmented polypeptides one after another, a local optimization might be reached whereas the global optimization is unable to be reached. Certainly, the computational load is heavy in order to reach the global optimization. For shorter polypeptide segments, the time required to complete the assignment can be significantly shorter.

The implementation of TSA was proved to be effective on a testing run of a 21 residue polypeptide [25]. For this relatively small polypeptide, the order of magnitude of the execution time to assign the entire polypeptide is minutes. However, for bigger proteins, such as the ones

having 70 or more residues, a "permutation explosion" problem makes the execution of TSA exceed an acceptable CPU time limit. To overcome this permutation explosion problem, one can attempt to reduce the number of candidate spin systems of each residue, i.e., the length of each row in the "residue to spin-systems" table. Fewer candidate spin systems implies that fewer assignments need to be traversed. Sometimes it is obvious to manually assign many spin systems. The "residue to spin-systems" table can be manually revised according to all available information (obtained from NMR and/or other sources) so as to reduce the possibility of having the permutation explosion. It is also suggested that TSA can be run segment by segment to save time, although this violates the principle of reaching global optimization. For a 70 residue protein, for example, one can assign residue 3-25, 20-45, 40-70 at three separate runs of TSA, making sure that the overlapped residues are assigned to the same spin systems.

The commercial version of TSA [64], bundled in SYBYL, Tripos Inc., made more revisions in both computational and methodological aspects.

Chapter 3

Determination of Protein Backbone Spin Systems

3.1 Introduction

This chapter reports computer algorithms that can extract a protein's backbone spin systems using heteronuclear 3D NMR. Because many heteronuclear 3D NMR experiments are able to record both intra- and interresidue correlations, the sequential information embedded in the spectra can also be derived at the same time. The algorithms presented in this study are not designed for any specific NMR experiment, so that any general data set can be used. Two sets of 3D NMR experiments are used to demonstrate how the the protein backbone is extracted by the algorithms. The first set of NMR data consists of 3D HNCO, HNCA, HN(CO)CA, HCACO and ^{15}N TOCSY-HMQC. The second set of NMR data is 3D CBCANH. Experimental data from the first set of NMR experiments were used to test the implemented algorithms. The target protein is the calcium loaded N-domain of chicken skeletal troponin-C(residue 1-90). Along with the sequence-specific resonance assignment protocol presented in chapter 5, it is possible to achieve the goal of developing a nearly fully automated resonance assignment package. This package is able to extract backbone spin systems; create dipeptide links from interresidue correlations observed in heteronuclear 3D NMR; obtain spin systems of protein side chain; merge backbone and side chains; identify amino acid types; and, finally, achieve sequence-specific assignment.

3.2 Identification of backbone spin patterns

Many heteronuclear 3D NMR experiments [5] have been designed for assigning backbone resonances of $^{15}N/^{13}C$ isotope enriched proteins. These experiments usually observe correlations between three or more nuclei on a protein's backbone. Both inter- and intraresidue correlations can be recorded therefore making it possible to assign the backbone resonances, along with their sequential connectivities, by applying heteronuclear 3D NMR exclusively.

Before illustrating how to make use of the information provided by 3D NMR experiments, a general description of using computer algorithms to assign NMR cross peaks is discussed here. In general NMR cross peaks from 3D spectra can be represented as $(\delta_i, \delta_j, \delta_k)$ where the three coordinates denote the three chemical shift values. For homonuclear 3D NMR all three coordinates represent proton chemical shifts. For heteronuclear 3D NMR, δ_i, δ_j and δ_k can be proton, carbon or nitrogen chemical shifts. To make use of the 3D NMR data, computer algorithms usually perform the following steps: for a starting peak $P_0(\delta_{i_0}, \delta_{j_0}, \delta_{k_0})$, a search is conducted on the same spectrum or other spectra to find one or more peak $P_1(\delta_{i_1}, \delta_{j_1}, \delta_{k_1})$, $P_2(\delta_{i_2}, \delta_{j_2}, \delta_{k_2})$ ... . $P_n(\delta_{i_n}, \delta_{j_n}, \delta_{k_n})$ from which two resonances are in common with P_0. For example, P_0 and P_1 may have the same resonances in the first two coordinates. That is, two resonances satisfy the relationships of $|\delta_{i_0} - \delta_{i_1}| \leq$ (a pre-defined tolerance) and $|\delta_{j_0} - \delta_{j_1}| \leq$ (another pre-defined tolerance). The next step involves the implementation of a ranking system to distinguish peaks P_1, P_2 ... P_n in such a way that a peak P_m is picked which is the most likely peak to be in the same spin coupling system with P_0. At this stage the target spin system expands its size from three resonances to four. This operation is shown in Figure 3.1. The ranking system usually involves searching for evidence in the way of peaks to confirm the merging of P_0 and P_m. In summary, to extract spin coupling systems out of 3D NMR peaks, computer algorithms must have the following features: (1) the algorithms must be able to merge cross peaks, (2) in order to merge two cross peaks, two of the three coordinates should overlap, (3) to verify the merge, other spectral evidence in the form of cross peaks is required.

The application of heteronuclear 3D NMR to protein backbone assignment is now discussed. Figure 3.2 shows a protein backbone segment. A typical triple resonance heteronuclear 3D NMR

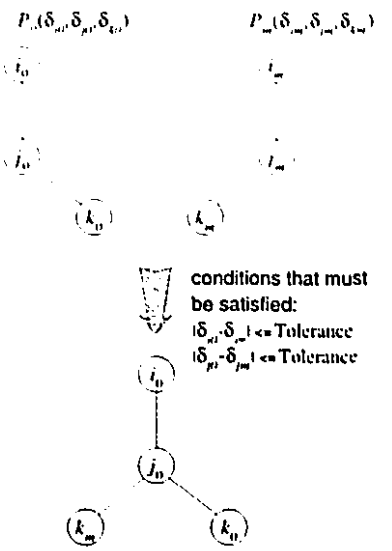


Figure 3.1: A 3D NMR cross peak $P_0(\delta_{i_0}, \delta_{j_0}, \delta_{k_0})$ can be merged with another peak $P_m(\delta_{i_m}, \delta_{j_m}, \delta_{k_m})$ provided that the two conditions shown are satisfied. The merge results in a spin system with four spins $\{i_0, j_0, k_0, k_m\}$.

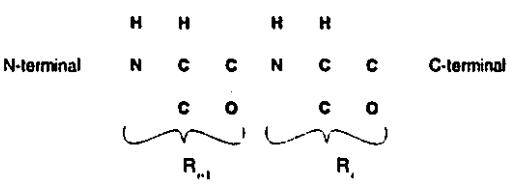


Figure 3.2: The chemical structure of a dipeptide with only the backbone atoms shown.

spectrum observes correlations of three resonances, a proton, a carbon and a nitrogen. For example, the 3D HNCA [65] experiment gives inter- and intraresidue correlations between NH, N and C_α. Some experiments can even observe correlations spanning more than three spins such as CBCANH [66], where inter- and intraresidue C_β, C_α, NH and N correlations are extracted in one single experiment. Since both inter- and intraresidue correlations are available in heteronuclear 3D NMR, individual amino acid residues and sequential connectivities can be obtained simultaneously. Suppose the general merging algorithm described above is applied, which means there must be at least three correlations available to construct the complete backbone spin system of an amino acid. Here complete backbone spin systems are the ones having their N, NH, αH, C_α and CO resonances assigned. Figure 3.3 shows two of the possible combinations from which the

backbone spin systems can be constructed. Note that these three correlations may come from three different experiments. However it is also possible that they all come from the same experiment

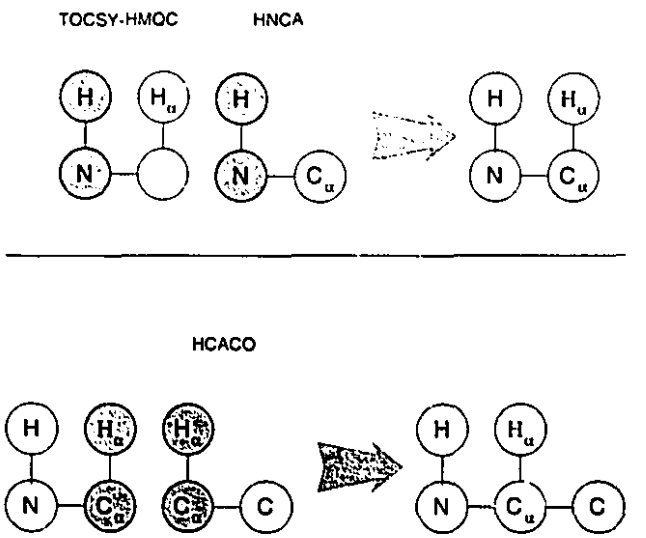


Figure 3.3: The construction of a backbone spin system is shown. Two possible approaches are listed. In the upper one, an HNCO peak, an HN(CO)CA peak and an HCACO peak are merged to form a spin system. In the lower one, a TOCSY-HMQC peak, an HNCA and an HCACO peaks are merged. The filled circles represent the overlapped resonances discovered by the computer algorithm in order to merge peaks.

which combines multiple information into one spectrum.

Recall in Figure 3.2 that the minimum peptide unit having inter- and intraresidue correlations is a dipeptide, i.e., two adjacent amino acid residues. It has been demonstrated that three NMR correlations are required to create an amino acid residue. To create a dipeptide, however, eight instead of six NMR correlations must be observed. The additional two correlations are necessary for establishing the interresidue connectivity. See Figure 3.4 for the pictorial illustration. In the next section the implementation of these ideas is described.

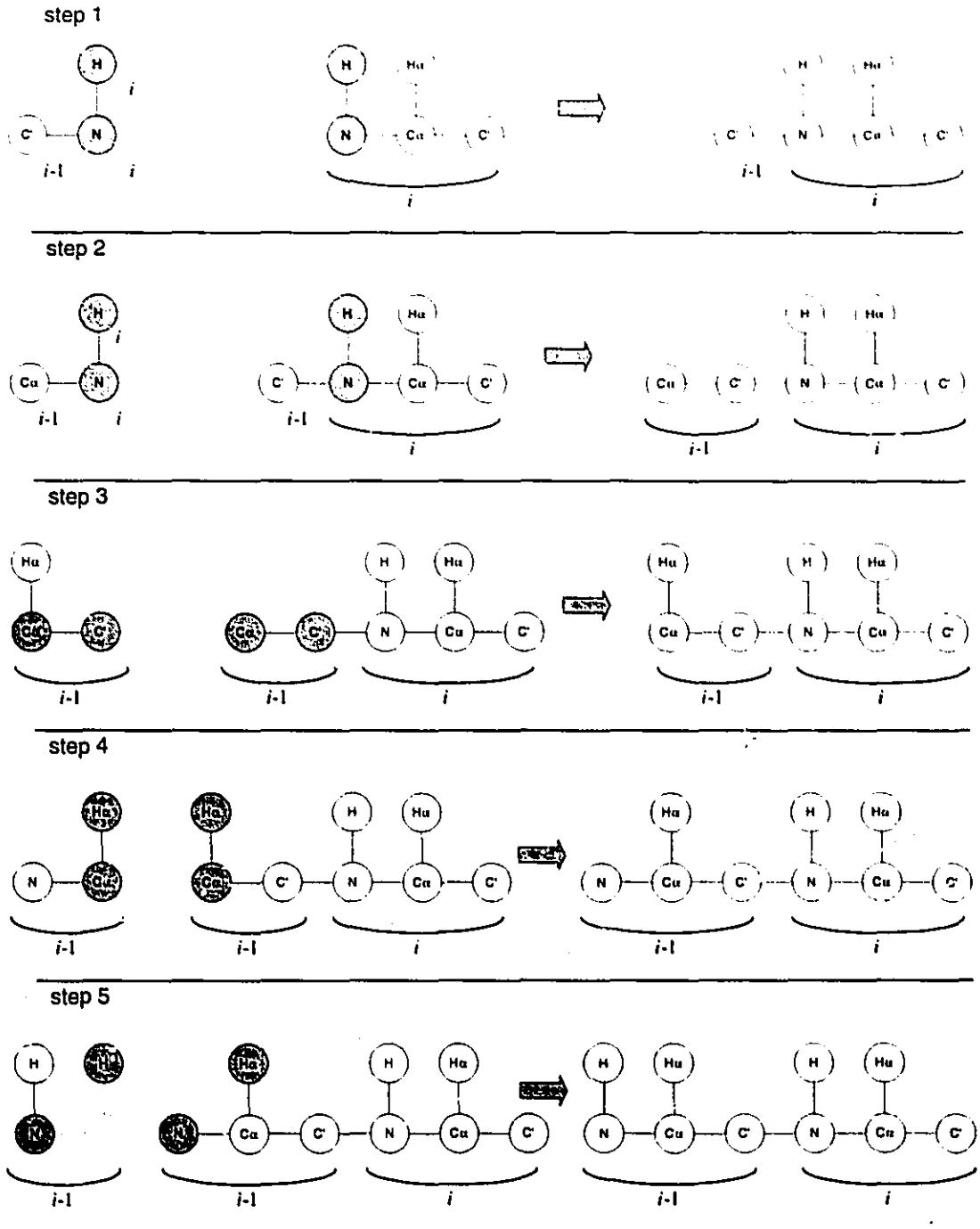
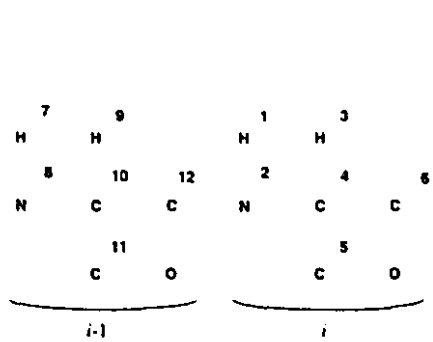


Figure 3.4: The formation of a dipeptide unit. In step 1, residue (i) is a determined spin system. A total of five peaks are required to extend the assignment from residue (i) to residue ($i-1$). Step 1 and 2 involve the interresidue correlations while step 3 to 5 use intraresidue correlations. Note that residue (i) needs three correlations to construct itself. Hence a total of eight correlations are required for the construction of a dipeptide unit.

3.2.1 *Description of backbone assignment strategy*

In this section examples from two sets of heteronuclear 3D NMR spectra are adopted to illustrate the general algorithm discussed in the previous section. Figure 3.5 shows the five 3D NMR experiments used in the first set of spectra.



	Correlated resonances
^{15}N TOCSY HMQC	(1,2,3) (7,8,9)
HNCA	(1,2,4) (1,2,10) (7,8,10)
HN(CO)CA	(1,2,10)
HCACO	(3,4,6) (9,10,12)
HNCO	(1,2,12)

Figure 3.5: Five triple resonance NMR experiments and the nuclei they correlate.

The algorithm for assigning protein backbone was designed in such a way to start the searching from any of the input NMR experiments. The advantage of choosing a specific experiment may sometimes be obvious. For example, spectroscopists may notice that a certain experiment is more sensitive, hence it is reasonable to start the assignment procedure from that experiment. However, it is emphasized that the complete assignment of a dipeptide can be achieved through more than one path. Figure 3.6 describes an eight steps scenario of assigning a dipeptide where cross peaks of 3D HNCO were chosen as the starting experiment. Each of the eight steps involved in the assignment procedure has an associated NMR cross peak. In step 1, the HNCO peak (1, 2, 3) is selected as the initial spin system. In step 2, the ^{15}N-HMQC-TOCSY peak (1, 2, 4), where the first two frequencies are in common with the previous HNCO peak (1,2,3), is added to the spin system. Similarly, by repeating the eight steps, the ten resonance dipeptide (N, NH, αH, C_α, CO)$_{i-1}$ – (N, NH, αH, C_α, CO)$_i$ can be constructed.

In the second example, a single 3D CBCANH experiment was chosen as the input data to illustrate how backbone assignment can be achieved through various approaches. Each of the 3D CBCANH peak may have four interpretations: NH-N-C_α(interresidue), NH-N-C_β(interresidue), NH-N-C_α(intraresidue) and NH-N-C_β(intraresidue). C_α resonances of glycine and C_β

	N	Cα	C'	NH	αH
Residue i-1	9	7	3	10	8
Residue i	1	5	6	2	4

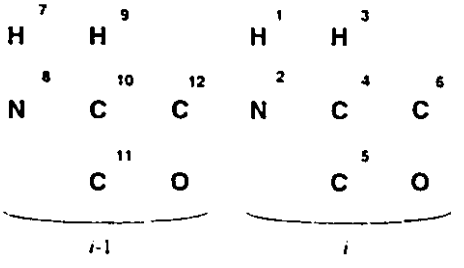
Steps	Experiment involved	cross peak	Results
1	HNCO	(1,2,3)	identify three resonances, 1,2 and 3
2	^{15}N TOCSY-HMQC	(1,2,4)	from 1,2, get resonance 4
3	HNCA	(1,2,5)	from 1,2, get resonance 5
4	HCACO	(4,5,6)	from 4,5, get resonance 6
5	HN(CO)CA	(1,2,7)	from 1,2, get resonance 7
6	HCACO	(3,7,8)	from 3,7, get resonance 8
7	^{15}N TOCSY-HMQC	(8,9,10)	from 7,8, get resonance 9 and 10
8	HNCA	(7,9,10)	same as above

Figure 3.6: The eight steps are listed for assigning the 10 resonances of a dipeptide. Starting from the 3D HNCO cross peak (1, 2, 3), each subsequent step adds one more resonance to the dipeptide, making a 10 resonance spin system.

resonances of all other residues are opposite in phase relative to the other C_α correlations [66]. To resolve the ambiguities between the inter- and intraresidue CBCANH peaks, another 3D experiment, CBCA(CO)NH [67], may be helpful. CBCANH has several advantages over the traditional heteronuclear 3D NMR experiments, for example, HNCA, in that CBCANH is able to distinguish inter- and intraresidue peaks in terms of the peak intensities [66]. Moreover, aliphatic C_α and C_β frequencies appear in opposite phases in CBCANH [66] making it possible to separate the C_α from the C_β in aliphatic region. Figure 3.7 shows a typical dipeptide and its corresponding cross peaks from 3D CBCANH spectrum. Figure 3.8 shows how the assignment procedure using CBCANH is accomplished. Note that additional spectra may be necessary in order to obtain the frequencies of αH, βH and CO.

3.2.2 Implementation of the algorithm

Our algorithm, Dipeptide Backbone Partitioning Algorithm (DBPA), is composed of two parts. In the first part all possible dipeptides are extracted from available spectra. Following this, the individual dipeptides are merged to form polypeptides in the second stage. The algorithm used



	Correlated resonances	
CBCANH	(1,2,10)	$NH_i, N_i, C\alpha_{i-1}$
	(1,2,11)	$NH_i, N_i, C\beta_{i-1}$
	(1,2,4)	$NH_i, N_i, C\alpha_i$
	(1,2,5)	$NH_i, N_i, C\beta_i$
	(7,8,10)	$NH_{i-1}, N_{i-1}, C\alpha_{i-1}$
	(7,8,11)	$NH_{i-1}, N_{i-1}, C\beta_{i-1}$

Figure 3.7: 3D CBCANH experiment provides three inter- and three intraresidue correlations of a dipeptide.

	NH	N	Cα	Cβ
Residue i-1	7	8	10	11
Residue i	1	2	4	5

Steps	cross peak	Results
1	(1,2,10)	Identify three resonances, 1,2 and 10
2	(1,2,11)	from 1,2, get resonance 11
3	(1,2,4)	from 1,2, get resonance 4
4	(1,2,5)	from 1,2, get resonance 5
5	(7,8,10)	from 10,11, get resonance 7
6	(7,8,11)	from 10,11, get resonance 8

Figure 3.8: The six correlations provided by the 3D CBCANH experiment can be used to create a dipeptide with 8 resonances.

in the extraction of backbone spin systems and creation of dipeptides is listed in the following pseudo codes.

```
void CreateDipeptide(PeakList_type, ... )
{
  StartingSpectrum=SelectStartingSpectrum(all of the input spectra);
  for each of the peak in StartingSpectrum {

    dipeptide=AddSpinsToDipeptide(the peak);

    for every possible two spin pair (i,j) combination in above dipeptide
    {
         In the entire spectrum database excluding the starting spectrum,
```

```
        look for peaks (i',j',k), (i',j,k') and (k,i',j')
        which have two frequencies in common with the
        initial spin pair (i,j);

        if many peaks satisfy the above condition
           BestPeak=RankingProcedure(all of the peaks
                    (i',j',k), (i',j,k') and (k,i',j'));

        dipeptide=AddSpinsToDipeptide(BestPeak);

    }
    if the number of spins in this dipeptide has reached ten
                              // (N,NH,αH,Cα,CO) for two peptides
        keep this dipeptide;
  }
}
```

The pseudo code is self-explanatory except for the ranking procedure which is responsible for choosing the most probable peak to be merged into the existing spin system out of many possible candidate peaks. The pseudo codes for this ranking procedure is outlined in the following:

```
peak_type RankingProcedure(const Peak_type *, ... )
{
  //Input: 1. two resonances i0 and j0
  //       2. all 3D NMR peaks with two frequencies in common with
  //          i0 and j0
  //Example:
  // peak 1 (i1, j1, k1) where |i0 − i1| ≤ tolerance, |j0 − j1| ≤ tolerance
  // peak 2 (i2, j2, k2) where |i0 − i2| ≤ tolerance, |j0 − j2| ≤ tolerance
  // peak 3 (i3, j3, k3) where |i0 − i3| ≤ tolerance, |j0 − j3| ≤ tolerance

  //Output: The most likely peak that can be merged with i0 and j0
        define a ranking parameter:
        for peak 1: A1≡1 − √(|i0 − i1| * |j0 − j1|)
        for peak 2: A2≡1 − √(|i0 − i2| * |j0 − j2|)
        for peak 3: A3≡1 − √(|i0 − i3| * |j0 − j3|)

        return peak n (in, jn, kn) with greatest A value;
}
```

The geometric mean $\sqrt{|i_0 - i_n| * |j_0 - j_n|}$ was adopted as the measure of the average deviation between peak n and peak 0. The geometric mean was chosen over the arithmetic mean because the former tends to reduce effects from extremes of large and small values.

DBPA has an option to handle two different searching operations. Both operations can be used in the construction of a dipeptide.

1. Given a dipeptide with m assigned frequencies, DBPA takes two frequencies, δ_i, δ_j, where $i, j \in \{1, 2, \cdots, m\}$ and $i \neq j$, and searches a candidate peak having two frequencies overlapped with δ_i, δ_j in the spectrum database. Suppose the third frequency of the candidate peak is δ_k. δ_k will be merged into the dipeptide and result in a dipeptide with $m+1$ assigned resonance. If many candidate peaks are found, a ranking system is implemented in DBPA to select a peak from the many candidates and merge this peak to the dipeptide. Alternatively, a user can tell DBPA to make a replication of the dipeptide for each of the candidate peaks and merge that candidate peak to the replicated dipeptide.

2. Given a dipeptide with m assigned frequencies, DBPA takes two frequencies, δ_i, δ_j, where $i, j \in \{1, 2, \cdots, m\}$ and $i \neq j$, and searches two candidate peaks in the input spectrum database. The first candidate has frequency δ_i and two other frequencies, suppose they are denoted as δ_k and δ_l. The second candidate peak has frequency δ_j, δ_k and δ_l. Note that two frequencies are overlapped between the two candidate peaks. DBPA would merge resonance δ_k and δ_l into the dipeptide. This procedure results in a dipeptide with $m+2$ frequencies.

These operations can both be seen in Figure 3.6. The first operation is used in step 1 to step 6 while the second operation is used in step 7 and 8.

Once the dipeptide database has been created, it is possible to merge these dipeptides into longer chains such as tripeptide, tetrapeptide ... etc. For example, a dipeptide $R10 - R28$ can be merged with $R28 - R35$ to make a tripeptide $R10 - R28 - R35$ where R_i simply indicates this is the ith residue retrieved by DBPA. The aim of constructing these polypeptides is to identify the amino acid type information of their component residues thereby mapping them to the primary sequence of the protein. The probability that an "amino-acid-type-recognized" polypeptide occurs only once in a protein depends on the length of the polypeptide [3]. A longer polypeptide has a higher probability of being mapped uniquely to its corresponding primary sequence. The algorithm PGA(Polypeptide Generating Algorithm)listed below shows how dipeptides can be merged together to form polypeptides. Details concerning the amino acid type recognition and primary sequence mapping are discussed in chapter 5.

```
void CreatePolypeptide(Dipeptide_type, ... )
{
  //Input: a set of dipeptides
  //Output: polypeptides
  //Examples:
  //          R3 – R5
  //          R5 – R29
  //          R29 – R18
  //          R18 – R16
  //          R18 – 38
  //          produce output R3 – R5 – R29 – R18 – R16 and R3 – R5 – R29 – R18 – R38

      for each dipeptide in the input  {
          copy this dipeptide into the polypeptide chain P;
          for each dipeptide in the input {
              if this dipeptide can be merged with chain P  {
                   push this dipeptide into stack S;
              }
          }
          append(P, S);     // append() function will increase the length
                                //  polypeptide P
      }
}

void append(Polypeptide_type p, Stack_type s)
{
  while stack s is not empty   {
    pop a dipeptide element out of s then merge it with polypeptide p;
      for each available dipeptide in the input of CreatePolypeptide()  {
        if this dipeptide can be merged with polypeptide p  {
            push this dipeptide into stack s2;
        }
      }
                          // An empty s2 implies that there is no
                          // dipeptide can be merged with polypeptide p
                          // If this is a nonempty s2, call append()
                          // recursively with argument polypeptide p
                          // and stack s2
      if stack s2 is empty  {
          store polypeptide p into output list;
      } else {
          append(p,s2);
      }
  }
}
```

3.2.3 *Applications and Results*

All of the algorithms are implemented in C computer language and were tested on a 90 residue globular protein. Figure 3.9 is a brief flowchart illustrating the relationships between the input data and various algorithms. The experimental data were provided by University of Alberta. All spectra were obtained on a Varian Unity 600 NMR spectrometer operating at 30 °C [68]. The sample protein is the calcium-loaded regulatory N-domain of chicken skeletal troponin-C (NTnC, residue 1-90). Uniformly enriched ^{15}N and ^{13}C NTnC were also prepared. Available heteronuclear 3D NMR experiments include 3D HNCA, 3D HNCO, 3D HNCOCA, 3D HCACO, 3D ^{15}N TOCSY-HMQC and NOESY-HMQC.

Cross peaks were automatically picked from the transformed 3D spectra using the CAPP peak picking program [56]. The CAPP program is run at the noise level, therefore a number of false peaks are unavoidably picked. Many of these false peaks can be removed by filtering the peak lists of the 3D spectra through high signal-to-noise 2D spectra [68]. The final peak lists were given to the authors by B. Sykes at the University of Alberta [68].

The 3D HNCO peak list contains 135 cross peaks compared with about 90 peaks predicted for a 90 residue protein. The 3D HCACO peak list has 125 peaks, 3D HNCA has 242 peaks, which include both interresidue $NH_i - N_i - C\alpha_{i-1}$ and intraresidue $NH_i - N_i - C\alpha_i$ peaks. 3D HN(CO)CA has 135 peaks and ^{15}N TOCSY-HMQC consists of 141 peaks. All peak lists were input into DBPA as shown in Figure 3.9.

To process peaks coming from different spectra, various tolerance values are introduced since the spectra were not perfectly aligned. The tolerance value for comparing proton frequencies was chosen to be 0.05 ppm. For the rest of the nuclei, tolerance values are 0.40 ppm for nitrogen, 0.30 ppm for CO and 0.47 ppm for C_α. These tolerance values are adjustable based on user's experience and spectra quality. The algorithm DBPA produced 161 dipeptides which in turns was input into the algorithm PGA. In PGA, the 161 dipeptides were compared against each other to eliminate redundant spin systems, finally resulting in 98 unique backbone spin systems. Theoretically 90 spin systems should be observed for the 90-residue NTnC.

According to Figure 3.6, eight 3D NMR cross peaks are required to construct a dipeptide.

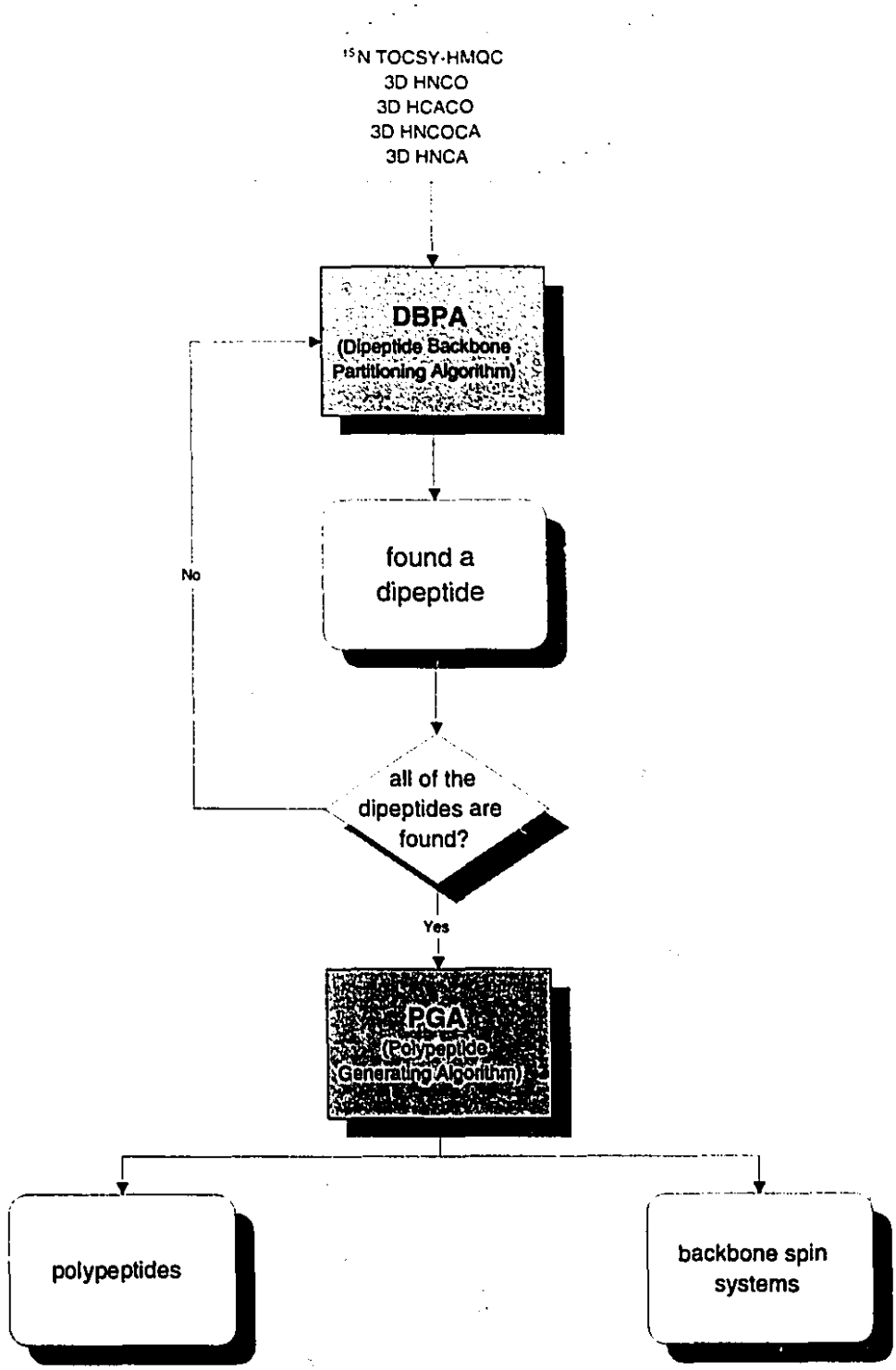


Figure 3.9: The flow diagram of the partitioning algorithm.

