University of Alberta

Automated Sequential Resonance Assignment in NMR Protein Structure Determination

by

Xiang Wan 

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

Department of Computing Science

Edmonton, Alberta
Fall 2006

Your file Votre référence
ISBN: 978-0-494-23123-4
Our file Notre référence
ISBN: 978-0-494-23123-4

Abstract

The success in sequential resonance assignment is fundamental to protein three dimensional structure determination via NMR spectroscopy. In general, the sequential resonance assignment consists of four components, namely, peak grouping, connectivity determination, string assignment and scoring scheme. The objective of peak grouping is to classify the detected resonance peaks from multiple NMR spectra into spin systems. Connectivity determination aims to find the true connectivity among the grouped spin systems in order to chain them into some strings. The goal of string assignment is to map the strings of spin systems to non-overlapping consecutive amino acid residues in the target protein. The task of scoring scheme is to measure the correlations between the amino acid types and the grouped spin systems. This thesis thoroughly addresses the computational issues that remain to be resolved in each component in the sequential resonance assignment process. Several novel computational models are developed to tackle those issues. We organize this thesis according to the issues we tackle in the development. First, we discuss the difficulties in scoring scheme learning, evaluate the existing learning methods with the string assignment algorithms, and identify the best one. Second, we provide our solutions to resolve the connectivity determination problem, which supplies valuable constraints for computing the reliable resonance assignment. A vital heuristic is disigned and applied to our solutions. Third, we reveal that the peak grouping, which is often assumed to be less important and neglected by most researchers, is the bottleneck in automated sequential resonance assignment. We present our graph-based solutions based on the improved automated assignment framework to resolve

the peak grouping and sequential resonance assignment simultaneously. The value of our approaches to solving the different issues is explored by conducting comparison experiments with many recently published similar methods. The experimental results show that this study has made a significant contribution to the field of NMR protein structure determination. The performance comparisons demonstrate the fact that our models would be more promising for practical use. We conclude this thesis with a discussion of the limitations in our models as well as related future work.

Acknowledgements

I am very grateful to all my thesis committee members, Dr. Guohui Lin, Dr. Russ Greiner, Dr. David Wishart, Dr. Randy Goebel, Dr. Robert Campbell and Dr. Patricia Evans, who made helpful comments on precious versions of this work. A special thank goes to my supervisor, Dr. Guohui Lin, whose encouragement and conceptual contributions greatly facilitated this work. Our research was primarily supported by an NSERC Discovery Grant to Dr. Guohui Lin.

To my parents

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis investigates the computational issues that remain to be resolved in NMR sequential resonance assignment (e.g., peak grouping and connectivity determination), and provides a number of corresponding solutions. Based upon the improved automated assignment protocol proposed in this thesis, we design a graph-based approach to automate the sequential resonance assignment process. This chapter starts with a brief introduction to our motivations, and then outlines the basic concepts and issues in the sequential resonance assignment process. Our research to date has made a significant contribution on NMR sequential resonance assignment and identified several issues for needed follow-up research.

1.1 Motivation

It is well known that proteins act as the most basic working units in life and understanding the functions of proteins requires the knowledge of their three dimensional structure. Protein structure determination is one of the most challenging topics in the area of structural biology. A variety of methods and techniques have been developed over the last several decades. Aside from the computer aided structure prediction through homology modeling and threading, Nuclear Magnetic Resonance (NMR) spectroscopy and X-Ray crystallography are still the dominant experimental techniques for protein structure determination. Researchers have now identified NMR as a superior approach to characterize the dynamics of proteins in solution because of its efficiency and low cost. Though NMR has not been able to achieve the same accuracy as X-Ray crystallography, enormous technological advances have brought NMR to the forefront of structural biology [28] since the publication of the first complete solution structure of a protein (bull seminal trypsin inhibitor) determined by NMR in 1985 [70].

The classical approach to protein structure determination involves three stages; namely, peak picking, sequential resonance assignment and structure determination. The objective of peak picking is to filter and identify the true resonance peaks from NMR spectral data. The task of sequential resonance assignment is to map the picked resonance peaks to the amino acid residues in the protein sequence. Such a sequential assignment labels the atoms in the target protein with their chemical shift values. This step provides the guidance for the structural constraint extraction from

NOESY, scalar coupling, and dipolar coupling spectra. Structure determination calculates the protein structure by using molecular dynamics and energy minimization under the identified structural constraints, which are extracted from the results of sequential resonance assignment. Manually conducting each of the three tasks is difficult and often takes a long time because of the problems frequently confronted by an NMR spectroscopist within the whole process. These problems involve difficulties in resolving correlations in crowded spectral regions, complications arising from dynamics, such as weak and missing peaks, or the fact that some atoms exhibit more than one peak [62]. Many efforts have been made to automate the whole process or at least part of it. In particular, peak picking and structure determination have been well studied over the past a few years. There are many software packages currently used in NMR labs, e.g., NMRView [43] for peak picking and X-Plor [17] for structure generation. Though the sequential resonance assignment problem is relatively easy for small proteins, it becomes more complicated and time-consuming for large ones. Since high-throughput NMR protein structure determination directly relies on high-throughput sequential resonance assignment, considerable research efforts have been dedicated to automate the sequential resonance assignment in the last two decades. To date, several programs have been developed. However, surprisingly, none of them has been widely used in NMR labs because of the unsatisfactory assignment accuracy in practice. Several observations can be made.

(1) The use of some tools are limited to small proteins with well resolved spectra. These tools often fail to produce assignments for datasets with a general degree of chemical shift degeneracy because of limitations with their exhaustive search and binary decision strategies.

(2) Some programs require a large number of NMR experiments for cross validation to resolve resonance or chemical shift ambiguities.

(3) Some programs require too much expertise to understand their internal designs and methods. Parameter setting are generally very hard to tune, and it seems that only designers can successfully apply it on real datasets.

(4) Some programs have been tested only on the experimental data generated in the designer's lab but do not generalize to the experimental data from other

labs. This may be due to differences in the experimental environment or differences in the signal-to-noise ratio in the spectra.

Given the above limitations, most NMR sequential resonance assignments are still done manually with the aid of some semi-automated software tools. It might even take months of manual work to produce a nearly complete assignment (*i.e.* all identified true peaks are assigned) because the tedious "undo – redo" process occurs fairly often if the data quality is poor. Therefore, designing a robust and user-friendly automated NMR sequential resonance assignment system can make an important contribution to the NMR protein structure determination.

1.2 Protein NMR Sequential Resonance Assignment

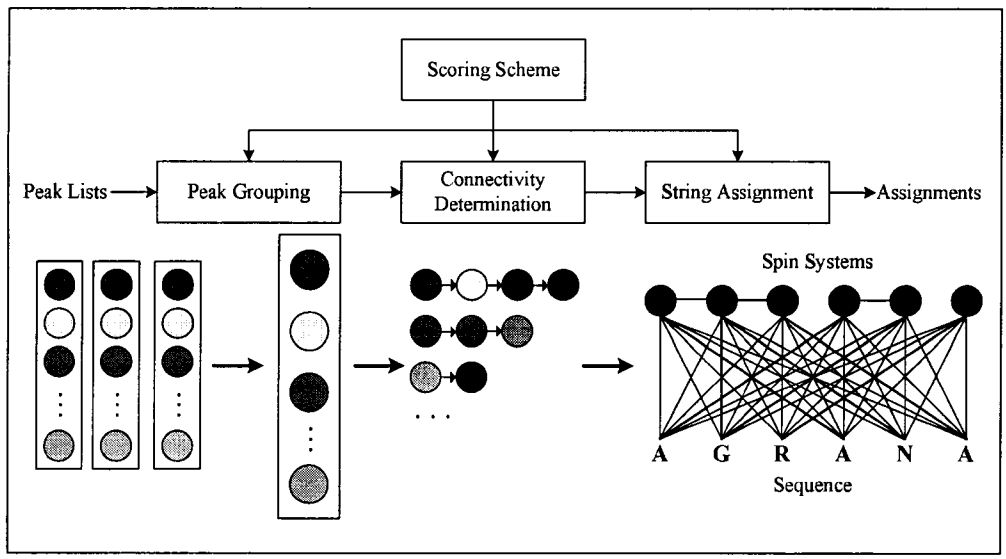


Figure 1.1: The framework in NMR sequential resonance assignment

Sequential resonance assignment is an essential step of data analysis before structure determination and structure refinement being conducted. This process consists of four components, namely, peak grouping, connectivity determination, string assignment and scoring scheme. Figure 1.1 illustrates the framework of a typical assignment program and the relationships between the different components.

Peak Grouping

The input to NMR sequential resonance assignment includes multiple peak lists from multi-dimensional NMR spectra. The goal is to map the intra-residue chemical shifts to their corresponding amino acid residues in the target protein. If each peak list is processed separately, the resonance assignment process would become very complicated because the chemical shifts contained in one peak list can not provide enough signature information to match to their host residues. Therefore, the first stage in sequential resonance assignment, peak grouping, will group together the resonance peaks produced by the same amino acid residue in the different peak lists to form a list of **spin systems**. The grouped spin systems, which contain collective signature information about their host residues, become the basis for completing the subsequent assignment steps.

Connectivity Determination

Resonance peaks from multi-dimensional NMR spectra record the nuclear correlations for atoms from a common residue and for atoms from the adjacent amino acid residues. Therefore, within a spin system, there are chemical shifts for the nuclei residing in the same amino acid residue and chemical shifts for the nuclei residing in the preceding residue. The inter-residue chemical shifts contained in most spin systems can be used as the evidences to determine whether some spin systems should be chained together and assigned to the adjacent residues in the protein sequence. This information is referred to as "connectivity information". The objective of connectivity determination is to identify the true connections among different spin systems, and to chain the spin systems into strings. These strings of spin systems will be assigned to non-overlapping polypeptide segments in the protein sequence during the string assignment process (see below).

String Assignment

The task of string assignment is to find the non-overlapping mapping between strings of spin systems and polypeptide segments in the protein sequence. This problem can be modeled as a *constrained* weighted bipartite matching problem on two disjoint groups, one group containing strings of spin systems and the

other containing a sequence of amino acids. The spin systems contained in one string must be matched only with the consecutive amino acids. The weight for each edge represents the probability of a corresponding mapping between the matched spin system and the amino acid, which could be computed through a defined scoring scheme.

The Scoring Scheme

A scoring scheme is used to estimate the likelihood of the mapping between a spin system and an amino acid type. Accurately quantifying the signature information of chemical shifts contained in the spin system provides a solid foundation for an accurate sequential resonance assignment. The performance of the string assignment directly relies on the discerning power of the scoring scheme.

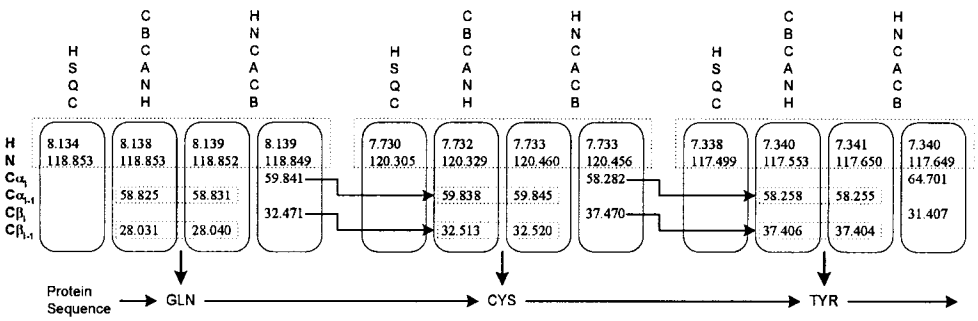


Figure 1.2: An example in NMR sequential resonance assignment. Three spin systems are constructed by grouping the peaks from three NMR experiments, which are HSQC, CBCA(CO)NH and HNCACB. The grouped peaks essentially share the same H and N chemical shifts. The C_i^{α} and C_i^{β} chemical shifts in one spin system are compared with C_{i-1}^{α} and C_{i-1}^{β} chemical shifts in other spin systems to build the possible directed connections.

To illustrate the sequential resonance assignment, Figure 1.2 provides an example of several peaks from the three NMR spectra: HSQC, CBCA(CO)NH, and HNCACB. The HSQC spectrum provides pairs of intra-residue chemical shifts (H_i, N_i), where i indexes the host residue to which the nuclei H and N belong. The CBCA(CO)NH spectrum provides triples of chemical shifts $(H_i, C_{i-1}^{\alpha}, N_i)$ and $(H_i, C_{i-1}^{\beta}, N_i)$; and the HNCACB spectrum provides triples of chemical shifts $(H_i, C_{i-1}^{\alpha}, N_i)$, $(H_i, C_{i-1}^{\beta}, N_i)$,

(H_i, C_i^{α}, N_i), and (H_i, C_i^{β}, N_i). Two special cases to be mentioned are (1) prolines do not have an H nucleus and (2) glycines do not have a C^{β} nucleus. The goal of the sequential resonance assignment is to assign each chemical shift to its host nucleus in the target protein. In the theoretically ideal case, the chemical shifts (and thus peaks) for any given nucleus (a set of nuclei) are identical across all three spectra and the number of spectral peaks read out of one spectrum exactly matches the number that should be observed. In other words, one peak in the HSQC spectrum matches exactly two peaks in the CBCA(CO)NH spectrum and four peaks in the HNCACB spectrum, through comparing the shared H and N chemical shifts. These seven peaks are grouped together to form a spin system, which is a multidimensional vector of the form $(H_i, N_i, C_i^{\alpha}, C_i^{\beta}, C_{i-1}^{\alpha}, C_{i-1}^{\beta})$, where chemical shifts indexed i are intra-residue chemical shifts and those indexed $i-1$ are inter-residue ones.

In the connectivity determination step, the inter-residue chemical shifts will be used as evidences to infer that two spin systems should be mapped to adjacent residues in the target protein, since they would appear as intra-residue chemical shifts in the other spin system. Assuming no ambiguity occurs, all spin systems could be connected in this way to form a string, which is required to be mapped to a segment of amino acid residues in the target protein by the string assignment algorithm. For the mapping to be done, the chemical shifts in a spin system are used either to determine the residue type or to provide a quantified score computed during the scoring process.

1.3 Issues

In the example shown in Figure 1.2, one can see that the sequential resonance assignment can be done straightforwardly. In practice, however, due to the problem of spectral noise and NMR data degeneracy, the chemical shifts observed for a nucleus are often not identical across the different spectra, some of them might not be observed, and many noise peaks will be present. Nevertheless, we still need the ability to compute the highly confident assignment because any minor error in the sequential resonance assignment would potentially feed erroneous structure constraints to the structure calculator and thus result in an erroneous structure.