In practice, it is unlikely to obtain NMR data set without missing peaks. Hence, the ability of handling missing peaks becomes an important criterion for automated assignment tools. In the Troponin-C spectral data, thirty-four of the 86 amino acid residues have at least one missing peak. In the first run of DBPA we defined a successfully assigned dipeptide as the one having all of the 10 resonance identified. This is a strict condition. As a result, the above 34 residues were not assigned in the first run of DBPA. The successful assignment percentage is approximately 60% (see Table 3.1).

Table 3.1: The extracted residues of protein NTnC using Dipeptide Backbone Partitioning Algorithm. See text for the definition of various runs of DBPA.

Observed residues in the first run of DBPA	Observed residues in the second run of DBPA	Observed residues in the third run of DBPA	Residues unable to assign without human inspection of data
			D5
Q6			
Q7			
A8			
E9			
A10			
R11			
A12			
F13			
L14			
S15			
E16			
E17			
M18			
I19			
A20			
E21			
F22			
K23			
A24			
A25			
		F26	
		D27	
F29			
D30			
A31			
D32			
G33			
G34			
			G35
			D36

			I37
			S38
		T39	
			K40
			E41
L42			
		G43	
		T44	
V45			
M46			
		R47	
		M48	
			L49
			G50
Q51			
N52			
			P53
			T54
K55			
E56			
	E57		
	L58		
D59			
A60			
I61			
		I62	
			E63
		E64	
			V65
			D66
			E67
D68			
	G69		
	S70		
			G71
			T72
I73			
D74			
			F75
		E76	
E77			
		F78	
		L79	
V80			
M81			
M82			
V83			
R84			
Q85			

M86			
K87			
E88			
D89			
A90			

Perhaps the best way to demonstrate how DBPA overcomes the peak missing problem is to use the example shown below. Residue E57 of Troponin-C misses a 3D NMR peak, the HCACO peak (αH, C_α, CO). HCACO and HNCO are the two experiments observing CO frequencies. While HNCO peaks, (CO($i-1$), HN(i), N(i)), in general determine the CO resonance of the first residue of a dipeptide, lack of HCACO peak makes DBPA unable to determine the CO frequency of E57 in dipeptide E56-E57. As a result, E56-E57 remained in the category of unassigned dipeptides in the first run of DBPA on Troponin-C data set because its CO frequency has not been determined yet. In order to identify E57, users have an option to relax the 10-resonance definition of a dipeptide. In other words, DBPA can think of E57 as the second residue of dipeptide E56-E57 even though E57 has an undetermined resonance. The relaxation of the definition of dipeptides must be conducted carefully, because the possibility of receiving multiple assignments for a dipeptide is increasing due to the fact that only seven instead of eight peaks are required for identifying a dipeptide. A compromised approach is to take out all the used peaks, i.e., peaks that have been used by DBPA to construct dipeptides in the first run, before the second run of DBPA. Using this approach, DBPA successfully assigned additional four dipeptides, E56-E57, E57-E58, D68-G69 and G69-S70 in the second run. Note that the CO frequencies of these residues are absent. Proper human assistance could help to retrieve the absent frequencies.

Sometimes a single missing peak may lead to two unassigned resonances in a dipeptide. Using Troponin-C as an example, the missing ^{15}N TOCSY-HMQC peak, (N, NH, αH), of F78 makes DBPA failing to determine the αH frequency of F78 in dipeptide E77-F78. The missing αH results in a missing CO of F78 because the CO frequency is supposed to comes from peak (αH, C_α, CO). To extract a dipeptide with two missing frequencies, in this example CO and αH, one needs to further relax the definition of a dipeptide, that is, eight assigned resonance can be considered as an assigned dipeptide in the third run of DBPA. Using the Troponin-C data, additional 12 dipeptides were determined after the third run of DBPA. This makes the percentage

of assigned residues to about 79% (67 of 85).

Eighteen residues remain unassigned after three runs of DBPA. Each of these residues has two or more missing peaks. Before conducting appropriate manual inspection on the spectral data, it is difficult to assign more residue at this stage.

3.3 Discussion

Computer algorithms are presented to automate the resonance assignment of protein backbone using heteronuclear NMR. The principle and implementation of the algorithm DBPA (Dipeptide Backbone Partitioning Algorithm) is described. The differences between DBPA and traditional heteronuclear NMR assignment strategy are illustrated as the following.

DBPA and manual assignment share a common strategy, namely they both make use of the scalar magnetization transfers through peptide bonds instead of using the through-space dipole-dipole interaction to establish the sequential connectivities. Figure 3.10 shows a typical manual assignment path from residue(i) to residue($i-1$) using heteronuclear 3D NMR. Initially, a 3D HNCO cross peak HN(i)-N(i)-CO($i-1$) was selected. Keeping the frequency of CO($i-1$) in mind, searches can be conducted on the 3D HCACO spectrum to locate a cross peak CO($i-1$)-Cα($i-1$)-αH($i-1$). Once the αH($i-1$) frequency has been determined, the following search on ^{15}N TOCSY-HMQC reveals the resonances of NH($i-1$) and N($i-1$). This terminates one iteration where seven resonances (NH(i), N(i), CO($i-1$), Cα($i-1$), αH($i-1$), NH($i-1$), N($i-1$)) are found. Figure 3.11 gives the summary of the procedures. Note that each search was performed based on the knowledge of one frequency. For example, both Cα($i-1$) and αH($i-1$) were found on the 3D HCACO spectrum based on the known CO($i-1$) chemical shifts. However, ambiguities resulting from overlapped CO($i-1$) may increase the difficulties of applying such manual assignment strategy. In the DBPA algorithm, the chance of the above overlapping is reduced by using two known frequencies to determine one unknown frequency. As shown in Figure 3.11 and 3.12, both of the NH(i) and N(i) contribute to the determination of the Cα($i-1$) using 3D HN(CO)CA. Moreover, in cases that it is ambiguous to determine the resonances of NH($i-1$) and N($i-1$) from the known frequency of Cα($i-1$) using 3D HNCA

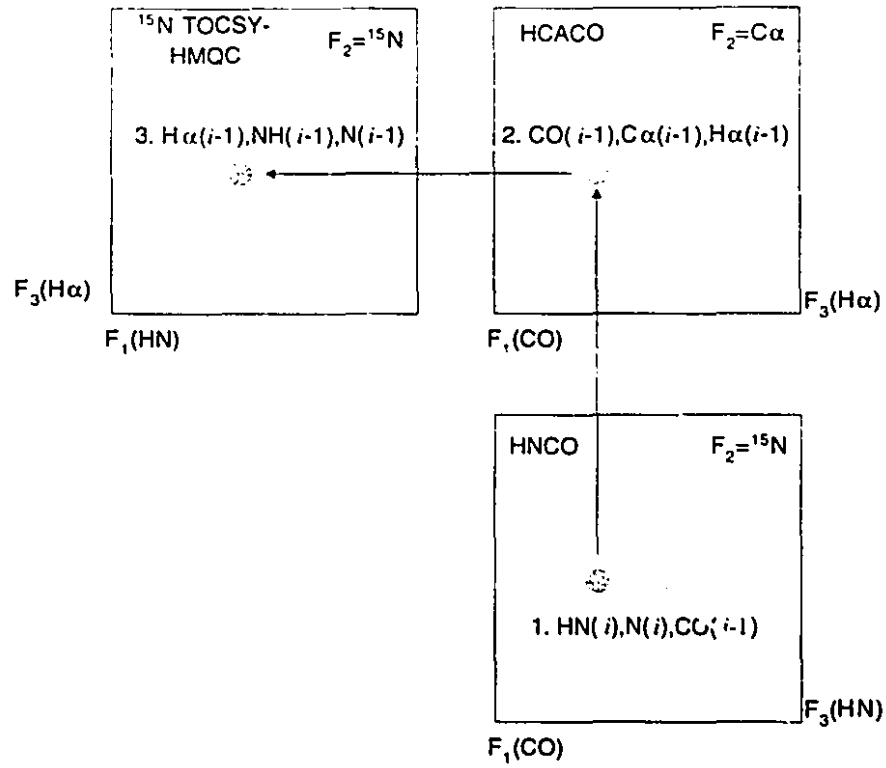


Figure 3.10: Schematic illustration of using three triple resonance correlation experiments to obtain the sequential assignment. Starting from peak 1 in 3D HNCO, N and NH of residue(i), and CO, Cα, αH, NH and N of residue($i-1$) can be obtained once the three peaks are merged. This is a typical strategy of manual assignment.

spectrum, DBPA automatically attempts to find another path to confirm the assignment of the NH($i-1$) and N($i-1$). In this particular case, DBPA looks for the frequencies of NH($i-1$) and N($i-1$) from the known αH($i-1$) using 3D ^{15}N TOCSY-HMQC. Figure 3.12 shows the connectivities determined by DBPA. Comparing Figure 3.12 to Figure 3.10, it is obvious that computer programs are good at taking more NMR evidences to resolve the possible ambiguities.

DBPA offers an option which affects the number of the output spin systems. The basic operation DBPA performs is searching. With respect to each starting peak, DBPA looks for all the candidate peaks which can be merged with the starting peak in available NMR spectra. In section 3.2 we described a ranking parameter using which it is possible to select the best candidate. The ranking procedure measures the chemical shift difference between the candidate and the starting peaks so as to decide which candidate is the most likely one to be partitioned with the starting

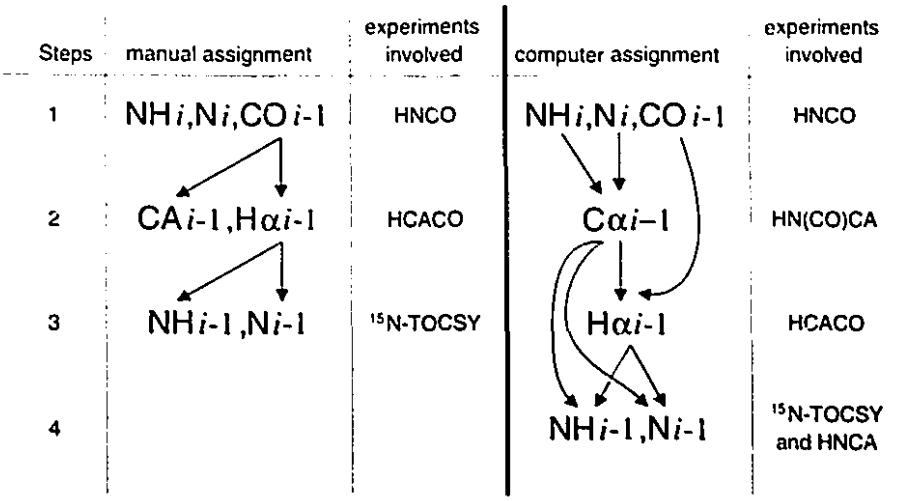


Figure 3.11: Comparison of the manual and automated assignment strategies. On the left, the manual assignment strategy assigns 7 resonances using three cross peaks (see Figure 3.10). On the right, DBPA assigns 7 resonances using five cross peaks (see Figure 3.12).

peak. However, correct merge doesn't always occur at the *best* candidate. Consider the following example: five candidates peaks, $(\delta_{i_1}, \delta_{j_1}, \delta_{l_1})$, $(\delta_{i_2}, \delta_{j_2}, \delta_{l_2})$, ... , $(\delta_{i_5}, \delta_{j_5}, \delta_{l_5})$, were observed and about to be merged with the starting peak $(\delta_{i_0}, \delta_{j_0}, \delta_{l_0})$. Suppose $(\delta_{i_2}, \delta_{j_2}, \delta_{l_2})$ is the one that should be partitioned into the spin system with the starting peak but coincidently the chemical shift differences $|\delta_{i_5} - \delta_{i_0}|$ and $|\delta_{j_5} - \delta_{j_0}|$ are smaller than $|\delta_{i_2} - \delta_{i_0}|$ and $|\delta_{j_2} - \delta_{j_0}|$, respectively. As a result, the best candidate will be determined as the fifth candidate instead of the correct one, the second candidate peak. To avoid this situation, DBPA implements an option by enabling which all of the above five candidates would be kept. In other words, the capability of choosing the best candidate will be disabled. The implication of this option is that there will be five independent four-spin systems, $\{\delta_{i_0}, \delta_{j_0}, \delta_{k_0}, \delta_{l_1}\}$, $\{\delta_{i_0}, \delta_{j_0}, \delta_{k_0}, \delta_{l_2}\}$, ... , $\{\delta_{i_0}, \delta_{j_0}, \delta_{k_0}, \delta_{l_5}\}$. Only one of the above merge is correct whereas the correct one is not necessary the one having the best partitioning parameter. It should be pointed out that once the multiple merging option is enabled, it affects all merging steps. The number of output spin systems could grow rapidly. Users should be able to determine when the option needs to be enabled depending upon the overlaps of the NMR spectra.

DBPA is not designed for specific NMR experiments. It can process many combinations of triple resonance heteronuclear 3D NMR experiments and give the backbone resonance assign-

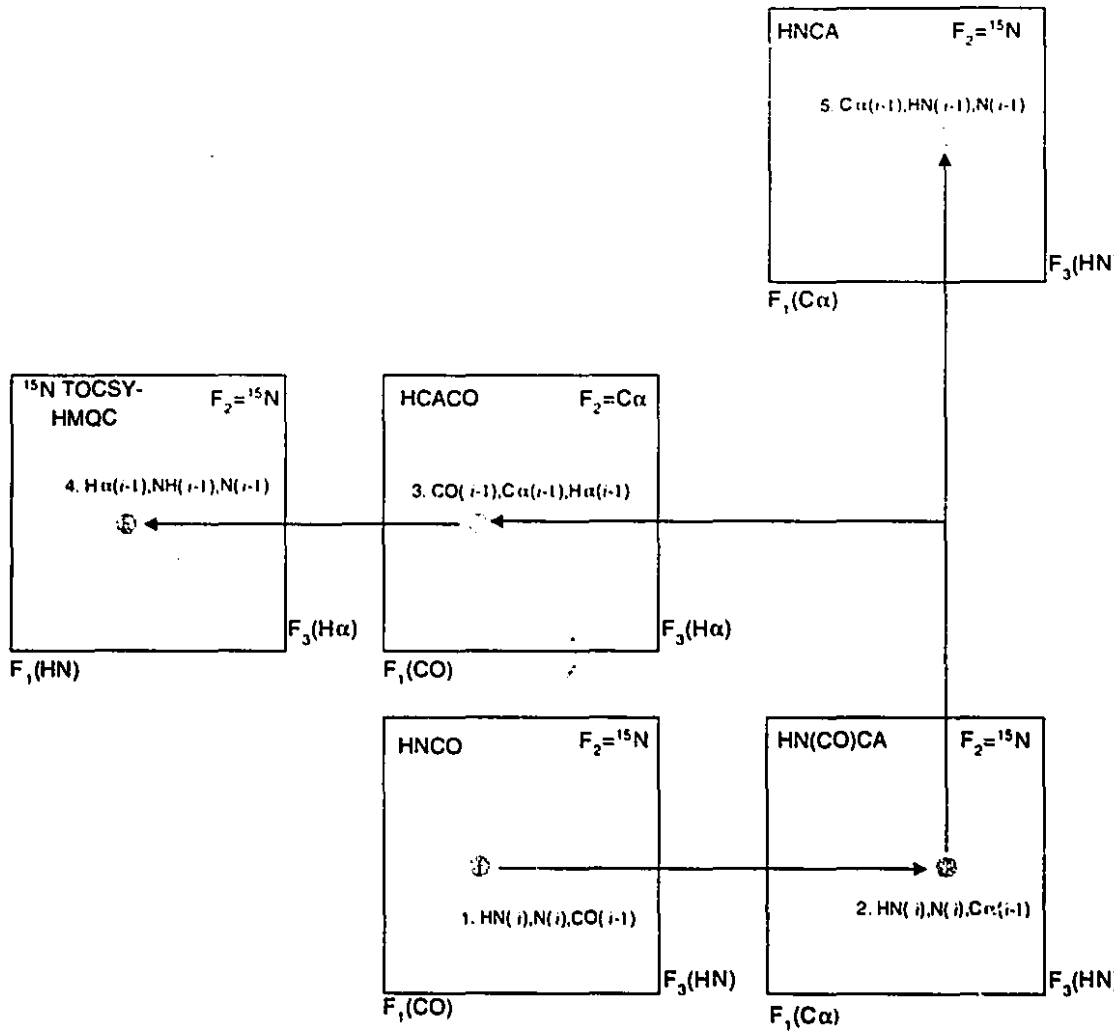


Figure 3.12: Schematic illustration of using five 3D triple resonance correlation experiments to obtain the sequential assignment. Seven resonances (NH and H of residue(i), N, NH, αH, Cα, CO of residue($i-1$)) can be obtained. This is the assignment path our computer algorithm uses.

ment. However, it is necessary to supply sufficient information to DBPA in order to accomplish complete dipeptide assignments. For example, a single 3D HNCO spectrum does not provide enough information to assign a dipeptide because only three resonances, NH_i, N_i, and CO_{i-1}, can be determined. Similarly, a 3D HNCO and a HNCA, giving four resonances, NH_i, N_i, $C\alpha_i$ and CO_{i-1}, don't provide enough information, either. Apparently we need to determine whether an NMR data set have sufficient information to assign the 10 resonances of a dipeptide. A simple algorithm was designed to verify the completeness of input NMR data set. The algorithm is listed

as follows:

```
void VerifyCompleteness(Heteronuclear3DNMR_type, ...)
{
//
// Input : All available heteronuclear 3D NMR spectra. Required
//         information includes the resonances observed in the experiment
//         and the correlations between the resonances.
//
// Example: For 3D HNCO spectrum, the input information is
//          ( NH_i, N_i, CO_{i-1} ).
//
// Output: All possible permutations of the input NMR experiments leading
//         to a complete dipeptide backbone assignment, i.e.,
//         ( N, NH,α H, Cα, CO)_{i-1}--( N, NH,α H, Cα, CO)_i

  suppose the number of input NMR experiments is N;
  compute all possible N! permutations for the N NMR experiments;
  for each of the permutation {
     fill the three observed resonances of the first experiment into an
     empty dipeptide backbone;

     for each of the remaining N - 1 experiments in this permutation {
        if two and only two of the three observed resonances overlaps
        with any other two resonances in the dipeptide backbone
            add the third resonance of this experiment into the
            dipeptide backbone;
        if all of the 10 resonances of the dipeptide backbone are filled
            a complete permutation is found, break the inner loop;
     }

     if the dipeptide backbone are filled with 10 resonances
        output this permutation;
     else
        this permutation does not provide sufficient information to assign
        10 backbone resonances;
  }
}
```

Essentially this approach follows the same concept of DBPA, namely, two overlapped resonances coming from two 3D NMR cross peaks confirm the merge of these two peaks. In the beginning all possible permutations of the supplied NMR experiments are computed. For a data set containing N spectra, there are $N!$ permutations. Here a permutation means a sequence of using NMR spectra to construct dipeptides. These $N!$ permutations are then examined to determine whether sufficient NMR correlations for dipeptide construction are present. Consider the following five NMR spectra: 3D HNCO, HNCA, HN(CO)CA, HCACO, ^{15}N TOCSY-HMQC. A total of $5! = 120$ possible

ways exist in terms of applying the five spectra sequentially. Not all the permutations result in a complete assignment of a dipeptide. It is possible that none of them provide sufficient information. Given a data set containing N NMR spectra, our program extracts all the permutations that produce complete dipeptides, i.e., all of the 10 resonances of a dipeptide are determined. Note that there might be more than one successful permutation. Currently our algorithm does not distinguish those permutations. In other words, the algorithm does not evaluate the permutations and determine the best assignment approach. This is simply due to the complexity of the information provided by the variety of NMR experiments. As new experiments are invented quickly, it is neither possible nor necessary to allow the algorithm to assess individual NMR experiments. This task is left to be done manually.

We mentioned that it is possible that more than one permutation of input NMR experiments can be adopted by DBPA to assign the dipeptide resonances. Here an example is given to illustrate two different approaches of using a five-experiment data set. Available experiments are 3D HNCO, HNCA, HN(CO)CA, HCACO and ^{15}N TOCSY-HMQC. Both of the assignment approaches start at a HNCO cross peak. The first approach assigns dipeptides from C-terminal to N-terminal. A total of 8 peaks are involved. The second approach assigns dipeptides in the reverse order, namely from N-terminal to C-terminal, and involves 9 peaks. Figure 3.13 lists the assignments and all of the involved peaks in the order they are used. The reported result in section 3.2.3 were produced using the first approach of Figure 3.13 simply because fewer involved peaks means less chance of having missing peaks. DBPA has an option to control the assignment direction. As illustrated in Figure 3.13 where dipeptides can be assigned from C- to N-terminal (residue(i) to ($i-1$)) or from N- to C-terminal (residue($i-1$) to (i)). Users can select either one as the assignment approach.

In this chapter we introduced the procedure that requires a minimum of eight correlations to assign the backbone resonances of a dipeptide. The minimum number is determined based on the fact that each residue's backbone has five resonances (N, αH, Cα, NH, CO), thus a dipeptide is composed of 10 resonances. Suppose these 10 resonances are denoted as $(a_1, b_1, c_1, d_1, e_1)$ and $(a_2, b_2, c_2, d_2, e_2)$ where the first five numbers represent the resonances of residue 1 while the last five numbers are the resonances of residue 2. One of the possible combinations of the eight necessary correlations are $\{a_1, b_1, c_1\}$, $\{b_1, c_1, d_1\}$, $\{c_1, d_1, e_1\}$, $\{d_1, e_1, a_2\}$, $\{c_1, d_1, b_2\}$, $\{a_2, b_2, c_2\}$,

```
/*** assignment from R(i) to R(i-1) ***
   N     CA     CO     NH     HA
124.40 58.60 178.30 7.66 4.01
114.00 53.00 177.90 8.10 4.62
step1: (114.00,178.30,8.10)  -- [N(i),CO(i-1),NH(i)] of HNCO
step2: (114.00,58.60,8.10)   -- [N(i),CA(i-1),NH(i)] of HN(CO)CA
step3: (114.00,8.10,4.62)    -- [N(i),NH(i),HA(i)] of TOCSY-HMQC
step4: (114.00,53.00,8.10)   -- [N(i),CA(i),NH(i)] of HNCA
step5: (58.60,178.30,4.01)   -- [CA(i-1),CO(i-1),HA(i-1)] of HCACO
step6: (53.00,177.90,4.62)   -- [CA(i),CO(i),HA(i)] of HCACO
step7: (124.40,58.60,7.66)   -- [N(i-1),CA(i-1),NH(i-1)] of HNCA
step8: (124.40,7.66,4.01)    -- [N(i-1),NH(i-1),HA(i-1)] of TOCSY-HMQC


/*** assignment from R(i) to R(i+1) ***/
   N     CA     CO     NH     HA
124.40 58.60 178.30 7.66 4.01
114.00 53.00 177.90 8.10 4.62
step1: (124.40,166.60,7.66) -- [N(i),CO(i-1),NH(i)] of HNCO
step2: (124.40,7.66,4.01)   -- [N(i),NH(i),HA(i)] of TOCSY-HMQC
step3: (124.40,58.60,7.66)  -- [N(i),CA(i),NH(i)] of HNCA
step4: (58.60,178.30,4.01)  -- [CA(i),CO(i),HA(i)] of HCACO
step5: (114.00,58.60,8.10)  -- [N(i+1),CA(i),NH(i+1)] of HN(CO)CA
step6: (114.00,178.30,8.10) -- [N(i+1),CO(i),NH(i+1)] of HNCO
step7: (114.00,8 10,4.62)   -- [N(i+1),NH(i+1),HA(i+1)] of TOCSY-HMQC
step8: (114.00,53.00,8.10)  -- [N(i+1),CA(i+1),NH(i+1)] of HNCA
step9: (53.00,177.90,4.62)  -- [CA(i+1),CO(i+1),HA(i+1)] of HCACO
```

Figure 3.13: Example showing two approaches for the assignment of a dipeptide.

$\{b_2, c_2, d_2\}$, $\{c_2, d_2, e_2\}$. In this case, correlation $\{a_1, b_1, c_1\}$ and $\{b_1, c_1, d_1\}$ give rise to four resonances, a_1, b_1, c_1 and d_1. Similarly, resonance e_1 can be determined by merging $\{b_1, c_1, d_1\}$ and $\{c_1, d_1, e_1\}$. Repeating this procedure, all the 10 resonance can be determined. It is generally not easy to declare a minimum set of required NMR experiments for automated assignment strategy like the one discussed here, nor is it necessary. There are many different NMR experiments, each provides one or more inter- or intraresidue correlations. What is relevant here is the minimum number of correlations between the nuclei, not the number of NMR spectra.

3.4 Summary of the spin system determination from triple resonance NMR

Algorithms are proposed to automate the resonance assignment of protein backbone using through-bond interresidue correlations. DBPA(Dipeptide Backbone Partitioning Algorithm) merges cross peaks among available NMR spectra and extracts the backbone spin systems. Ev-

ery merge is confirmed by two pieces of evidences, i.e., two overlapped frequencies of a 3D cross peak. To fulfill this requirement, six intraresidue and two interresidue correlations are needed to construct the spin systems of a dipeptide. Once all the possible dipeptides are obtained, PGA(Polypeptide Generating Algorithm) links the dipeptides to form polypeptides. Each of the polypeptides in turn can be manually or automatically assigned to the primary sequence of the protein. DBPA can be applied to many different types of NMR experiments. The five-experiment set (3D HNCO, HNCA, HN(CO)CA, HCACO and ^{15}N TOCSY-HMQC) along with 3D CBCANH were chosen to demonstrate the generality of DBPA.

3.5 Using double resonance heteronuclear 3D NMR

3.5.1 Introduction

TOCSY type NMR experiments play important roles in protein resonance assignment. TOCSY cross peaks have absorption peak shape, thereby simplifying the peak identification and picking procedure. Most available automated peak picking software can process TOCSY type spectra while some of them have difficulties processing COSY type experiments. TOCSY experiment observes neighboring as well as distant correlations between protons. In other words, the TOCSY spectrum generally consists of all the information available on the COSY spectrum. In practice, by setting a short mixing time, the 2D homonuclear TOCSY spectrum provides almost the same cross peak information 2D DQF-COSY does. Moreover, by setting an appropriate long mixing time, the TOCSY experiment is able to provide long range cross peaks between amide proton and αH, βH, γH, even δH.

In this section we demonstrate the way a single TOCSY type experiment can be used to construct amino acid spin systems. A computer algorithm called NCPA(Nitrogen Constrained Partitioning Algorithm) was proposed. The implementation of the algorithm was tested on a ^{15}N TOCSY-HMQC spectrum of the 90-residue protein NTnC. Once the amino acid spin systems are created by NCPA, the amino acid pattern recognition program determines the amino acid types of observed spin systems. In the final stage, the sequential assignment protocol described in chapter 5

takes responsibility of placing the spin systems within protein primary sequence.

Algorithm NCPA shares the same assignment strategy as its predecessor, CPA (Constrained Partitioning Algorithm) [23, 24]. The main task CPA performs is to merge as many NMR cross peaks as possible in order to form spin systems. Each merging operation has to be strictly confirmed by constraints, which could be another cross peak in the same spectrum or in a supplementary spectrum. Spin systems are created in the form of graphs, a combination of nodes (spins) and edges (cross peaks) and represented by adjacency lists. The extracted graphs contain information of chemical shifts as well as inter-resonance connections which make the design of an automated algorithm for amino acid type identification easier. There is, however, a major difference between NCPA and CPA. CPA takes COSY as its primary input spectrum while NCPA takes TOCSY spectrum as the only input. As is seen in the next section, in principle ^{15}N TOCSY-HMQC provides all correlations between side chain protons and amide NH. Correlations between side chain protons themselves are not observed in the spectrum, however. Spin systems derived by NCPA are therefore different from those derived by CPA due to the lacks of correlations between side chain protons. NCPA's spin systems requires a revised database of the standard amino acid patterns to carry out the spin pattern recognition.

3.5.2 Concept

For larger proteins, the NH-αH fingerprint region, where most resonance assignment strategies start from, may have severe overlap of multiple cross peaks. To solve this problem, Marion *et al.* [69] proposed two 3D NMR experiments, the ^{1}H-^{15}N TOCSY-HMQC and NOESY-HMQC, to provide the through-bond and through-space connectivities necessary for the sequential assignment procedure. In the above experiments, the ^{1}H and ^{15}N resonances are recorded in F_1 and F_2 dimensions, respectively. The NH resonances are recorded in F_3 dimension. The $F_1(^{1}\mathrm{H}) - F_3(^{1}\mathrm{H})$ projection corresponds to the $F_1(^{1}\mathrm{H}) - F_2(^{1}\mathrm{NH})$ region of a regular ^{1}H-^{1}H NOESY or TOCSY spectrum and thus ensures that the NH-αH connectivities can be easily observed. Figure 3.14 shows that residues having different ^{15}N chemical shifts appear on different $F_1 - F_3$ planes. All protons within an amino acid residue are observed in a straight line intersected with the F_3 axis at

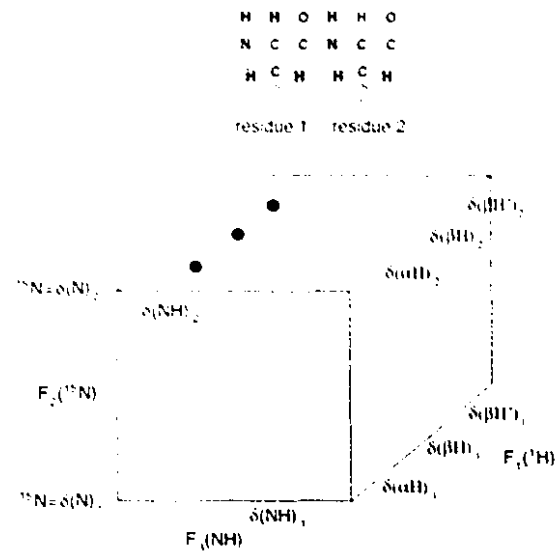


Figure 3.14: A simulated 3D ^{15}N TOCSY-HMQC spectrum. Cross peaks belonging to the same residue are observed in an F_1-F_3 plane. The corresponding F_2 coordinate is the chemical shift of that residue's amide nitrogen nucleus.

the chemical shift of that residue's amide proton. Spectra overlap is resolved by projecting the regular 2D ^{1}H-^{1}H TOCSY or NOESY into many $F_1(^1\mathrm{H}) - F_3(^1\mathrm{H})$ planes. One possible limitation for the two 3D NMR experiments is that the spectral ambiguities occur in case that two residues have common ^{15}N and NH resonance frequencies. Another problem for the ^{1}H-^{15}N TOCSY-HMQC and NOESY-HMQC experiments involves the relatively smaller $^3J_{\mathrm{NH}-\alpha\mathrm{H}}$ couplings for α-helix based proteins. The small J couplings might give rise to weak γH, δH, ... , etc., cross peaks.

3.5.3 The Constrained Partitioning Algorithm using Nitrogen chemical shifts

The algorithm takes the only input from the 3D ^{1}H-^{15}N TOCSY-HMQC spectrum and output the individual amino acid spin systems. The basic concept of the algorithm is simple: to merge two 3D ^{15}N TOCSY-HMQC cross peaks ($\delta\mathrm{H}_1$, $\delta\mathrm{N}_1$, $\delta(\mathrm{NH})_1$) and ($\delta\mathrm{H}_2$, $\delta\mathrm{N}_2$, $\delta(\mathrm{NH})_2$), the chemical shift differences of $|\delta\mathrm{N}_1 - \delta\mathrm{N}_2|$ and $|\delta(\mathrm{NH})_1 - \delta(\mathrm{NH})_2|$ must be observed within their corresponding chemical shift tolerance values. For the comparison of $|\delta\mathrm{N}_1 - \delta\mathrm{N}_2|$, the tolerance of nitrogen chemical shift is set to be 0.20 ppm by default, while the tolerance of proton chemical shift for the comparison of $|\delta(\mathrm{NH})_1 - \delta(\mathrm{NH})_2|$ is set to 0.02 ppm. If both of the above comparisons

are satisfied, a spin system with three protons and one nitrogen is constructed as the one shown in Figure 3.15. Note in Figure 3.15 that the protons αH and βH have their own connections to

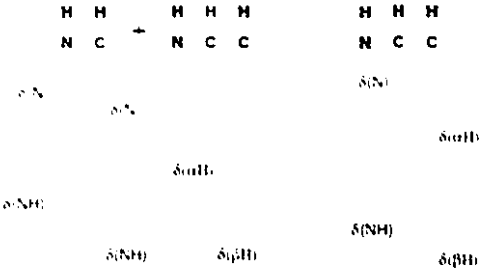


Figure 3.15: The merge of two 3D ^{15}N TOCSY-HMQC cross peaks. Two resonances, NH and N in this particular case, are required to be overlapped in order to conduct the merge. A four-spin system will be created.

$(NH)_1$. However, they don't have a connection between each another. This is the feature for the spin coupling patterns generated from TOCSY type experiments. Since the correlations between aliphatic side chain protons are not observed, it is generally not possible to establish the connectivities between side chain protons. Figure 3.16 lists a few example spin systems generated from the 3D ^{1}H-^{15}N TOCSY-HMQC experiment.

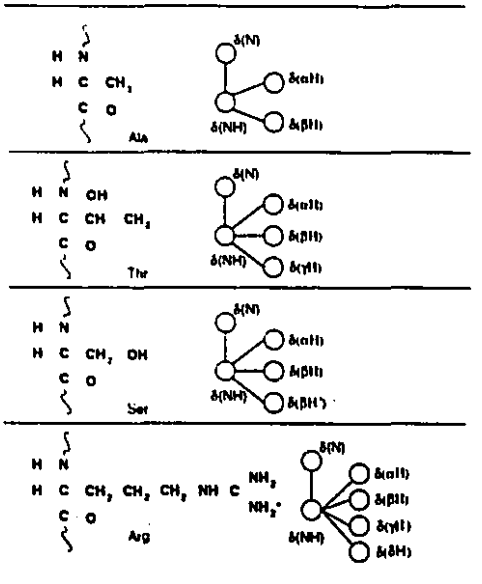


Figure 3.16: Some sample spin systems deduced from the 3D ^{15}N TOCSY-HMQC spectrum.

The following codes explain the detailed partitioning operations.

```
SpinSystem_type NCPA(PeakList' type 3D 15N TOCSY-HMQC)
{
//
// Input : 3D 15N TOCSY-HMQC peak list.
// Output: amino acid spin systems.
//

  for each of the peak i (i=1 to N) in
              the 15N TOCSY-HMQC peak list {
     search a cross peak n in the input peak list such that
     n is the most likely peak to be in the same spin system with
     peak i;

     record pair (i, n) in a temporary table;
  }

  for each input peak i (i=1 to N) {
    add peak i into a new spin system Si;
    for each of the peak j (j=1 to N) in
    the 15N TOCSY-HMQC peak list {

    find the most likely partner peak for peak j from the above temporary
    table, suppose the partner peak is peak n;

    if peak j is a member of the spin system Si
          add peak n into Si;
       else if peak n is a member of the spin system Si
          add peak j into Si;
    }
  }

  get rid of the redundant spin systems;
  output all Si;
}
```

In principle N input peaks give rise to N output spin systems. However, a number of them are redundant spin systems. For instance, starting from the cross peak (αH, N, NH) of an alanine, the spin system {N, NH, αH, βH} can be created; on the other hand, the same spin system can also be derived from the cross peak (βH, N, NH) of the same alanine. One of the above two spin systems are redundant and must be removed from the output. This is why NCPA conducts a purging operation before giving the spin system output. One may also notice that many spin systems are composed of only one peak in the output list of an NCPA running. Falsely picked peaks are the common reason for those one-peak spin systems, because a false peak generally can not be merged with other cross peaks.

Once the spin systems are generated from the ^{15}N TOCSY-HMQC data, the information

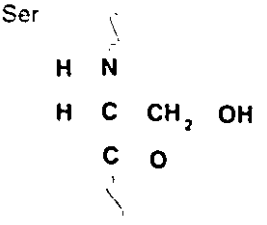
of the amino acid types is required for the eventual placement of the spin systems within the primary sequence. In chapter 3 we described an amino acid pattern recognition algorithm which determines the spin system types automatically. The original algorithm deals with the spin systems extracted from homonuclear 2D NMR data. As we've seen earlier, the spin systems extracted from 3D ^{15}N TOCSY-HMQC consist of both protons and amide nitrogen atoms. This indicates that a standard chemical shift database for amide nitrogens is needed to perform the automated pattern recognition. In addition, the standard patterns of the 20 amino acids must be revised to reflect the fact that no connectivity between side chain protons is established in the ^{15}N TOCSY-HMQC spectra. The expected chemical shifts of the amide nitrogens for the 20 commonly seen amino acids are listed in Table 3.2. The data was provided by Choy [70]. The chemical shifts are listed

Table 3.2: The expected chemical shifts of amide nitrogen nuclei for the three protein conformations. Numbers are in ppm. The standard deviations are also given.

Amino	Helix		Sheet		Coil	
Acid	mean	std	mean	std	mean	std
Ala	122.36	2.82	124.72	5.22	124.47	4.37
Arg	119.79	2.87	124.50	4.16	120.56	5.25
Asn	117.23	3.49	122.43	5.51	118.81	4.49
Asp	119.81	2.90	122.73	4.65	120.27	4.28
Cys	118.06	3.36	119.15	3.57	118.90	4.08
Gln	119.28	3.91	122.12	3.85	120.43	3.97
Glu	119.22	2.62	123.21	3.74	121.58	4.08
Gly	107.48	3.93	109.73	4.45	109.84	3.80
His	117.45	1.99	121.48	4.49	118.70	4.72
Ile	120.20	3.44	124.73	4.20	120.80	6.88
Leu	120.42	3.18	125.39	4.27	122.95	3.93
Lys	120.16	2.55	123.27	4.82	121.06	4.48
Met	118.19	3.06	122.44	5.38	120.61	3.92
Phe	119.60	3.45	121.97	4.22	121.68	6.53
Pro	133.14	3.96	N/A	N/A	136.83	1.76
Ser	115.91	3.42	118.03	3.61	117.18	4.88
Thr	115.73	4.89	117.47	5.15	115.51	6.20
Trp	119.99	1.76	124.95	3.96	120.29	5.72
Tyr	120.19	3.31	122.75	4.83	120.19	5.07
Val	119.74	4.47	123.33	4.80	120.61	5.91

according to three major structural components: helix, sheet and coil.

In Figure 3.17 the expected spin coupling patterns for a serine are listed. The spin systems derived from the 3D ^{15}N TOCSY-HMQC spectrum differs from the one derived from 2D COSY.



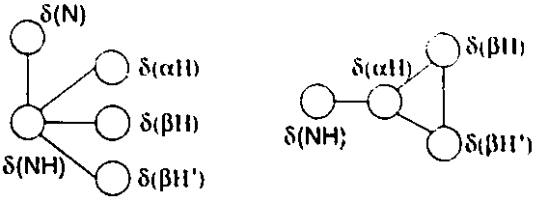


Figure 3.17: Comparison of the spin systems deduced from 3D ^{15}N TOCSY-HMQC and from 2D COSY/TOCSY spectra. The former has an amide nitrogen resonance and is lack of the connectivity between side chain protons.

To complete the sequence-specific assignment, the individual amino acid spin systems must be placed in their corresponding positions within the protein primary sequence. Up to this point the only NMR experiment used is 3D ^{15}N TOCSY-HMQC which does not provide any interresidue information. A similar 3D NMR experiment, ^{15}N NOESY-HMQC, provides the through-space correlations needed for the sequential assignment. The 3D ^{15}N NOESY-HMQC experiment resolves spectral ambiguities which limit the analysis of the conventional 2D NMR spectra. The absence of overlapping cross peaks in 3D NOESY-HMQC allows the unambiguous identification of $d_{\alpha N}(i, i+1)$ and $d_{NN}(i, i+1)$ through space nuclear Overhauser connectivities which are necessary for connecting spin systems sequentially. Our strategy of applying the 3D ^{15}N NOESY-HMQC experiment is similar to the one described in chapter 5. Using the interresidue correlations provided by ^{15}N NOESY-HMQC, the individual amino acid spin systems can be connected to form many dipeptides. Those dipeptides are used as the building blocks of polypeptide chains which in turn are to be mapped to the proper positions within the primary sequence. The actual mapping task involves the use of an algorithm called PMA which is described in chapter 5.

A $d_{\alpha N}(i, i+1)$ cross peak in NOESY connects the αH of a residue and the NH of the

following residue while a $d_{NN}(i, i+1)$ cross peak connects the amide protons of two sequentially neighboring residues. The $d_{\alpha N}(i, i+1)$ and $d_{NN}(i, i+1)$ cross peaks are the two commonly used interresidue correlations in identifying sequentially connected amino acid spin systems [2]. A simple program was designed to create dipeptides from the deduced spin systems. The required interresidue information is adopted from the $d_{\alpha N}(i, i+1)$ and $d_{NN}(i, i+1)$ peaks of the 3D ^{15}N NOESY-HMQC spectrum. The codes for the establishment of dipeptides from 3D NOESY-HMQC are listed below:

```
void annn(PeakList_type 3D 15N NOESY-HMQC, SpinSystem_type, ... )
{
//
// Input : 1. 3D 15N NOESY-HMQC peak list.
//         2. all of the spin systems derived by the algorithm NCPA.
//            Totally N spin systems.
// Output: dipeptides connected through dnn(i,i+1) and dan(i,i+1)
//

  for each of the spin system pair |i, j|, (i, j = 1 to N, i≠j) {
    if both dnn(i,i+1) and dan(i,i+1) are observed in
    the input peak list
       link spin system S_i and S_j to a dipeptide S_i − S_j;
  }
  output all discovered dipeptides;
}
```

The discovered dipeptides along with the available amino acid type information makes it possible to use our sequential assignment protocol to complete the sequential assignment.

3.5.4 Applications and Results

The ^{15}N TOCSY-HMQC and NOESY-HMQC spectra were provided by University of Alberta [68]. Sample protein is the calcium-loaded regulatory N-domain of chicken skeletal troponin-C (1-90). Both experiments were carried out on a Varian Unity-600 NMR spectrometer operating at 30 °C. The mixing times for ^{15}N TOCSY-HMQC and NOESY-HMQC are 70 and 150 ms, respectively. The ^{15}N carrier frequency is 117.44 ppm and the spectra width is 23.03 ppm. The ^{15}N chemical shifts are reported relative to external acidic NH_4Cl (24.93 ppm). Automatic peak picking of the transformed 3D spectra was achieved using the CAPP program [56]. A total

of 241 ^{15}N TOCSY-HMQC and 675 ^{15}N NOESY-HMQC cross peaks were reported. The peak list was then given the the authors by B. Sykes [68]

The algorithm NCPA was implemented using C language. On a 75 MHz Pentium computer, the typical execution time for the entire execution is about 3 minutes.

Using the 241 ^{15}N TOCSY-HMQC peaks, NCPA program produced 82 spin systems. The tolerance value for nitrogen and proton chemical shifts are set to 0.20 and 0.02 ppm, respectively. Each deduced spin system consists of an amide nitrogen, amide proton and some protons. A sample output of NCPA is listed here:

```
/*9th G/        Total Peaks= 3
//Peak 25 (8.660 , 5.140 , 117.860)
//Peak 26 (8.660 , 2.790 , 117.850)
//Peak 27 (8.660 , 2.540 , 117.860)
//TOCSY-HMQC 1.00 25(5.140, 8.660,117.860)+26(2.790, 8.660 , 117.850)
//TOCSY-HMQC 1.00 27(2.540, 8.660,117.860)+25(5.140, 8.660 , 117.860)
//Spin Coupling Topological Graph:
 N,117.857
1H,8.660,2,3,4
2H,5.140,1
3H,2.790,1
4H,2.540,1
```

In the listing, a spin system with four protons and one nitrogen was created from the ^{15}N TOCSY-HMQC peak 25, 26 and 27. The adjacency list of the spin system is also shown. For example, proton 1H (8.660 ppm) has three neighbors: 2H, 3H and 4H. Among a total of 82 output spin systems, seventy-four of them can be verified against the independently done manual assignment [68]. Figure 5.9 summarizes the result of the NCPA run.

As a subsequent test, we examined the interresidue $d_{\alpha N}(i, i+1)$ and $d_{NN}(i, i+1)$ correlations. Upon the 675 ^{15}N NOESY-HMQC peaks and 82 deduced spin systems, a total of 77 dipeptides were generated. Those dipeptides were linked to one another to form the 174 polypeptides with the length from 3 to 10 residues. On an earlier run of the amino acid pattern recognition program, the amino acid types of the 82 deduced spin systems were determined. The output is digested in the following listing where each spin system has a candidate list showing the possible amino acids.

```
G1(1st G): Ile/0.793 Leu/0.766 Arg/0.741 Lys/0.741 Ser/0.614 ......
G2(2nd G): Ile/0.658 Arg/0.618 Lys/0.618 Met/0.543 Gln/0.536 ......
```

In the last sequential assignment stage, the algorithm PMA successfully assigned residue 4 to 10, 15 to 20, 21 to 24, 27 to 30 and 79 to 86.

3.5.5 Discussion

The sample protein NTnC (1-90) has five major helix segments [68]. In those segments, most of the $^3J_{NH-\alpha H}$ are less than 6 Hz. The small couplings often result in shorter TOCSY transfer. In other words, the ^{15}N TOCSY-HMQC spectrum doesn't provide a sufficient number of long range through-bond cross peaks. This can be verified from the output of the NCPA program. Many of the extracted spin systems contain N, NH, αH and βH only. The additional side chain resonances are unable to be determined as the spectral data is insufficient. The short side chain effects the accuracy of the determination of amino acid types, because it is the side chain that makes the 20 amino acids distinct to one another. The low percentage of successfully assigned residues are due to the incomplete TOCSY connections.

Without using ^{15}N NOESY-HMQC experiment, ^{15}N TOCSY-HMQC itself provides an alternate approach to determine the amino acid side chain resonances. The more detailed side chain information is available, the more accurate determination of amino acid types can be anticipated. The ^{15}N TOCSY-HMQC alone might not be able to provide sufficient data for a complete resonance assignment. However the extracted spin system information does play an important role in the overall sequential assignment process. See section 5.2 for further discussion of the NCPA algorithm.

3.6 Summary of spin system determination from double resonance 3D NMR

The algorithm NCPA, (Nitrogen Constrained Partitioning Algorithm), was proposed to automate the determination of amino acid spin systems. The algorithm is a direct extension of the 2D CPA algorithm described in chapter 2. The algorithm has the feature that it can accept a single TOCSY type NMR experiment as the input and identify the individual spin systems from 2D TOCSY or 3D ^{15}N TOCSY-HMQC spectrum. Along with the sequential connectivities provided by 3D ^{15}N NOESY-HMQC, we demonstrated the possibility of using a minimum number of NMR experiments to conduct the automated sequential assignment.

Chapter 4

Automated Extraction of Aliphatic Side-chain Spin Systems

4.1 Introduction

The aim of this chapter is to extend the CPA algorithm to 3D NMR and present a computer assisted spin system extraction procedure based on 3D HCCH-COSY and HCCH-TOCSY NMR spectra.

Resonance assignment of a protein's backbone can be achieved by a combination of several triple resonance 3D NMR experiments [5]. Furthermore, to obtain the detailed structure of a protein, the NOE cross peaks of the chain nuclei must be unambiguously assigned so that enough distance constraints can be produced to construct the protein side chain orientation. The analysis of NOE cross peaks usually requires the side chain resonance assignment to be completed. Several 3D NMR experiments have been proposed for the resonance assignment of protein side chain, such as 3D HCCH-COSY [71–73], HCCH-TOCSY [74], HCC(CO)NH-TOCSY [75, 76] and HCCNH-TOCSY [75, 77].

Among the several attempts for the automated analysis of 3D NMR, two of them [9, 16] studied the applications of homonuclear 3D NMR to protein proton resonance assignments. The rest of the approaches use triple resonance heteronuclear 3D NMR to obtain the assignment of protein backbone [17] and to establish the sequential connectivities of amino acid spin systems

[18, 19]. The availability of the information about spin systems, including backbone and side chain resonances, as well as the amino acid types, is crucial in all these methods. However, in all of the heteronuclear 3D NMR approaches mentioned above, the information of side chain spin systems has to be manually obtained elsewhere. This chapter is directed in this regard to design an automatic strategy to obtain the information of protein spin systems. In this chapter an algorithm is proposed to extract aliphatic side chain spin systems from heteronuclear 3D NMR data of proteins. The algorithm merges cross peaks from 3D NMR data, such as 3D HCCH-COSY, to form spin coupling systems. At each merging step at least two constraints are required to assure the validity of the merge. Thus an additional NMR spectrum, such as 3D HCCH-TOCSY can be used to supply these constraints. The output spin coupling systems are given as a series of graphs represented as adjacency lists which can be processed by the subsequent graph pattern recognition algorithm, which is described in chapter 5, to perform the amino acid identification.

4.2 Methods for extracting side-chain spin systems

4.2.1 Concept of the peak merging process

The central idea of the algorithm is to extract amino acid spin systems from NMR spectra. To illustrate how this approach works, a simple three-spin system is first considered(see Figure 4.1). On a 2D DQF-COSY NMR spectrum, such a three-spin system gives two cross peaks on each

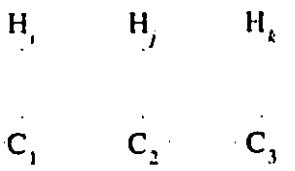


Figure 4.1: Example of a chemical structure fragment with three hydrogen atoms.

side of the diagonal, while in a 2D TOCSY spectrum, an extra peak is observed on each side(see Figure 4.2). To construct this three-spin system from the cross peak data, conventional assignment procedure probably picks the starting point from the peak 1 (see Figure 4.1 and 4.2), then observes the peak 2 in a subsequent searching in the peak list. In terms of an automated computer procedure,

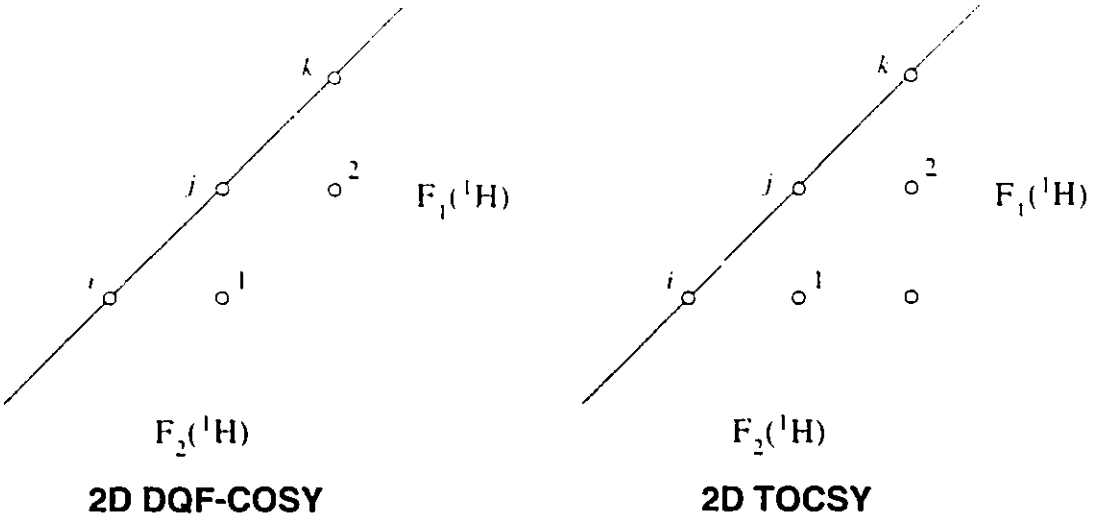


Figure 4.2: 2D DQF-COSY and TOCSY spectra of the chemical structure shown in Figure 4.1. The peaks on other sides are not displayed for convenience.

for peak 1 (δ_i, δ_j), and peak 2 $(\delta_{j'}, \delta_k)$, if δ_j and $\delta_{j'}$ are close enough (controlled by a pre-defined tolerance value), the three-spin system, $\{i, j, k\}$, can be constructed. Applying this procedure to the entire peak list enables, in principle, all the amino acid spin systems to be extracted. However, in certain regions of the spectrum, heavy overlap makes this kind of merging process unreliable.

Suppose, for example, we have two three-spin systems, $\{\delta_i, \delta_j, \delta_k\}$,$\{\delta_l, \delta_m, \delta_n\}$, and coincidentally two spins, j and m, have resonance frequencies which are similar in values (see Figure 4.3). The COSY cross peaks produced by the two systems are (δ_i, δ_j), $(\delta_{j'}, \delta_k)$, (δ_l, δ_m) and $(\delta_{m'}, \delta_n)$ where δ_j, $\delta_{j'}$, δ_m and $\delta_{m'}$ are difficult to distinguish in terms of chemical shifts. In the analysis of the peak merging procedure, it is necessary to determine that the cross peak (δ_i, δ_j) should be merged with $(\delta_{j'}, \delta_k)$ or $(\delta_{m'}, \delta_n)$. Since j and m have similar resonance frequencies, an extra constraint is needed to remove the ambiguity. One way is to look at the TOCSY spectrum. If spin i, j, k are indeed in the same spin system, i.e., δ_j and $\delta_{j'}$ come from the same spin, the TOCSY cross peak (δ_i, δ_k) should be observed. Similarly, if i, j, n are in the same spin system, namely δ_j and $\delta_{m'}$ come from the same spin, another TOCSY cross peak (δ_i, δ_n) should be observed. Hence by cross referencing with such TOCSY constraints, one can reduce the possibility of the ambiguities caused by spectral overlap, making it possible to design an automated spin

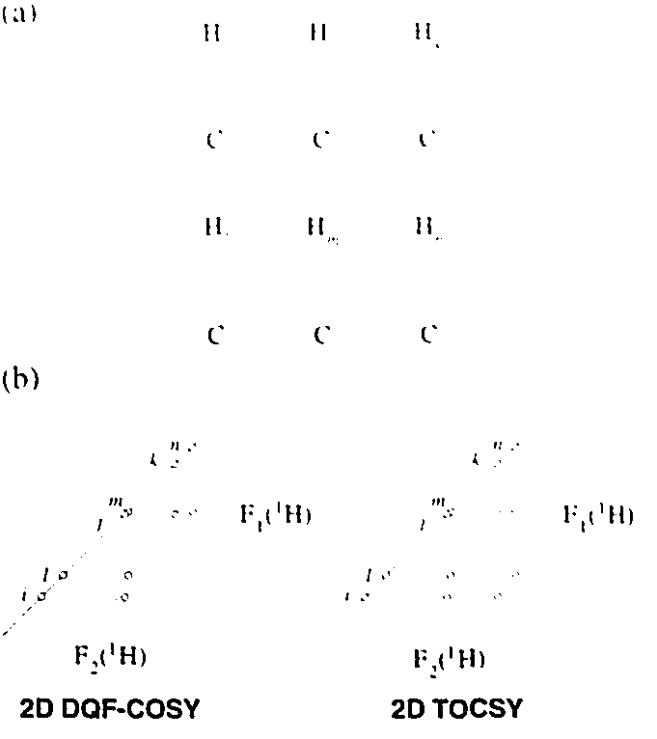


Figure 4.3: (a) Three-spin systems $\{\delta_i, \delta_j, \delta_k\}$ and $\{\delta_l, \delta_m, \delta_n\}$ where δ_j and δ_m are within a chemical shift tolerance. (b) 2D DQF-COSY and TOCSY spectra of the above spin systems.

system extraction algorithm.

As the size of the target protein increases, the corresponding 2D NMR spectrum becomes more crowded. It is unlikely that one constraint alone can resolve the overlap when doing peak merging. One solution is to acquire another 2D NMR spectrum which may provide additional information to resolve the overlap. Another way is to introduce the third dimension in which another nucleus can be used as the additional constraint. The former was treated previously in chapter 2 while we discuss the latter in this chapter.

The complete amino acid spin systems of a protein's side chain can be determined by 3D HCCH-COSY and HCCH-TOCSY experiments [6, 71, 73, 74]. Both experiments make use of the one bond ^{1}H$-^{13}$C ($\sim$ 140Hz) and ^{13}C$-^{13}$C ($\sim 30-40$Hz) J couplings to transfer magnetization along the side chain via the pathway

$$^1\mathrm{H} \xrightarrow{^1J_{\mathrm{CH}}} {}^{13}\mathrm{C} \xrightarrow{^1J_{\mathrm{CC}}} {}^{13}\mathrm{C} \xrightarrow{^1J_{\mathrm{CH}}} {}^1\mathrm{H}$$

To interpret the 3D HCCH COSY/TOCSY spectra, consider first a 2D TOCSY segment. Figure 4.4 shows the spectrum that corresponds to the chemical structure shown on the left of the figure.

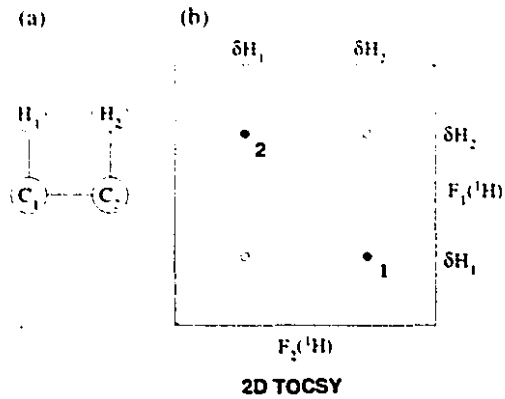


Figure 4.4: (a) Structure of a CH-CH fragment. (b) The corresponding 2D TOCSY spectrum. Cross peak 1 has chemical shifts (δH_1, δH_2), cross peak 2 has chemical shifts (δH_2, δH_1).

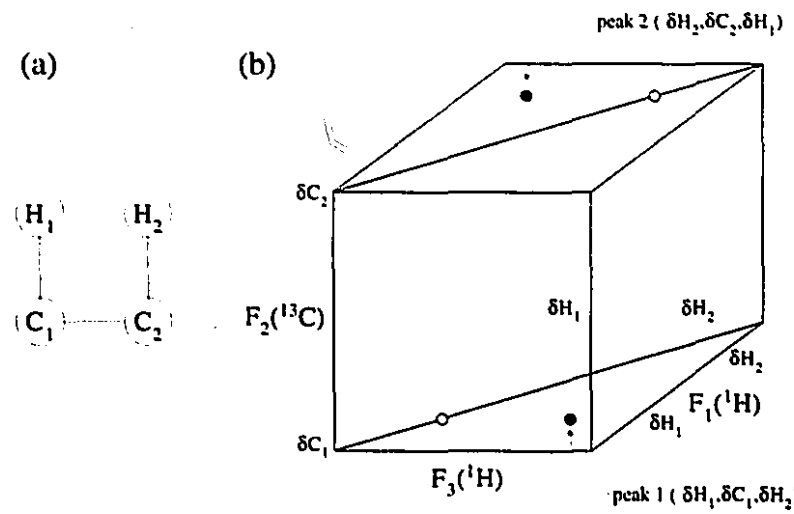


Figure 4.5: (a) The same structure as in Figure 4.4(a). (b) The corresponding 3D HCCH-TOCSY spectrum. The $^1H(F_1) - {}^1H(F_3)$ planes are similar to that of 2D $^1H - {}^1H$ COSY or TOCSY experiment, except that the $^1H(F_1) - {}^1H(F_3)$ are edited by the chemical shift of the ^{13}C nuclei. Note that peak 1 and 2 do not occur symmetrically on both sides of the diagonal on the same plane.

Figure 4.5 shows the 3D HCCH-TOCSY spectrum of the same CH-CH fragment as in Figure 4.4. The $^1H(F_1) - {}^1H(F_3)$ planes are similar to that of 2D $^1H - {}^1H$ COSY or TOCSY experiment,

except that these planes are edited by the chemical shifts of the ^{13}C nuclei. Hence off-diagonal peaks in a $^{1}H - {}^{1}H$ plane at the ^{13}C frequency arise from protons directly bonded to that ^{13}C. For example, in Figure 4.5, the magnetization transfer pathway of cross peak 1 ($\delta H_1, \delta C_1, \delta H_2$) follows the path $H_1 \rightarrow C_1 \rightarrow C_2 \rightarrow H_2$, while the transfer pathway of cross peak 2 ($\delta H_2, \delta C_2, \delta H_1$) has path $H_2 \rightarrow C_2 \rightarrow C_1 \rightarrow H_1$. The cross peaks 1 and 2 in 3D HCCH experiments do not occur symmetrically on both sides of the diagonal of the same plane, but rather, occur on different $F_1 - F_3$ planes as shown in Figure 4.5.

4.2.2 Concept of the algorithm

The NMR data set used in the present algorithm are 3D HCCH-COSY and 3D HCCH-TOCSY. Currently the implemented computer program is designed to process peak lists. That is, cross peaks in the spectra must have been previously picked by a reliable peak picking procedure. In the peak list, cross peaks are represented by three chemical shift coordinate points, e.g., (3.52, 58.17, 1.46), where the first coordinate denotes the resonance frequency of the proton which is directly bonded to the carbon. The frequency of that carbon is the second coordinate, while the third coordinate is the frequency of another proton which can be reached by the transfer of magnetization along the side chain via the HCCH pathway. In the following context, a generic 3D cross peak is represented as (H_i, C_1, H_j). The corresponding chemical structure of the 3D HCCH-COSY cross peak (H_i, C_1, H_j) are shown in Figure 4.6.

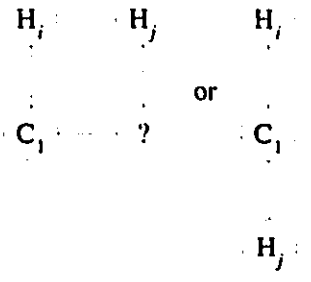


Figure 4.6: The possible chemical structures corresponding to the 3D HCCH-COSY cross peak (H_i, C_1, H_j). In the left, the chemical shift of the carbon to which H_j bonds is undetermined.

The algorithm, called ASPA(Aliphatic Side-chain Partitioning Algorithm) , starts with the

entire HCCH-COSY data set being searched to find pairs of cross peaks, (H_i, C_1, H_j) and ($H_{i'}$, $C_{1'}$, H_k), which have one proton and one carbon resonance frequencies in common. In the algorithm, H_i ,$H_{i'}$ and C_1 ,$C_{1'}$ are tested to determine whether they are within the user-defined chemical shift tolerance values , such as 0.02 ppm for proton and 0.20 ppm for carbon. There are three different situations regarding the connectivities between protons and carbons to be considered in merging cross peaks into spin systems. The first is that all of the three protons, H_i, H_j and H_k bond to different carbons. A schematic view in Figure 4.7 shows the two HCCH-COSY cross peaks ,along with various constraint peaks, can arrive at a merged spin system. Figure 4.7(d) is the first possible merged spin system which is formed from Figure 4.7(a) along with the two constraint peaks shown in Figure 4.7(b). Similarly, the spin system in Figure 4.7(e) can be obtained from the two cross peaks shown in Figure 4.7(a) along with the two constraint peaks in Figure 4.7(c).