In peak grouping, most existing methods use resonance peaks from the HSQC

spectra as anchors and map resonance peaks from other spectra to HSQC peaks by using a binary decision strategy with setting the tolerance thresholds on H and N chemical shift (the H and N differences between the anchor peak and the mapped peaks must fall within the given tolerance thresholds). However, the binary decision strategy is often not effective enough to resolve the ambiguities in the grouping. A typical issue is that the grouped spin systems might contain extra peaks. As a result, the judgement work has to be done manually, which might take a long time for the proteins having more than 100 amino acid residues.

The classic method for determining the connectivity information between spin systems in most existing methods is to compare the differences between chemical shifts for common nuclei and use the given tolerance thresholds to decide the correct connections, which again involves a binary decision strategy. However, due to noise and data degeneracy, connectivity determination is no longer a binary decision but probabilistic. Subsequently, one spin system could start more than one connectivity pair and could end more than one connectivity pair. Better computational models are needed to efficiently and effectively resolve this issue.

The string assignment problem can be modeled as a *constrained* weighted bipartite matching problem on two disjoint groups with one group containing strings of spin systems and the other containing a sequence of amino acids. Unfortunately, the constrained bipartite matching problem is NP-hard, even if the edges are unweighted [76]. Hence the efficient and effective algorithms are needed to solve this problem.

The ideal scoring scheme, if it existed, could directly identify the correct assignment. But it is almost impossible to find such a scoring scheme because the chemical shifts generated from some types of amino acids are close to each other, and the variances of the measured chemical shifts depend on the experimental environment and many other factors. The same types of chemical shifts generated from the same residues might vary among different NMR labs or sometimes in different experiments conducted in the same lab. Therefore, an effective learning process is necessary to score the preferences as accurately as possible. Most published methods for automated sequential resonance assignment make an assumption that for one residue type, the chemical shift values of a nucleus follow a normal (Gaussian)

distribution. In the BioMagResBank (BMRB, http://www.bmrb.wisc.edu/), which is the central repository for known protein NMR data, the means and standard deviations for H, N, C^{α}, C^{β}, C, and H^{α} (and more) chemical shifts in all 20 amino acid residues are collected. With these parameters at hand, a typical procedure is to use the density functions of the corresponding normal distributions to estimate a probability for mapping a spin system to a residue. Although the scoring scheme that assumes the Gaussian distribution is frequently adopted, we suspect that such an assumption is correct. We believe more work is needed to tackle this problem by using advanced learning techniques.

1.4 Structure of the Document

Chapter 2 highlights some basic concepts in NMR spectroscopy. It also briefly describes the NMR protein structure determination procedure, and introduces the NMR experiments used in the thesis. It serves to help the descriptions of the computational models in the succeeding chapters.

Chapter 3 reviews the previous works on the protein NMR sequential resonance assignment. In particular, a variety of approaches are examined and their strengths and weaknesses discussed.

Chapter 4 deals with the scoring scheme and string assignment algorithms. The existing scoring methods, as well as our histogram-based scoring scheme, are evaluated with the string assignment algorithms and the best one is identified.

Chapter 5 presents an algorithm, CISA, for connectivity determination by combining chemical shift signature information. The performance of this algorithm is evaluated by comparing it with another assignment program, PACES [22].

Chapter 6 discusses the issues in peak grouping, and describes a novel computation model, GASA, for resolving the peak grouping and conducting the sequential resonance assignment simultaneously. This model separates the assignment procedure not into physical steps but only virtual steps, and uses their output to cross validate each other. Our approach is compared with several recently developed tools, RANDOM [47], MARS [45], and RIBRA [73].

Chapter 7 concludes this thesis with a discussion of the limitations of our models as well as a discussion of future work.

Chapter 2

Background

Nuclear magnetic resonance (NMR) spectroscopy is a biophysical method that can provide high resolution structures of biological molecules such as proteins and nucleic acids at atomic resolution [11]. In this chapter, we will introduce some basic concepts in the NMR area, and briefly describe the NMR protein structure determination procedure and the NMR experiments used in the thesis.

2.1 Nuclear Magnetic Resonance Phenomenon

Atoms are basic building blocks of matter, and cannot be chemically subdivided by ordinary means. Atoms are composed of three types of particles: protons, neutrons, and electrons. Each proton has a positive charge and each electron has a negative charge, while neutrons have no charge. The number of protons in an atom is the *atomic number*, which determines the type of the atom. Both protons and neutrons reside in the nucleus. The same type of atoms or elements may contain different numbers of neutrons, and they are called *isotopes*.

A nucleus often acts as if it is a single entity with intrinsic total angular momentum I, the nuclear *spin*, which is the overall effect of the imaginary spinning protons and neutrons. Despite many spin-pairing rules, one characteristic is that a nucleus of odd mass number (which is the sum of the numbers of protons and neutrons) will have a half-integer spin and a nucleus of even mass number but odd numbers of protons and neutrons will have an integer spin. For a nucleus of spin I, there are $2I + 1$ spin states (or orientations) ranging from $-I$ to $+I$. In NMR spectroscopy for protein structure determination, the most important nuclei with spin $I = 1/2$ are ^{1}H (Hydrogen), ^{13}C (Carbon), ^{15}N (Nitrogen), ^{19}F (Fluorine), and ^{31}P (Phosphorus), each of which has two spin states. An example of a nucleus with spin $I = 1$ is deuterium ^{2}H (Hydrogen); Examples of isotopes with no spin (i.e., $I = 0$) are ^{12}C, ^{14}N, and ^{16}O (Oxygen).

Nuclear Magnetic Resonance (NMR) is a phenomenon which occurs when nuclei with non-zero spins are immersed in a static magnetic field and then exposed to a second oscillating magnetic field (which is created by radio frequency (r.f.) pulse). In the absence of an external magnetic field, for nuclei of spin I, those $2I + 1$ states are of equal energy. When an external magnetic field is applied, the energy levels split. In an external magnetic field of strength B_0, the spinning rotation axis of a

nucleus will *precess* about the magnetic field with angular frequency $\omega_0 = \gamma B_0$. ω_0 is called *Larmor Frequency*, where the *gyromagnetic ratio* γ is different for distinct types of nuclei. For nuclei of spin $I = 1/2$, there will be two possible spinning orientations/states in the external magnetic field, i.e., parallel to the external field (low energy state) and opposite to the external field (high energy state). At the time the external magnetic field is applied, the initial populations of nuclei in the energy levels are determined by thermodynamics, described by the Boltzmann distribution. This means that the lower energy level will contain slightly more nuclei than the higher energy level. It is possible to incite the low energy level nuclei into the high energy level with electromagnetic radiation. In fact, if these aligned nuclei are irradiated with an r.f. pulse of a proper frequency, the nuclei will spin-flip from the low energy state to the high energy state or from the high energy state to the low energy state by absorbing or emitting a quantum of energy, respectively. The frequency of radiation needed is determined by the difference in energy between the two energy levels and when such a spin transition occurs the nuclei are said to be *in resonance* with this radiation. The electromagnetic radiation supplied by the second oscillating magnetic field must be equal to the frequency of the oscillating electric field generated by nucleus precession, which is $\frac{\omega_0}{2\pi}$. This is because only under that circumstance, the energy needed in resonance can be transferred from electromagnetic radiation to precession nucleus. It is possible that by absorbing energy, the nuclei will reach a state with equal populations in both states. In such a case, the system is *saturated.* If the electromagnetic radiation supplied by the second oscillating magnetic field is then switched off, some of the nuclei at the high energy state will fall back to the low energy state and the system will return to thermal equilibrium. Such a process is the *relaxation* process. The relaxation process produces a measurable amount of r.f. signal at the resonant frequency associated with the spin-flip. This frequency is received and amplified to display the NMR signal.

2.1.1 Chemical Shift

The resonance frequencies of individual nuclei are not only relevant to the strength of the applied external magnetic field B_0, but also are dependent on their local

chemical environments. The magnetic field generated by a nucleus itself tends to contradict the effect of the external magnetic field. This contradiction effect is defined as *shielding*. The strength of this shielding effect increases with the local electron density. This effect is called the *Chemical Shift* phenomenon. The actual field present at the nucleus is not B_0 but $B_{\text{local}} = B_0(1 - \sigma)$, where σB_0 is the shielding effect (σ is the shielding factor, which is small — typically 10^{-5} for protons and 10^{-3} for other nuclei [11]). Chemical shift in parts per million (ppm) is defined as

$$\delta = \frac{(\omega_0 - \omega_{\text{reference}}) \times 10^6}{\omega_{\text{reference}}} \approx (\sigma_{\text{reference}} - \sigma) \times 10^6, \tag{2.1}$$

where $\omega_{\text{reference}}$ is the reference frequency and $\sigma_{\text{reference}}$ is the reference shielding factor. For both protons and carbons, the reference material is often perdeuterated 3−(trimethylsilyl) propionate sodium salt (TSP) or 2, 2-Dimethyl−2−silapentane−5−sulfonate sodium salt (DSS). The chemical shift effect is small but it is a very sensitive probe of the chemical environment of the resonating nucleus. Using chemical shift values, it is possible to distinguish among nuclei in different chemical environments. Once the chemical shifts of all the atoms of amino acids are collected from an NMR spectrum, the sequential resonance assignment can be conducted to map the chemical shifts back to their host amino acid residues in the protein sequence. After the sequential resonance assignment is finished, experimental parameters that define the three-dimensional structure are measured. The most important structural information derived from the NMR spectra is based on the *Nuclear Overhauser Effect (NOE)*.

2.1.2 Nuclear Overhauser Effect (NOE)

The *Nuclear Overhauser Effect* (NOE) is the result of cross-relaxation between dipolar coupled spins as a result of spin/spin interactions through space. The NOE allows the nuclear magnetization to transfer from one spin to another through space and scales with the distance between two spins. The NOE-derived distance is one of the most important sources of structural information for protein structure determination. In an NOESY (Nuclear Overhauser Effect Spectroscopy) spectrum, NOE interactions between pairs of nuclei are shown as NOE peaks. Each dimension of the spectrum is the chemical shift of one type of nucleus. For example, a peak at

(4.5ppm, 4.6ppm) in an ^{1}H-^{1}H NOESY spectrum records an interaction between a proton with chemical shift of 4.5ppm and another proton with chemical shift of 4.6ppm. The intensity of the NOE is related to the distance r between these two protons by an equation of the general form which is defined in [74] as

$$NOE \propto \frac{1}{\langle r \rangle^6} f(\tau_c), \tag{2.2}$$

where the second term $f(\cdot)$ is a correlation function that accounts for the modulation of the spin-spin coupling by stochastic rate processes with an effective correlation time τ_c.

NOEs are generally only observed between protons that are separated within 5Å. *J-Coupling* constants, which are mediated through chemical bonds, provide information about dihedral angles, and thereby can define the peptide backbone and side chain conformations.

2.1.3 J-Coupling

J-coupling (or spin-spin coupling) is the interaction between nuclear spins transferred through the electrons of the chemical bonds. The energy levels of each spin are slightly altered depending on the spin state of a scalar coupled spin. This gives rise to a splitting of the resonance lines. There are a few factors in a J-coupling, which affect the coupling constant. These factors are the nuclei involved, the distance between the two nuclei, the angle of interaction between the two nuclei, and the nuclear spins of the nuclei.

Both homonuclear and heteronuclear J-couplings can provide information about internuclear distance (the smaller the number of chemical bonds between a pair of nuclei, the stronger the coupling constant is) and the covalent chemical bonds angle (the smaller the angle, the bigger the coupling constant). Among them, one of the most commonly employed coupling constant is *Vicinal* (or three-bond, or ^{3}J) coupling that is dependent upon the dihedral angle θ between the nuclei.

2.1.4 NMR Spectroscopy

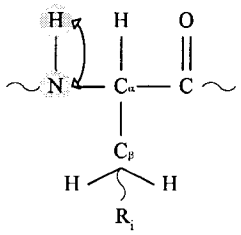
NMR spectroscopy is the use of the NMR phenomenon to study physical, chemical, and biological properties of matter. As a consequence, NMR spectroscopy finds applications in several areas of science. For example, NMR spectroscopy is routinely

applied by chemists to study the chemical structure of small organic molecules using simple one-dimensional techniques. Two and higher dimensional techniques are used to determine the structure of more complicated molecules. These techniques are continually improved and are replacing X-ray crystallography for the determination of protein structure. The protein structural information obtained from NMR spectroscopy includes a network of distance restraints between spatially close (i.e., < 5Å) hydrogen atoms extracted from the NOEs, dihedral-angle restraints calculated from scalar coupling constants and chemical shifts, and other various geometric restraints including orientation information from the residual dipolar coupling.

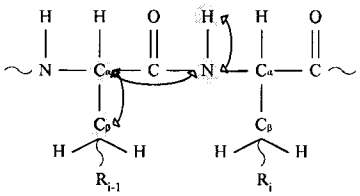
2.1.5 NMR Experiments

In general, all contemporary NMR studies on protein structure determination are done with two-dimensional (2D) or three-dimensional (3D) NMR experiments. The H-N coupling in the peptide bond is typically the starting point for the heteronuclear NMR analysis of proteins. This bond is present in every amino acid residue in a protein except the N-terminal and the proline residues. The HSQC spectrum measures the correlation between N and the directly attached H (See Figure 2.1(a)). It provides pairs of intra-residue chemical shifts $(\mathrm{H}_i, \mathrm{N}_i)$, where i indexes the residue to which the nuclei H and N belong. The three dimensional NMR spectrum can be used to identify couplings between the three nuclei in amino acid residues. In the previous resonance assignment strategy using homonuclear 2D NMR, the inter-residue connections were established from NOESY data. Recently, heteronuclear 3D NMR has been shown to provide inter-residue connectivity through a series of triple resonance experiments that overcome the peak overlap problem in homonuclear 2D NMR by introducing the third dimension and separating overlapped peaks into a number of 2D planes.