A second case occurs when H_j and H_k bond to the same carbon as shown in Figure 4.8. One of two possible constraint peak sets, Figure 4.8(b) or Figure 4.8(c), is required to confirm that the spin system shown in Figure 4.8(d) can be constructed. A third case has H_i and H_j bonded to the same carbon as shown in Figure 4.8(e). The presence of two constraint peaks, Figure 4.8(f), confirms the spin system shown in Figure 4.8(g).

To summarize the above pictorial representations, Figure 4.9 shows the control flow of the partitioning algorithm.

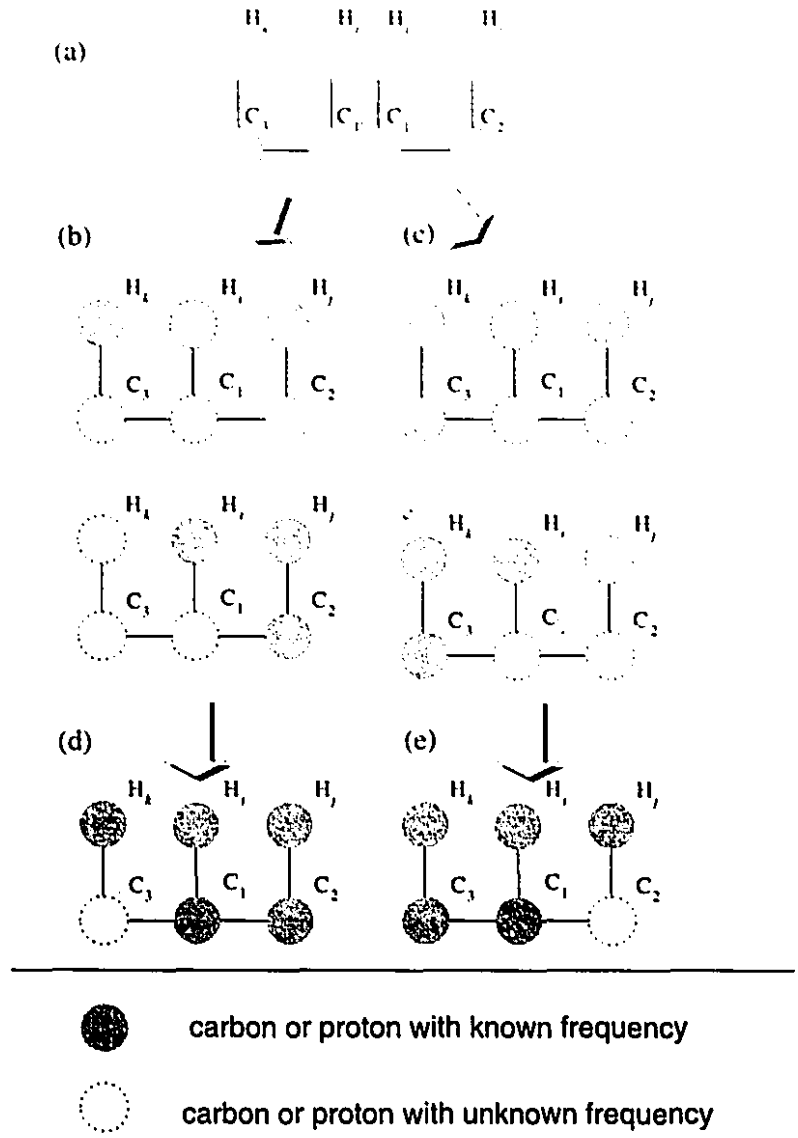


Figure 4.7: Schematic representation showing the 3D HCCH-COSY cross peaks, (H_i, C_1, H_j) and ($H_{i'}$, $C_{1'}$, H_k), are merged to form a spin system. Each cross peak contains three frequencies depicted by filled circles, while the open circles indicate the frequencies are unknown from the cross peak data. (a) Two cross peaks (H_i, C_1, H_j) and ($H_{i'}$, $C_{1'}$, H_k), where H_i, $H_{i'}$ and C_1, $C_{1'}$ are within the specified tolerance. (b) The two possible cross peaks (H_j, C_2, H_k) and (H_j, C_2, H_i) as the constraints. (c) Another two possible peaks (H_k, C_3, H_j) and (H_k, C_3, H_i) as the constraints. (d) The possible merged spin system with three protons, H_i, H_j, H_k, and three carbons, C_1, C_2 ,C_3. (e) Another possible merged spin system. The two peaks in (a) along with the two constraint peaks in (b) lead to the spin system in (d). The two peaks in (a) along with the two constraint peaks in (c) lead to the spin system shown in (e). In summary, (a)-(b)-(d) is a possible pathway to merge two cross peaks while (a)-(c)-(e) is another one.

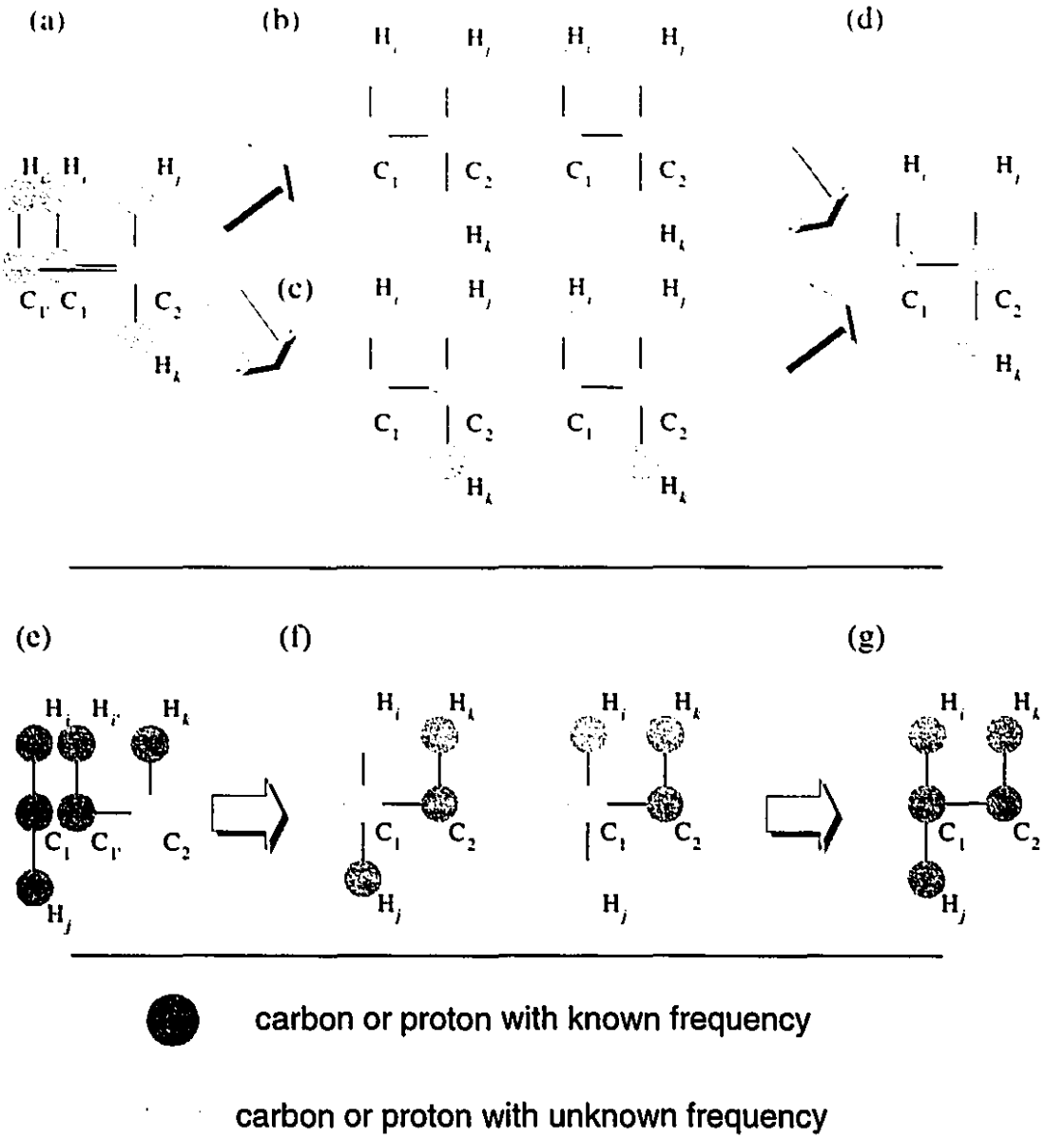


Figure 4.8: (a) The two 3D HCCH-COSY cross peaks (H_i, C_1, H_j) and ($H_{i'}$, $C_{1'}$, H_k). (b) The two peaks (H_j, C_2, H_k) and (H_j, C_2, H_i) as the constraints. (c) Another two peaks (H_k, C_2, H_j) and (H_k, C_2, H_i) as the constraints. (d) The merged spin system with three protons, H_i, H_j ,H_k, and two carbons, C_1 and C_2. The two peaks in (a) along with the two constraint peaks either in (b) or (c) lead to the spin system in (d). (e) The two 3D HCCH-COSY cross peaks. (f) The two peaks as the constraints. (g) The merged spin system with three protons, H_i, H_j ,H_k, and two carbons, C_1 and C_2. The two peaks in (e) along with the two constraint peaks in (f) give rise to the spin system in (g).

Step1 Search the HCCH-COSY cross peak list for pairs of (H_i, C_1, H_j) and ($H_{i'}$, $C_{1'}$, H_k), where H_i and $H_{i'}$ are within the ^1H chemical shift tolerance range, and C_1 and $C_{1'}$ are within the ^{13}C chemical shift tolerance range. Do the following steps to test if H_i, C_1, H_j and H_k can be added to a spin system.

Step2 If a HCCH-TOCSY (H_j, C_2, H_k) is found

and a HCCH-COSY (H_j, C_2, H_i) or HCCH-TOCSY (H_j, C_2, H_i) is found

then add H_i, C_1, H_j, H_k and C_2 to a spin system.

Step3 else if a HCCH-TOCSY (H_k, C_2, H_j) is found

and a HCCH-COSY (H_k, C_2, H_i) or HCCH-TOCSY (H_k, C_2, H_i) is found

then add H_i, C_1, H_j, H_k and C_2 to a spin system.

Step4 else if a HCCH-COSY (H_j, C_2, H_k) is found

and a HCCH-COSY (H_j, C_2, H_i) or HCCH-TOCSY (H_j, C_2, H_i) is found

then add H_i, C_1, H_j, H_k and C_2 to a spin system.

Step5 else if a HCCH-COSY (H_k, C_2, H_j) is found

and a HCCH-COSY (H_k, C_2, H_i) or HCCH-TOCSY (H_k, C_2, H_i) is found

then add H_i, C_1, H_j, H_k and C_2 to a spin system.

Step6 Back to **Step1** until no more COSY cross peak pair fulfilled the condition of **Step1** remain in the data set.

Figure 4.9: Control flow of the partitioning algorithm.

4.2.3 *Detailed description of the algorithm*

The algorithm takes input from the 3D HCCH-COSY and HCCH-TOCSY peak lists then conducts partitioning operations to extract the aliphatic side chain spin systems. Suppose that N peaks are picked in the 3D HCCH-COSY spectrum. For each peak i in the peak list, the algorithm attempts to merge all other peaks that possibly reside in the same spin system with the peak i. For a peak list with N peaks, the output of N spin systems are expected. The details of the partitioning operations are listed in the following code segments.

```
SpinSystem_type partitioning(PeakList_type 3D HCCH-COSY, HCCH-TOCSY)
{
// This function is the kernel of the Aliphatic Side chain
// Partitioning Algorithm
//
// Input : 3D HCCH-COSY and HCCH-TOCSY peak lists
// Output: For N 3D HCCH-COSY peaks, the output will be N spin systems.
//         Those systems are not the final output. A merging
//         procedure is to be applied to obtain the final side chain
//         spin systems.

  for each of the peak i (i=1 to N) in the HCCH-COSY peak list {
    put peak i into spin system Si;
    for each of the peak j (j=1 to N) in the HCCH-COSY peak list {

       n=best_partition(j);
                                    // Find the peak n in the
                                    // HCCH-COSY peak list
                                    // which is the most likely peak
                                    // to be merged with peak j ;
       if peak j is a member of the spin system Si
          add peak n into Si;
       else if peak n is a member of the spin system Si
          add peak j into Si;
    }
  }
  output all Si;
}
```

A function called 'best_partition() is invoked within the partitioning(). The former is responsible for the actual searching and merging tasks and is listed in the following.

```
Peak_type best_partition(Peak_type m, PeakList_type 3D HCCH-COSY,
                         HCCH-TOCSY)
{
// Input : 1. 3D HCCH-COSY and HCCH-TOCSY peak lists.
//         2. the cross peak m in the HCCH-COSY peak list.
```

```
// Output: Return the peak which is considered to have the best
           chance to be partitioned with the peak m.

  for each of the peak i (i=1 to N) in the HCCH-COSY peak list {
      if peak i can be merged with peak m {
         call merge1(), merge2() or merge3() depending on the
         overlapped resonances between the peak i and m;
         compute the ranking parameter A_i;
      }
  }
  return the peak with the highest ranking parameter;
}
```

Three merging functions are invoked within the the function of `best_partition()`. `merge1()`, `merge2()` and `merge3()` perform the operations illustrated in Figure 4.7, Figure 4.8(a)-(d) and Figure 4.8(e)-(g), respectively.

```
void merge1(Peak_type m, Peak_type n, PeakList_type 3D HCCH-COSY/TOCSY)
{
// Input two peaks m and n. They are overlapped in the first and
// second coordinates.

   if the peak (H_j, C_2, H_k) can be observed in the peak list
      of 3D HCCH-TOCSY and
      the peak (H_j, C_2, H_i) can be observed in the peak list
      of 3D HCCH-COSY or TOCSY {

      The peak m and n are allowed to merge.
      Note that the following calculation decides if m and n can be
      actually merged.

      compute the ranking parameter A;

   } else if the peak (H_k, C_3, H_j) can be observed in the peak list
                  of 3D HCCH-TOCSY and
                  the peak (H_k, C_3, H_i) can be observed in the peak list
                  of 3D HCCH-COSY or TOCSY {

                  the peak m and n are allowed to merge;
                  compute the ranking parameter A;

                  } else
                              the peak m and n are not allowed to merge;
}

void merge2(Peak_type m, Peak_type n, PeakList_type 3D HCCH-COSY/TOCSY)
{

   if the peak (H_j, C_2, H_k) can be observed in the peak list
```

```
      of 3D HCCH-TOCSY or COSY and
      the peak (H_j, C_2, H_i) can be observed in the peak list
      of 3D HCCH-TOCSY or COSY{

      The peak m and n are allowed to merge.
      Note that the following calculation decides if m and n can
      be actually merged.

      compute the ranking parameter A;
   } else if the peak (H_k, C_2, H_j) can be observed in the peak list
                of 3D HCCH-TOCSY and
                the peak (H_k, C_2, H_i) can be observed in the peak list
                of 3D HCCH-COSY or TOCSY {

                the peak m and n are allowed to merge;
                compute the ranking parameter A;

                } else
                          the peak m and n are not allowed to merge;
}

void merge3(Peak_type m, Peak_type n, PeakList_type 3D HCCH-COSY/TOCSY)
{

   if the peak (H_k, C_2, H_j) can be observed in the peak list
      of 3D HCCH-TOCSY or COSY and
      the peak (H_k, C_2, H_i) can be observed in the peak list
      of 3D HCCH-TOCSY or COSY{

      The peak m and n are allowed to merge.
      Note that the following calculation decides if m and n can
      be actually merged.

      compute the ranking parameter A;

    } else
              the peak m and n are not allowed to merge;
}
```

Note that some of the output spin system S_i might contain only one peak. This indicates that none of the peak can be merged with the peak i therefore peak i retains its single status. For example, due the lack of side chain hydrogens, glycines always give rise to one-peak spin systems (αH, Cα, αH').

Figure 4.10 is a fragment of the output spin system from ASPA. Note that the resonance frequencies of protons and carbons are both determined. The connectivity relationship between the protons is also displayed using the adjacency list.

```
//8th G/        Total Peaks= 2
//Peak 19 (4.652 , 70.400 , 4.438)
//Peak 20 (4.652 , 70.400 , 1.176)
//Spin Coupling Topological Graph:
1H,4.652(70.400),2,3
2H,4.438(61.085),1
3H,1.176(21.310),1
```

Figure 4.10: An example of the extracted spin system represented by the adjacency list. In this case, the two HCCH-COSY cross peaks (No.19 and No.20) were merged into a three-proton spin system. Proton 1 (4.652 ppm) bonds to a carbon (70.400 ppm), couples to proton 2 (4.438 ppm) and proton 3 (21.310 ppm). Proton 2(21.310 ppm) bonds to a carbon (61.085) and couples to proton 1 (4.652 ppm).

As the number of peaks and the complexity of spectra increase, the uniqueness of the merging process is compromised. In other words, for a specific peak with which it is common that more than one candidate peak can be merged. This is mainly due to the spectral overlap, making it necessary to design a strategy to rank the candidate peaks, in other words, to select the most likely merging from the many possibilities.

A scoring parameter in the partitioning algorithm is introduced to rank all the candidate peaks. Consider the cross peak (H_i, C_1, H_j), with which the candidate peak, ($H_{i'}$, $C_{1'}$, H_k) can be merged based on the presence of the constraints already discussed(see Figure 4.9). The two constraints might be the presence of peaks ($H_{k'}$, C_2, $H_{j'}$) and ($H_{k''}$, $C_{2'}$, $H_{i''}$). The scoring parameter A is defined as

$$A = 1 - \sqrt[3]{\left(\frac{w_0}{T_H}\right)\left(\frac{D}{2T_H}\right)\left(\frac{w_1}{T_C}\right)} \tag{4.1}$$

where

$$w_1 = |\delta_{C_1} - \delta_{C_{1'}}|$$

$$w_0 = |\delta_{H_i} - \delta_{H_{i'}}|$$

T_H = the tolerance value for comparing proton chemical shifts

T_C = the tolerance value for comparing carbon chemical shifts

$$D = \frac{(D_1 + D_2)}{2}$$

with D_1 and D_2 depending on the constraining peaks as

$$D_1 = |\delta_{H_k} - \delta_{H_{k'}}| + |\delta_{H_j} - \delta_{H_{j'}}|$$

$$D_2 = |\delta_{H_i} - \delta_{H_{i'}}| + |\frac{(\delta_{H_j} + \delta_{H_{j'}})}{2} - \delta_{H_{j''}}|$$

Figure 4.11 illustrates all the distance used in equation 4.1.

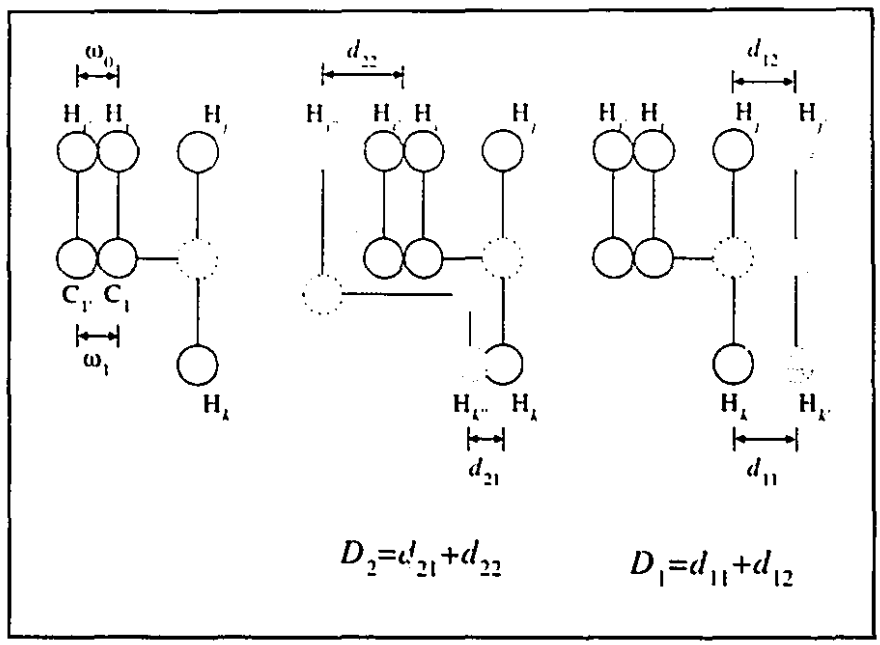


Figure 4.11: Pictorial representation of the variables used in calculating the scoring parameter. The solid circles represent the observed resonances. The dashed-line circles represent the undetermined resonances. The filled circles are the constraint peaks.

w_0 measures the difference of the chemical shifts between the original and candidate peaks in the first coordinate of the involved 3D cross peaks. T_H is the user-defined tolerance value to compare the proton chemical shifts. Candidate peaks which make w_0 greater than T_H are discarded, thus w_0 is always less than or equal to T_H, or $w_0/T_H \leq 1$.

w_1 measures the difference of the carbon chemical shifts between the original and candidate peaks. T_C is the tolerance value for comparing carbons. Similarly, $w_1/T_C \leq 1$.

D measures how well the two constraint peaks match the original and candidate peaks. A smaller D corresponds to a better match.

The above three factors are used to decide how good a candidate peak is. In terms of the first factor w_0, a smaller proton chemical shift difference between H_i and $H_{i'}$ indicates a better match of the cross peak (H_i, C_1, H_j) and ($H_{i'}$, $C_{1'}$, H_k). Secondly, a smaller carbon chemical shift

difference, i.e., a smaller w_1, between C_1 and C_1' also indicates a better match. Finally, D uses the constraint peaks to evaluate the two to-be-merged peaks.

The computer program calculates the scoring parameter for each of the merging pair giving a score from 0 to 1. A higher value of A is taken as a better match. Under such a scoring strategy, the candidate peak with the largest value of A is chosen to merge with the original peak.

4.3 Results

The algorithm was implemented in C programming languages. A simple GUI (graphical user interface) has been built for the implemented program based on X11 MOTIF library. Figure 4.12 shows the snapshot of the running program. The program was tested on both real and simulated 3D

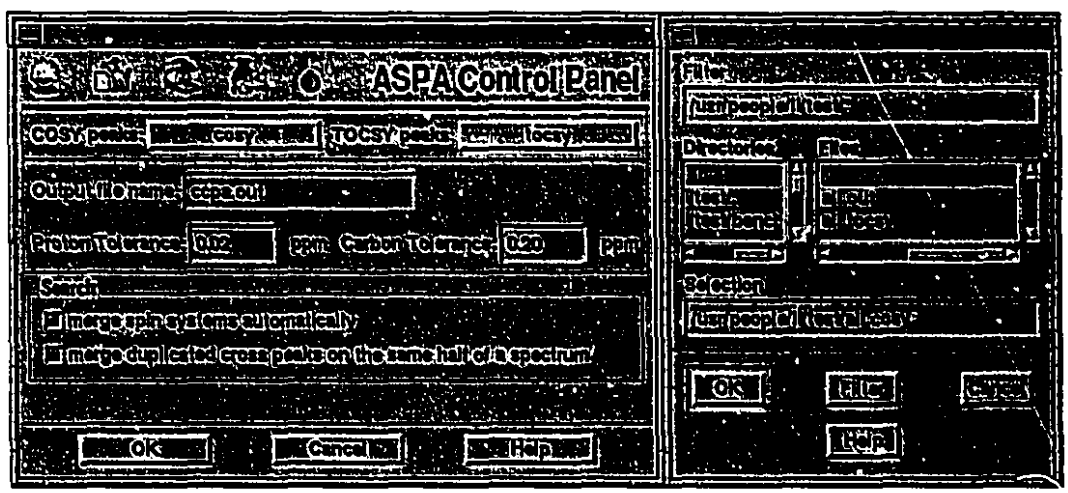


Figure 4.12: The snapshot of the implemented computer program.

HCCH-COSY/TOCSY data for the 90-residue protein N-domain of chicken skeletal troponin-C (1-90).

The experimental spectra and manual assignments were provided by University of Alberta [68]. The simulated data were generated based upon the manual assignments. Both exact and dispersive (with respect to chemical shifts, described later) simulations were used. The testing procedures and results are described below.

4.3.1 Analysis of simulated 3D HCCH-COSY/TOCSY data

Simulations were generated based upon the manual assignments conducted previously at University of Alberta [68]. Here an example is given to illustrate how the simulation were done. Figure 4.13 shows the manual assignment for Met3 and Thr4 which are used to generate the COSY and TOCSY peaks that should exist for these residues. The generated peaks are also shown in the

3 Met		4 Thr				
N		N	116.090	COSY		
HN		HN	8.013			
CA	55.950	CA	61.085	4.438	61.085	4.652
HA	3.840	HA	4.438	4.652	70.400	4.438
CB		CB	70.400	4.652	70.400	1.176
HB1		HB	4.652	1.176	21.310	4.652
HB2		CG2	21.310			
HB2		HG2	1.176	TOCSY		
CG		C	175.000	4.438	61.085	1.176
HG1				1.176	21.310	4.438
CE	16.600					
HE	2.070					
C	177.100					

Figure 4.13: Fragment from the manual assignment listing of the N-domain of chicken skeletal troponin-C (1-90). Met3 and Thr4 are shown here. Some resonances were not assigned, for example, C_β and H_β of Met3. For Met3, the assigned resonances are not sufficient to simulate COSY cross peak. The simulated cross peaks for Thr 4 are shown on the right of Thr4's manual assignment.

figure. Resonance frequencies from C_β, H_β, C_γ and H_γ are missing for Met3. Therefore no cross peak can be simulated from the manual assignment for this residue. For Thr4, four HCCH-COSY cross peaks can be generated, among them two are symmetrical cross peaks. Similarly, six HCCH-TOCSY cross peaks can be generated as there are two additional peaks of (H_α, C_α, H_γ) and (H_γ, C_γ, H_α).

At the first stage of testing, no chemical shift dispersion was introduced in the simulated data set. That is, two cross peaks are allowed to be partitioned into a spin system as long as they share exactly same chemical shift value. The value of the chemical shift tolerance is therefore zero. The purpose of simulating the exactly data is to confirm that the algorithm works as designed. A total of 674 HCCH-COSY cross peaks and 1014 HCCH-TOCSY were simulated for the protein NTnC. Note that among all of the amino acid residues, glycines are considered to be two-spin systems.

Each has two H_α protons. The amide proton is not detectable in HCCH spectrum. Similarly, alanines, each of which has one H_α and three methyl H_β, are also two-spin systems. The algorithm was designed to extract the amino acid spin systems with three or more spins, hence alanines and glycines are excluded in this particular test. Glycines and alanines are considered during the real data testing presented later in this chapter. Another point of note is that the chemical shift data of aromatic carbons are not available since their resonance frequencies are much higher ($\sim$ 130ppm) than that of aliphatic carbons. As a consequence of the above, and due to several residues not being detected in the manual assignments, only 63 residues of the 90 were simulated.

The test results are summarized in Table 4.1. Note that all the spin systems that were included in the simulated data 63 residues are detected. The execution time for this running is about 5 minutes on a 75 MHz Pentium PC.

Table 4.1: Results for the test of simulated data 1. See text for details

Residues	No. of occurrence of a residue	No. of S.S. simulated as input	No. of S.S. obtained from output	Remarks
Gly	7	N/A	N/A	spin systems with 2 spins are not tested[a]
Ala	10	N/A	N/A	spin systems with 2 spins are not tested[b]
Asp	10	10	10	
Glu	13	9	9	E41,57,67,77 were not simulated due to incomplete data
Lys	4	4	4	
Met	8	7	7	
Gln	4	3	3	
Arg	3	3	3	
Val	4	4	4	
Leu	5	4	4	
Phe	6	4	4	
Ile	5	5	5	
Thr	5	5	5	
Ser	4	3	3	
Pro	1	1	1	
Asn	1	1	1	
Total	90	63	63	

[a]Gly has two H_α which produces only one cross peak pair. This is excluded from the simulation.
[b]For the same reason as Gly.

In the second test, the manual assignment, which results in 63 spin systems, were modified by the introduction of chemical shift dispersion. That is, to better simulate real experimental data,

systematic dispersion less than the pre-defined tolerance was introduced for every frequencies. The main aim of this test is to inspect the algorithm's capability of handling ill-aligned cross peaks. To better explain the dispersion, consider a three-spin system AMX. In principle there should be three cross peaks occurring on either side of the diagonal of the COSY or TOCSY spectrum. These three peaks are represented as (δ_A, δ_X), (δ_A, δ_M) and (δ_M, δ_X). The simulated dispersion involves a pseudo random number generator which gives random numbers R_i between -0.5 and +0.5. The simulated cross peaks are thus modified to $(\delta_A + R_1T, \delta_X + R_2T)$, $(\delta_A + R_3T, \delta_M + R_4T)$ and $(\delta_M + R_5T, \delta_X + R_6T)$, where T is the tolerance value. For this particular testing, T is set to 0.02 ppm for protons and 0.20 ppm for carbons.

An example of a spin system and its simulated COSY/TOCSY cross peaks are listed in Figure 4.14 which can be compared with Figure 4.13.

```
4 THR                                  ---> COSY
N                116.090                      4.436 61.102 4.645
HN                 8.013                      4.645 70.403 4.433
CA                61.085                      4.644 70.338 1.168
HA                 4.438                      1.183 21.309 4.656
CB                70.400
HB                 4.652               ---> TOCSY
CG2               21.310                      4.445 61.106 1.171
HG2               1.176                       1.175 21.390 4.445
C                 175.000
```

Figure 4.14: Fragment from the manual assignment listing of the N-domain troponin-C (1-90). Thr4 is shown. The simulated cross peaks for Thr 4 are shown on the right. Note that a small chemical shift dispersion is introduced in the simulation, for example, 4.446 vs. 4.433.

The result of applying the algorithm to the randomly distributed data set is listed in Table 4.2. Fifty-six of the 63 residues are successfully partitioned and no missing assignment was found. Of the residues that are not successfully separated, 4 are glutamines, one is methionine and 2 are isoleucines. These residues have severely overlapped resonance frequencies. For example, Figure 4.15 shows that E16 has 4 spins which are overlapped with E21. The inability to resolve such overlapped spins is discussed in the discussion section.

Table 4.2: Results for the test of simulated data II. See text for details

Residues	No. of occurrence of a residue	No. of S.S. simulated as input	No. of S.S. obtained from output	Remarks
Gly	7	N/A	N/A	spin systems with 2 spins are not tested[a]
Ala	10	N/A	N/A	spin systems with 2 spins are not tested[b]
Asp	10	10	10	
Glu	13	9	5	E9,16,21,63 were not separated
Lys	4	4	4	
Met	8	7	6	M46 were not separated with Glutamine
Gln	4	3	3	
Arg	3	3	3	
Val	4	4	4	
Leu	5	4	4	
Phe	6	4	4	
Ile	5	5	3	I19 and I62 are not separated.
Thr	5	5	5	
Ser	4	3	3	
Pro	1	1	1	
Asn	1	1	1	
Total	90	63	56	

[a]Gly has two H_α which produces only one cross peak pair. This is excluded from the simulation.
[b]For the same reason as Gly.

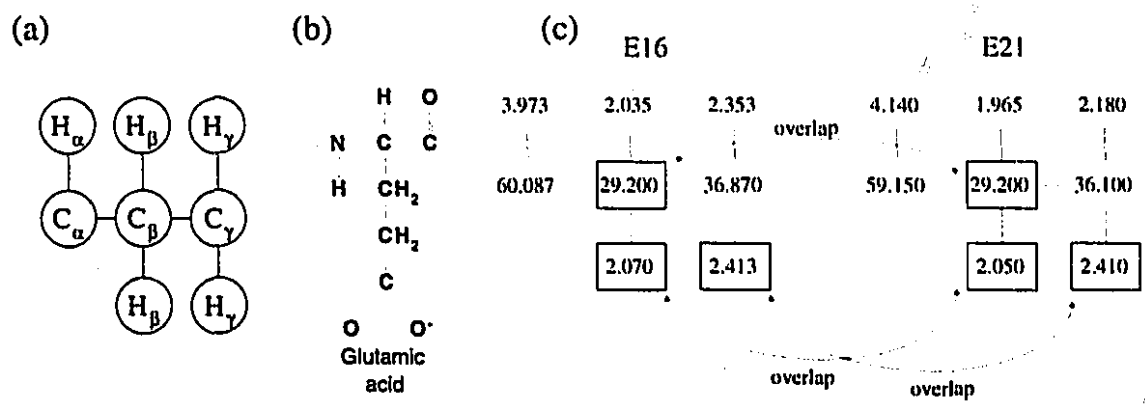


Figure 4.15: An example of two residues with three degenerate resonances. (a) The graph representation of the glutamic acid. (b) The chemical structure of the glutamic acid. (c) Glu16 and Glu21 are shown with their chemical shifts. Resonances in the boxes overlap.

4.3.2 Analysis of experimental 3D HCCH-COSY/TOCSY data

The success of the test on the simulated data indicates that the problem of chemical shift degeneracy can be successfully resolved by the algorithm. The capability of handling missing peaks and spectrum artifacts is however inadequately tested by the simulated data. Therefore, it is still necessary to conduct a test using the real data.

3D HCCH-COSY/TOCSY spectra of the test protein troponin-C were obtained from University of Alberta [68]. Cross peaks in these spectra were picked automatically from a quick run of the CAPP software [56]. No refinement in terms of peak picking were done since the original spectra and the peak picking program were not available to the authors. A total of 915 HCCH-COSY and 710 TOCSY cross peaks were picked by the CAPP software. 321 of the 915 COSY peaks and 225 of the 710 TOCSY peaks can be verified as real peaks by comparing with the previously conducted manual assignment.

Since extensive spectrum folding is employed in the multidimensional NMR experiments, the actual ^{13}C chemical shifts are given by $x \pm nSW$, where x is the ppm value of a carbon obtained from the spectrum, n is an integer and SW is the spectral width. It is necessary to unfold the ^{13}C chemical shifts so that our program can work on the real ^{13}C chemical shift data. A ^{13}C 2D HMQC peak list is available from the same source for this unfolding purpose. The unfolding procedure is divided into two stages. First each of the HCCH-COSY and TOCSY cross peaks (H_i, C_i, H_j) are examined against the ^{13}C HMQC peak list. If the 2D ^{13}C HMQC cross peak (H_i, $C_i - SW$) is found, the 3D cross peak is corrected to (H_i, $C_i - SW$, H_j). The same procedure is also applied to the HMQC peaks (H_i, C_i) and (H_i, $C_i + SW$). Secondly, for each 3D cross peak (H_i, C_i, H_j), if no corresponding 2D ^{13}C HMQC (H_i, $C_i \pm nSW$) is found, a statistical ^{13}C chemical shift database [78] is used to empirically determine the unfolded value of carbon chemical shifts. Following this the 915 HCCH-COSY peaks and 710 TOCSY peaks are used as the input for our program. Various proton and carbon chemical shifts tolerance values are checked to get good partitioning. Essentially, a small tolerance generates more reliable results. In practice, however, small tolerance is unable to find all the spin systems due to the experimentally inconsistent chemical shift values, i.e., the same spin could have different chemical shifts in different spectra. A large tolerance might

incorrectly merge the independent spin systems together. Compromise should be chosen carefully. Table 4.3 shows the partitioning results of the 915 COSY and 710 TOCSY peaks based upon the proton chemical shift tolerance 0.03 ppm and ^{13}C tolerance 0.40 ppm.

Table 4.3: Results for the test of real data

Residues	No. of occurrence of a residue	No. of A.A. obtained from output	Remarks
Gly	7	5	G33,43,50,69,71
Ala	10	10	A1,8,10,12,25,31,90, (A20,24,60 not separated)
Asp	10	4	D89, (D5,27,59 not separated)
Glu	13	4	E9,16,21, (E17,M18,V65 not separated
Lys	4	3	K40,55 (K87, Q85 not separated)
Met	8	3	M3,18,86
Gln	4	2	Q51,85
Arg	3	3	R11,47,84
Val	4	2	V65,80
Leu	5	4	L14,42,58,79
Phe	6	1	F13
Ile	5	5	I19,37,61,62,73
Thr	5	4	T4,39,44,54
Ser	4	3	S2,38,70
Pro	1	1	P53
Asn	1	1	N52
Total	90	55	

As can be seen from Table 4.3 some of the amino acid spin systems are incorrectly merged together, e.g., A20, A24 and A60 were given as a large spin system. This is because all of their resonance frequencies overlap. By checking the manual assignment, those three alanines share common H_α, H_β, C_α and C_β frequencies. (see Figure 4.16) Resolving such cases, after the automated assignment is done, is a relatively simple manual task.

Another point of note from Table 4.3 is that some expected spin systems are missing. For example, out of the 10 aspartic acids, only 4 can be found. This is mainly due to the missing of crucial peaks in the experimental data. Aspartic acid is an AMX spin system and therefore should have one αH and two βH's. According to our algorithm, all of the correlations between (H_α, $H_{\beta 1}$), (H_α, $H_{\beta 2}$) and ($H_{\beta 1}$, $H_{\beta 2}$) must be observed in order to place H_α, $H_{\beta 1}$ and $H_{\beta 2}$ into a spin system. The algorithm's condition is stricter than regular manual assignments procedure

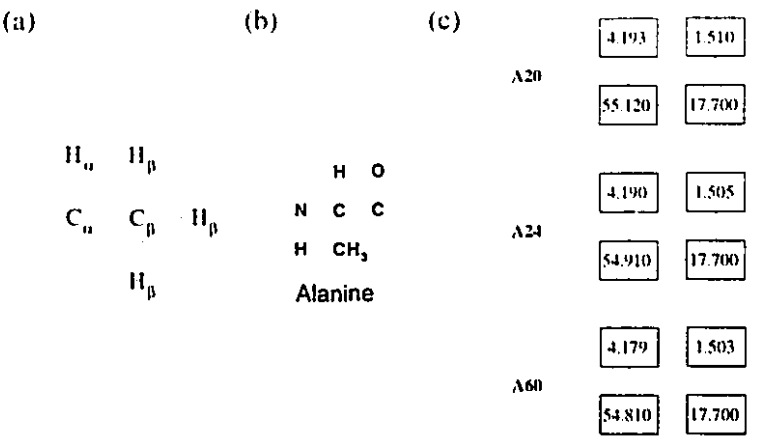


Figure 4.16: (a) The graph representation of an alanine. (b) The chemical structure of an alanine. (c) A20, A24 and A60 are shown with their chemical shifts. Resonances in the boxes overlap. It can be seen that these three alanines have nearly degenerated chemical shifts.

since avoiding incorrect merge is essential for computer-assisted assignment tool. By carefully checking the peak lists, for D30, D32, D36, D59, D66, D68 and D89, the correlations between $H_{\beta 1}$ and $H_{\beta 2}$ are all missing, i.e., neither the COSY ($H_{\beta 1}$, C, $H_{\beta 2}$) nor the TOCSY ($H_{\beta 1}$, C, $H_{\beta 2}$) cross peak was found in the peak lists. This is probably due to the fact that these βH cross peaks are too close to the diagonal to be unambiguously identified.

4.4 Discussion

The advantage of using 3D HCCH-COSY/TOCSY experiments to resolve the chemical shift degeneracy is discussed in this section. Comparisons are made to the conventional 2D COSY/TOCSY method.

In Figure 4.17, two amino acid residues whose H_α and H_β have close resonance frequencies are illustrated. In the 2D COSY/TOCSY approach, two cross peaks can be merged into a spin system as long as they share a common resonance frequency and there is a constraint to confirm that these two cross peaks belong to the same spin system. In the above example, the cross peak (4.073, 2.049) and (4.073, 2.175) belong to one spin system, while (4.062, 2.057) and (4.062, 2.943) belong to another spin system. The problem is that 4.073 and 4.062, as well as 2.049 and

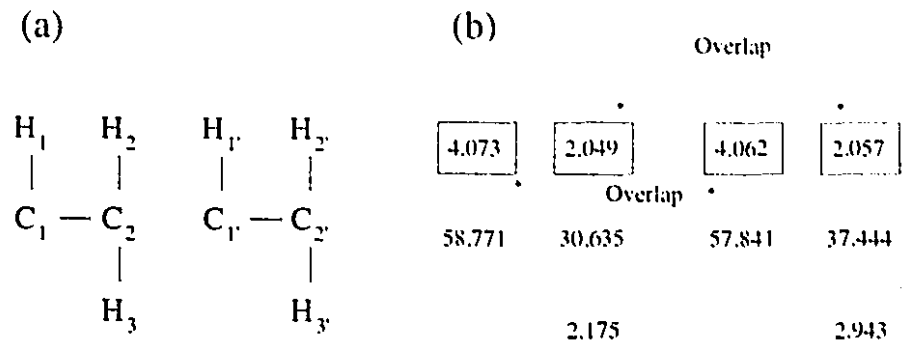


Figure 4.17: Schematic illustration explaining how overlapped resonances are resolved. See text for details. (a) Fragments from two molecules are shown. (b) The chemical shifts of the protons and carbons are displayed. Resonances in boxes are those having significantly overlapped chemical shifts.

2.057, are too close to be distinguished computationally using 2D data alone. As a consequence, all four cross peaks (4.073, 2.049), (4.073, 2.175), (4.062, 2.057) and (4.062 2.943) are incorrectly merged into a large spin system, which is apparently wrong because this large spin system contains three H_β's and as many as four H_γ's. In other words, from 2D NMR, cross peaks (4.073, 2.049) and (4.062, 2.943) are put into the same spin system since they have one frequency in common, 4.073 vs. 4.062. Besides, the presence of the TOCSY peak (2.057, 2.943) incorrectly confirms the merging. In contrast, if 3D NMR cross peaks are available, the computer algorithm will verify if 4.073 and 4.062 bond to the same carbon. If not, these two resonances, 4.073 and 4.062, are put into different spin systems and thus the degeneracy problem is solved. In case that the carbon bonded to 4.073 overlaps with the carbon bonded to 4.062, (see Figure 4.17, if 58.771 and 57.841 cannot be distinguished,) even 3D NMR cannot solve this triple degeneracy situation.

Table 4.4 summarizes the limitations of the present algorithm of handling overlap ambiguities. It should be noticed that Table 4.4 simply lists the theoretical limitations of the algorithm, while in practice, certain overlap can be resolved by using the scoring parameter introduced in equation 4.1.

In general, two factors effect the efficiency of the algorithm. They are the chemical shift degeneracy and the missing peaks. Degenerate chemical shifts usually result in large spin systems which are formed by incorrect merging of two of more spins systems. On the other side, missing of crucial peaks is the major cause of the absence of expected spin systems.

Table 4.4: Summary of the overlap resolution. See Figure 4.17 for notation.

	3D	2D
H_1 overlaps with $H_{1'}$ H_2 overlaps with $H_{2'}$	resolved by checking C_1	unable to resolve
H_2 overlaps with $H_{2'}$ H_3 overlaps with $H_{3'}$	resolved by checking C_2	unable to resolve
H_1 overlaps with $H_{1'}$ H_2 overlaps with $H_{2'}$ C_1 overlaps with $C_{1'}$	unable to resolve	unable to resolve
H_1 overlaps with $H_{1'}$ H_2 overlaps with $H_{2'}$ H_3 overlaps with $H_{3'}$	unable to resolve	unable to resolve

The tests of this algorithm on both simulated and experimental data show that if there is no missing peak, the algorithm correctly produces all the desired spin systems that can be extracted from 3D data. Nevertheless, in the case where critical cross peaks are missing, expected spin systems may not be extracted. To cope with this problem, one can relax some merging conditions, described in Figure 4.9. However, less stringent merging conditions may risk getting incorrect results.

Another feature of our algorithm is that the number of input experiments is flexible. To obtain the complete spin system of an amino acid including all the resonance frequencies and their connectivity relationships, COSY type experiments, which observe three-bond scalar couplings, and TOCSY type experiments, which record long range relay couplings, are required. A sole COSY experiment, can still provide much information about resonance frequencies and connectivity between spins. Because CPA and ASPA both need long range couplings to confirm merge of some cross peaks, the lack of the TOCSY type cross peaks may cause incomplete extraction of certain amino acids, such as the threonines. A sole TOCSY type experiment, on the other hand, provides sufficient information concerning all the resonance frequencies, but fails to provide complete connectivities between spins.

Although ASPA was designed for 3D HCCH-COSY/TOCSY NMR spectra, the idea can be extended to other 3D NMR experiments. The basic concept behind this algorithm is to take advantage of the third dimension as an additional constraint so as to reduce the ambiguities causing

by heavy overlap.

Under certain manual assignment situations, side chain spin systems are investigated after the backbone spins have been successfully assigned. Therefore the backbone H_α, C_α frequencies can be taken as the starting points for the side chain assignment. In the design of ASPA, however, the traditional protein resonance assignment strategy was adopted, i.e., the spin system identification is accomplished prior to sequential assignment. This implies that sequential information of amino acid residues is not incorporated into the algorithm. A possible improvement of the algorithm includes adding an option to supply H_α, C_α frequencies from earlier backbone assignments so that more efficient searches can be achieved due to a resulting smaller searching space. Furthermore, an integrated computer assisted environment for protein resonance assignment using 3D heteronuclear NMR is described in chapter 5. This environment includes complete identification of the protein backbone and side chain resonances, the pattern recognition of the deduced amino acid spin systems, and the creation of the sequential connectivity.

4.4.1 Options of the implemented computer program

The implemented program provides an option to remove the duplicated peaks occurring at the same half of the spectrum. Here duplicated peaks are referred to those peaks picked by the automated picking program as separate peaks but are close in ppm. These peaks might be attributed to the noise level of the spectrum. However, it is also possible that the peaks considered to be duplicated are actually arising from distinct correlations. The algorithm is in a dilemma. On the one hand, close positioned peaks, e.g., (4.29, 35.43, 2.98) and (4.28, 35.38, 3.00), might easily produce unreasonable large spin systems such as the one with two αH's at 4.29 and 4.28 ppm. On the other hand, to merge the close positioned peaks prior to the partitioning process increases the risk of losing significant peaks. If the option of removal of duplicated peaks is enabled, a set of chemical shift tolerance will be used to judge the removal. The default setting for this option is to enable the removal. It might be necessary to disable this setting if a crowded spectrum is processed and the falsely picked peaks have been manipulated by other means in earlier stages.

The algorithm not only merges NMR peaks to form spin systems, it also merges the small,

fragmented spin systems to become bigger ones. Once all of the N initial spin systems are generated, the algorithm merges them and constructs the bigger, less redundant spin systems. Figure 4.18 shows the redundancy and how the corresponding merge can be performed to resolve the redundancy. It should be noticed that the merge in Figure 4.18 is not always safe. In crowded

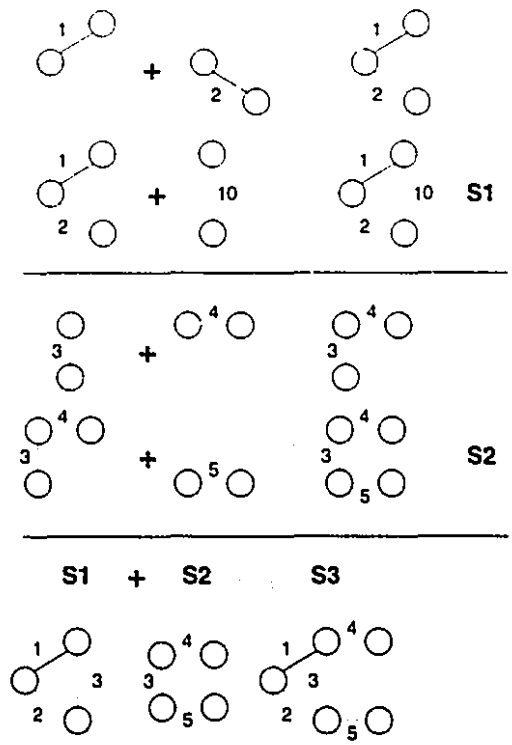


Figure 4.18: The merging of two spin systems. Spin system S1 is constructed from the cross peak 1, 2 and 10. Spin system S2 is constructed from cross peak 3, 4 and 5. Suppose peak 3 and peak 10 are symmetrical cross peaks, i.e., they represent the correlations between the same two protons. It is possible to construct another spin system S3 by merging S1 and S2.

spectra, it might be difficult to verify two peaks are symmetrical ones or not. If peak i and peak j are incorrectly considered as symmetrical peaks, the partitioning algorithm will merge the spin systems originating from peak i and from peak j. This incorrect merge usually gives rise to large spin systems. In other words, the merge operation described in Figure 4.18 has the risk of producing unreasonably large spin systems. The default setting of the option is to enable the spin system merge. The redundant spin systems usually can be effectively eliminated while overlapped spin systems can also be properly merged. In severely crowded spectra, the option of merging spin

system might need to be disabled otherwise many large spin systems will be constructed. By skipping the automated spin system merging, one must manually examine all the output spin systems and determine which of them should be merged or deleted.

4.4.2 Peak unfolding problem

The chemical shifts of carbon nuclei usually span the range from $10 \sim 80$ ppm. To save experimental time, the practical spectral width on the dimension observing carbon is set to around 30 ppm. Apparently extensive spectrum folding is applied. It is introduced earlier this chapter that one can unfold the carbon chemical shifts using the chemical shifts of the directly bonded hydrogen atoms. For example, the carbon in a methyl group, which is easily determined by the small ^{1}H chemical shift, must has relatively small chemical shift. Therefore a 50 ppm chemical shift for the carbon in a methyl group should be unfolded to 50 − (spectral width) ppm.

Usually the experimental spectral width SW is chosen in such a way that the aliphatic carbon resonances are folded no more than once into the observed spectral width. This can be achieved by setting the experimental spectral width equal to 1/3 the aliphatic carbon frequency range. Suppose the aliphatic carbon chemical shifts range from 15 to 75 ppm. The corresponding spectral width can be set to 30 ppm. If the phase ramp for the folded dimension, the carbon dimension, is chosen to be 180°, the folded cross peaks have the opposite sign of non-folded peaks [17]. Given this experimental condition, the carbon resonances can be unfolded using the sign of the corresponding cross peaks.

4.5 Summary

The Aliphatic Side-chain Partitioning Algorithm, ASPA, is proposed in this chapter to automatically extract amino acid spin systems from three dimensional COSY and TOCSY type experiments. This algorithm is extended from the 2D Constrained Partitioning Algorithm, whose main feature is that all the merging steps are accomplished by imposing various constraints. Another distinct feature of ASPA is that by supplying both COSY and TOCSY type experiments, not only

the resonance frequencies of all the spin systems can be determined, but their connectivity relationships are also extracted. This makes the design of subsequent pattern recognition procedure easier.

The extracted amino acid spin systems can be used in various sequential assignment approaches. A number of sequential assignment strategies [7, 9, 18, 25, 79] can be applied to the deduced spin systems. The algorithm described in this chapter provides a strategy to obtain the side chain resonance of proteins. By properly incorporating the backbone and side chain information, an integrated sequential assignment protocol is introduced in the next chapter.

Chapter 5

Development of an Integrated Software Environment for the Sequential Assignment

5.1 Introduction

Resonance assignment is tedious work in protein structure determination from NMR. To develop a computer-assisted resonance assignment package, several steps have to be accomplished.

1. The spin coupling systems of all the residues must be determined.
2. The sequential connectivities between these spin systems must be established based on available interresidue correlations.
3. The spin system identification, i.e., which amino acid each determined spin system actually is, must be conducted.
4. The sequence-specific mapping between the spin systems and the primary sequence of the protein must be created.

In chapter 3, we present a computer algorithm to extract spin systems of the protein backbone. This chapter reports a complete resonance assignment protocol covering the above four steps using heteronuclear 3D NMR. Initially an algorithm was developed to merge data from the protein

backbone and aliphatic side chain spin systems. Secondly, a spin system pattern recognition algorithm is extended to automatically determine all the possible amino acids each spin system may be assigned to. Finally, a mapping algorithm maps the deduced spin systems to their proper positions within the protein primary sequence. The protocol of sequence-specific assignment and the implementation of the algorithms are described in this chapter. Application of all the proposed computer algorithms to a 90-residue protein is reported. The heteronuclear 3D NMR experiments involved in the application include 3D HNCO, HNCA, HCACO, HN(CO)CA, ^{15}N TOCSY-HMQC, HCCH-COSY and HCCH-TOCSY.

5.2 Toward the sequential assignment

As mentioned in chapter 3, the spin systems of individual amino acid residues and the sequential connectivities between these patterns can be derived from heteronuclear 3D NMR. The remaining problem of the protein resonance assignment is to match the derived polypeptides onto the known protein primary sequence. This task can be done manually using human expertise. For example, spectroscopists may notice that one of the spin systems in a polypeptide might be a leucine. Moreover, another spin system three residues away from the leucine may be identified as a glycine. Provided that the leucine-X-X-glycine pattern occurs only once in the primary sequence, it is easy to match the target polypeptide to the correct primary sequence.

To automate this "polypeptide to primary sequence" mapping, it is necessary to have sufficient information about each spin coupling system, i.e., one must know all the possible amino acids each spin system could be. Suppose a polypeptide is composed of 5 spin systems, $S1 - S2 - S3 - S4 - S5$. Spin system $S1$ is identified to be one of the following amino acids: leucine, isoleucine or valine. Similarly, $S2$ can be one of serine, phenylalanine ... , etc., see Figure 5.1. Knowing the amino acids each spin system may be assigned to, it is possible to construct a set of primary sequence combinations. In Figure 5.1 these combinations include Leu-Phe-Arg-Gly-Glu, Leu-Ser-Arg-Gly-Glu, Leu-Asp-Arg-Gly-Glu, ... , etc. If the polypeptide is long enough and the number of possible amino acids each spin system may be assigned to is not too large, a unique mapping between the polypeptide and the primary sequence can be achieved. This is shown Fig-

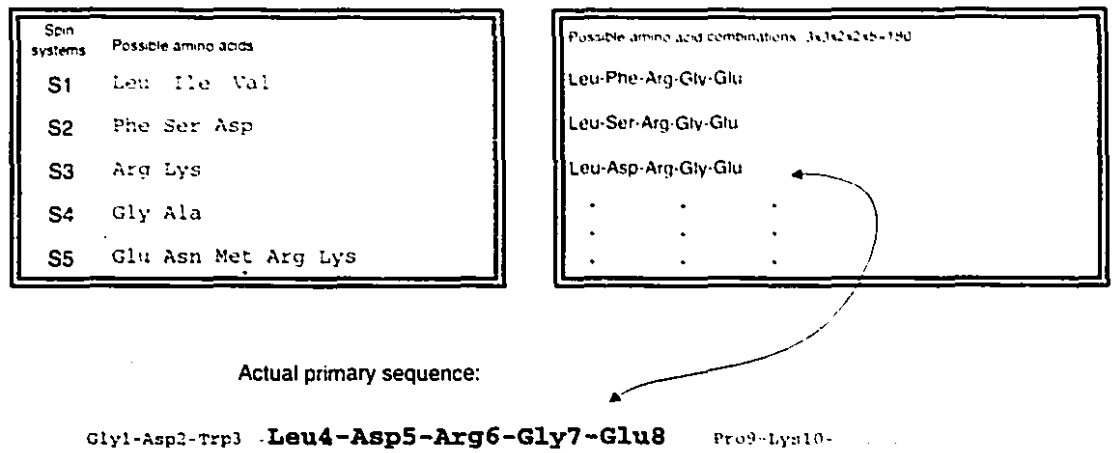


Figure 5.1: Schematic representation of the mapping of a polypeptide $S1$-$S2$-$S3$-$S4$-$S5$ to Leu4-Asp5-Arg6-Gly7-Glu8. Residue $S1$ can be assigned to one of Leu, Ile and Val. There are 180 possible combinations of amino acid sequences for this polypeptide. In this example, the sequence Leu-Asp-Arg-Gly-Glu is the correct mapping on the actual primary sequence.

ure 5.1. Only Leu-Asp-Arg-Gly-Glu has a matching position within the primary sequence, that is residue 4 to residue 8 on the protein's primary sequence, while all the other combinations fail to find a match. Thus it is reasonable to assign the polypeptide $S1 - S2 - S3 - S4 - S5$ to residue 14-15-16-17-18. In the case that a unique mapping is not possible, a ranking parameter can be introduced based on the mathematical similarities between each spin system of the polypeptide and its possible amino acid identity.

The amino acid pattern recognition algorithm(AAPR) was designed to achieve the goal of mapping individual spin pattern to possible amino acids residues. AAPR gives all possible amino acids each of the spin patterns may be assigned to. Every possible assignment has an associated similarity value measuring the likeness between the amino acid and the spin system. In general, it is not easy for computer algorithms to determine the amino acid types for deduced spin systems based on the backbone frequencies exclusively. Several database of protein chemical shifts were published [62, 78]. Although it is possible to classify the backbone spin systems using one of the database, the accuracy of the amino acid type recognition will be higher if the side chain information of each spin system is also available. The more details available of a spin system leads to a more accurate spin pattern recognition. For this reason, the algorithm ASPA [27](Aliphatic

Side-chain Partitioning Algorithm) was designed to retrieve the aliphatic side chain resonances of proteins from heteronuclear 3D NMR. Combining the protein backbone with the side chain information, an amino acid pattern recognition procedure can provide sufficient information about each spin pattern thereby making it possible to automate the mapping between polypeptides and protein primary sequence.

In summary, DBPA was developed to retrieve a protein's backbone resonances and establish parts of the sequential connectivities in the forms of dipeptides. PGA is then responsible for merging retrieved dipeptides to polypeptides. ASPA was designed to extract a proteins' aliphatic side chain information. Having the information of backbone and side chain spin systems, AAPR gives knowledge about the amino acid types of each spin pattern. PBSMA(Protein Backbone Side chain Merging Algorithm) then is required to merge backbone and side chain frequencies. The final step involves an algorithm called PMA(Polypeptide Mapping Algorithm) which maps the polypeptides to the primary sequence. Figure 5.2 shows the relationships between these algorithms.

5.2.1 Integration of backbone and aliphatic side chains

Many 3D NMR experiments have been proposed for protein side chain resonance assignment, such as 3D HCCH-COSY [71–73], HCCH-TOCSY [74], HCC(CO) NH-TOCSY [75, 76] and HCCNH-TOCSY [75, 77]. These experiments resolve the crowded side chain proton regions of traditional 2D DQF-COSY and TOCSY by introducing the third dimension. Therefore the overlapped 2D spectrum can be split into a series of less overlapped 2D planes in the 3D experiments. For example, the ^{1}H-^{1}H planes in 3D HCCH-COSY experiment resemble the 2D ^{1}H-^{1}H COSY spectrum except that these planes are edited by the chemical shifts of the ^{13}C nuclei bonded to the ^{1}H resonance observed in the F_1 dimension of 3D HCCH experiment.

The algorithm ASPA was proposed to automatically extract amino acid spin systems from three dimensional HCCH-COSY and TOCSY experiments.

Side chain spin systems are usually investigated after the backbone spins are successfully assigned provided that the ^{15}N/^{13}C labeled protein samples are available thereby triple resonance 3D NMR data can be acquired. The backbone αH and C_α frequencies can then be taken into

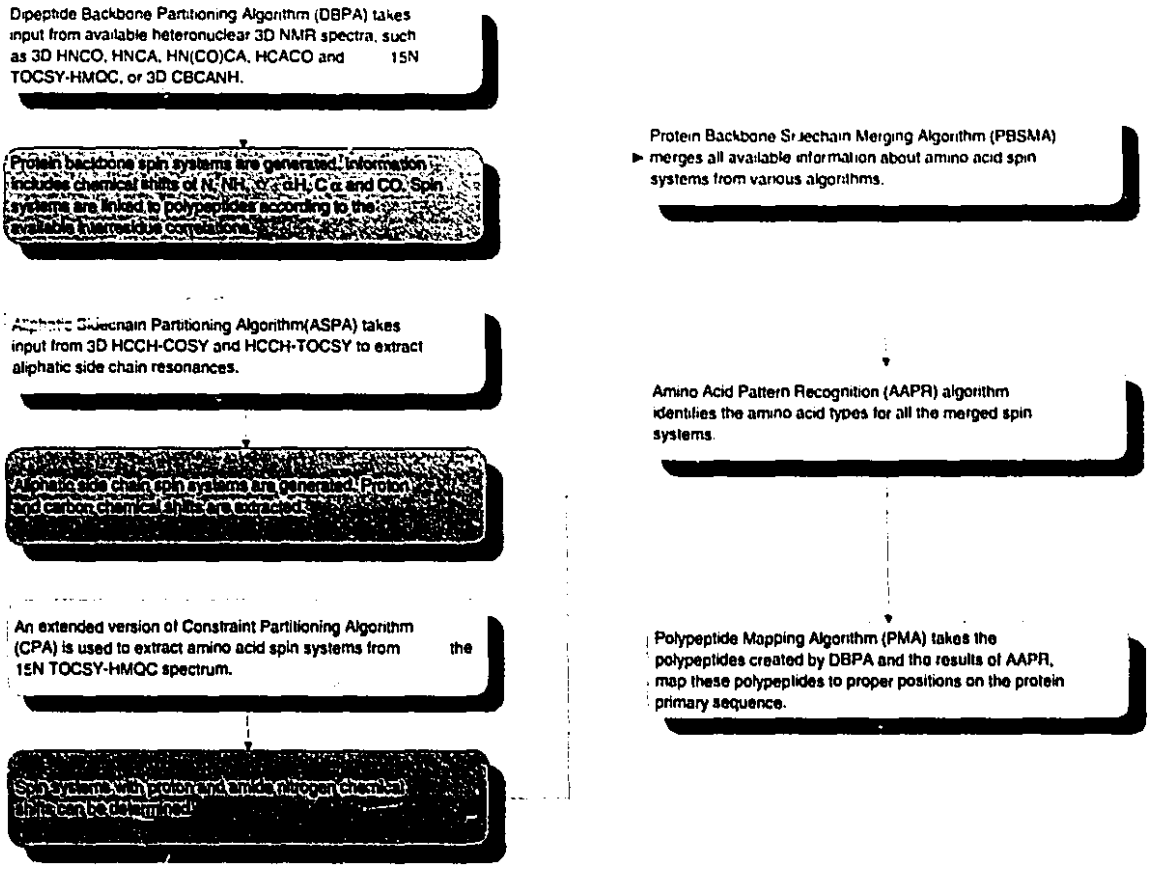


Figure 5.2: A flow diagram of the sequential assignment protocol using heteronuclear 3D NMR.

consideration in creating side chain spin systems. For example, DBPA produces backbone spin systems, the αH and C_α chemical shifts of these spin systems can be taken as the starting points for side chain resonance assignment using ASPA. Thus more efficient searches can be accomplished due to a resulting smaller searching space.

The side chain resonance frequencies can also be revealed by experiments observing long range couplings between protons, such as 2D TOCSY and 3D ^{15}N TOCSY-HMQC. In principle a sole 2D TOCSY or 3D ^{15}N TOCSY-HMQC spectrum has sufficient information to assign a protein's entire side chain and backbone spins. In practice, however, not all spin systems can be identified in a TOCSY experiment, especially in the case of α-helix based proteins which have small $^3J_{NH-\alpha H}$ coupling constants.

Despite the fact that a sole TOCSY experiment sometimes fails to provide sufficient infor-

mation for long spin systems, it is still useful to examine these TOCSY experiments as they have simpler cross peak patterns compared to DQF-COSY. NCPA(Nitrogen Constraint Partitioning Algorithm) was proposed to extract the amino acid spin coupling systems from 2D TOCSY or 3D ^{15}N TOCSY-HMQC experiment. NCPA is complimentary to ASPA as both of them provide side chain information but using different approaches(see Figure 5.3).