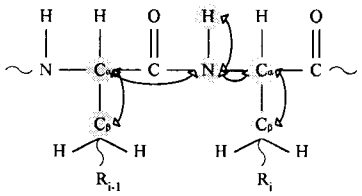
The CBCA(CO)NH experiment especially measures the heteronuclear coupling between H and N in one residue and the coupling across C to the C_α and C_β in the preceding residue (See Figure 2.1(b)). It provides triples of chemical shifts $(\mathrm{H}_i, \mathrm{C}^{\alpha}_{i-1}, \mathrm{N}_i)$ and $(\mathrm{H}_i, \mathrm{C}^{\beta}_{i-1}, \mathrm{N}_i)$. The HNCACB spectrum records two different heteronuclear correlation spectra. One records couplings between H, N and C_α and C_β in the same residue, and the other between H and N in one residue and the



(a) HSQC experiment



(b) CBCA(CO)NH experiment



(c) HNCACB experiment

Figure 2.1: Schematic illustration of the correlations in NMR experiments.

coupling across C to the C_α and C_β in the preceding residue (See Figure 2.1(c)). It provides triples of chemical shifts $(H_i, C^\alpha_{i-1}, N_i)$, $(H_i, C^\beta_{i-1}, N_i)$, (H_i, C^α_i, N_i), and (H_i, C^β_i, N_i). In a combined analysis of these two types of three dimensional NMR spectra, it is possible for each individual H-N pair in an HSQC spectrum, to be used to identify the C_α and C_β chemical shifts in the same residue and the preceding residue.

2.1.6 Spectral Data Acquisition and Processing

A wide variety of NMR instrumentation is available for NMR experiments to produce the data for protein structure determination. The common components of NMR spectrometers (see Figure 2.2) include (a) superconducting magnet for supplying an external magnetic field, (b) a pulse programmer and r.f. transmitter to generate and control r.f. pulses, (c) a probe for holding the sample in the magnet, (d) receiver for receiving the resulting NMR signals, and (e) computers for data acquisition and processing. Superconducting magnets can provide a wide range of

frequencies from 60 to 800 MHz. A higher frequency implies the higher sensitivity and stability of the NMR spectrometer because the differences between the chemical shifts are amplified with the increase of magnetic field strength, which produces a better separation between different nuclei.

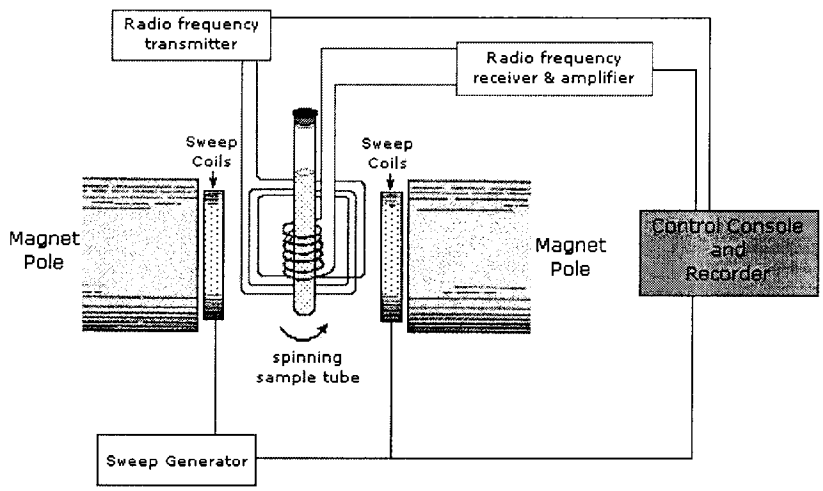


Figure 2.2: The structure of NMR spectrometer [63].

In NMR spectrometers, the superconducting magnet provides the external static magnetic field. The transverse magnetic field is generated by a series of r.f. pulses coming from the probe. During the relaxation process of the nuclei in the probe, the time-varying current is amplified and digitized by preamplifier and analog-to-digital converter (ADC), respectively, and then is recorded by the spectrometer. This time domain signal is sent to computer for further processing that transforms the time domain signals into the frequency domain signals. The main step of such a processing is the Fourier transformation, ahead of which multiple processing methods including *zero filling*, *apodization*, and *linear prediction* are applied to prevent information loss. After Fourier transformation, a post-processing method *phase correlation* is applied to optimize the appearance of the frequency domain spectrum. The frequency domain signals are the chemical shift values that will be analyzed next.

2.2 NMR Protein Structure Determination

The classical approach to protein structure determination via NMR can be summarized in three steps, *peak picking, resonance assignment* and *structure determination.*

2.2.1 Peak Picking

The objective of peak picking is to filter and identify the resonance peaks from the NMR spectral data. Each resonance peak indicates a particular magnetic interaction within a group of atoms (could be intra- or inter-residue) in the target protein. The measured values of resonance peaks are the resonance frequencies, or *chemical shifts*, of the interacting atoms. The peak intensities provide geometric relationships (*e.g.* distances and angles) among the interacting atoms. Figure 2.3 shows a sample one dimensional chemical shift spectrum which is a sketch of a proton NMR spectrum for the diacetone alcohol molecule [11]. In this spectrum, the x-axis is the chemical shift in ppm and the y-axis is the intensity. In the spectrum, the peak at 0 ppm is

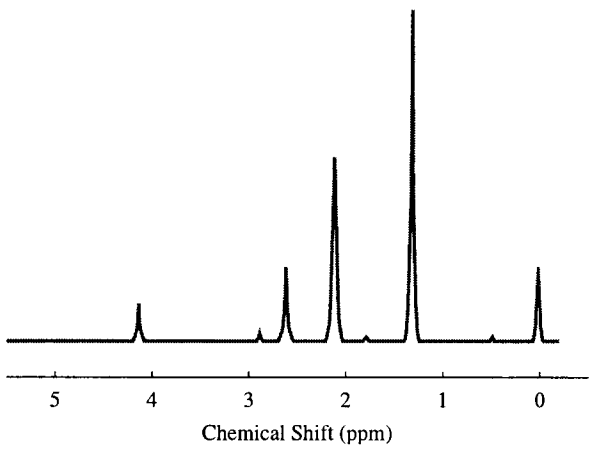


Figure 2.3: One dimensional NMR proton spectrum for diacetone alcohol molecule.

the reference peak and there are some other low intensity peaks which are considered as noise peaks.

For protein structure determination, two and higher dimensional NMR spectra are used, where each axis is the chemical shift in ppm for a certain type of nuclei. Because of strongly overlapped peaks and spectral distortions due to noise peaks,

the robust recognition methods are needed to identify true resonance peaks. There are a number of existing methods available for peak picking, such as neural networks [24, 19], statistical approaches [61, 3], and numerical analysis [48, 29].

2.2.2 Sequential Resonance Assignment

NMR spectra contain sufficient information to determine biomolecular structures in solution. However, none of the embedded information can be used without having the peaks assigned. In other words, it must be first determined which peaks come from which nuclear spins. Then the distance information in the NOESY spectrum can be analyzed. Therefore, the sequential resonance assignment process plays a vital role in the structure determination process. Resonance peaks from multi-dimensional NMR spectra contain the chemical shifts for atoms from a common residue and for atoms from its adjacent residues. In a sequential assignment step, the resonance peaks extracted from peak picking are mapped to host residues in the protein sequence. The method first groups the chemical shifts for atoms from a common residue into a spin system and then uses the identified inter-residue chemical shifts to determine the connectivity among the grouped spin systems. This helps constrain which pairs of spin systems should map to adjacent residues in the protein sequence. The mapping between spin systems and residues in the protein sequence is evaluated by using both the signature information of the spin system and the connectivity information. The signature information of a spin system in our work is defined as the likelihood that a particular amino acid type residing in some type of secondary structure could produce the spin system. There are four components involved in the sequential resonance assignment, which are peak grouping, connectivity determination, string assignment, and scoring scheme. The details about these four components have been discussed in Chapter 1.

2.2.3 Structure Determination

Based on the results of sequential resonance assignment, we could fully interpret the NOESY spectrum to provide many distance constraints between the hydrogen atoms in a protein. The inter-proton distance can be calculated from the intensity of the NOE cross peaks. In general, 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. As

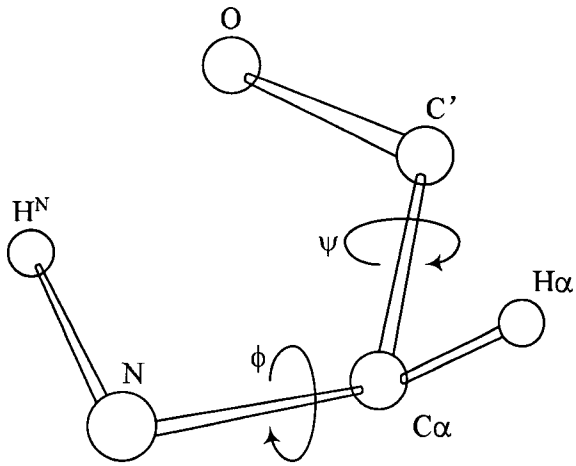


Figure 2.4: The torsion angles of an amino acid residue

shown in Figure 2.4, two dihedral angles are associated with each peptide bond. The ϕ angle is the torsion angle between the $N - H^N$ bond and $C^\alpha - H^\alpha$ bond and the ψ angle is another torsion angle between the $C^\alpha - H^\alpha$ bond and $C - O$ bond. The dihedral angle ϕ can be calculated from the spin-spin couplings $J_{H_\alpha - NH}$ using the Karplus equation which is defined as

$$J_{H^\alpha - NH} = 6.4\cos^2\theta - 1.4\cos\theta + 1.9, \tag{2.3}$$

where $\theta = |\phi - 60|$ [46]. With the use of the above equation, measurement of $J_{H^\alpha - N}$ provides complementary information to the NOE distance constraints for calculating the initial structure of a protein.

The next step is to determine an initial protein structure that is consistent with the thousands of NOE constraints and any other conformational constraints. Distance geometry is the most commonly used mathematical procedure in which the NOE distance constraints are converted into a three dimensional structure [39, 40]. The distance geometry procedure is essentially a projection from a high-dimensional space into ordinary three-dimensional space. The initial structure calculated from distance geometry may violate a number of experimental constraints. The subse-

quent structure refinement is required to obtain a high resolution protein structure with no constraint violations.

Chapter 3

Related Work

The sequential resonance assignment is one of the key tasks in NMR protein structure determination. Many researchers have perceived for a long while that the laboriously manual work has to be substituted by computer programs to better exploit the power of NMR in protein structure determination. Considerable efforts have been devoted to the automated assignment programs and several software tools have been developed. Nonetheless, most methods essentially use the same procedure, although different programs might have different focuses and start from different positions.

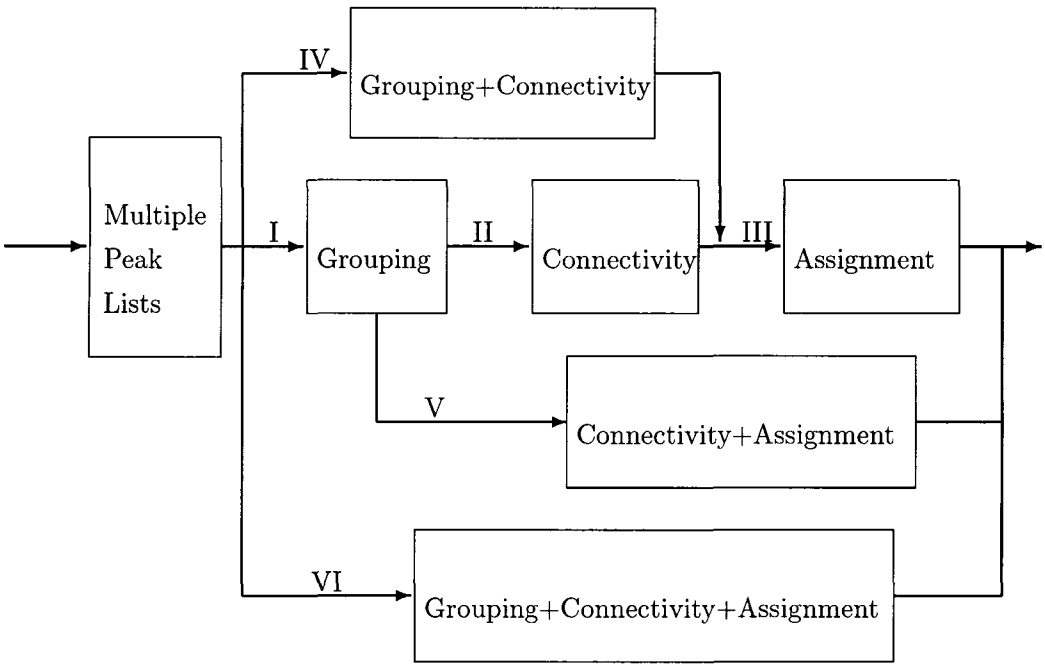


Figure 3.1: The flow chart of the resonance assignment process: different works assume different starting positions. Phase I includes AutoAssign [78], RIBRA [73], PASTA [50]; Phase II includes AutoAssign [78], RIBRA [73], PASTA [50], RANDOM [47], CISA [67], MARS [45]; Phase III includes AutoAssign [78], RIBRA [73] MAPPER [38], CBM [76]; Phase IV includes SmartNoteBook [62]; Phase V includes PACES [22], MARS [45], CISA [67]; Phase VI includes GARANT [9, 10].

In Figure 3.1, we classify most of the assignment methods in the literature. To name a few, GARANT [9, 10] uses a genetic algorithm, PASTA [50] uses threshold accepting algorithms, AutoAssign [78] uses heuristic best-first algorithms, MAPPER [38] and PACES [22] use exhaustive search algorithms, RANDOM [47] applies

a randomized algorithm, and RIBRA [73] applies a weighted maximum independent set algorithm for the sequential resonance assignment, MARS [45] first applies an exhaustive search for all the legal paths with length 5 and then conducts a bidirectional validation.

3.1 GARANT

GARANT [9, 10] is an automated resonance assignment program that combines a genetic algorithm with a local optimization routine. GARANT consists of three main components. The first one is the representation of a resonance assignment, which considers the resonance assignment as an optimal matching of two graphs. One graph represents the correlation between the atoms of the protein and expected cross peaks and the other the correlation between the chemical shifts and observed cross peaks. The second one is a scoring scheme that evaluates the matching between two constructed graphs. The last one is a genetic algorithm with a local optimization strategy that computes an optimal matching between two graphs, which corresponds to the optimal assignment.