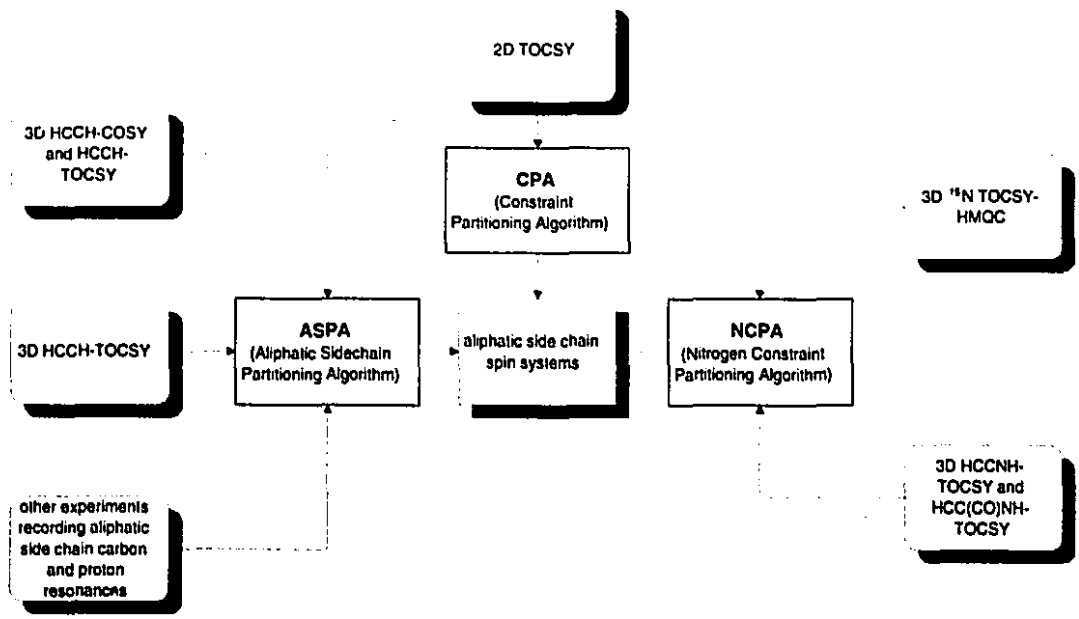


Figure 5.3: Many approaches can be used to obtain protein's side chain resonances. In this example, three algorithms were designed to extract side chain spin systems from 2D and 3D NMR spectra.

The actual procedure to merge the backbone and side chain spin systems are described in the following pseudo codes:

```
void  MergeBackboneSidechain(BackboneSpinsystem_type, ... ,
                             SidechainSpinsystem_type, ... )
{
   //Input: 1. a set of backbone spin systems B1, B2, B3,...
   //       2. a set of side chain spin systems S1, S2, S3,...
   //       3. if available, another set of side chain spin
   //          systems T1, T2,...
   //Examples: Bi were derived from algorithm DBPA, Bi contains
   //                    (N,NH,αH,Cα,CO).
   //          Sj were derived from algorithm NCPA, Sj contains
   //                    (N,NH,αH,βH, ... ).
   //          Tk were derived from algorithm ASPA, Tk contains
   //                    (αH,βH,γH,Cα,Cβ, ... ).
```

```
//
//Output: a set of amino acid spin systems A1,A2...Ai with
//          backbone and side chain information.

for each of the backbone spin systems Bi {
  for each of the side chain spin system Sj {
    compare Bi and Sj;
      if Bi and Sj share several
           common resonances, e.g., αH,NH,N  {
           if another set of side chain spin
           systems Tk are available {
             if((one or more resonance in Bi can be found in Tk) &&
                (one or more resonance in Sj can be found in Tk)) {
                     Al = Bj + Sj + Tk;
              }
           } else
              Al = Bi + Sj;
       }
    }
  }
}
```

To merge a backbone and a side chain spin systems, PBSMA requires that they share several common frequencies. Suppose a backbone amino acid contains five frequencies (NH, N, αH, C$_\alpha$, CO), and a side chain spin system is composed of four spins (NH, αH, βH$_1$, βH$_2$). Depending on the NMR experiments used to construct these spin systems, some resonances may be present in both the backbone and the side chain spin systems. In the above example, NH and αH are the two overlapped resonances. The more overlapped resonances found, the more reliable the merge. In some cases, another experimental data set provides additional information which can be used as extra constraints to confirm the merge of a backbone and a side chain spin system. 3D HCCH-COSY/TOCSY provides aliphatic side chain resonances including αH, Cα, βH, Cβ, ... , etc., these frequencies can be treated as the additional constraints for merging backbone and side chain resonances. In other words, to merge a backbone spin system, which has the resonances of NH, N, αH, Cα, CO, and a side-chain spin system, which has the resonances of NH, αH, βH$_1$, βH$_2$, one can check the spin system output from 3D HCCH-COSY/TOCSY to seek evidence such as the spin system (αH, Cα, βH$_1$, Cβ, βH$_2$, ...) where two frequencies (αH and Cα) can be found in the backbone candidate while another two (αH and βH$_1$) can be found in the side chain candidate.

Once the backbone and side chain spin systems are properly merged, it is possible to perform the amino acid identification process, that is, to recognize these spin systems according to their

spin coupling patterns and chemical shifts. The aim of spin pattern recognition is to obtain all possible amino acids that a spin system may be assigned to. The spin pattern recognition algorithm described in section 2.6 conducts the identification of the deduced spin systems. This algorithm makes use of fuzzy mathematics to recognize the distinct pattern of each amino acid. Many spin system recognition algorithms(e.g., the one by Kleywegt [8]) utilize chemical shift information exclusively. However, our algorithm is able to recognize amino acids' spin topologies based on the fact the each topology has different connectivities between its components. Along with the chemical shift information, the graph theory and fuzzy mathematics based pattern recognition algorithm provides more accurate results in terms of determining the possible amino acids that a spin system corresponds to.

The backbone and side chain spin systems can be extracted from various NMR experiments. Backbone spin systems may come from 3D HNCO, HNCA , HCACO, HN(CO)CA and ^{15}N TOCSY-HMQC, they may also come from 3D CBCANH experiment. Similarly, side chain spin systems may be derived from 3D HCCH type experiments as well as from HCC(CO)NH-TOCSY. Even 2D DQF-COSY and TOCSY NMR spectra provide valuable information for the determination of spin systems. The spin system candidates therefore may consist of various information. Those spin systems from 2D COSY/TOCSY may contain proton frequencies whereas those spin systems derived from 3D HCCH COSY/TOCSY may be composed of carbon and proton frequencies. Moreover, the spin systems may differ from each other in terms of the connectivity relationships. Spin systems from TOCSY type experiments may not contain the details of side chain connectivities. For example, TOCSY type experiments are unable to distinguish spin systems 4.53(αH), 2.25(βH), 1.93(βH) from system 4.53(αH), 1.93(βH), 2.25(γH) as it is not generally easy to determine if a specific peak is arising from short or long range coupling. Figure 5.4 provides a summary of the three different kinds of spin systems described above and several experimentally observed spin systems are given as examples.

Figure 5.5 illustrates how an experimentally observed amino acid spin system is mapped to various amino acid residues. The standard amino acid patterns may contain protons only; protons and nitrogens; protons and carbons; or protons, carbons and nitrogens, depending on the available NMR experiments. The proton database of the standard 20 amino acid was adopted

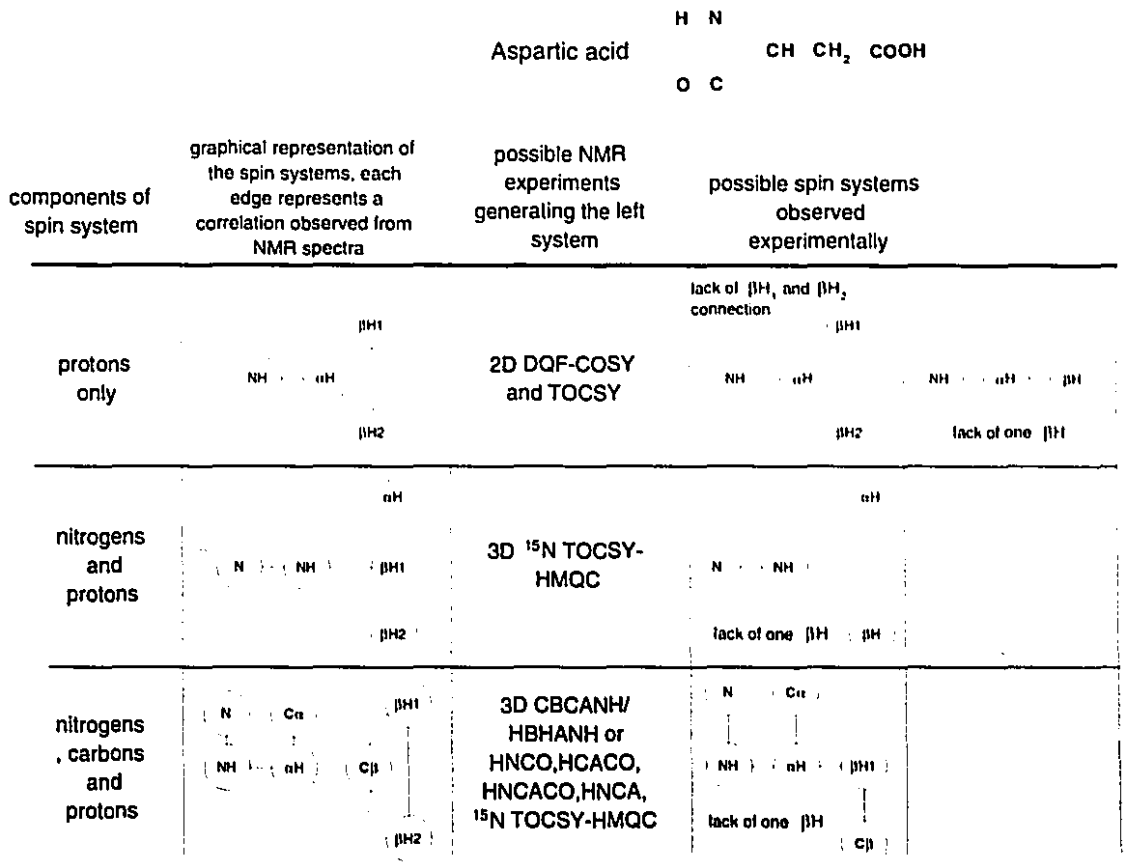


Figure 5.4: Aspartic acid is used to show a spin coupling systems can have various types of nuclei. The possible experiments generating these systems also listed.

from Groß [62]. The nitrogen database was adopted from Choy [70] and the carbon chemical shift database was adopted from Wishart [78]. Note that in Figure 5.5 there might be more than one mapping from an observed spin system to a standard amino acid. For each of the mapping there is an associated value representing the similarity between the observed and the standard spin systems. Details about the similarity values is presented in section 2.7. After performing the pattern recognition on all of the extracted amino acid spin systems, a "spin pattern to residues" table can be created where one can locate all the possible amino acids that each spin system can be assigned to. Figure 5.6 shows a small segment of such a table. Note that amino acids with low similarity values were eliminated to shorten the table.

A brief summary is presented for the topics described up to this point. Amino acid spin sys-

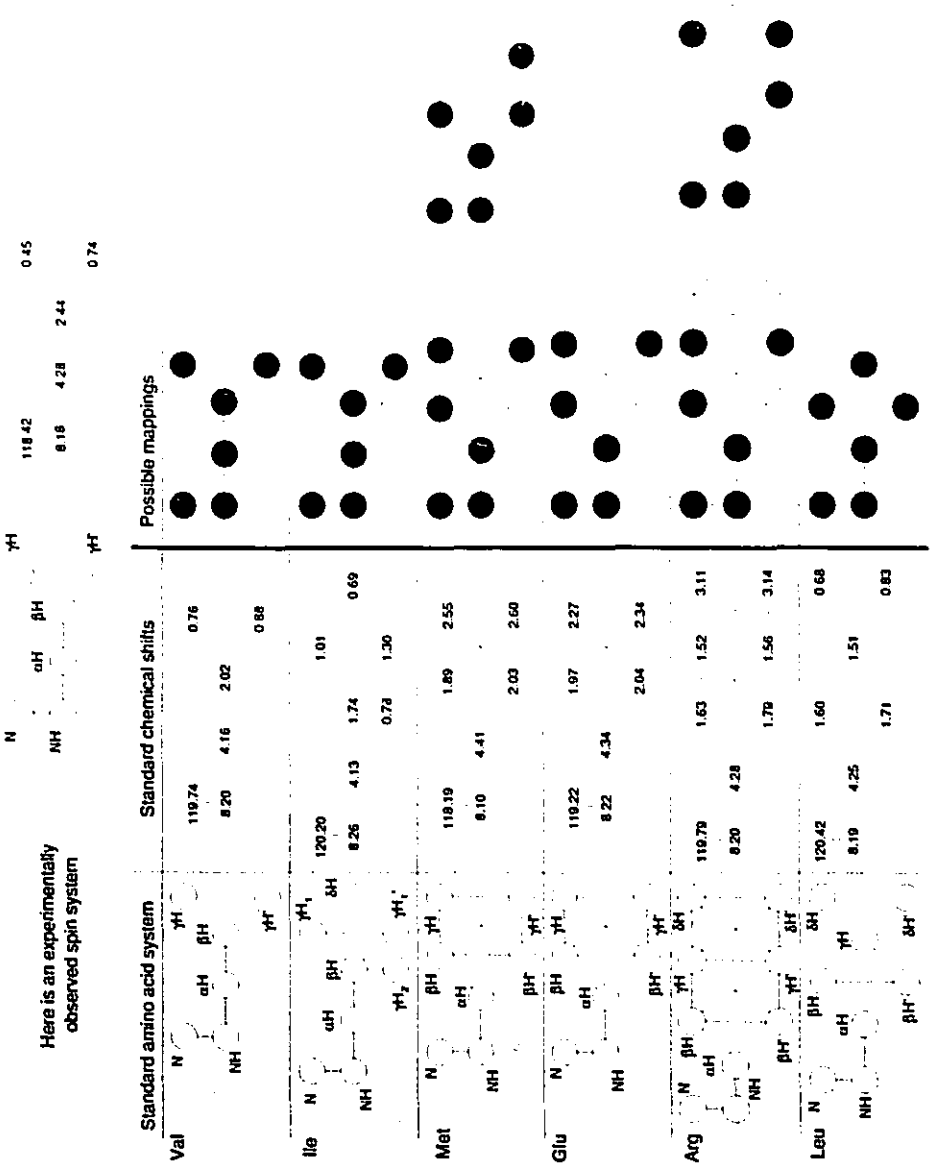


Figure 5.5: Schematic representation of the mappings between an observed spin system and its amino acid candidates: Val, Ile, Met, Glu, Arg and Leu. Note that there could be more than one mapping for the same amino acid, such as the cases of Met and Arg.

tems with backbone and side chain information are derived. The identities of these spin systems are examined, that is, a table, such as the one shown in Figure 5.6, will be given so that all the possible amino acids that a spin system may be assigned to will be listed. The sequential assignment

15	Gln 0.877	**Glu 0.854**	Met 0.738	Ile 0.627	Arg 0.615	Lys 0.615	Leu 0.595					
11	**Ala 0.898**	Leu 0.819	Arg 0.794	Ile 0.781	Val 0.725	Met 0.620	Glu 0.616	Gln 0.610	Thr 0.545	Ser 0.535	Phe 0.488	Asn 0.415
75	**Arg 0.820**	Leu 0.799	Ile 0.702	Met 0.702	Val 0.665	Gln 0.636	Glu 0.584					
77	Arg 0.931	Leu 0.908	Ile 0.868	**Ala 0.841**	Val 0.779	Glu 0.747	Gln 0.738	Met 0.692	Phe 0.654	Thr 0.619	Asp 0.568	Gly 0.565
81	Arg 0.856	Lys 0.856	**Phe 0.726**	Ser 0.617	Glu 0.591	Thr 0.591	Leu 0.585	Met 0.585	Gln 0.581	Val 0.568		
82	Ile 0.650	Arg 0.636	Lys 0.636	**Leu 0.589**	Gln 0.586	Met 0.570	Glu 0.570	Pro 0.441				
88	Thr 0.972	Asn 0.774	Met 0.771	Gln 0.730	Phe 0.715	Val 0.685	Glu 0.671	Asp 0.666	Arg 0.636	Leu 0.629	Ile 0.621	**Ser 0.595**
66	Val 0.819	Ile 0.763	Gln 0.705	**Glu 0.671**	Leu 0.666	Arg 0.662	Ala 0.657	Met 0.632	Phe 0.547	Gly 0.468	Ser 0.436	Thr 0.420
32	**Glu 0.813**	Gln 0.804	Val 0.728	Ile 0.728	Thr 0.702	Ser 0.697	Lys 0.673	Arg 0.673	Gly 0.669	Leu 0.631		

Figure 5.6: A "spin-system to amino-acids" table. Spin system No. 15 can be assigned to one of Gln, Glu, Met, Ile, Arg, Lys or Leu. This table was generated by the Amino Acid Pattern Recognition algorithm. The number below each amino acid denotes the similarity between that amino acid and the spin system on the very left. A higher similarity indicates a closer match. The values range from 0 to 1.

problem is partially solved by using triple resonance 3D NMR since these experiments provide the interresidue correlations from which polypeptides can be built. The rest of the resonance assignment task is to map these polypeptides to their actual positions within the primary sequence with the help of the "spin system to amino acids table". This task can be achieved manually since spectroscopists usually have additional information at hand to guide them through the mapping of the polypeptides. Here a general purpose sequential assignment protocol was proposed to automate the mapping. This protocol aims at giving an additional tool to help spectroscopists to handle tedious assignment work. The first step of the sequential assignment protocol involves a conversion of the "spin-system to amino-acids" table to an "amino acid residue to spin systems" table. Figure 5.7 illustrates such a conversion. Once the conversion is done, the remaining work is to check each of the polypeptides against the "amino-acid-residue to spin-systems" table. If a polypeptide can be located in the table, the corresponding assignment is immediately determined. In Figure 5.8, a nine-residue polypeptide is used to explain the assignment procedure. The Polypeptide Mapping Algorithm, PMA, was designed to carry out the mapping. The pseudo codes of PMA are listed here.

spin systems	possible amino acids
S1	Thr Asn Met Gln .....
S2	Val Ile Gln Glu .....
S11	Ala Leu Arg Ile .....
S15	Gln Glu Met Ile .....
S75	Arg Leu Ile Met .....
S77	Arg Leu Ile Ala .....

amino acid residues	possible spin systems
Leu79	S11,S75,S77, ...
Val80	S2, ...
Met81	S1,S15,S75, ...
Met82	S1,S15,S75, ...
Val83	S2, ...
Arg84	S11,S75,S77, ...
Gln85	S1,S2,S15, ...
Met86	S1,S15,S75, ...

Figure 5.7: Conversion between a "spin-system to amino-acids" table to the "amino-acid-residue to spin-systems" table.

```
void MapPolypeptide(primary_sequence, polypeptides,
                                    SpinSystemToAminoAcid_table )
{
  //Input: 1. protein's primary sequence R1 - R2 - R3 -...- Rm.
  //       e.g.: Glu9-Ala10-Arg11-Ala12-Phe13-Leu14-Ser15-Gly16-Glu17- ...
  //
  //       2. a set of polypeptides: P1, P2, P3, ... ...
  //       e.g.: P1 =S15-S11-S75-S77-S81-S82-S88-S66-S32
  //             where S stands for spin systems.
  //
  //       3. spin-systems to amino-acids table which maps each spin
  //          system to the possible amino acids.
  //       e.g.:
  //       Spin system          Possible amino acids
  //      ------------------------------------------------------------
  //          S15               Gln,Glu,Met,Ile, ... ...
  //          S11               Ala,Leu,Arg,Ile, ... ...
  //          S75               Arg,Leu,Ile,Met, ... ...
  //          S77               Arg,Leu,Ile,Ala, ... ...
  //
  //
  //
     Known the protein's primary sequence, it is possible to convert
       the above table to the "amino-acid-residue to spin-systems" table ;
  //       e.g.:
  //       Residue              Possible spin system candidates
  //      ------------------------------------------------------------
  //          Glu9              ... ... ,S25,S15,S12, ... ...
  //          Ala10             ... ,S54,S11,S13, ... ...
  //          Arg11             ... ..,S74,S75,S5, ... ... .
  //          Ala12             ... ... ..,S49,S77,S95, ... ...
```

```
  //              :
  //

     for each of the polypeptide P_i = S_i1 - S_i2 - S_i3 - ... - S_in  {
        for each of the amino acid residue R_j in the primary sequence {
           check(1,j);  // to see if S_i1 can be found in the candidate
                        // list of R_j ;
        }
     }
}
void check(integer p,integer q)
{
     if spin system S_ip can be found in the candidate list
      of residue R_q
     {
         if (p≤n)                  // spin system S_iq is
                                   //  the end of polypeptide P_i
         and (q+(n-p))≤m{ // assure that there are enough number
                                   // of residues remaining
                                   // in the primary sequence to be
                                   // mapped to polypeptide P_i
                 check(p+1,q+1);
                                   // call itself recursively
         } else if (p==n) {
                    a mapping is found; // S_i1--->R_j
                                   // S_i2--->R_j+1
                                   // S_i3--->R_j+2

                                   // S_in--->R_j+n
                  }
      }
}
```

In the pseudo codes the function `check()` is called recursively to compare each element of a polypeptide with a residue of the primary sequence. If `check()` reaches the end of the polypeptide, a proper mapping is located as shown in Figure 5.8.

5.2.2 *Applications*

A sequential assignment protocol is describe in the previous section. The protocol involves two major steps. In the first step amino acid spin systems are extracted from NMR spectra, then linked to form polypeptides. In the second step, all amino acid spin systems are identified according to their spin topological patterns. As a result, polypeptides can be mapped to the primary sequence automatically. Each of these tasks can be achieved through various strategies, both

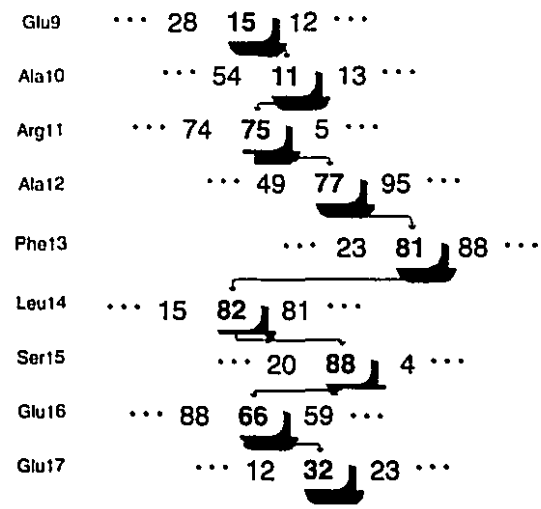


Figure 5.8: Illustration of a possible sequential assignment of the polypeptide 15-11-75-77-81-82-88-66 to Glu9-Ala10-Arg11-Ala12-Phe13-Leu14-Ser15-Glu16-Glu17. The numbers on the right are the spin system numbers.

manually or using computer algorithms. To illustrate the effectiveness of the sequential assignment protocol, several computer algorithms were implemented to accomplish all the mentioned tasks. The details of these algorithms are already described in previous sections while this section presents the application of these computer programs to a real case.

Sample protein is the calcium-loaded regulatory N-domain of chicken skeletal troponin-C (NTnC) residue 1-90. Uniformly enriched ^{15}N and ^{13}C NTnC were also prepared. Available heteronuclear 3D NMR experiments include 3D HNCA, 3D HNCO, 3D HNCOCA, 3D HCACO, 3D ^{15}N TOCSY-HMQC and NOESY. Peak lists of the above NMR experiments were given to the authors by the University of Alberta [68]. Peaks were picked using the CAPP pick peaking program [56], then processed by a filter program to remove some of the false peaks [68].

The amino acid spin systems can be derived from three separated algorithms each using a different set of NMR experiments. Algorithm DBPA involves several triple resonance heteronuclear 3D NMR experiments and is able to deduce the backbone spin systems. In addition, polypeptides can be created since the interresidue information can also be observed from some triple resonance NMR experiments. The details of DBPA are presented in chapter 3. DBPA gave 98 output protein backbone spin systems, 58 of which can be verified against the separately conducted manual

assignment [68]. Using the interresidue information embedded in the NMR cross peaks, 161 dipeptides can be created based on the 98 spin systems. Further, a total of 5432 polypeptides with length from 3 to 26 were built from this 161 dipeptides.

Besides triple resonance experiments, spin systems can also be determined from TOCSY type experiment exclusively as long as sufficient long range couplings can be observed. Algorithm NCPA was used to extract spin systems composed of amide nitrogen and protons from ^{15}N TOCSY-HMQC. Application of NCPA to the 90-residue NTnC gives a total of 83 spin systems of which 73 can be verified against manual assignment [68]. The tolerance values for comparing proton and nitrogen chemical shifts were chosen to be 0.02 ppm and 0.20 ppm, respectively.

Side chain resonances occur in crowded aliphatic regions of NMR spectra. Therefore complete assignment of side chain resonances is a challenging undertaking especially for large proteins. The algorithm ASPA was designed for the 3D HCCH-COSY/TOCSY NMR spectra. For protein NTnC, nine hundred and fifteen HCCH-COSY peaks and 710 HCCH-TOCSY peaks were automatically picked by CAPP. The output of ASPA includes 60 spin systems among which 55 can be verified against the manual assignment. However there are 395 unpartitioned cross peaks which may arise from the falsely picked peaks by the automaticaliy peak picking program. Figure 5.9 summarizes the spin systems information retrieved up to this point.

The remaining task, that is, the second part of the sequential assignment protocol involves the integration of available spin system information, the recognition of amino acid types and the mapping of polypeptides to their anticipated position on the protein primary sequence.

Three types of information are available for the spin systems.

1. The backbone spin systems containing sequential information from triple resonance NMR.

2. The spin systems derived from TOCSY type correlations.

3. The side chain spin systems determined from 3D HCCH type experiments.

Algorithm PBSMA analyzed these data and gave 40 spin systems with detailed side chain correlations and 32 spin systems with TOCSY correlations on the side chain. Figure 5.10 is the schematic representation of these two types of spin systems and their corresponding building blocks. Once

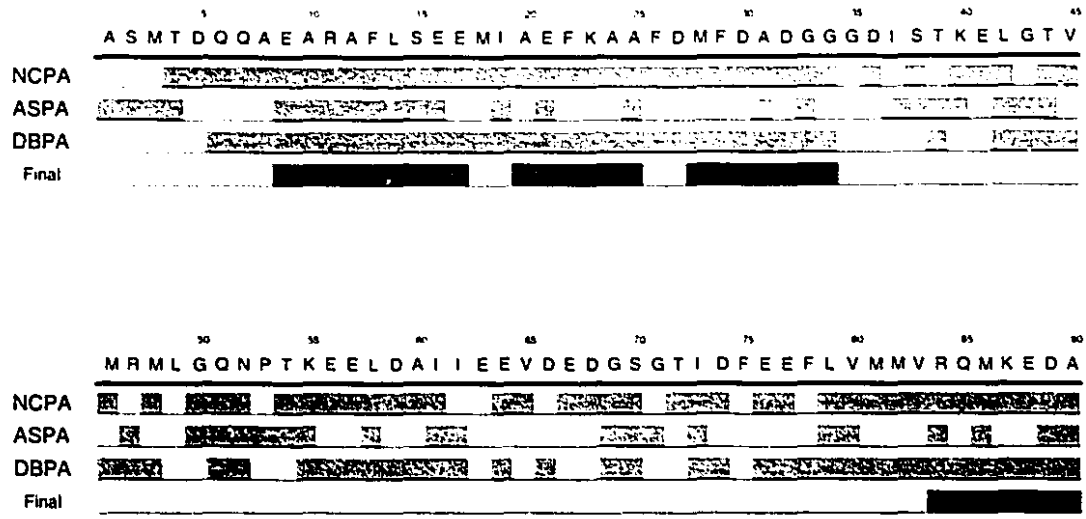


Figure 5.9: The results of the sequential assignment protocol for the 90-residue protein NTnC. NCPA represents the extracted residues using 3D ^{15}N TOCSY-HMQC and Nitrogen Constraint Partitioning Algorithm. ASPA represents the extracted side chain spin systems using 3D HCCH-COSY, HCCH-TOCSY and Aliphatic Side-chain Partitioning Algorithm. DBPA represents the extracted backbone spin systems using 3D HNCO, HCACO, HNCO, HN(CO)CA, ^{15}N TOCSY-HMQC and Dipeptide Backbone Partitioning Algorithm. "Final" represents the sequence-specific assigned residues. Lack of sufficiently long backbone polypeptides between residue 35 and 80 prevents automated sequence-specific assignment in that region. However, individual residue's assignment is still obtained.

the complete amino acid spin systems, that is, the backbone and side chain, are constructed as shown in Figure 5.10, they can be identified using algorithm AAPR. Figure 5.6 shows a fragment of the output from AAPR. In the final stage, PMA mapped the 5432 polypeptide candidates to the primary sequence based on the similar information shown in Figure 5.6. PMA gave a total of 2161 mappings. Of these, many are redundant. For example, polypeptide 8-9-49-15-11 (where the numbers denote spin systems numbers) was assigned to Gln6-Gln7-Ala8-Glu9-Ala10, while simultaneously the polypeptide 8-9-49-15-11-75 was assigned to Gln6-Gln7-Ala8-Glu9-Ala10-Arg11. It is obvious that the former is a redundant mapping of the latter. A set of rules were introduced to remove such redundancies. In addition, human expertise and intuition can also be applied to reduce the number of mapping. Details about these rules are described in the next section.

The final assignment includes mapping of a 14-residue polypeptide to "Gln7 Ala8 Glu9

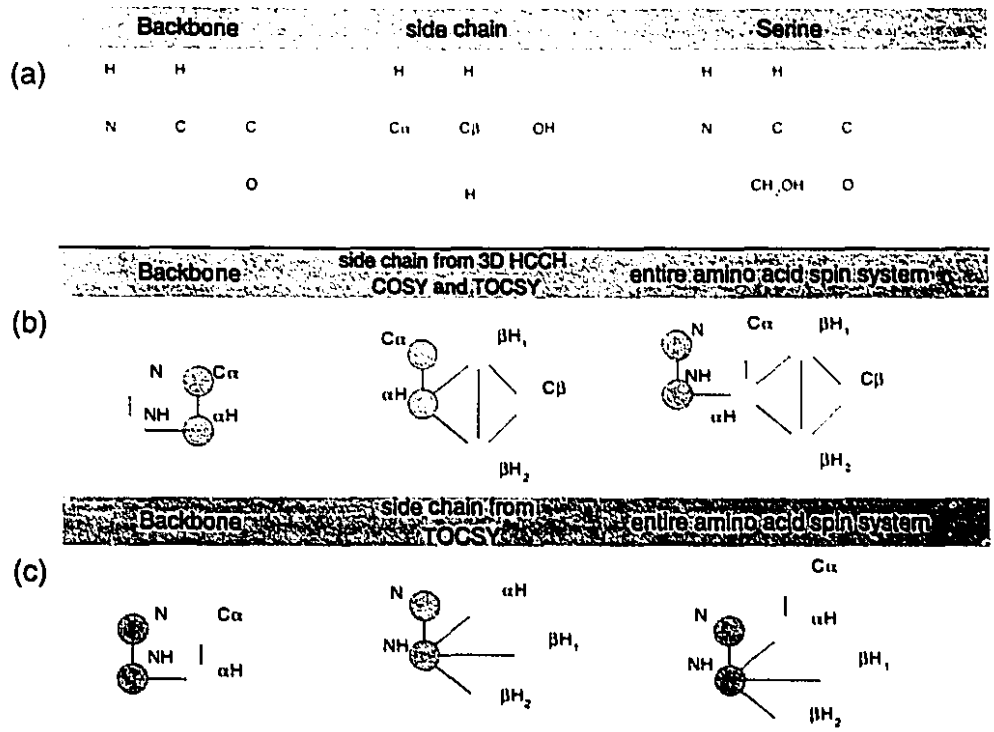


Figure 5.10: Illustration of the merging of backbone and side chain spin systems. Filled circles represent overlapped resonances. (a) Chemical structure of serine's backbone and side chain. (b) Using 3D HCCH-COSY and HCCH-TOCSY, it is possible to obtain the carbon frequencies of side chains. Thus the merged spin system contains proton and carbon frequencies. (c) Using 3D ^{15}N TOCSY-HMQC, the side chain spin system contains a nitrogen frequency.

Ala10 Arg11 Ala12 Phe13 Leu14 Ser15 Glu16 Glu17 Met18 Ile19 Ala20", a 7-residue polypeptide to "Ile19 Ala20 Glu21 Phe22 Lys23 Ala24 Ala25", a 7-residue polypeptide to "Met28 Phe29 Asp30 Ala31 Asp32 Gly33 Gly34", a 7-residue polypeptide to "Arg84 Gln85 Met86 Lys87 Glu88 Asp89 Ala90". Figure 5.9 lists the summary of the results.

5.3 Discussion

The algorithm PBSMA provides a way to integrate the backbone and side chain data of proteins. The detailed information of the backbone and side chain can be determined independently using different NMR data. PBSMA does not limit itself to certain types of experiments. On the contrary, PBSMA accepts a wide variety of spin systems including spin systems composed of

protons, spin systems composed of protons and carbons, in addition to spin systems composed of protons, carbons and nitrogens. As examples to illustrate the effectiveness of PBSMA, two sets of experimental data were used. The first set of NMR data includes 3D HNCO, HNCA, HCACO, HN(CO)CA and ^{15}N TOCSY-HMQC. The spin systems of the backbone and parts of the sequential connectivities can be obtained from those five experiments. Furthermore, ^{15}N TOCSY-HMQC alone provides another set of spin systems based on the long range couplings between protons. PBSMA merges the backbone and side chain data by overlapping each backbone spin system with its side chain counterparts. They can be merged if reasonable overlapping between these two can be verified. The second sets of NMR data to test PBSMA includes two more experiments, 3D HCCH-COSY and HCCH-TOCSY. These two NMR experiments give an additional set of side chain spin systems which in turn act as constraints to increase the accuracy of PBSMA. The more experimental data available, the more accurate the backbone and side chain merging can be anticipated.

The second algorithm discussed in this chapter is the Amino Acid Pattern Recognition algorithm(AAPR). Originally this pattern recognition algorithm was designed for spin systems containing protons only. The extended version is presented where other atoms can be included in the spin patterns. The availability of hetero atoms (carbon and nitrogen) mainly depends on experimental data. Spin patterns with carbon resonances can be derived provided that the NMR data set which correlates carbon and proton frequencies is available. Here the flexibility of the resonance assignment protocol is evident, since the accepted types of experimental data are almost unlimited.

The third and the most important algorithm is PMA(Polypeptide Mapping Algorithm). It is responsible for mapping all the polypeptides to their proper positions on the protein primary sequence. In principle, unique mapping can be determined provided that the polypeptide is sufficiently long. For example, a 10-residue polypeptide could end up being mapped uniquely to residue 18-27 on the primary sequence. However, in practice, this kind of uniqueness is not likely since each component residue of a polypeptide could be assigned to many amino acids (although only one of them can be correct). This usually leads to multiple possibilities. A set of rules was introduced to manipulate such kind of multiple possibilities. The first rule is the simplest one and depends on human experience. Recall in conducting amino acid pattern recognition, each spin sys-

tem is assigned a similarity value with respect to an amino acid. This value is calculated according to a mathematical similarity between the query spin system and the standard one. Both topological and chemical shift similarities are considered during the process. The similarity values range from 0 to 1, a higher value indicating a closer match. Having obtained each residue's similarity, an overall score of each mapping can be given. Suppose a polypeptide $S_1 - S_2 - S_3 - \ldots - S_n$ is mapped to the primary sequence between residue R_p and $R_{(p+n-1)}$. The similarity value between S_i and $R_{(p+i-1)}$ is denoted as r_i. The overall score of this mapping is defined as

$$\sqrt[n]{\prod_{i=1}^{n} r_i} \tag{5.1}$$

Because all of r_i's range between 0 to 1, the overall score also ranges from 0 to 1. A higher score indicates a more likely mapping. The first rule to reduce the number of multiple mapping is to simply set a threshold for the overall scores from all the mapping. Only those mapping with scores higher than this threshold remains. A typical threshold value is between 0.6 to 0.7 and is determined by the quality of all spectra and individual user's experience. This threshold of mapping score can eliminate a large number of multiple mapping.

The second rule deals with the redundant mapping. Suppose polypeptide P_i can be mapped to S_i, and another polypeptide P_j can be mapped to S_j, where S are segments on the primary sequence. Suppose P_i is a subset of P_j and S_i is a subset of S_j. Mapping P_i-S_i will be discarded since it is a subset of mapping P_j-S_j. For example, polypeptide ($S5 - S4 - S91 - S94 - S95$) is mapped to residue 30-34 while polypeptide ($S21 - S78 - S5 - S4 - S91 - S94 - S95$) is mapped to residue 28-34. It is obvious that the former is a redundant mapping with respect to the latter. In cases that more than one polypeptide can be mapped to residue 28-34, a third rule is used which suggests that the polypeptide with the highest mapping score is picked. Similarly, if a polypeptide can be mapped to more than one position, the mapping with the highest score is kept.

By employing these rules, the number of mapping can be reduced to a reasonable figure whereby users are able to manually select the final assignment.

The efficiency of the Polypeptide Mapping Algorithm is a considerable improvement over its predecessor, the Tree Search Algorithm(TSA). Consider the following example. A polypeptide with N spin systems is to be assigned. In Figure 5.11, suppose each amino acid residue has 10

possible spin system candidates, only one of them can be assigned to the corresponding residue. In

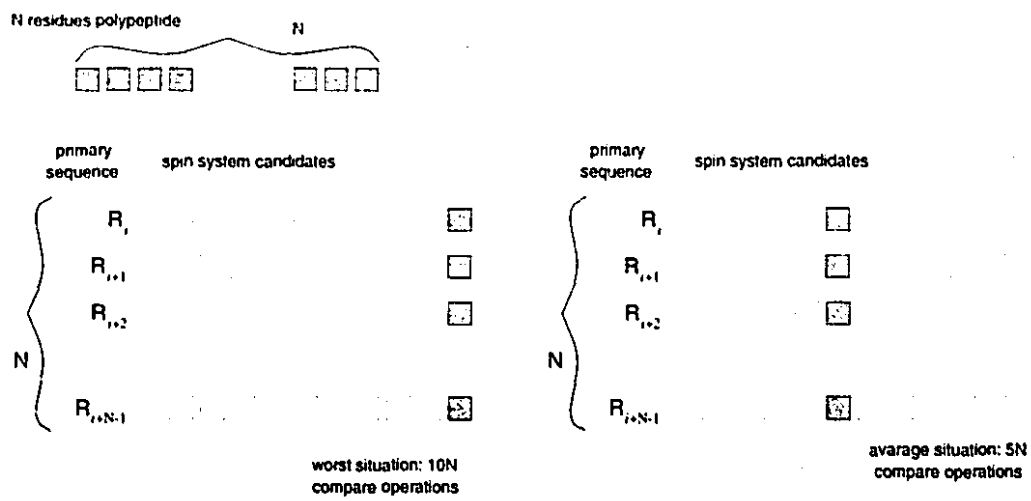
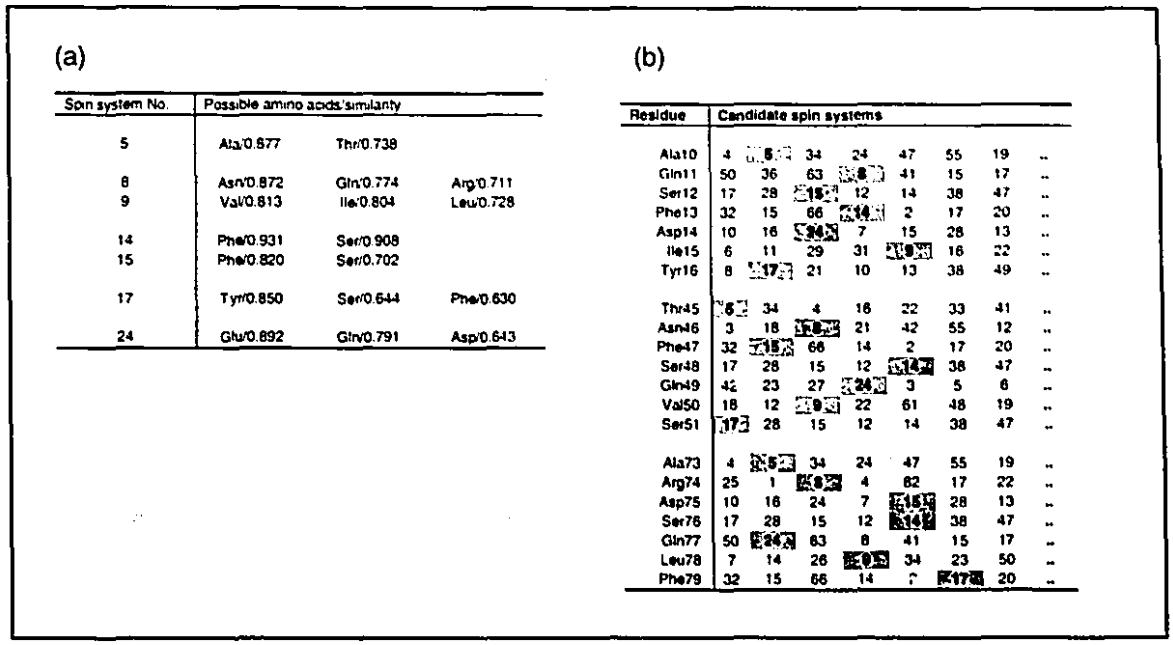


Figure 5.11: Performance analysis of Polypeptide Mapping Algorithm. An N-residue polypeptide is to be assigned. In the worst situation, the correct spin systems all occur at the end of the spin system candidate lists. $10N$ comparisons are expected in this case. In the average situation, the correct spin systems occur in the middle of the spin system candidate lists thus a total of $5N$ comparisons can be expected.

the worst situation the correct mapping occurs at the last spin systems of each residue, thus a total of $10N$ comparison operations must be conducted in order to assign this N residues polypeptide. In the average situation, however, only $5N$ comparisons are needed.

5.3.1 Options of the implemented computer program

The Polypeptide Mapping Algorithm provides several options by adjusting which one can fine tune the sequential mapping procedure. The first option deals with the multiple mapping of a polypeptide. This is illustrated in Figure 5.12. Recall that before PMA starts the actual mapping actions, amino acid types of each observed spin system must have been obtained. In the sequential assignment protocol, the amino acid types are determined by the Amino Acid Pattern Recognition algorithm. As an example, the amino acid types of several spin systems and are shown in Figure 5.12(a). The mathematical similarity between each spin system and its amino acid candidate is also shown. The similarities are calculated by comparing the query spin system with the statis-



(a)

Spin system No.	Possible amino acids/similarity		
5	Ala/0.877	Thr/0.738	
8	Asn/0.872	Gln/0.774	Arg/0.711
9	Val/0.813	Ile/0.804	Leu/0.728
14	Phe/0.931	Ser/0.908	
15	Phe/0.820	Ser/0.702	
17	Tyr/0.850	Ser/0.644	Phe/0.630
24	Glu/0.892	Gln/0.791	Asp/0.643

(b)

Residue	Candidate spin systems							
Ala10	4	5	34	24	47	55	19	..
Gln11	50	36	63	8	41	15	17	..
Ser12	17	28	15	12	14	38	47	..
Phe13	32	15	66	14	2	17	20	..
Asp14	10	16	24	7	15	28	13	..
Ile15	6	11	29	31	9	16	22	..
Tyr16	8	17	21	10	13	38	49	..
Thr45	6	34	4	16	22	33	41	..
Asn46	3	18	58	21	42	55	12	..
Phe47	32	15	66	14	2	17	20	..
Ser48	17	28	15	12	14	38	47	..
Gln49	42	23	27	24	3	5	6	..
Val50	18	12	9	22	61	48	19	..
Ser51	17	28	15	12	14	38	47	..
Ala73	4	5	34	24	47	55	19	..
Arg74	25	1	8	4	82	17	22	..
Asp75	10	16	24	7	15	28	13	..
Ser76	17	28	15	12	14	38	47	..
Gln77	50	24	63	8	41	15	17	..
Leu78	7	14	26	0	34	23	50	..
Phe79	32	15	66	14	?	17	20	..

Figure 5.12: An example showing multiple assignments of a polypeptide. The polypeptide contains 7 spin systems, S5-S8-S15-S14-S24-S9-S17. (a) The possible amino acids each deduced spin system can be assigned to. The associated similarity values are also shown. (b) The polypeptide has three mapping positions within the primary sequence: Ala10-···-Tyr16, Thr45-···-Ser51 and Ala73-···-Phe79.

tically determined standard amino acids. It is possible that an amino acid with low similarity is assigned eventually. In Figure 5.12(b) a simulated sequential assignment is listed. For example, the candidate spin systems for residue Ala10 includes S4, S5, S34, S24, ... , etc. Each of them has an associated similarity which is directly translated from Figure 5.12(a). Consider the polypeptide S5-S8-S15-S14-S24-S9-S17, three different assignments can be located from Figure 5.12(b). The polypeptide might be assigned to Ala10-Glu11-Ser12-···-Tyr16, Thr45-Asn46-Phe47-···-Ser51 or Ala73-Arg74-Asp75-···-Phe79. Each assignment has its overall assigning score which is calculated using equation 5.1. The assignment bearing with the highest score is considered more likely to be the correct one. However, users have the option to output all valid assignments or the one with the greatest assigning score. If all valid assignments are chosen to be printed out, the users must manually verify them. With respect to each polypeptide, PMA outputs the assignment having the greatest assigning score by default.

The second option provided by PMA is best explained by an example. Consider a 10-residue polypeptide S54-S45-S8-S9-S49-S58-S68-S34-S35-S97 as shown in Figure 5.13. To make the

Residue	Candidate spin systems	Comments
Gly20	..... 54	
Asp21	.... 45	
Thr22	..... 8	
Ile23	..... 9	
Ser24	..... 49 .	
Gln25	.... 58	
Arg26	... ...	68 is not in the candidate list of Arg26
Lys27	... ...	34 is not in the candidate list of Lys27
Ala28	... ...	35 is not in the candidate list of Ala28
Phe29	... ...	97 is not in the candidate list of Phe29

Figure 5.13: An example showing that different lengths of polypeptides might lead to different assignment results. If the assigning polypeptide is chosen to be S54-S45-S8-S9-S49-S58-S68-S34-S35-S97, there is no corresponding assignment within the known primary sequence. An assignment, however, can be determined once the assigning polypeptide is chosen to be a shorter one, S54-S45-S8-S9-S49-S58.

sequential assignment, PMA attempts to locate the query polypeptide in the "residue to spin-systems" table. If the polypeptide appears in the table, the corresponding assignment can be determined immediately from the left column of the table. If the query polypeptide doesn't have a corresponding position in the table, it is considered that the assignment for the polypeptide on this particular protein segment is unavailable. However, although the assignment for the entire polypeptide is not available, there might be chances to assign part of the polypeptide. To investigate the possibility for such a "partial" assignment, it is possible to customize PMA so that one or more residue can be subtracted from the either end of the query polypeptide. Attempts then are addressed toward the assignment of that shorter polypeptide. This procedure can be conducted iteratively until an assignment is reached. In the example in Figure 5.13, the assignment is determined for the polypeptide S54-S45-S8-S9-S49-S58 which is four-residue shorter than the original one. The implication of the above iterative subtraction procedure falls on the fact that the sequential connectivity between S58 and S68 might be incorrectly established in earlier stage. In other words, S68-S34-S35-S97 should not be connected with the S54-S45-S8-S9-S49-S58 during the polypeptide generation period. This possible mistake resulted in a 10-residue polypeptide which apparently is too long to be successfully assigned. Finally, it should be noticed that by turning on the iterative mapping option of PMA, there are risks that more assignments will be output and

needed to be analyzed. A reasonable compromise is to set a lower limit of permitted length of the assigning polypeptides, for example, four or five residues. By restricting the length of the polypeptides, the output mappings will remain manageable.

5.4 Summary

The sequential assignment protocol presented in this chapter is the first one using amino acid pattern recognition and heteronuclear 3D NMR. Detected spin patterns are compared with the 20 standard amino acid patterns to determine their amino acid types. The comparison is twofold. First, the similarities of chemical shifts are calculated. Secondly, the topological consistency between the query pattern and standard pattern is checked. Using heteronuclear 3D NMR, the chemical shifts can include nitrogen, carbon and proton nuclei. DBPA(Dipeptide Backbone Partitioning Algorithm), ASPA(Aliphatic Side-chain Partitioning Algorithm) and NCPA(Nitrogen Constraint Partitioning Algorithm) are introduced to extract the backbone and side chain spin systems from heteronuclear 3D NMR spectra. PBSMA(Protein Backbone Side-chain Merging Algorithm) is introduced to incorporate all the spin system information and prepare spin patterns for amino acid type determination. These "amino-acid-type-determined" spin systems then become the input of PMA(Polypeptide Mapping Algorithm). Along with the sequential connectivities extracted in DBPA, PMA completes the final assignment.

A complete resonance assignment protocol is presented. It is fully automated and generic, i.e., not limited to any particular NMR experiment. However, the automated assignment protocol is not designed to entirely replace the manual assignment. Proper human intervention still plays an important role in the computer-assisted protein resonance assignment.

Chapter 6

Conclusion

This thesis presents automated approaches for doing resonance assignment of proteins from heteronuclear 3D NMR spectra. Algorithms for extraction of spin systems and establishment of sequential connectivities are described in the contexts of a constrained partitioning mechanism and a graph theory based pattern recognition procedure. The proposed algorithms are validated with simulated and experimental data based on implemented computer programs.

6.1 Contributions to original research

The research described in this thesis represents contributions to the development of automated NMR resonance assignment tools. The specific contributions to original research may be stated as follows:

1. An automated spin system extraction algorithm is proposed. The algorithm has the following features:

   (a) The input data can be taken from a wide variety of triple resonance heteronuclear 3D NMR spectra. No specific type of NMR experiment is required for the input.

   (b) The algorithm is able to determine if the input data provide sufficient information to accomplish the complete backbone resonance assignment.

(c) The backbone spin systems are determined based on strict merging rules to overcome spectral overlap. The sequential connectivities are established in the form of dipeptides which subsequently can be converted into polypeptides.

(d) The extraction algorithm is flexible so that users can control the behavior of the algorithm through various options. The deduced polypeptides can be placed to corresponding protein primary sequence manually or by the automated approach discussed in chapter 5.

2. An algorithm for determining the side chain spin systems of proteins has been formulated. The implemented computer program is applied to the 3D HCCH-COSY and HCCH-TOCSY experiments. Use of heteronuclear correlation experiments can resolve certain chemical shift degeneracy problems which can't be otherwise handled by the conventional 2D COSY and TOCSY experiments. The available carbon chemical shifts are able to separate potential spectral overlap. Previous conducted backbone assignment provides αH and C_α resonances which can be incorporated into the side chain extraction algorithm to further separate the crowded aliphatic side chain region. The deduced aliphatic side chain spin systems can be integrated with the independently determined protein backbone spin systems thus making a fully automated sequential assignment protocol possible.

3. An automated sequential assignment protocol is applied to the information of spin systems determined in the above two stages. The protocol is centered around a spin pattern recognition algorithm. The algorithm determines the amino acid types for the deduced amino acid spin systems using mathematical graph theory and fuzzy subset theory. The determined amino acid types along with the detailed spin system information are sent into a mapping algorithm to complete the sequence-specific resonance assignment. In most available automated assignment packages, the determination of amino acid types and the mapping of deduced polypeptides are not completely automated. The proposed protocol presents the possibility of developing a fully automated assignment package although the complexity of the experimental data make the complete automation not realistic at the present time.

4. In addition to the above three studies, the possibility of using fewer NMR experiments, in our study, 3D ^{15}N TOCSY-HMQC and NOESY-HMQC, to conduct the sequence assignment is investigated. An algorithm for determining spin systems from the 3D ^{15}N TOCSY-HMQC experiment is presented. Despite the fact that a sole TOCSY experiment might not be able to provide all the long range correlations, the TOCSY data contain sufficient information for constructing the backbone and part of the side chain spin systems. With some extension, the spin pattern recognition algorithm is able to determine the possible amino acid types for all the deduced spin systems. Along with the 3D ^{15}N NOESY-HMQC spectrum, which provides through-space sequential connectivities, the deduced spin systems can be placed to the corresponding primary sequence.

6.2 Practical application

The implemented computer programs have been applied to a real-life situation: the automated assignment of a 90-residue protein. In general, available NMR experiments maybe different from the ones demonstrated. In planning resonance assignment of proteins using computer-assisted methods, the current studies may be useful in the following ways:

1. *Determination of protein aliphatic side chain resonances.* Given correlation spectra of the side chain resonances, our algorithm can determine aliphatic side chain spin systems automatically. If the αH and C_α resonances have been independently assigned prior to the determination of side chain resonances, the αH and C_α information can assist the partitioning algorithm in such a way that every merging of a spin system must be initiated from an available $\alpha H/C_\alpha$ node.

2. *Extraction of protein backbone spin systems.* Our algorithm offers flexibility in this aspect. The input NMR experiments can be a single 3D CBCANH spectrum or it can be a set of many triple resonance NMR experiments. The algorithm is able to inform users whether the input data is a complete set or not. Moreover, the through-bond sequential connectivities are established at the same time of the deduction of individual backbone spin systems.

3. *Creation of polypeptide chains from already established dipeptides.* Dipeptides consist of two amino acid spin systems which are already determined either manually or through an automated approach. Once all the dipeptides are ready, our algorithm is able to merge the individual dipeptides into longer stretches which can be further assigned onto the protein primary sequence.

4. In the case where triple resonance NMR experiments are not available, our assignment package is able to take the input from 2D COSY, TOCSY and 3D ^{15}N TOCSY-HMQC spectra and output the individual amino acid spin systems. The sequential connectivities can be determined from the through-space correlations obtained from 2D NOESY or 3D ^{15}N NOESY-HMQC.

5. *Determination of amino acid types.* The amino acid types of the deduced spin systems can be determined automatically through the pattern recognition technique. The input spin systems can be composed of proton, carbon and nitrogen nuclei and can be derived either manually or by computer-assisted methods.

6. *The sequence-specific assignment can be determined automatically.* In this case, the deduced spin systems, the information about amino acid types along with the established polypeptide chains act as the input of the automated mapping procedure. The corresponding positions of the deduced spin systems within the primary sequence can be determined.

The above operations can be conducted independently, that is, users can manually conduct any part of the assignment and then integrate the result into the automated assignment approach.

6.3 Future work

This study has presented several opportunities for future research. In the long term, the possibilities of using various advanced computing methods, such as artificial neural networks, genetic algorithms, parallel algorithms, to automate the protein resonance assignment remain to be explored. In the short term, several related extensions from the current work should be further investigated. They are described in the following.

6.3.1 *Automation of spectrum analysis*

All of the algorithms described in this thesis require the input data to be presented in the form of peak lists. Therefore, a reliable automated peak picking procedure becomes crucial. Unfortunately, due to the complexity of actual spectra, a perfect automated peak picking program remains to be developed. A fully automated resonance assignment package cannot be realized without a robust peak picking program. Current peak picking algorithms are mostly focused on the analysis of peak shapes by comparing the shapes of real and false peaks. A possible extension from our studies is to develop an intelligent peak picking algorithm which considers not only the peak shapes but also the logical relationships between the suspicious peaks and their surroundings. For example, a genuine peak should have coupled partners whereas a false peak should not. By implementing these types of logical constraints, along with the investigation of peak shapes, it should be possible to improve the reliability of the current peak picking procedures.

6.3.2 *Assignment of the aromatic protons*

A direct extension of the aliphatic side chain extraction algorithm is to include the aromatic protons into the assignment target. To cope with the aromatic proton assignments, the algorithm should extract aromatic spin systems as well as create proper relationships between the aromatic ring and its aliphatic partner. The selection of experiments is also important because some NMR experiments don't record aromatic resonances, especially for aromatic carbons.

6.3.3 *Use of information not determined from NMR*

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 software. For example, the protein secondary structures can be roughly determined from various approaches including chemical and computational ones [3]. The availability of secondary structures provides information about the distribution of backbone chemical shifts, especially αH's. This is a useful criterion which should be considered when doing

the sequential mapping of spin systems. Currently there is no systematic approach developed or implemented in our assignment protocol. A well-designed expert system might be necessary to make use of all such types of miscellaneous information.

6.3.4 *Nucleic acids and carbohydrates*

The resonance assignments between proteins, nucleic acids and polysaccharides have fundamental similarities. It is necessary to identify NOE correlations between neighboring residues, which enable one to step along the backbone of the polymer. When degeneracy occurs in the chemical shifts of an assigning residue, it can be resolved through correct identification of the type of that residue and a knowledge of the primary sequence. The techniques developed for protein resonance assignment in principle can be applied to nucleic acids and polysaccharides. Although the details remains to be defined, the development of an automated approach for nucleic acids and polysaccharides resonance assignment is a feasible long term goal.

Appendix A

Derivation of the cross and diagonal peaks of 2D COSY and DQF-COSY experiments

This appendix presents the 2D COSY and double quantum filtered COSY experiments using a more theoretical approach.

First consider the evolution of a density operator under the unperturbed weak coupling Hamiltonian

$$H = \sum_k \Omega_k I_{kz} + \sum_{k<l} \sum 2\pi J_{kl} I_{kz} I_{lz} \tag{A.1}$$

The shift frequency of nucleus k in the rotating frame is defined by $\Omega_k = \omega_{0k} - \omega_{rf}$, with the Larmor frequency ω_{0k} and the rf frequency ω_{rf}, J_{kl} is the scalar coupling between nucleus k and l.

Since all terms in equation A.1 commute, the evolution caused by the individual terms can be computed separately in arbitrary order:

$$\begin{aligned} \sigma(t+\tau) &= \prod_k \exp(-i\Omega\tau I_{kz}) \prod_{k<l} \exp(-i\pi J_{kl}\tau 2 I_{kz} I_{lz}) \sigma(t) \\ &\quad \times \prod_{k<l} \exp(i\pi J_{kl}\tau 2 I_{kz} I_{lz}) \prod_k \exp(i\Omega\tau I_{kz}) \end{aligned} \tag{A.2}$$

or symbolically:

$$\sigma(t) \xrightarrow{\Omega\tau I_{1z}} \xrightarrow{\Omega\tau I_{2z}} \cdots \xrightarrow{\pi J_{12}\tau 2 I_{1z} I_{2z}} \xrightarrow{\pi J_{13}\tau 2 I_{1z} I_{3z}} \cdots \sigma(t+\tau) \tag{A.3}$$

By expanding the exponential, it is straightforward to prove the following relations for I=1/2 spins [80].

$$
\begin{aligned}
e^{-i\varphi I_z} I_x e^{i\varphi I_z} &= I_x \cos\varphi + I_y \sin\varphi \\
e^{-i\varphi I_z} I_y e^{i\varphi I_z} &= I_y \cos\varphi - I_x \sin\varphi \\
e^{-i\varphi I_x} I_z e^{i\varphi I_x} &= I_z \cos\varphi - I_y \sin\varphi \\
e^{-i\varphi I_x} I_y e^{i\varphi I_x} &= I_y \cos\varphi + I_z \sin\varphi \\
e^{-i\varphi I_y} I_z e^{i\varphi I_y} &= I_z \cos\varphi + I_x \sin\varphi \\
e^{-i\varphi I_y} I_x e^{i\varphi I_y} &= I_x \cos\varphi - I_z \sin\varphi
\end{aligned}
\tag{A.4}
$$

Thus, the effects of the chemical shifts, scalar couplings and radio-frequency pulses can be treated as rotations of the angular momentum operators. The effects of some 90° pulses with different phases are summarized.

$$
\begin{array}{cccc}
I_z \xrightarrow{(\frac{\pi}{2})_x} -I_y & I_z \xrightarrow{(\frac{\pi}{2})_y} I_x & I_z \xrightarrow{(\frac{\pi}{2})_{-x}} I_y & I_z \xrightarrow{(\frac{\pi}{2})_{-y}} -I_x \\
I_x \xrightarrow{(\frac{\pi}{2})_x} I_x & I_x \xrightarrow{(\frac{\pi}{2})_y} -I_z & I_x \xrightarrow{(\frac{\pi}{2})_{-x}} I_x & I_x \xrightarrow{(\frac{\pi}{2})_{-y}} I_z \\
I_y \xrightarrow{(\frac{\pi}{2})_x} I_z & I_y \xrightarrow{(\frac{\pi}{2})_y} I_y & I_y \xrightarrow{(\frac{\pi}{2})_{-x}} -I_z & I_y \xrightarrow{(\frac{\pi}{2})_{-y}} I_y
\end{array}
\tag{A.5}
$$

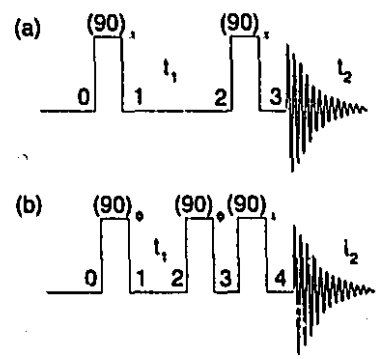


Figure A.1: (a) the pulse sequence of 2D-COSY, (b) the pulse sequence of 2D COSY with the double quantum filter. The numbers denote the points of time.

For a spin system with two I=1/2 spins, the density operator of the basic 2D COSY experiment can be described as follows: (the lower indices of σ_i refer to the points of time in Figure A.1)

the original system is

$$\sigma_0 = I_{kz} + I_{lz} \tag{A.6}$$

after the first 90 degree pulse,

$$\sigma_1 = -I_{ky} - I_{ly}$$

after the evolution time t_1,

$$\begin{aligned}
\sigma_2 \;=\; & -[I_{ky}\cos\pi J_{kl}t_1 - 2I_{kx}I_{lz}\sin\pi J_{kl}t_1]\cos\Omega_k t_1 \\
& +[I_{kx}\cos\pi J_{kl}t_1 + 2I_{ky}I_{lz}\sin\pi J_{kl}t_1]\sin\Omega_k t_1 \\
& -[I_{lx}\cos\pi J_{kl}t_1 + 2I_{kz}I_{lx}\sin\pi J_{kl}t_1]\cos\Omega_l t_1 \\
& +[I_{lx}\cos\pi J_{kl}t_1 + 2I_{kz}I_{ly}\sin\pi J_{kl}t_1]\sin\Omega_l t_1
\end{aligned} \tag{A.7}$$

after the second 90 degree pulse,

$$\begin{aligned}
\sigma_3 \;=\; & -[I_{kz}\cos\pi J_{kl}t_1 + 2I_{kx}I_{ly}\sin\pi J_{kl}t_1]\cos\Omega_k t_1 \\
& +[I_{kx}\cos\pi J_{kl}t_1 - 2I_{kz}I_{ly}\sin\pi J_{kl}t_1]\sin\Omega_k t_1 \\
& -[I_{lz}\cos\pi J_{kl}t_1 + 2I_{ky}I_{lx}\sin\pi J_{kl}t_1]\cos\Omega_l t_1 \\
& +[I_{lx}\cos\pi J_{kl}t_1 - 2I_{ky}I_{lz}\sin\pi J_{kl}t_1]\sin\Omega_l t_1
\end{aligned} \tag{A.8}$$

The third term of σ_3, $I_{kx}\cos\pi J_{kl}t_1\sin\Omega_k t_1$, leads to the diagonal peak at $\omega_1 = \omega_2 = \Omega_k$ while the other diagonal peak at $\omega_1 = \omega_2 = \Omega_l$ is contributed by the seventh term. The fourth term of σ_3, $2I_{kz}I_{ly}\sin\pi J_{kl}t_1\sin\Omega_k t_1$, will resume precession at $\Omega_l \pm \pi J_{kl}$ in the detection period and therefore lead to a cross peak multiplet at $\omega_1 = \Omega_k$, $\omega_2 = \Omega_l$ with antiphase doublet structure. The other cross peak, at $\omega_1 = \Omega_l$, $\omega_2 = \Omega_k$ is contributed by the eighth term, $2I_{ky}I_{lz}\sin\pi J_{kl}t_1\sin\Omega_l t_1$.