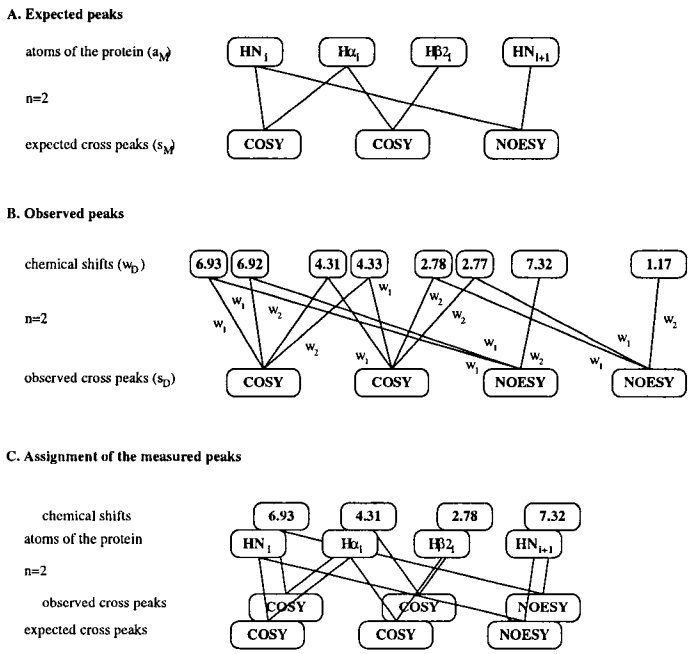


Figure 3.2: A schematic representation of expected (A) and observed cross peaks (B), and the mapping used to describe possible resonance assignments (C) .

Figure 3.2 [9] gives an example on how to represent the resonance assignment as a graph matching problem. In this figure, peaks, atoms and chemical shifts are represented by vertices, and the relations between them are represented by edges. The goal is to find an optimal matching between two graphs.

The matching between two graphs is evaluated by the matching scores between expected and observed peaks, which is referred to as "mutual information". Let $I_R(D;M)$ denote the mutual information between the observed graph D and the expected graph M, which is calculated by

$$\begin{aligned} I_R(D;M) &= \sum_k I_R(a_D^{(k)}; a_M^{(k)}) \\ &= \sum_k \log \frac{p(a_D^{(k)}, a_M^{(k)})}{p(a_D^{(k)}) \cdot p(a_M^{(k)})} \\ &= \sum_k \log \frac{p(a_D^{(k)}|a_M^{(k)})}{p(a_D^{(k)})} \\ &= \sum_k \log \frac{p(a_D^{(k)}|a_M^{(k)})}{\sum_l p(a_D^{(k)}|a_{M,l}^{(k)}) p(a_{M,l}^{(k)})}. \end{aligned}$$

where a_D represents the observed peaks in NMR experiments, a_M represents the expected peaks, k runs over all types of atoms, l runs over all possible observed values that could be assigned to atom k, $p(a_D^{(k)}|a_M^{(k)})$ denotes the conditional probability that, for atom type k, the value $a_D^{(k)}$ is observed when its expected value is known to be $a_M^{(k)}$, $p(a_D^{(k)})$ denotes the prior probability that the value $a_D^{(k)}$ is observed for atom k, and $p(a_{M,l}^{(k)})$ denotes the probability that the expected peak a_M is assigned to value l.

A general genetic algorithm is used in conjunction with a specific local optimization procedure to find an optimal matching between two graphs. However, a limitation of genetic algorithms is their slow convergence. For large proteins, the solution space grows exponentially with the number of residues and, in practical time scales, searching this huge space is intractable unless some heuristics are used to prune the search space.

3.2 PASTA

PASTA [50], Protein ASsignment by Threshold Accepting, uses **threshold accepting** that is a combinatorial optimization strategy and is superior to the genetic algorithm used in GARANT in terms of convergence time. In Figure 3.3, the steps of the PASTA assignment process is shown with a flow chart.

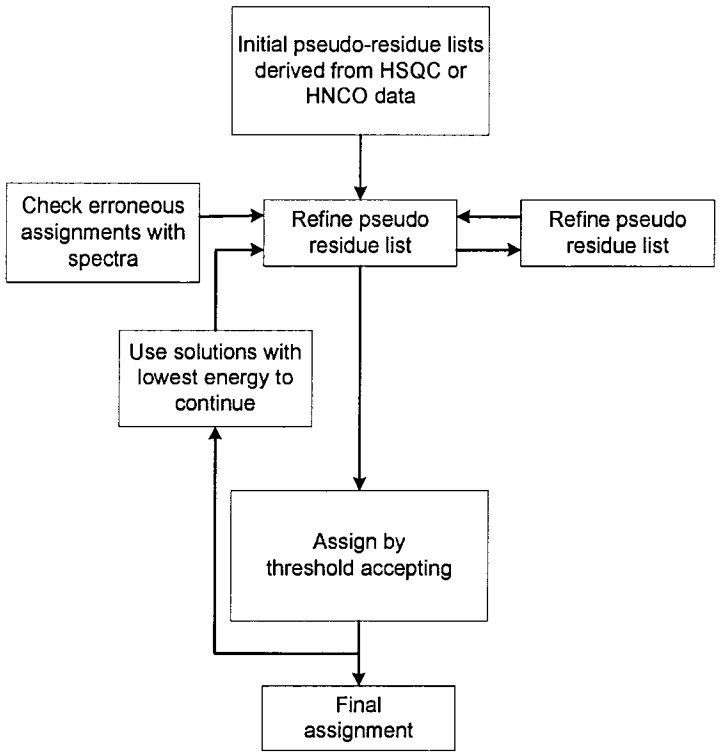


Figure 3.3: The assignment cycle of PASTA. An initial pseudo-residue list is created from the peak list of the HSQC or HNCO spectra. Additional information is added by searching the peak lists of the appropriate 3D experiments. The refinement of the list is done iteratively with the use of the assignment routine.

To start, an initial pseudo-residue list is created from the peak list of either the HSQC or HNCO spectra. The additional inter-residue chemical shifts and other intra-residue chemical shifts contained in the triple-resonance spectra are then added by finding the matched triplets for every H and N pair in HSQC or HNCO spectra. The matching between the peaks is determined by comparing their H and N chemical shifts under the tolerance thresholds. For the ambiguous pseudo-residues, such as some containing too many chemical shifts or too few chemical shifts, manual

work has to be conducted to refine the pseudo-residue list. The amino acid type identification for each pseudo-residue is based on the published random coil chemical shifts in [74] and [71]. For the connectivity determination and string assignment, a combinatorial minimization strategy, **threshold accepting**, is applied instead of the deterministic approach. The algorithm of threshold accepting consists of four basic steps outlined below.

(1) Start at a random solution x_i.

(2) Generate a new solution x_{i+1} via a random local change of the solution x_i.

(3) Compare the quality of both solutions with a penalty function f. If the value of the penalty function for solution x_{i+1} is not larger than the penalty function for the solution x_i plus a user-defined threshold T, proceed to solution x_{i+1}; else discard solution x_{i+1}.

(4) Repeat steps (2) and (3). If for a user-given number of steps, there is no improvement of the current minimum, the threshold T is reduced stepwise to zero. The solution with minimum value of penalty function during the whole run is the final result.

The penalty function to be minimized in the algorithm is defined as

$$E_{TOT} = E_{MATCH} + E_{SEQ}, \tag{3.1}$$

where E_{MATCH} describes the fit between two adjacent residues. The optional term E_{SEQ} is an additional feature resulted from the initially obtained assignment. To obtain a new solution x_{i+1} from x_i, two strategies are chosen:

(1) An interchange of two randomly chosen residues.

(2) A 'cut and paste' of a larger fragment. The starting point, length and new position of the fragment are determined by a random number generator.

3.3 AutoAssign

AutoAssign [78] is a constraint-based expert system for determining resonance assignments from many NMR spectra. The spin systems are firstly identified by

matching the peaks in various spectra to the peaks in the HSQC spectrum. The C^{α} and C^{β} chemical shifts contained in the spin systems are used to obtain the probabilities, which are used to determine the set of residue types consistent with the spin systems. The probability score is defined as the Bayesian posterior probability and the likelihoods are calculated using the expected C^{α} and C^{β} chemical shift values and standard deviations. Given an observed pair of C^{α} and C^{β} chemical shifts, the probability score with respect to amino acid residue type R is computed as

$$p(R|C^{\alpha}, C^{\beta}) = p(C^{\alpha}, C^{\beta}|R)P(R)/\sum_{R} p(C^{\alpha}, C^{\beta}|R)P(R) \tag{3.2}$$

where $p(C^{\alpha}, C^{\beta}|R)$ is the probability of observing chemical shift values C^{α} and C^{β}, given the residue type R, and $P(R)$ is the frequency of occurrence of residue type R in the protein sequence. The C^{α} and C^{β} chemical shifts are assumed to be independent and to follow Gaussian distributions. For each residue in the protein sequence, AutoAssign defines a list of spin systems that the nuclei in the residue may generate with high scores.

In the next stage, the pairwise relationships between spin systems are built by using the Euclidean distance. Specifically, for each spin system, a vector of normalized values is computed.

$$v_i = \langle \frac{C_i' - \mu(C')}{s(C')}, \frac{C_i^{\alpha} - \mu(C^{\alpha})}{s(C^{\alpha})}, \frac{C_i^{\beta} - \mu(C^{\beta})}{s(C^{\beta})}, \frac{H_i^{\alpha} - \mu(H^{\alpha})}{s(H^{\alpha})} \rangle \tag{3.3}$$

where the means μ and standard deviations s for each chemical shift dimension are collected over all available assigned chemical shifts. The Euclidean distance between the associated vectors is computed as the distance between two spin systems. For each pair of spin systems, AutoAssign checks if they reside in the two lists of spin systems for two adjacent residues or not. If they do, then the pair is considered to be a valid adjacent pair. At the same time, its mapping location can be confirmed if the pair of adjacent residues in the protein sequence is unique. AutoAssign extends the assignment of two spin systems to more spin systems by using an exhaustive search to find all valid combinations.

AutoAssign combines the connectivity determination and the string assignment to validate each other, which reduces the total number of possible connections. However, the number of combinations increases exponentially as the length of strings

increases. Even worse, the ambiguities of the connections among the spin systems also increase the complexity of this approach, even if the list for a residue only contains 2 spin systems on average. Autoassign's exhaustive search strategy with its constraint propagation might fail because of a search tree explosion if the data quality is poor and thereby a big number of possible connections are created. As a matter of fact, AutoAssign requires the redundant information from extra NMR spectra in order to reduce the complexity. In general, AutoAssign needs seven to eight three-dimensional NMR spectra in order to produce meaningful assignments.

3.4 MAPPER

MAPPER [38] is a semi-automatic sequence-specific NMR assignment program. Basically, MAPPER only performs the string assignment. The input of MAPPER contains the primary protein sequence and the strings of sequentially connected spin systems with information on the C_α and C_β chemical shifts and/or identification of amino acid types for the spin systems. MAPPER first treats each string separately to find its legal locations in the protein sequence. To determine the possible mapping positions for a given string i with the length $n(i)$, the sum of the squared deviations of all chemical shift values contained in the string is computed by using the reference values at the mapping positions k, which is

$$\chi^2(i;k) = \sum_{j=0}^{n(i)} \sum_{a \in A_j(i)} \left[\frac{\omega_j^\alpha(i) - \tilde{\omega}^\alpha_{R(k+j)}}{\Delta\tilde{\omega}^\alpha_{R(k+j)}} \right]^2 \tag{3.4}$$

where $A_j(i)$ denotes the set of atoms at position j in the fragment i, $\omega_j^\alpha(i)$ denotes the experimental chemical shift for the atom $a \in A_j(i)$ at the residue position j, $\tilde{\omega}^\alpha_R$ and $\Delta\tilde{\omega}^\alpha_R$ are the reference chemical shift value and its standard deviation for the atom a of the amino acid type R. The chemical shifts are assumed to follow a Gaussian distribution. For the correct mapping, the probability that the magnitude of the sum of the squared relative chemical shift deviation exceeds the value computed in Equation 3.4 is given by the χ^2 probability function $Q(\chi^2(i;k)|v_i)$ where $v_i = \sum_{j=0}^{n(i)} |A_j(i)|$ is the number of known chemical shifts in the fragment F_i. Acceptable individual mappings have a value of $Q(\chi^2(i;k)|v_i)$ above a user-defined threshold Q_0.

In the second step, MAPPER applies an exhaustive search to enumerate all consistent global mappings. The global mappings found by MAPPER are ranked with $\chi^2(global)$ that is the sum of the individual χ^2 values of all fragments. The $Q(global)$ is defined in the same way and a value of $Q(global)$ close to 100% indicates that a global mapping is confident in the sense that the chemical shift deviations are within the range expected statistically on the basis of their standard deviations.

3.5 PACES

PACES [22] is an interactive program for sequential resonance assignment. It uses an exhaustive search algorithm to establish the sequential connectivity and then perform the string assignment. The input data for PACES could be peak lists or a list of assembled spin systems. If the input is peak lists, PACES will first group them into the spin systems in a semi automatic manner, in which the users have to specify the tolerance thresholds and manually conduct the adjustment to resolve the ambiguities. The method for grouping in PACES is essentially the same as what most programs do, which anchors the peaks in other NMR spectra to the HSQC peaks by using the tolerance thresholds.

After the spin systems are correctly compiled, PACES starts by building a directed network to represent the connectivity relationship among the spin systems. For two spin systems j and k,

$$j = C_j^{\alpha}, C_j^{\beta}, C_j', H_j^{\alpha}, C_{j-1}^{\alpha}, C_{j-1}^{\beta}, C_{j-1}', H_{j-1}^{\alpha},$$
$$k = C_k^{\alpha}, C_k^{\beta}, C_k', H_k^{\alpha}, C_{k-1}^{\alpha}, C_{k-1}^{\beta}, C_{k-1}', H_{k-1}^{\alpha},$$

a directed connection from spin system j to spin system k will be established if

$$\begin{aligned} |C_j^{\alpha} - C_{k-1}^{\alpha}| &\leq \delta_{C^{\alpha}}, \\ |C_j^{\beta} - C_{k-1}^{\beta}| &\leq \delta_{C^{\beta}}, \\ |C_j' - C_{k-1}'| &\leq \delta_{C'}, \\ |H_j^{\alpha} - H_{k-1}^{\alpha}| &\leq \delta_{H^{\alpha}}, \end{aligned}$$

where $\delta_{C^{\alpha}}, \delta_{C^{\beta}}, \delta_{C'}$ and $\delta_{H^{\alpha}}$ are the user-specified threshold tolerances for C^{α}, C^{β}, carbonyl and H_{α} chemical shifts respectively.

The exhaustive search is first applied to enumerate all possible paths in the directed network, and each path represents a possible string of involved spin systems, which has to be assigned to a polypeptide in the protein sequence. During the process of the path enumeration, the encountered cycles will be broken at the last visited vertices, and the back edges between the last visited vertices and the upstream vertices are considered to be false connections. In the next step, a mapping process is invoked to validate each path by aligning it to every valid location in the target protein. The possible residue type for each spin system in the path is determined by using the chemical shift ranges of each amino acid type, which is derived from BioMagResBank [13]. PACES does not weigh the different mappings but equally treats all possible residue types for each spin system. In the ideal case, each path only contains the correct connections, and it should match some portion of protein sequence completely at the correct position. However, due to chemical shift degeneracy, some wrong connections might be chained in the path and thus create an illegal assignment. PACES cuts off the longest contiguous matching portions in the path and the non-matching portions are recycled to be validated in the succeeding iterations.

Without the manual finalization, PACES claims from its simulation study to provide the unambiguous mappings for 80% residues of any target protein. However, we remark that this approach is only suitable for the simple networks constructed with the high resolution experimental data because it is almost impossible to enumerate all paths in a graph with an average out-degree above 2. In fact, most directed networks tested in PACES have the average out-degrees below 2. For the low resolution datasets, PACES either runs out of memory or fails to compute a meaningful assignment. Another drawback in PACES is that the last visited connections in its cycles are always considered to be false connections. Therefore, the order of vertex visiting decides which edges represent the wrong connections. From our point of view, this is not a good strategy to identify the false connections in complex cases. It is highly possible that more correct connections are mistakenly considered to be wrong ones and thus removed from the directed network.

3.6 Random Graph Approach

Random graph approach [47] provides a novel method to find the connectivity information by using a randomized algorithm. It again models the relationship among the spin systems as a directed graph, where the vertices represent the spin systems and the weighted edges between the spin systems represent their connections with probabilities. The weight of each connection between two spin systems is derived from a function of distance on their chemical shifts. The connectivity determination problem is reduced to the path cover problem that is a classic NP-hard problem. A natural randomized algorithm is designed to find the optimum path cover that contains the minimum number of paths to cover all vertices in the directed graph. The randomized algorithm consists of two phases, of which the first phase performs the initialization and the second phase explores the connection choices.

In the first phase, a path cover is initiated with all unambiguous edges that start from the vertices with out-degree 1 to the vertices with in-degree 1. Those unambiguous edges are assumed to represent the correct connections in the random graph approach. In the second phase, the unambiguous edges in the path cover are extended by randomly choosing edges from the remaining graph with probabilities proportional to their weights. The extending process in the second phase is iteratively run until the path cover contains all vertices. The basic two-phase procedure is presented in Figure 3.4. To resolve the errors produced in the randomized algorithm, the random graph approach runs its algorithm for $20,000$ iterations to produce an ensemble set of path covers and the paths agreed on by most path covers are collected. The randomized strategy in this approach guarantees that this algorithm will terminate in a reasonable time period. For a graph with n vertices and average out-degree d, the algorithm will stop with a high probability in expected $O(n^{4+log(d-1)})$.

However, it is doubtful that the above approach would output the correct connectivity information in real applications because the noise and chemical shift degeneracy might cause some unambiguous edges to represent the wrong connections, and these wrong connections would lead to the wrong assignments in the output. Furthermore, an edge with a high probability does not always indicate a good con-

nection. Since the random graph approach only performs connectivity determination, it has to use other programs to finish the assignment. The results reported in the random graph approach is based on its combination with MAPPER. In general, 50% of residues in a protein are correctly mapped without manual work, which thwarts its use in real applications.

Given $G = (V, E)$
Let initial cover $C = V$
Let vertices with successors $W = \emptyset$
Choose vertex u from V

$Phase - 1$:
Let visited vertices $U = \emptyset$
While $U \neq V$ do
  Add u to U
  If u has single out-edge $e = (u, v)$ and v has a single in-edge then
    Add e to C
    Add u to W
    Set u to v
  Else Choose u from $V - U$
Endwhile

$Phase - 2$:
While C is not a Hamiltonian path or cycle do
  Choose u from $V - W$
  Choose an edge (u, v) with probability proportional to its weight
  If pred(v, C) is null then
    Join the two fragments in C
    Add u to W
  Else
    Create two fragments in $C : \langle \cdots u, v \cdots \rangle$,
    and $\langle \cdots, pred(v, C) \rangle$
    Add u to W
    Remove pred(v, C) from W
Endwhile

Figure 3.4: The randomized algorithm in Random Graph Approach.

3.7 MARS

MARS [45] is an automatic backbone assignment program that only performs the connectivity determination and string assignment. The input spin systems should be generated using other programs, such as NMRView [43], and they must be inspected manually to guarantee the high quality of input. The key features of MARS include, (1) exhaustive search for all strings with 5 spin systems, (2) bidirectional validation of each possible string, (3) best-first strategy for both linking and mapping, (4) combination of secondary structure, and (5) evaluation and assessment by performing multiple assignment. MARS applies a Z-score to compute the score of mapping i_{th} spin system to j_{th} residue, which is defined as

$$S(i,j) = \sum_{k=1}^{N_{cs}} \left\{ \frac{\delta(i)_k^{exp} - \delta(j)_k}{\delta_k} \right\}^2,$$

where $\delta(i)_k^{exp}$ is the measured chemical shift of type k of i_{th} spin system, $\delta(j)_k$ is the predicted (expected) chemical shift of type k of j_{th} residue, N_{cs} is the number of chemical shift types and ${\delta_k}^2$ is the variance of the statistical chemical shift distribution that is used for calculating $\delta(j)_k$. If type k is missing, $\delta(i)_k^{exp} - \delta(j)_k$ is set to 0.

To reduce the impact of chemical shift deviation, the score $S(i,j)$ is converted into a pseudoenergy $U(i,j)$ by ranking all residues j with respect to the spin system i. The score that one string belongs to a specific position in the protein sequence is computed according to

$$U_i^m(j) = \sum_{k=i}^{i+n} U(k, j_i),$$

where i is the index of the first spin system in the string and n is the length (in general, $n = 5$), m is the index of the string starting from spin system i, and j_i are the residue numbers to which spin systems i to $i+n$ are tentatively assigned. All $U_i^m(j)$ are ranked and the string with the best value is the target string for spin system i to $i+n$.

A major factor influencing the performance of MARS is the quality of spin systems and the quality of the chemical shifts contained in the spin systems because MARS limits the length of the longest string in order to make the exhaustive search feasible but may not be robust for lower quality NMR spectral.

3.8 RIBRA

RIBRA [73], Relaxation and Iterative Backbone Resonance Assignment, is a recently developed work, published in 2005, on fully automated sequential resonance assignment. In RIBRA, the sequential resonance assignment problem is reduced to the weighted maximum independent set problem in a graph and a relaxation approach is designed to solve this graph problem in an iterated fashion. The peaks with top-level quality are first identified to produce a partial assignment with high confidence, and then the peaks with middle-level and low-level quality are used to generate more assignments. There are two main operations in RIBRA, which are called RGT and LM. RGT performs the grouping and spin system identification while LM involves the connectivity determination and string assignment.

RGT:

The input of RIBRA is HSQC, CBCA(CO)NH and HNCACB spectral data. RGT first maps all peaks in CBCA(CO)NH and HNCACB to the peaks in HSQC to form a set of spin systems by comparing their shared H and N chemical shifts. Then it uses an extended version of the classification result (see Table 3.1) in TATAPRO [5] to attach each spin system with a list of amino acid types.

Carbon chemical shift	Amino acid
Absence of C^{β}	Gly
$14 < C^{\beta} < 24$	Ala
$56 < C^{\beta} < 67$	Ser
$24 < C^{\beta} < 36$ and $C^{\alpha} < 64$	Lys, Arg, Gln, Glu, His, Trp, Cys^{red} , Val and Met
$24 < C^{\beta} < 36$ and $C^{\alpha} \geq 64$	Val
$36 < C^{\beta} < 52$ and $C^{\alpha} < 64$	Asp, Asn, Phe, Tyr, Cys^{oxd}, Ile and Leu
$36 < C^{\beta} < 52$ and $C^{\alpha} \geq 64$	Ile
-	Pro
$C^{\beta} > 67$	Thr

Table 3.1: TATAPRO II residue typing scheme.

LM:

LM starts with all possible pairs of the grouped spin systems and tries to link them to form longer segments. During the expansion, LM validates each

segment by mapping it to the protein sequence. All possible segments will be generated in the LM operation. To resolve the conflicts of mapping among the generated segments, an undirected graph $G(V, E)$ is defined to create a graph optimization instance. Each node in V represents a possible mapping for one segment. If a segment has n possible mappings, there will be n nodes in the graph G. Each edge between two nodes represents a conflict between two possible mapping if 1) they share the same spin systems. 2) they overlap in the target protein. To favor the longer segment, each node v is given a weight defined as

$$w(v) = \frac{|v| + \sum_{x \in v} \frac{1}{N(x)}}{fre(v)} \tag{3.5}$$

where $|v|$ is the length of v, x is a spin system in v, $N(x)$ is the number of spin systems having the same H and N chemical shifts as x, and $fre(v)$ is the number of v's possible mapping positions. The modified heuristic algorithm proposed in [14] is applied to find several independent sets from G, which represents some possible assignments.

The difference between the grouping model applied in RIBRA and the previous works is that the ambiguities appearing in the grouping could be automatically resolved to some extent by trying all possible scenarios. Nonetheless, we argue that the grouping model in RIBRA is still susceptible to the change of pre-selected tolerance thresholds because high tolerance thresholds will make RIBRA produce a huge number of legal spin systems while low tolerance thresholds will lead to too few spin systems to complete the assignment. Furthermore, the spin system identification is not considered using probability but derived with the fixed list in Table 3.1, which is constrained by the quality of input spectral data.

Chapter 4

Scoring Schemes

A scoring scheme is required in NMR resonance assignment to assess the likelihood of the mapping between an identified spin system and an amino acid residue in the protein sequence. Accurately quantifying the signature information contained in chemical shifts provides a foundation for the precise and complete sequential resonance assignment in protein NMR spectroscopy. In this chapter, we describe our histogram-based learning method, and evaluate several different scoring schemes.

4.1 Overview

A spin system in NMR contains an array of intra-residue chemical shifts and inter-residue chemical shifts that are generated by a specific amino acid and its preceding amino acid in NMR experiments. It can be represented by a vector of chemical shifts, such as (H_i, N_i, C_i^{α}, C_i^{β}; C_{i-1}^{α}, C_{i-1}^{β}), where H_i, N_i, C_i^{α} and C_i^{β} are intra-residue chemical shifts generated by their host amino acid and C_{i-1}^{α} and C_{i-1}^{β} are inter-residue chemical shifts generated by the preceding amino acid. To measure the correlation between a given spin system and a given amino acid type, we need to quantify the signature information of each type of chemical shift contained in the spin system. Many published methods assume that for one residue type, the chemical shift value of a nucleus follows a normal (Gaussian) distribution. In the BioMagResBank (BMRB, http://www.bmrb.wisc.edu/), which is a repository for the known protein NMR data, the means and standard deviations have been collected for H, N, C^{α}, C^{β}, C, and H^{α} (and more) chemical shifts in all 20 types of amino acid residues. With these parameters available, a typical procedure to estimate the probability for mapping a spin system to a residue is to use the density functions of the corresponding normal distributions for the intra-residue chemical shifts in the spin system. Mathematically, for every intra-residue chemical shift (cs) in a spin system, the density function of the corresponding normal distribution is used to estimate a probability $p(cs \mid aa)$ that the host nucleus is in residue aa, where

$$p(cs \mid aa) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(cs-\mu)^2}{2\sigma^2}},$$

$\mu = \mu(aa)$ is the mean, and $\sigma = \sigma(aa)$ is the standard deviation. Subsequently, the product of the probabilities for all the intra-residue chemical shifts in the spin system is taken as the probability that aa is the host residue of the spin system.

The scoring scheme assuming a normal distribution is adopted by many groups working on automated NMR resonance assignment. However, we believe that this scoring scheme is biased on very simple statistics, and we also conjecture that other chemical environmental factors might affect the chemical shift values. In our investigations, we found that a minor improvement in the scoring scheme might have a significant effect on the accuracy of assignment. Therefore, we sought to design a better scoring scheme by combining more domain knowledge.

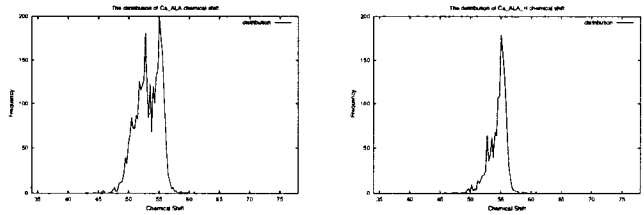
4.2 Histogram-Based Scoring Scheme

To avoid the bias arising from any specific assumption, we have designed a histogram-based scoring scheme. One of the most important elements in our scoring scheme is the chemical shift classification based on **Protein Secondary Structure**.

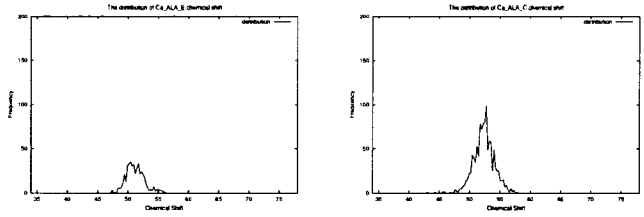
4.2.1 Protein Secondary Structure Prediction

Protein secondary structure refers to certain common repeating structures found in proteins. There are three types of secondary structures, which are α-helix, β-sheet and loops. It is well accepted in NMR work that for the same atom inside the same type of amino acid, the measured chemical shifts depend on the types of secondary structures where the amino acids lie. Statistics tells us that most amino acids display this dependency to some extent. For example, for alanines, the dot plots of chemical shifts of C_α's in α-helices, β-sheets, and loop regions show a marked difference.

Figure 4.1(a) is the sum of these 3 dot plots as shown in Figures 4.1(b), 4.1(c), and 4.1(d). Our scoring scheme accounts for this structural information by incorporating secondary structure information. The chemical shifts of each amino acid type in our training set is further divided into three categories according to the secondary structure type. Therefore, our training set has 60 classes in total. Each class is denoted by a couple (aa, ss), where aa represents one amino acid type and ss represents one secondary structure type. The secondary structure information is obtained in two ways. Given a protein sequence, we first check the Protein Data Bank (PDB) [12] to extract its secondary structure. If there is no entry for this sequence in PDB, then we predict its secondary structure by running the PsiPred program [44]. Psipred is one of the best known secondary structure prediction programs



(a) sum (b) in α-helices



(c) in β-sheets (d) in random coils

Figure 4.1: Distribution of C_α chemical shifts from alanines in the training data set.

with approximately 80 percent of accuracy for assigning a residue to an α-helix, a β-strand, or a loop.