Multiple quantum filtering can be achieved by the sequence [81] 90°(φ)–t_1–90°(φ)–90°(x)–acquisition. For the double quantum filter, the phase φ is cycled through the values $\varphi = 0, \pi/2, \pi, 3\pi/2$. The resulting signals are alternately added and subtracted to eliminate all the terms but the pure double quantum coherence. Table A.1 shows one of the possible phase cycling schemes. The pure double quantum state can be represented as

Table A.1: The evolution of density operators for a two-spin system through a phase cycled COSY pulse sequence. The indices of σ are the points of time in Figure A.1.

pulse		density operator for the two-spin system (k,l), $\sigma_0 = I_{kz} + I_{lz}$
$(\frac{\pi}{2})_x$	σ_1	$-I_{ky} - I_{ly}$
	σ_2	$-[I_{ky}\cos\pi J_{kl}t_1 - 2I_{kx}I_{lz}\sin\pi J_{kl}t_1]\cos\Omega_k t_1 + [I_{kx}\cos\pi J_{kl}t_1 + 2I_{ky}I_{lz}\sin\pi J_{kl}t_1]\cos\Omega_l t_1$ $-[I_{ly}\cos\pi J_{kl}t_1 + 2I_{kz}I_{lx}\sin\pi J_{kl}t_1]\cos\Omega_l t_1 + [I_{lx}\cos\pi J_{kl}t_1 + 2I_{kz}I_{ly}\sin\pi J_{kl}t_1]\sin\Omega_l t_1$
	σ_3	$-[I_{kz}\cos\pi J_{kl}t_1 + 2I_{kx}I_{ly}\sin\pi J_{kl}t_1]\cos\Omega_k t_1 + [I_{kx}\cos\pi J_{kl}t_1 - 2I_{kz}I_{ly}\sin\pi J_{kl}t_1]\sin\Omega_k t_1$ $-[I_{lz}\cos\pi J_{kl}t_1 + 2I_{ky}I_{lx}\sin\pi J_{kl}t_1]\cos\Omega_l t_1 + [I_{lx}\cos\pi J_{kl}t_1 - 2I_{ky}I_{lz}\sin\pi J_{kl}t_1]\sin\Omega_l t_1$
$(\frac{\pi}{2})_y$	σ_1	$I_{kx} + I_{lx}$
	σ_2	$[I_{kx}\cos\pi J_{kl}t_1 + 2I_{ky}I_{lz}\sin\pi J_{kl}t_1]\cos\Omega_k t_1 + [I_{ky}\cos\pi J_{kl}t_1 - 2I_{kx}I_{lz}\sin\pi J_{kl}t_1]\cos\Omega_l t_1$ $+[I_{lx}\cos\pi J_{kl}t_1 + 2I_{kz}I_{ly}\sin\pi J_{kl}t_1]\cos\Omega_l t_1 + [I_{ly}\cos\pi J_{kl}t_1 - 2I_{kz}I_{lx}\sin\pi J_{kl}t_1]\sin\Omega_l t_1$
	σ_3	$-[I_{kz}\cos\pi J_{kl}t_1 - 2I_{ky}I_{lx}\sin\pi J_{kl}t_1]\cos\Omega_k t_1 + [I_{ky}\cos\pi J_{kl}t_1 + 2I_{kz}I_{lx}\sin\pi J_{kl}t_1]\sin\Omega_k t_1$ $-[I_{lz}\cos\pi J_{kl}t_1 - 2I_{kx}I_{ly}\sin\pi J_{kl}t_1]\cos\Omega_l t_1 + [I_{ly}\cos\pi J_{kl}t_1 + 2I_{kx}I_{lz}\sin\pi J_{kl}t_1]\sin\Omega_l t_1$
$(\frac{\pi}{2})_{-x}$	σ_1	$I_{ky} + I_{ly}$
	σ_2	$[I_{ky}\cos\pi J_{kl}t_1 - 2I_{kx}I_{lz}\sin\pi J_{kl}t_1]\cos\Omega_k t_1 - [I_{kx}\cos\pi J_{kl}t_1 + 2I_{ky}I_{lz}\sin\pi J_{kl}t_1]\cos\Omega_l t_1$ $+[I_{ly}\cos\pi J_{kl}t_1 - 2I_{kz}I_{lx}\sin\pi J_{kl}t_1]\cos\Omega_l t_1 - [I_{lx}\cos\pi J_{kl}t_1 + 2I_{kz}I_{ly}\sin\pi J_{kl}t_1]\sin\Omega_l t_1$
	σ_3	$-[I_{kz}\cos\pi J_{kl}t_1 + 2I_{kx}I_{ly}\sin\pi J_{kl}t_1]\cos\Omega_k t_1 - [I_{kx}\cos\pi J_{kl}t_1 - 2I_{kz}I_{ly}\sin\pi J_{kl}t_1]\sin\Omega_k t_1$ $-[I_{lz}\cos\pi J_{kl}t_1 + 2I_{ky}I_{lx}\sin\pi J_{kl}t_1]\cos\Omega_l t_1 - [I_{lx}\cos\pi J_{kl}t_1 - 2I_{ky}I_{lz}\sin\pi J_{kl}t_1]\sin\Omega_l t_1$
$(\frac{\pi}{2})_{-y}$	σ_1	$-I_{kx} - I_{lx}$
	σ_2	$-[I_{kx}\cos\pi J_{kl}t_1 + 2I_{ky}I_{lz}\sin\pi J_{kl}t_1]\cos\Omega_k t_1 - [I_{ky}\cos\pi J_{kl}t_1 - 2I_{kx}I_{lz}\sin\pi J_{kl}t_1]\cos\Omega_l t_1$ $-[I_{lx}\cos\pi J_{kl}t_1 + 2I_{kz}I_{ly}\sin\pi J_{kl}t_1]\cos\Omega_l t_1 - [I_{ly}\cos\pi J_{kl}t_1 - 2I_{kz}I_{lx}\sin\pi J_{kl}t_1]\sin\Omega_l t_1$
	σ_3	$-[I_{kz}\cos\pi J_{kl}t_1 - 2I_{ky}I_{lx}\sin\pi J_{kl}t_1]\cos\Omega_k t_1 - [I_{ky}\cos\pi J_{kl}t_1 + 2I_{kz}I_{lx}\sin\pi J_{kl}t_1]\sin\Omega_k t_1$ $-[I_{lz}\cos\pi J_{kl}t_1 - 2I_{kx}I_{ly}\sin\pi J_{kl}t_1]\cos\Omega_l t_1 - [I_{ly}\cos\pi J_{kl}t_1 + 2I_{kx}I_{lz}\sin\pi J_{kl}t_1]\sin\Omega_l t_1$

$$\sigma_3^{2QT} = \frac{1}{4}\left[-\sigma_3(\frac{\pi}{2})_x + \sigma_3(\frac{\pi}{2})_y - \sigma_3(\frac{\pi}{2})_{-x} + \sigma_3(\frac{\pi}{2})_{-y}\right] \tag{A.9}$$

$$= \frac{1}{2}\left[2I_{kx}I_{ly}\sin\pi J_{kl}t_1\cos\Omega_k t_1 + 2I_{lx}I_{ky}\sin\pi J_{kl}t_1\cos\Omega_l t_1 + 2I_{ky}I_{lx}\sin\pi J_{kl}t_1\cos\Omega_k t_1 + 2I_{ly}I_{kx}\sin\pi J_{kl}t_1\cos\Omega_l t_1\right]\sin\pi J_{kl}t_1 \tag{A.10}$$

The third pulse (with a constant phase) generates the single quantum coherence to be detected:

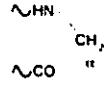
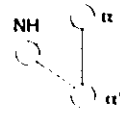
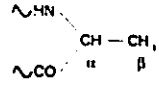
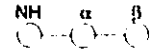
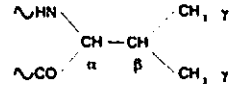
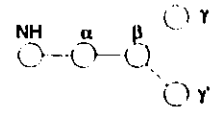
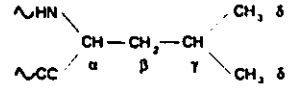
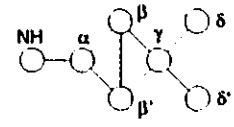
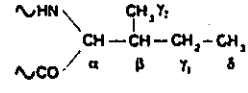
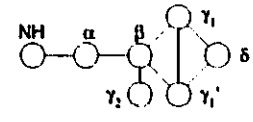
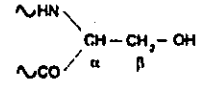
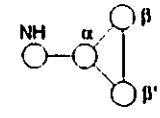
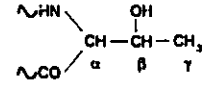
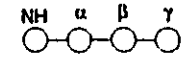
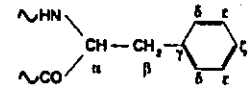
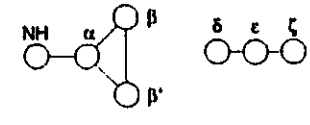
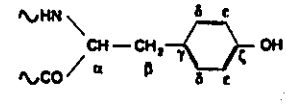
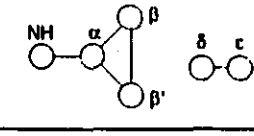
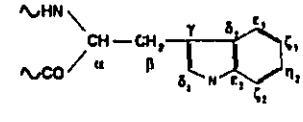
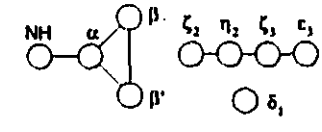
$$\sigma_4 = \frac{1}{2}\left[(2I_{kx}I_{lz} + 2I_{kz}I_{lx})\cos\Omega_k t_1 + (2I_{kx}I_{lz} + 2I_{kz}I_{lx})\cos\Omega_l t_1\right]\sin\pi J_{kl}t_1 \tag{A.11}$$

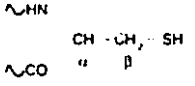
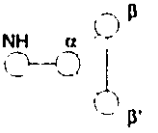
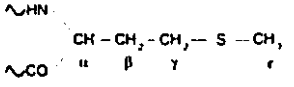
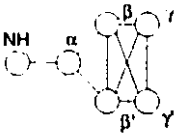
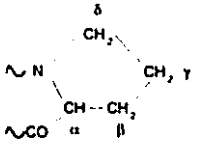
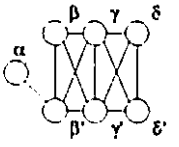
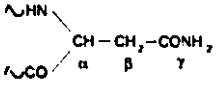
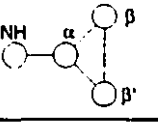
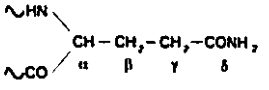
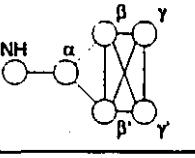
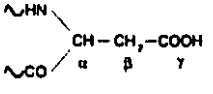
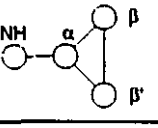
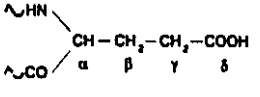
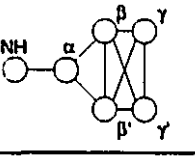
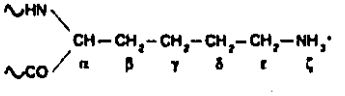
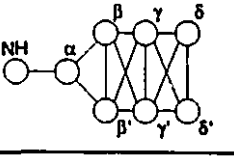
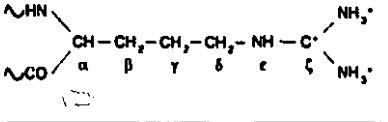
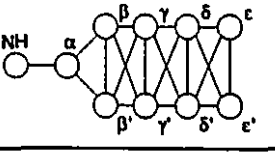
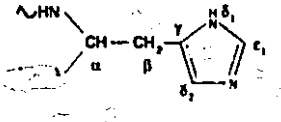
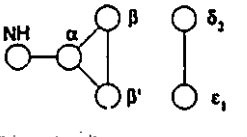
The first and the fourth terms give rise to the diagonal peaks while the second and the third terms lead to the cross peaks. All diagonal and cross peaks consist of antiphase multiplets with almost pure 2D absorption peak shapes. Thus broad diagonal lines can be eliminated. Besides, all the single spin signals are suppressed, particularly those stemming from solvent.

Appendix B

The 20 common amino acids and their spin coupling graphs

This appendix lists the chemical structures and the proton-proton spin coupling graphs of the 20 common amino acids.

Residue	Chemical structure	Spin coupling graph
Gly		
Ala		
Val		
Leu		
Ile		
Ser		
Thr		
Phe		
Tyr		
Trp		

Residue	Chemical structure	Spin coupling graph
Cys		
Met		
Pro		
Asn		
Gln		
Asp		
Glu		
Arg		
Lys		
His		

Bibliography

[1] Wider, G.; Lee, K.; Wüthrich, K. Sequential resonance assignments in protein ^{1}H nuclear magnetic resonance spectra: Glucagon bound to perdeuterated dodecylphosphocoline micelles. *J. Mol. Biol.*, **1982**, *155*, 367–388.

[2] Wüthrich, K. Sequential individual resonance assignments in the proton-NMR spectra of polypeptides and proteins. *Biopolymers*, **1982**, *22*, 131–138.

[3] Wüthrich, K., *NMR of Proteins and Nucleic Acids*. Wiley, New York, NY, 1986.

[4] Wüthrich, K.; Wider, G.; Wagner, G.; Braun, W. Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. *J. Mol. Biol.*, **1982**, *155*, 311–319.

[5] Bax, A.; Grzesiek, S. Methodological advances in protein NMR. *Acc. Chem. Res.*, **1993**, *26*, 131–138.

[6] Clore, G. M.; Gronenborn, A. M. Application of three- and four-dimensional heteronuclear NMR spectroscopy to protein structure determination. *Prog. NMR Spectrosc.*, **1991**, *23*, 43–92.

[7] Cieslar, C.; Clore, A. M.; Gronenborn, A. M. Comupter-aided sequential assignment of protein 1H NMR spectra. *J. Magn. Reson.*, **1988**, *80*, 119–127.

[8] Kleywegt, G. J.; Lamerichs, R. M. J. N.; Boelens, R.; Kaptein, R. Toward automatic assignment of protein 1H NMR spectra. *J. Magn. Reson.*, **1989**, *85*, 186–197.

[9] Kleywegt, G. J.; Vuister, G. W.; Padilla, A.; Knegtel, R. M. A.; Boelens, R.; Kaptein, R. Computer-assisted assignment of homonuclear 3D NMR spectra of proteins. Application to Pike Parvalbumin III. *J. Magn. Reson. B*, **1993**, *102*, 166–176.

[10] Kleywegt, G. J.; Boelens, R.; Cox, M.; Llinás, M.; Kaptein, R. Computer-assisted assignment of 2D 1H NMR spectra of proteins: Basic algorithms and application to phoratoxin B. *J. Biomol. NMR*, **1991**, *1*, 23–47.

[11] Van de Ven, F. J. M. PROSPECT, a program for automated interpretation of 2D NMR spectra of Proteins. *J. Magn. Reson.*, **1990**, *86*, 633–644.

[12] Yu, C.; Hwang, J.-F.; Chen, T.-B.; Soo, V.-W. RUBIDIUM, a program for computer-aided assignment of two-dimensional NMR spectra of polypeptides. *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 183–187.

[13] Catasti, P.; Carrara, E.; Nicolini, C. Pepto: an expert system for automatic peak assignment of two-dimensional nuclear magnetic resonance spectra of proteins. *J. Comput. Chem.*, **1990**, *11*, 805–818.

[14] Eads, C. D.; Kuntz, I. D. Programs for computer-assisted sequential assignment of proteins. *J. Magn. Reson.*, **1989**, *82*, 467–482.

[15] Billeter, M.; Basus, V. J.; Kuntz, I. D. A program for semi-automatic sequential resonance assignments in protein 1H nuclear magnetic resonance spectra. *J. Magn. Reson.*, **1988**, *76*, 400–415.

[16] Oschkinat, H.; Holak, T. A.; Cieslar, C. Assignment of protein NMR spectra in the light of homonuclear 3D spectroscopy: An automatable procedure based on 3D TOCSY-TOCSY and 3D TOCSY-NOESY. *Biopolymers*, **1991**, *31*, 699–712.

[17] Meadows, R. P.; Olejniczak, E. T.; Fesik, S. W. A computer-based protocol for semiautomated assignments and 3D structure determination of proteins. *J. Biomol. NMR*, **1994**, *4*, 79–96.

[18] Bernstein, R.; Cieslar, C.; Ross, A.; Oschkinat, H.; Freund, J.; Holak, T. A. Computer-assisted assignment of multidimensional NMR spectra of proteins: Application to 3D NOESY-HMQC and TOCSY-HMQC spectra. *J. Biomol. NMR*, **1993**, *3*, 245–251.

[19] Zimmerman, D.; Kulikowski, C.; Wang, L.; Lyons, B.; Montelione, G. T. 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*, **1994**, *4*, 241–256.

[20] Oschkinat, H.; Croft, D. Automated assignment of multidimensional nuclear magnetic resonance spectra. *Methods in Enzymology*, **1994**, *239*, 308–318.

[21] Vuister, G. W.; Boelens, R.; Padilla, A.; Kleywegt, G. J.; Kaptein, R. Assignment strategies in homonuclear three-dimensional ^{1}H NMR spectra of proteins. *Biochemistry*, **1990**, *29*, 1829–1839.

[22] Morelle, N.; Brutscher, B.; J.-P., S.; Marion, D. Computer assignment of the backbone resonances of labelled proteins using two-dimensional correlation experiments. *J. Biomol. NMR*, **1995**, *5*, 154–160.

[23] Xu, J.; Sanctuary, B. C. CPA: Constrained Partitioning Algorithm for initial assignment of protein ^{1}H resonances from MQF-COSY. *J. Chem. Inf. Comput. Sci.*, **1993**, *33*, 490–500.

[24] Xu, J.; Sanctuary, B. C.; Gray, B. N. Automated extraction of spin coupling topologies from 2D NMR correlation spectra for protein ^{1}H resonance assignment. *J. Chem. Inf. Comput. Sci.*, **1993**, *33*, 475–489.

[25] Xu, J.; Straus, S. K.; Sanctuary, B. C.; Trimble, L. Use of fuzzy mathematics for complete automated assignment of peptide ^{1}H 2D NMR spectra. *J. Magn. Reson. B*, **1994**, *103*, 53–58.

[26] Xu, J.; Weber, P. L.; Borer, P. N. Computer-assisted assignment of peptides with non-standard amino acids. *J. Biomol. NMR*, **1995**, *5*, 183–192.

[27] Li, K.-B.; Sanctuary, B. C. Automated extracting of amino acid spin systems in protein using 3D HCCH-COSY/TOCSY spectroscopy and Constrained Partitioning Algorithm(CPA). *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 585–593.

[28] Li, K.-B.; Sanctuary, B. C. Automated assignment of proteins using 3D heteronuclear NMR. Part I: Backbone spin systems extraction and creation of polypeptides. *J. Chem. Inf. Comput. Sci.*, **1996**, *accepted*.

[29] Li, K.-B.; Sanctuary, B. C. Automated assignment of proteins using 3D heteronuclear NMR. Part II: side chain and sequence-specific assignment. *J. Chem. Inf. Comput. Sci.*, **1996**, *submitted*.

[30] Jeener, J., *Ampere International Summer School*. Basko polje, Yugoslavia, 1971.

[31] Aue, W. P.; Bartholdi, E.; Ernst, R. R. Two-dimensional spectroscopy. Application to nuclear magnetic resonance. *J. Chem. Phys.*, **1976**, *64*, 2229–2246.

[32] Rance, M.; Sørensen, O. W.; Bodenhausen, G.; Wagner, G.; Ernst, R. R.; Wüthrich, K. Improved spectral resolution in COSY ^{1}H NMR spectra of proteins via double quantum filtering. *Biochem. Biophys. Res. Commun.*, **1983**, *117*, 479–485.

[33] Braunschweiler, L.; Ernst, R. R. Coherence transfer by isotropic mixing: application to proton correlation spectroscopy. *J. Magn. Reson.*, **1983**, *53*, 521–528.

[34] Davis, D. G.; Bax, A. Assignment of complex ^{1}H NMR spectra via two-dimensional homonuclear Hartmann-Hahn spectroscopy. *J. Am. Chem. Soc.*, **1985**, *107*, 2820–2821.

[35] Bax, A.; Davis, D. G. MLEV-17 based two-dimensional homonuclear magnetization transfer spectroscopy. *J. Magn. Reson.*, **1985**, *65*, 355–360.

[36] Jeener, J.; Meier, B. H.; Bachmann, P.; Ernst, R. R. Investigation of exchange process by two-dimensional NMR spectroscopy. *J. Chem. Phys.*, **1979**, *71*, 4546–4553.

[37] Kumar, A.; Ernst, R. R.; Wüthrich, K. A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. *Biochem. Biophys. Res. Commun.*, **1980**, *95*, 1–6.

[38] Macura, S.; Ernst, R. R. Elucidation of cross relaxation in liquids by two-dimensional NMR spectroscopy. *Mol. Phys.*, **1980**, *41*, 95–117.

[39] Müller, L. Sensitivity enhanced detection of weak nuclei using heteronuclear multiple quantum coherence. *J. Am. Chem. Soc.*, **1979**, *101*, 4481–4484.

[40] Griesinger, C.; Sørensen, O. W.; Ernst, R. R. Three-dimensional Fourier spectroscopy. Application to high-resolution NMR. *J. Magn. Reson.*, **1989**, *84*, 14–63.

[41] James, T. L.; Basus, V. J. Generation of high-resolution protein structures in solution from multidimensional NMR. *Annu. Rev. Phys. Chem.*, **1991**, *42*, 501–542.

[42] Markley, J. L. Two-dimensional nuclear magnetic resonance spectroscopy of proteins: an overview. *Methods in Enzymology*, **1989**, *176*, 12–63.

[43] Karplus, M. Contact electron-spin coupling of nuclear magnetic moments. *J. Chem. Phys.*, **1959**, *30*, 11–15.

[44] Pardi, A.; Billeter, M.; Wüthrich, K. Calibration of the angular dependence of the amide proton-C_α proton coupling constants, $^3J_{HN\alpha}$, in a globular protein: use of $^3J_{HN\alpha}$ for identification of helical secondary structure. *J. Mol. Biol.*, **1984**, *180*, 741–751.

[45] Harvel, T. F.; Kuntz, I. D.; Crippen, G. M. Theory and practice of distance geometry. *Bull. Math. Biol.*, **1983**, *45*, 665–720.

[46] Englander, S. W.; Wand, A. J. Main-chain-directed strategy for the assignment of ^{1}H NMR spectra of proteins. *Biochemistry*, **1987**, *26*, 5953–5958.

[47] Thomsen, J. U.; Meyer, B. Pattern recognition of the ^{1}H NMR spectra of sugar alditols using a neural network. *J. Magn. Reson.*, **1989**, *84*, 212–217.

[48] Hare, B. J.; Prestegard, J. H. Application of neural networks to automated assignment of NMR spectra of proteins. *J. Biomol. NMR*, **1994**, *4*, 35–46.

[49] Wehrens, R.; Lucasius, C.; Buydens, L.; Kateman, G. Sequential assignment of 2D-NMR spectra of proteins using genetic algorithm. *J. Chem. Inf. Comput. Sci.*, **1993**, *33*, 245–251.

[50] Redfield, C.; Dobson, C. M. Sequential ^{1}H NMR assignment and secondary structure of hen egg white lysozyme in solution. *Biochemistry*, **1988**, *27*, 122–136.

[51] Redfield, C.; Dobson, C. M. ^{1}H NMR studies of human lysozyme: spectral assignment and comparison with hen lysozyme. *Biochemistry*, **1990**, *29*, 7201–7214.

[52] Barkhuijsen, H.; de Beer, R.; Bovée, W. M. M. J.; van Ormondt, D. Retrieval of frequencies, amplitudes, damping factors, and phases from time-domain signals using a linear least-square procedure. *J. Magn. Reson.*, **1985**, *61*, 465–481.

[53] Mitschang, L.; Cieslar, C.; Holak, T. A.; Oschkinat, H. Application of the Karhunen-Loéve transformation to the suppression of undesired resonances in three-dimensional NMR. *J. Magn. Reson.*, **1991**, *92*, 208–217.

[54] Glaser, S.; Kalbitzer, H. R. Improvement of two-dimensional NMR spectra by weighted mean t_1-ridge subtraction and antidiagonal reduction. *J. Magn. Reson.*, **1986**, *68*, 350–354.

[55] Neidig, K.; Kalbitzer, H. R. Improved representation of two-dimensional NMR spectra by local rescaling. *J. Magn. Reson.*, **1990**, *88*, 155–160.

[56] Garret, D. S.; Powers, R.; Gronenborn, A. M.; Clore, G. M. A common sense approach to peak picking in two-, three-, and four-dimensional spectra using automatic computer analysis of contour diagrams. *J. Magn. Reson.*, **1991**, *95*, 214–220.

[57] Corne, S. A.; Johnson, P.; Fisher, J. An artificial neural network for classifying cross peaks in two-dimensional NMR spectra. *J. Magn. Reson.*, **1992**, *100*, 256–266.

[58] Gersting, J. L., *Mathematical structures for computer science.* Computer Science Press, New York, NY, 1993.

[59] Bonchev, D.; Rouvray, D. H., *Chemical Graph Theory: Introduction and fundamentals.* Abacus Press, Amsterdam, 1991.

[60] Zadeh, L. A., "Fuzzy sets" Information and control, vol 8., **1965**.

[61] Kaufmann, A., *Theory of fuzzy subsets.* Academic Press, London, 1975.

[62] Groß, K.-H.; Kalbitzer, H. R. Distribution of chemical shifts in ^{1}H nuclear magnetic resonance spectra of proteins. *J. Magn. Reson.*, **1988**, *76*, 87–99.

[63] Xu, J.; Zhang, M. HBA: New algorithm for structural match and applications. *Tetrahedron Comput. Methodol.*, **1989**, *2*, 75–83.

[64] SYBYL Version 6.2, Tripos Inc., July, **1995**.

[65] Grzesiek, S.; Bax, A. Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. *J. Magn. Reson.*, **1992**, *96*, 432–440.

[66] Grzesiek, S.; Bax, A. An efficient experiment fro sequential backbone assignment of medium-sized isotopically enriched proteins. *J. Magn. Reson.*, **1992**, *99*, 201–207.

[67] Grzesiek, S.; Bax, A. Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. *J. Am. Chem. Soc.*, **1992**, *114*, 6291–6293.

[68] Gagné, S. M.; Tsuda, S.; Li, M. X.; Chandra, M.; Smillie, L. B.; Sykes, B. D. Quantification of the calcium-induced secondary structural changes in the regulatory domain of troponin-C. *Protein Science*, **1994**, *3*, 1961–1974.

[69] Marion, D.; Driscoll, P. C.; Kay, L. E.; Wingfield, P.; Bax, A.; Gronenborn, A. M.; Clore, G. M. Overcoming the overlap problem in the assignment of ^{1}H NMR spectra of larger proteins by use of three-dimensional heteronuclear ^{1}H-^{15}N Hartmann-Hahn-multiple quantum coherence and nuclear overhauser-multiple quantum coherence spectroscopy:Application to interleukin 1 β. *Biochemistry*, **1989**, *28*, 6150–6156.

[70] Choy, W.-Y.; Sanctuary, B. C. Protein ^{15}N chemical shift database. private communication, **1995**.

[71] Bax, A.; Clore, G. M.; Driscoll, P. C.; Gronenborn, A. M.; Ikura, M.; Kay, L. E. Practical aspect of proton-carbon-carbon-proton three dimensional correlation spectroscopy of ^{13}C-labeled proteins. *J. Magn. Reson.*, **1990**, *87*, 620–627.

[72] Kay, L. E.; Ikura, M.; Bax, A. Proton-proton correlation via carbon-carbon couplings: a three-dimensional NMR approach for the assignment of aliphatic resonances in proteins labeled with carbon-13. *J. Am. Chem. Soc.*, **1990**, *112*, 888–889.

[73] Clore, G. M.; Bax, A.; Driscoll, P. C.; Wingfield, P. T.; Gronenborn, A. M. Assignment of the side-chain ^{1}H and ^{13}C resonances of interleukin-1B using double and triple-resonance heteronuclear three-dimensional NMR spectroscopy. *Biochemistry*, **1990**, *29*, 8172–8184.

[74] Bax, A.; Clore, G. M.; Gronenborn, A. M. ^{1}H-^{1}H correlation via isotropic mixing of ^{13}C magnetization, a new three-dimensional approach for assigning ^{1}H and ^{13}C spectra of ^{13}C-enriched proteins. *J. Magn. Reson.*, **1990**, *88*, 425–431.

[75] Lyons, B. A.; Tashiro, M.; Cedergren, L.; Nilsson, B.; Montelione, G. T. An improved strategy for determining resonance assignments for isotopically enriched proteins and its application to an engineered domain of staphylococcal protein A. *Biochemistry*, **1993**, *32*, 7839–7845.

[76] Montelione, G. T.; Lyons, B. A.; Emerson, S. D.; Tashiro, M. An efficient triple resonance experiment using carbon-13 isotropic mixing for determining sequence-specific resonance assignments of isotopically-enriched proteins. *J. Am. Chem. Soc.*, **1992**, *114*, 10974–10975.

[77] Lyons, B. A.; Montelione, G. T. An HCCNH triple-resonance experiment using carbon-13 isotropic mixing for correlating backbone amide and side-chain aliphatic resonances in isotopically enriched proteins. *J. Magn. Reson. B*, **1993**, *101*, 206–209.

[78] Wishart, D. S.; Bigam, C. G.; Holm, A.; Hodges, R. S.; Sykes, B. D. ^{1}H, ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. *J. Biomol. NMR*, **1995**, *5*, 67–81.

[79] Van de Ven, F. J. M.; Lycksell, P.-O.; Van Kammen, A.; Hilbers, C. W. Computer-aided assignment of the ^{1}H-NMR spectrum of the viral-protein-genome-linked polypeptide from cowpea mosaic virus. *Eur. J. Biochem.*, **1990**, *109*, 583–591.

[80] Slichter, C. P., *Principles of magnetic resonance*. Springer-Verlag, Berlin, 1990 page 344.

[81] Piantini, U.; Sørensen, O. W.; Ernst, R. R. Multiple quantum filters for elucidating NMR coupling networks. *J. Am. Chem. Soc.*, **1982**, *104*, 6800–6801.

Index