4.2.2 Training Datasets

We built two training datasets using known protein NMR data deposited in the BMRB and corrected by the RefDB [77], and the secondary structures extracted from the PDB [12].

The initial set of protein NMR data was obtained from the BioMagResBank and included all protein entries present in the databank as of May 30, 2005. We applied several filtering steps to remove potential noise and bias from the dataset so as to make it as clean as possible. Firstly, proteins containing less than 50 amino acids or containing amino acids not part of the standard twenty were eliminated. Secondly, corrected NMR protein entries were obtained from the RefDB and these proteins overwrote any BMRB proteins present in the dataset. In the resultant dataset, every protein entry (a single file) was parsed in order to obtain the primary amino acid sequence, the chemical shift value for each nucleus, as well as the PDB

accession number(s). In the third filtering step, the PDB number was used to retrieve the sequence and secondary structure information related to that protein. The final dataset only contains those proteins that contain PDB accession numbers where the corresponding PDB protein sequence is a subsequence of the BMRB protein sequence or the BMRB protein sequence is a subsequence of the PDB protein sequence. The secondary structure information from the PDB protein entry were obtained for that protein. The PDB secondary structure notation has eight different letters; we translated this into a notation system of three letters to match up with the PsiPred secondary structure format (namely, G, H, and I from PDB became H in PsiPred, E from PDB remained as E, and S, T, B, and non-annotated positions in PDB became C in PsiPred). Such a translation is necessary since we would use PsiPred as the secondary structure predictor in our testing. Nonetheless, a suitable adjustment can always be made if other secondary structure predictors are applied. A total of $1,493$ protein entries and $165,122$ amino acid residues were obtained in the final dataset, denoted as ALL (cf. Figure 4.2 for more detailed statistics); 456 of these proteins and $45,964$ amino acid residues were from the RefDB corrected data. A total of 6 files were created with each corresponding to a nucleus from H, N, C^{α}, C^{β}, C, and H^{α}. For those protein entries in the final dataset, chemical shifts were placed into these 6 files. Each chemical shift is represented as a triplet of amino acid type, secondary structure type, and the chemical shift value.

We observed that a tiny number of chemical shift values should be treated as outliers because they diverge far from the main stream significantly. Since the abnormal behavior of a single outlier could disrupt the scoring scheme, an efficient statistical method, namely "`boxplot`" [27] with parameter set at 1.5, was applied to remove the outliers — the fourth chemical shift filtering step.

In order to reduce the bias that could be caused by multiple homologous sequences, a second dataset was generated. "BLAST 2 sequences (bl2seq)" [64] was run between every pair of sequences. Any protein having greater than 50% identity against another protein already included was removed from this dataset (though order dependent). The resulting dataset, denoted as HOMO, contains 822 proteins and $91,382$ residues, among which 336 proteins and $34,225$ residues were from the RefDB. The `boxplot` was also applied on HOMO to get rid of chemical shift outliers.

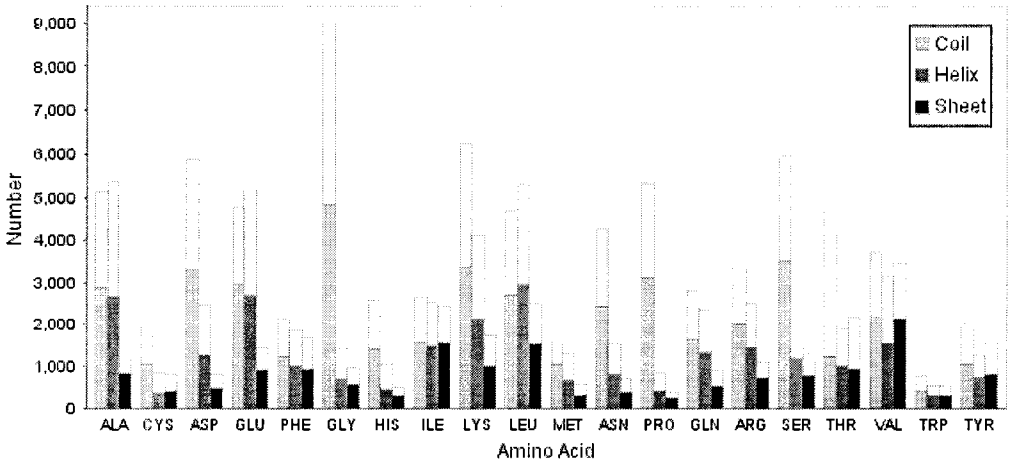


Figure 4.2: A detailed amino acid composition of the two training datasets ALL and HOMO: the height of each bar corresponds to the number of amino acid residues in that amino acid and secondary structure couple in dataset ALL. The height of the shaded region records the number in the reduced dataset HOMO.

4.2.3 Histogram-Based Scoring Scheme

For every amino acid (aa) and secondary structure (ss) combination, we do not assume there is any specific pattern that the distribution follows, but use the chemical shift values directly. For every type of chemical shift, we associate with it an error bound (ϵ), which is different for different types of chemical shifts and is learned from our training set. To estimate a probability for mapping a chemical shift (cs) to an amino acid residue (aa) and a secondary structure (ss) couple,

- let N denote the total number of the same type of chemical shifts in the training dataset;
- let $N(aa, ss)$ denote the number of (aa, ss) couples (which is typically in thousands) within N;
- let $N(cs)$ denote the number of chemical shifts in N that fall in the chemical shift window $(cs - \epsilon_{cs}, cs + \epsilon_{cs})$;
- let $N(cs \mid aa, ss)$ denote the number of chemical shifts in $N(aa, ss)$ that fall in the same chemical shift window.

Then, we employed the Naive Bayes method (see Figure 4.3) to derive the score for mapping the spin system (H_i, N_i, C_i^{α}, C_i^{β}; C_{i-1}^{α}, C_{i-1}^{β}) to an amino acid type aa residing in secondary structure ss:

$$\frac{1}{4} \sum_{cs \in \{H_i, N_i, C_i^{\alpha}, C_i^{\beta}\}} \log\Big(p(cs \mid aa, ss)\Big), \tag{4.1}$$

where

$$p(cs \mid aa, ss) = \frac{\frac{N(cs|aa,ss)}{N(cs)} \frac{N(cs)}{N}}{\frac{N(aa,ss)}{N}} = \frac{N(cs \mid aa, ss)}{N(aa, ss)}.$$

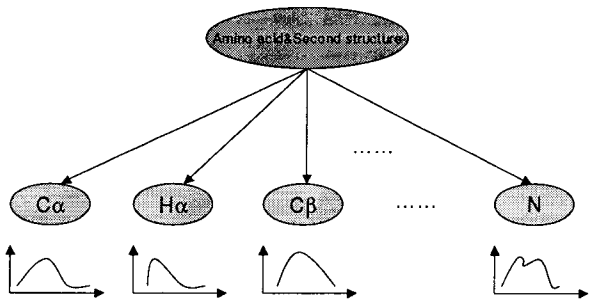


Figure 4.3: The naive bayes scoring scheme learning

4.2.4 Scoring Scheme Enhancement

Most methods in automated NMR resonance assignment only take the intra-residue chemical shifts in the spin system into account in the scoring scheme. We propose to add the signature information of inter-residue chemical shifts, since in practice the spin systems from peak grouping process do contain the inter-residue chemical shifts and these inter-residue chemical shifts also contain signature information for the preceding residue. A subsequent simulation experiment demonstrated that using inter-residue chemical shifts can significantly improve the scoring scheme performance, which is measured by the quality of the resultant assignment. Our scoring scheme took advantage of a few special features of chemical shifts. To name a few, since there is no C^{β} nucleus in glycine, no C^{β} chemical shift can be observed for the glycine spin system. Consequently, when a spin system does contain a non-zero C^{β} chemical shift value, then it should not be mapped to glycine. In this case, we associated with the mapping a score `maximum`, which was set at 9999.99 and it tells

the assignment algorithm that such a mapping is *illegal*. Similarly, since proline does not have a H^N nucleus, a spin system containing a non-zero H^N chemical shift value gets a score `maximum` when mapping to proline.

4.3 Assignment Algorithm

The general weighted bipartite matching problem is to find a one-to-one matching between elements of two groups that maximizes the total weight, where each matched pair of elements has a pre-specified weight. The NMR sequential resonance assignment process can be naturally modeled as a weighted bipartite matching problem, where each weighted edge has a confidence value representing a possible mapping of a spin system to an amino acid in the protein sequence. Nevertheless, the quality of assignment from such a weighted bipartite matching is poor because frequently there are multiple amino acid residues of the same type in a protein sequence. To differentiate the mapping between spin systems and the same type residues, we have to explore more constraints. The most important one is the connectivity information. In practice, a string of connected spin systems typically have a much better score at the "correct" assignment position (*i.e.* the matching between a spin system of NMR peaks and the residue that generates the peaks) than almost all other (incorrect) assignment positions, especially as the size of the string increases.

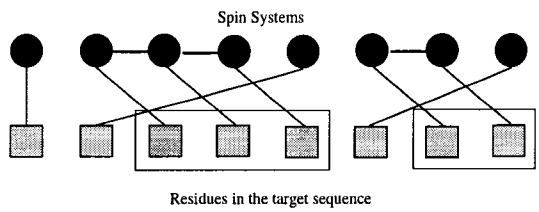


Figure 4.4: The problem of constrained bipartite matching

To incorporate the connectivity information, the general matching problem is extended to a *constrained* weighted bipartite matching problem [76] on two disjoint groups, one group containing strings of spin systems and the other containing a sequence of amino acids. The spin systems in one string must be matched only with neighbors of the other group (an example is shown in Figure 4.4). Unfortu-

nately, the constrained bipartite matching problem is NP-hard, even if the edges are unweighted [76]. Many heuristics have been proposed very recently, including some approximation algorithms [20], our fast greedy and filtering algorithm and our integer programming solver.

Our fast greedy filtering assignment algorithm can be described as a two-phase procedure: in the first phase, a *greedy filtering* is conducted to select some number of best possible mappings for the identified strings; in the second phase, for every combination of string mappings, an efficient maximum weighted bipartite matching algorithm is used to complete the assignment by mapping isolated spin systems to the rest of the residues. The algorithm reports the best assignment from all combinations in terms of the assignment confidence (the total weights of all individual mappings). The heuristics applied in a greedy filtering algorithm is fairly intuitive, and is very close to what is currently manually done in an NMR laboratory. The main difference between the algorithm and manual work is that we employ efficient computational methods to automate the assignment process at a global view, which produce an assignment within seconds on a Pentium IV PC. The global view also helps avoid the tedious "undo-redo" operations which occurs very often through manual efforts.

However, the greedy filtering algorithm can not guarantee an optimal solution though it runs very fast. To make a fair comparison between different scoring schemes, we adopted Cplex, one of the best integer programming solvers, to compute the exact solutions for CBM instances. The formulation of CBM instance in integer programming is defined as follows:

- let $\{a_1, a_2, \ldots, a_n\}$ denote a protein sequence of length n, $\{(s_1 s_2 \cdots s_{i-1}), (s_i s_{i+1} \cdots s_{i+k}), \ldots, (s_j s_{j+1} \cdots s_n)\}$ denote a set of chained spin systems, W denote a score matrix in which each entry w_{ij} measures the likelihood that the i-th spin system is mapped to j-th amino acid residue, and X denote an assignment matrix in which the entry x_{ij} with 1 value means that the i-th spin system is assigned to j-th amino acid.

- For each string $s_i s_{i+1} \cdots s_{i+k}, k \geq 0$, if $x_{ij} = 1$, then $x_{i+l,j+l} = 1$ for every $l = 1, 2, \ldots, k$.

- The goal is to minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_{ij}$$

under the constraints,

$$\begin{aligned}
\sum_{j=1}^{n} x_{ij} = 1, &\quad \text{for } i = 1, 2, \ldots, n; \\
\sum_{j=1}^{n} x_{ij} = 1, &\quad \text{for } j = 1, 2, \ldots, n; \\
x_{i+l,j+l} = x_{ij}, &\quad \text{for string } i+1 \cdots s_{i+k}; \\
&\quad l = 1, 2, \ldots, k; \\
&\quad j = 1, 2, \ldots, n-k; \\
x_{ij} \in 0, 1, &\quad \text{for } i, j = 1, 2, \ldots, n;
\end{aligned}$$

4.4 Evaluation

4.4.1 Test Dataset Simulation

An instance of CBM consists of an edge-weighted bipartite graph $G = (A, S, E)$, where A consists of the amino acid residues linearly ordered as they show up in the target protein, S consists of the spin systems, and every edge (a_i, s_j) indicates a mapping between residue a_i and spin system s_j with its weight recording the mapping score. Without any extra information for spin systems, the above CBM instance expects a minimum-weight perfect matching, which can be computed efficiently. The number of correctly assigned spin systems divided by the total number of assigned spin systems is defined as the *assignment accuracy*. If the scoring scheme were ideal, then the assignment accuracy would reach 100%. Therefore, we can use the assignment accuracy to measure the quality of the scoring scheme.

We chose a total of 470 sets of protein NMR data for our simulation study, each of which contains all H, N, C^{α}, and C^{β} chemical shifts. For every protein, the primary sequence was retrieved, and the secondary structures were predicted using PsiPred. For every amino acid residue, the chemical shifts for H, N, C^{α}, and C^{β} were retrieved from the BMRB entry, which formed an initial spin system. Subsequently, the chemical shifts for C^{α} and C^{β} in the preceding residue were appended to form the second spin system.

For every chemical shift in a second spin system, we perturbed it by adding to it a value that follows a zero-mean normal distribution, for which the standard deviation was set to the standard deviation we collected out of the training dataset. This gave a third spin system, which was finalized by randomly throwing away some C^{α} and C^{β} chemical shifts. The probability of throwing away chemical shifts was set to 5%.

In our simulation study, since we have all connectivity information for every protein, we randomly added some portion back to generate a few instances for every protein. The different instances have different levels of connectivity abundance. More precisely, an instance of $k\%$ connectivity contains $k\%$ connectivity that was randomly added. We have set k in tens and are interested in reasonable amounts of connectivity, namely, $k = 0, 10, 20, 30, 40, 50, 60, 70, 80, 90$.

4.4.2 Score Generation

For the purposes of comparison, we also designed the scoring scheme based on the assumption of normal distributions, using the means and standard deviations collected in our two training datasets ALL and HOMO. More specifically, in our implementation, we took the absolute logarithm of a probability divided by the number of intra-residue chemical shifts in the spin system, and multiplied by 100, to be the score for mapping the spin system. The factor 100 is solely for computational precision purpose and taking the average is for score normalization purpose. Clearly, the smaller the score, the higher confidence we have for the mapping. For ease of presentation, the scoring schemes assuming normal distributions for chemical shifts are denoted as *Normal*. Furthermore, if it uses only intra-residue chemical shifts for score evaluation, then it is denoted as Normal-*Intra*; if it uses both intra-residue and inter-residue chemical shifts for score evaluations, then it is denoted as Normal-*Both*.

The histogram-based naive Bayes scoring schemes using the chemical shift statistics in the training datasets directly (as described in the last section), are denoted as *Bayes*. In these scoring schemes, the chemical shift thresholds have to be learned, and they were set as follows. For triplet (aa, ss, nu), let (ϵ_{nu}) denote the window-size associated with this triplet such that exactly 20 intervals of length (ϵ_{nu}) cover the whole range of the chemical shifts. The value 20 was set so that these window-sizes

map closely to the standard deviations, collected as described. For every observed chemical shift value (cs) for each nucleus (nu), using (cs) as the midpoint, the number of chemical shifts in the training dataset that fall into the window of size (ϵ_{nu}) is $N(cs \mid aa, ss)$. Similarly as in the last paragraph, Bayes scoring schemes using only intra-residue chemical shifts are denoted as Bayes-*Intra* and those using both intra-residue and inter-residue chemical shifts are denoted as Bayes-*Both*.

To obtain the secondary structures for the protein sequence, we adopted PsiPred to predict the secondary structures. The PsiPred secondary structure format consists three notations, H for alpha helix, E for beta sheet, and C for coil. In addition to each predicted secondary structure for an amino acid, PsiPred also provides a confidence score, which is a single digit in the range of 0 to 9. We find that such a confidence value is a post-treatment of the neural network output, which are three values associated with three output units (helix, sheet, and coil). All three values for every amino acid residue in the target protein are stored in an intermediate PsiPred output file with suffix "ss2". These values can be regarded as the "prediction probabilities" for an individual secondary structure, and our second idea is to take in the predicted secondary structures together with their probabilities into the scoring schemes. Such scoring schemes are classified to have index 2. More specifically, when one amino acid residue aa is predicted to be in a helix with probability 0.55, to be in a sheet with probability 0.25, and to be in a coil with probability 0.40, then $\frac{0.55}{0.55+0.25+0.40}$ of the final score comes from mapping the spin system to (aa, H), $\frac{0.25}{0.55+0.25+0.40}$ from mapping the spin system to (aa, E), and $\frac{0.40}{0.55+0.25+0.40}$ from mapping the spin system to (aa, C). In this way, the scoring scheme Normal-Both-2 denotes the normal scoring scheme using both intra-residue and inter-residue chemical shifts in the spin system and using the prediction probabilities from PsiPred output for the score evaluation. To summarize, we have two training datasets ALL and HOMO and a total of eight scoring schemes Normal/Bayes-Intra/Both-1/2.

4.4.3 Results

Table 4.1 summarizes the average assignment accuracies of the eight different scoring schemes that are based on the training dataset ALL. Table 4.2 summarizes the average assignment accuracies of the eight different scoring schemes that are based

on the training dataset HOMO.

Scoring Scheme	Connectivity Percentage				
	0%	10%	20%	30%	40%
Normal-Intra-1	0.103	0.177	0.263	0.365	0.489
Normal-Both-1	0.509	0.575	0.650	0.718	0.783
Normal-Intra-2	0.130	0.209	0.301	0.412	0.541
Normal-Both-2	0.540	0.609	0.684	0.751	0.816
Bayes-Intra-1	0.140	0.232	0.342	0.465	0.591
Bayes-Both-1	0.553	0.621	0.696	0.760	0.823
Bayes-Intra-2	0.172	0.264	0.375	0.494	0.624
Bayes-Both-2	**0.583**	**0.650**	**0.721**	**0.787**	**0.844**
Scoring Scheme	Connectivity Percentage				
	0%	10%	20%	30%	40%
Normal-Intra-1	0.619	0.753	0.875	0.958	0.992
Normal-Both-1	0.844	0.900	0.946	0.976	0.993
Normal-Intra-2	0.676	0.798	0.903	0.967	0.991
Normal-Both-2	0.872	0.917	0.955	0.979	0.992
Bayes-Intra-1	0.724	0.832	0.922	0.972	0.993
Bayes-Both-1	0.879	0.922	0.958	0.982	0.993
Bayes-Intra-2	0.749	0.853	0.930	0.975	0.993
Bayes-Both-2	**0.895**	**0.932**	**0.963**	**0.983**	**0.995**

Table 4.1: Assignment accuracies of scoring schemes based on the dataset ALL.

From these results, we see that the Bayes scoring schemes performed uniformly significantly better than the Normal scoring schemes. Their average performances are plotted in Figure 4.5, where each average is taken over 470 proteins. The average difference between the two is about 4% and it is as much as 6% in the instances with 70% connectivity information. We consider this as no surprise for two reasons: one reason is that the assumption of normal distributions for chemical shifts is very rough and there might be other structural factors that affect the chemical shift values; the other reason is even if the assumption makes sense, the estimate of means and standard deviations could differ from the true values.

Along with the boosting concept, for spin systems that do contain inter-residue chemical shifts, using them into the scoring schemes must be beneficial. We implemented this idea and we found that inter-residue chemical shifts indeed help distinguishing the residues. The above results demonstrate that using them can im-

Scoring Scheme	Connectivity Percentage				
	0%	10%	20%	30%	40%
Normal-Intra-1	0.103	0.177	0.263	0.365	0.489
Normal-Both-1	0.495	0.562	0.637	0.706	0.775
Normal-Intra-2	0.128	0.205	0.300	0.407	0.529
Normal-Both-2	0.531	0.599	0.675	0.744	0.810
Bayes-Intra-1	0.140	0.232	0.341	0.464	0.590
Bayes-Both-1	0.551	0.617	0.692	0.757	0.822
Bayes-Intra-2	0.172	0.264	0.373	0.494	0.622
Bayes-Both-2	**0.580**	**0.649**	**0.720**	**0.783**	**0.843**
Scoring Scheme	Connectivity Percentage				
	0%	10%	20%	30%	40%
Normal-Intra-1	0.619	0.753	0.875	0.958	0.992
Normal-Both-1	0.837	0.894	0.945	0.977	0.993
Normal-Intra-2	0.671	0.791	0.902	0.966	0.992
Normal-Both-2	0.868	0.915	0.955	0.980	0.994
Bayes-Intra-1	0.721	0.829	0.921	0.972	0.992
Bayes-Both-1	0.880	0.922	0.958	0.982	0.992
Bayes-Intra-2	0.747	0.850	0.929	0.976	0.993
Bayes-Both-2	**0.895**	**0.931**	**0.962**	**0.984**	**0.994**

Table 4.2: Assignment accuracies of scoring schemes based on HOMO.

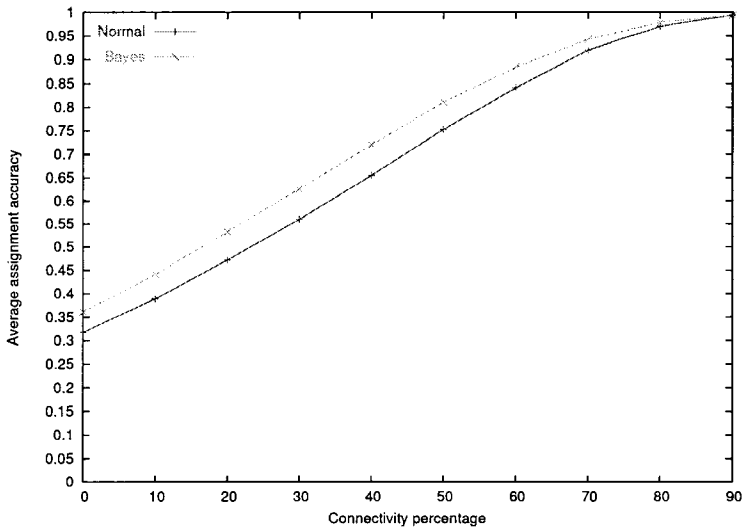


Figure 4.5: A comparison between the Bayesian scoring schemes and the scoring schemes based on normal assumptions: each assignment accuracy is taken as the average of 4 scoring schemes, namely, Intra/Both-1/2, on two training datasets ALL and HOMO.

	Connectivity Percentage									
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
Normal	0.318	0.389	0.472	0.560	0.655	0.753	0.841	0.920	0.970	0.993
Bayes	0.361	0.441	0.533	0.626	0.720	0.811	0.884	0.943	0.978	0.993
HOMO	0.337	0.413	0.500	0.590	0.685	0.780	0.861	0.931	0.974	0.993
ALL	0.342	0.417	0.505	0.595	0.690	0.784	0.864	0.932	0.974	0.993
Intra	0.136	0.220	0.321	0.434	0.560	0.693	0.808	0.908	0.968	0.992
Both	0.543	0.610	0.684	0.751	0.814	0.871	0.917	0.955	0.980	0.993
1	0.325	0.399	0.486	0.576	0.671	0.767	0.852	0.926	0.972	0.993
2	0.354	0.431	0.519	0.609	0.704	0.797	0.873	0.937	0.976	0.993

Table 4.3: The comparison of assignment accuracies of different types of scoring schemes.

prove the performance on average significantly, for example by 12% and 10% in the instances with 50% and 60% connectivity information respectively. Moreover, when no connectivity is used, using inter-residue chemical shifts can improve the assignment accuracy by as much as 35%. Figure 4.6 shows their average performances.

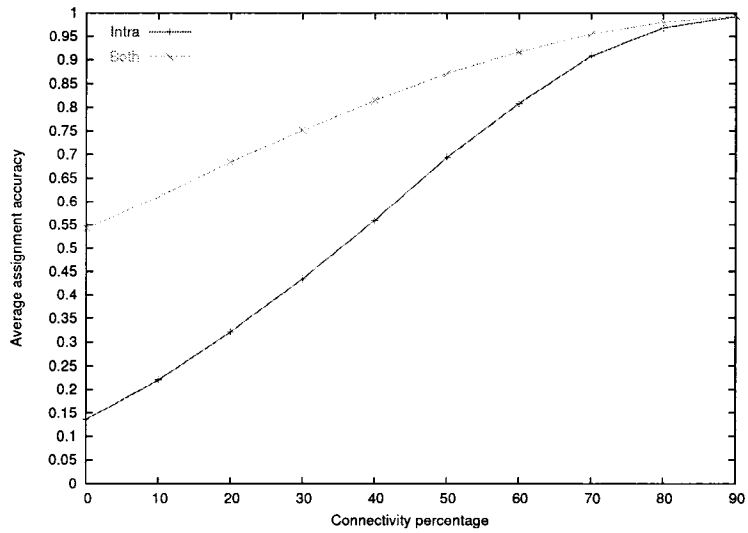


Figure 4.6: A comparison between using both chemical shifts and using only intra-residue chemical shifts: each assignment accuracy is taken as the average of 4 scoring schemes, namely, Normal/Bayes-1/2, on two training datasets ALL and HOMO.

Theoretically, training datasets for scoring scheme development should not be biased on any typical portion and hence NMR data for homologous proteins should be removed. Though our two training datasets ALL and HOMO vary quite a bit

in the numbers of all types of chemical shifts, their composition percentages are close to each other. This might explain the fact that we did not see much difference in the assignment accuracies by using different training datasets. By examining, in detail, the proteins that were removed from ALL to obtain HOMO, we found that the numbers of homologous proteins for different proteins are not large, but usually only a few. Figure 4.7 shows the average performances over the eight scoring schemes, where one could not really see the difference.

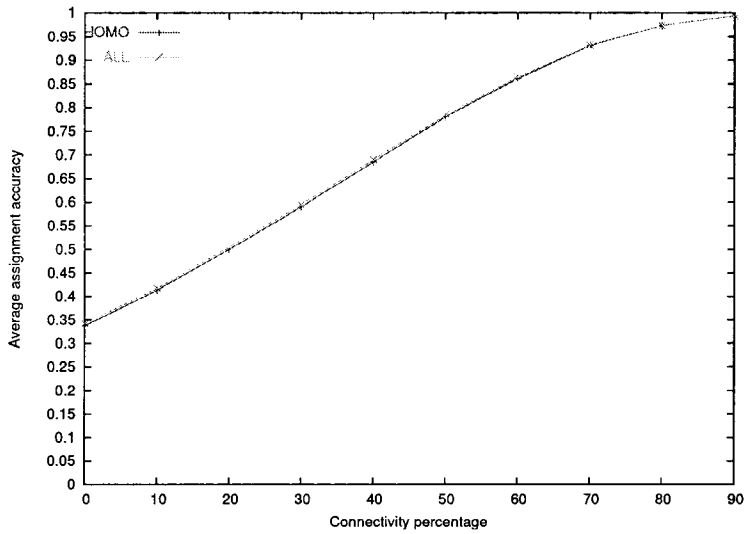


Figure 4.7: A comparison between the two datasets HOMO and ALL: each assignment accuracy is taken as the average of 8 scoring schemes, namely, Normal/Bayes-Intra/Both-1/2, on the two training datasets.

Since we know ahead of time that the secondary structures predicted by PsiPred come from a neural network where the secondary structures with the largest probability are reported, using them naively might introduce errors to the sequential assignment. We conjectured that using the accompanied probabilities of PsiPred might be helpful in reducing the prediction errors. We have tested a scheme to take advantage of the probabilities and the experimental results demonstrated that using them does improve the performance significantly. Figure 4.8 shows the average performances of scoring schemes using and not using the prediction probabilities, where we can see that using the accompanied probabilities is always a better choice, and it could improve the assignment accuracy as much as 5% (in the instances with

50% connectivity information).

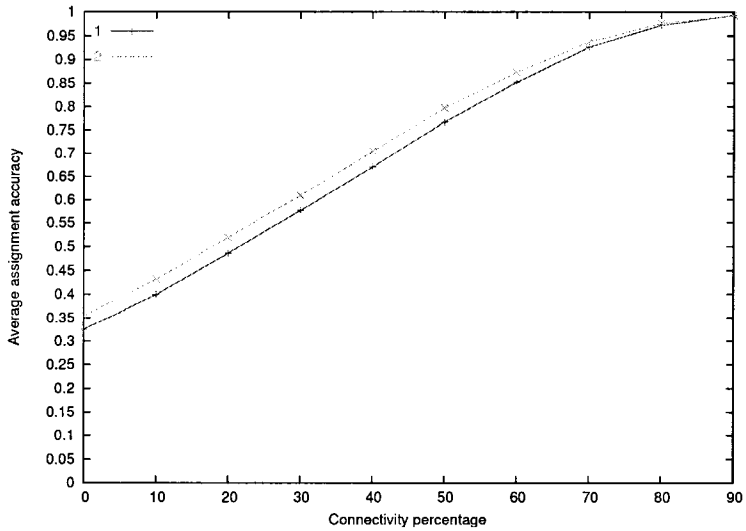


Figure 4.8: A comparison between using the prediction confidences by PsiPred and without using them: each assignment accuracy is taken as the average of 4 scoring schemes, namely, Normal/Bayes-Intra/Both, on two training datasets ALL and HOMO.

To summarize, we are able to claim that, according to our simulation study, the scoring schemes Bayes-Both-2 learned using both training datasets ALL and HOMO perform the best among all 16 scoring schemes. The scoring scheme trained using HOMO is provided freely as a web server [53] that is accessible through http://www.cs.ualberta.ca/~ghlin/src/WebTools/score.php, where the training dataset HOMO is also available. The web server contains two main functions, one is "single testing" that returns a score for mapping an input spin system to an amino acid residue and a secondary structure couple, and the other is "batch function" that accepts a protein sequence together with its secondary structures in PsiPred format and a file containing the spin systems, and returns an edge-weighted bipartite graph file, which can be readily fed to an integer programming solver, or any other algorithms for the CBM problem, together with some (or empty) connectivity information. Figure 4.9 shows a snapshot of the web server.

Our current work on developing the better scoring scheme focuses mainly on the scoring scheme training for backbone resonance assignment. This is a crucial

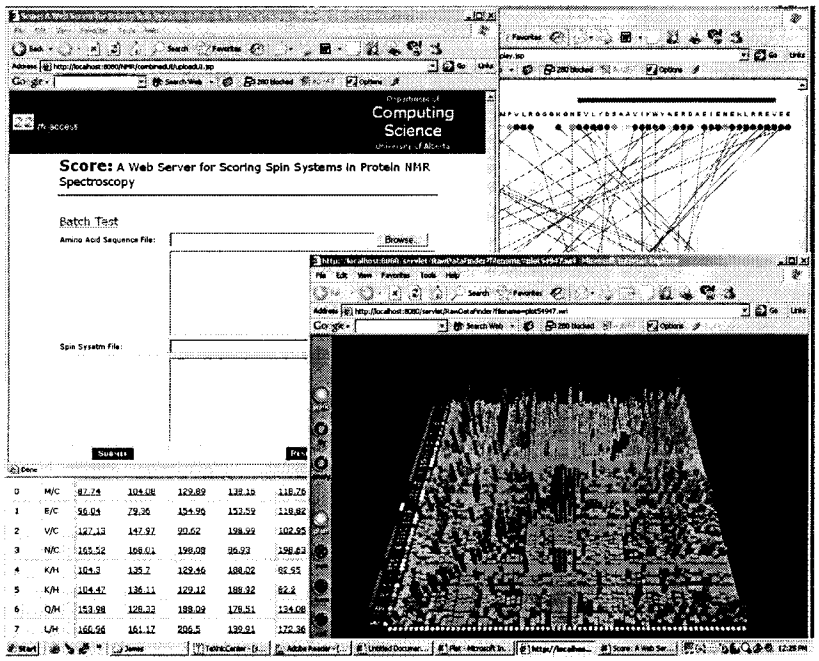


Figure 4.9: A snapshot of the Score web server using "batch function". Top left: two windows expecting a file of protein sequence together with secondary structures in Psi-Pred format and a file of spin systems. Bottom left: a window showing the score matrix (the complete bipartite graph). Top right: a bipartite graph with one side containing the spin systems and the other containing the linearly ordered amino acid residues in the target protein, where an edge indicates the best mappings for the residues. Bottom right: a graphical view of the score matrix, where the heights of the colored bars are proportional to the inverse of scores.

step towards our objective to develop a fully automated tool for protein NMR backbone resonance assignment that would be both robust and efficient. The scoring schemes we have developed here can be adopted in any existing assignment frameworks besides the CBM model, such as AutoAssign [78], Mapper [38], and MARS [45]. We expect the automated assignment tool to considerably speed up the protein structure determination process via NMR spectroscopy, and to transform the resonance assignment from a time-consuming task to a high-throughput process. For the scoring scheme itself, it can be extended into a more general one oriented towards full protein structure determination to include side-chain nuclei into the backbone assignment, as well as J-coupling constants and residual dipolar coupling constants. Such an integration not only fulfills the assignment of other structural factors, but also improves the assignment accuracy altogether as they can be used to cross validate each other.

Chapter 5

CISA: Combined NMR Resonance Connectivity Information Determination and Sequential Assignment

Many researchers realize that the determination of spin system connectivity plays a vital role in NMR sequential resonance assignment. The accuracy of connectivity information has a direct impact on the performance of automated sequential assignment process. In this chapter, we describe our solution to this issue, which is a heuristic algorithm called "CISA".

5.1 Overview

Resonance peaks from multi-dimensional NMR spectra contain chemical shifts for atoms from a common residue and for atoms from adjacent residues. For example, the CBCA(CO)NH spectrum records the heteronuclear coupling between H and N in one residue and the C^{α} and C^{β} in the preceding residue (see Figure 2.1(b)), which provides triples of chemical shifts $(H_i, C^{\alpha}_{i-1}, N_i)$ and$(H_i, C^{\beta}_{i-1}, N_i)$. Therefore, the inter-residue chemical shifts contained in the grouped spin systems can be used as evidence to determine whether some spin systems should assigned to adjacent residues in the protein sequence. This is what we refer to as "connectivity information". The objective of connectivity determination is to identify the true connections among the spin systems, and to chain the spin systems into strings. Then in the string assignment process, these strings of spin systems can be assigned to the non-overlapping polypeptides in the protein sequence. The quality and quantity of the connectivity information (or the identified connections among the spin systems) directly impact the success of any sequential resonance assignment. Once the connectivity determination is done with a certain high level of confidence, the string assignment problem could become trivial. Our simulation study in Chapter 4 supports the conclusion that if 80% correct connectivity information is available, the sequential resonance assignment problem can be solved efficiently and accurately.

Among the sequential assignment programs that use connectivity information, some of them [10, 18, 78, 38, 52] assume the availability of connectivity information and only focus on the string assignment problem. In the other models proposed for the sequential assignment [7, 22, 42, 62, 6], the connectivity information is determined along the way to assignment. These programs first use the differences between chemical shift values for the same nuclei in any pair of spin systems to find the connectivity information and then use these connectivities as constraints

to compute a sequential assignment. At various levels of success, these algorithms typically generate a large number of potential connectivity constraints, which grow exponentially as the spectral data quality decreases.

One big issue in connectivity determination is how to identify the true connections from multiple choices. Mostly due to the noise and data degeneracy, the connectivity determination is no longer a binary decision but a probabilistic one. As a result, one spin system could start more than one connectivity pair (many in general) and could end with more than one connectivity pair. A desirable way to describe the relationship among the spin systems is to use the graph, which we call "connectivity graph", where the vertices represent the spin systems, the directed edges represent the possible connections, and edge weights represent the probabilities associated with the connections. A path cover of the connectivity graph G, which is a set of disjoint paths that contain all vertices of G, indicates one set of potential connectivity constraints. The cost of a path cover is the sum of all edge weights in it. A minimum path cover contains the least paths among all path covers. The goal of connectivity determination is to find the minimum path cover with the minimum cost (or maximum cost depending on the definition of edge weight) in the connectivity graph. This problem is NP-hard because it is at least as hard as the NP-complete problem of finding a path cover of size 1 in an unit-weighted directed graph, which is referred to as a *Hamiltonian* path [23].

After the connectivity graph is constructed, PACES [22], a recently proposed sequential resonance assignment program, enumerates all the paths in the graph. A final set of non-conflicting paths are picked as identified connectivities. These identified connectivities are then used as constraints to finish the sequential assignment. One disadvantage in PACES is that it enumerates all paths in the connectivity graph without using the edge probability values and the enumeration might not be feasible if the graph is not sparse enough (see Experimental Results section for more information).

We proposed to perform the Connectivity Determination and Sequential Assignment simultaneously (acronym CISA, pronounced as 'kiss-a') by incorporating the chemical shift (or spin system) signature information into the connectivity determination. A key idea used in our sequential assignment program is that the chemical

shift signature information can be used to validate the connectivity constraint determination, and thus to dramatically decrease the number of constraints. In our development, we found that a string of connected spin systems typically has a much better score at the correct mapping position in the protein sequence than almost all the other (incorrect) mapping positions. This appears quite obviously when the size of the string increases. Such an observation leads to our conclusion that a string of spin systems having an outstanding mapping score has a high probability of being correctly chained. In other words, the connectivity determination and string assignment support each other.

Our algorithm starts with an *Open List* of strings and seeks to expand the string with the best mapping score. The subsequently generated descendant (longer) strings are appended to the Open List only if their mapping scores are better than their ancestor's. Another list, *Complete List*, kept in the algorithm, saves strings not further expandable. At the time Open List becomes empty, the high confident strings with their mapping positions are filtered out from Complete List; meanwhile, the conflicts among them are resolved in a greedy fashion.

The main distinction between CISA and PACES is the use of spin system signature information to progressively grow and validate the paths (the strings of spin systems) in the connectivity graph. In this way, a large number of connectivity edges could be filtered out according to the low quality of their resultant assignments. Therefore, the paths found in our output assignment might not necessarily be maximal paths in the connectivity graph, but they all have the outstanding mapping positions in the protein sequence. The extensive simulation studies on various test datasets demonstrated that our proposal of combining chemical shift signature information into connectivity determination is effective, and the combining improves the assignment accuracy significantly in comparison to PACES.

5.2 Connectivity Graph

The relationships between spin systems are formulated into an edge-weighted directed graph referred to as a *connectivity graph*. For every spin system, there is a vertex in the graph (in the rest of the chapter, vertex and spin system are used interchangeably). In this section, we describe the use of C^{α} and C^{β} chemical shift

differences to determine the connectivities between spin systems. In our experiments, we used another combination that contains C^{α}, C^{β}, and C chemical shift differences. Other combinations of chemical shifts are possible and their connectivity graphs can be built similarly. For two spin systems $v_i = (H_i, N_i, C_i^{\alpha}, C_i^{\beta}, C_{i-1}^{\alpha}, C_{i-1}^{\beta})$ and $v_j = (H_j, N_j, C_j^{\alpha}, C_j^{\beta}, C_{j-1}^{\alpha}, C_{j-1}^{\beta})$, if both $|C_i^{\alpha} - C_{j-1}^{\alpha}| \leq \delta_{\alpha}$ and $|C_i^{\beta} - C_{j-1}^{\beta}| \leq \delta_{\beta}$ hold, then there is an edge from v_i to v_j with its weight calculated as

$$\frac{1}{2}\left(\frac{|C_i^{\alpha} - C_{j-1}^{\alpha}|}{\delta_{\alpha}} + \frac{|C_i^{\beta} - C_{j-1}^{\beta}|}{\delta_{\beta}}\right); \tag{5.1}$$

Similarly, if both $|C_j^{\alpha} - C_{i-1}^{\alpha}| \leq \delta_{\alpha}$ and $|C_j^{\beta} - C_{i-1}^{\beta}| \leq \delta_{\beta}$ hold, then there is an edge from v_j to v_i with its weight calculated as

$$\frac{1}{2}\left(\frac{|C_j^{\alpha} - C_{i-1}^{\alpha}|}{\delta_{\alpha}} + \frac{|C_j^{\beta} - C_{i-1}^{\beta}|}{\delta_{\beta}}\right).$$

Here both δ_{α} and δ_{β} are pre-determined tolerance thresholds, which are typically set to 0.2 ppm and 0.4 ppm [22, 47], though minor adjustments are sometimes necessary to ensure a sufficient amount of connectivity. If neither case occurs, then there is no edge between v_i and v_j. Equation (5.1) is not the only weighting function, and some other functions as suggested in [47] on the chemical shift differences could be adopted to weigh the edges.

In the other combination that contains C^{α}, C^{β}, and C chemical shift differences, it is required that at least 2 out of the following 3 conditions hold: $|C_i^{\alpha} - C_{j-1}^{\alpha}| \leq \delta_{\alpha}$, $|C_i^{\beta} - C_{j-1}^{\beta}| \leq \delta_{\beta}$, and $|C_i - C_{j-1}| \leq \delta$, and the weight of edge from v_i to v_j is evaluated analogously as in Equation (5.1).

After every pair of vertices (spin systems) has been examined to have an edge or not, we finish the construction of the connectivity graph. However, some true connectivities might not be present in the connectivity graph while some wrong ones might be present.

5.3 String Growing

With the connectivity graph constructed, PACES proceeds to enumerate all the (simple, directed) paths in the graph without using the detailed edge weights. We choose another approach to grow a path using the edge weights. The growth is

guided by the quality of the path mapping to the protein sequence. Given a path, its quality is measured by the mapping score of the path at the best mapping position on the target protein. All the edges, coming out of the ending spin system in the current path to be expanded, are sorted in a non-increasing order of their weights. For the edge at the head of the order, the temporary extended path (called *child path*) is formed and its best mapping position on the target protein can be found via a linear search. The mapping score of this child path is calculated and compared with the mapping scores of its parent path to decide whether to accept it or not. It has been observed that a sufficiently long path is able to detect the succeeding spin system by taking advantage of the discerning power of the scoring scheme [68]. Therefore, it is expected that using mapping scores to filter the extended paths would give rise to much fewer potential paths for further consideration and eventually avoid exhaustive search as done in PACES.

In each iteration, CISA starts with an *Open List* (OL) of paths and seeks to expand the one with the best mapping score. The OL has a fixed size S (in our experiments, $S = 60$) and the detailed value set for S depends on computer memory size. In our case, the experiments were done on a typical desktop with a 1Gb RAM. We found that S can be chosen from a value in the range between 40 and 80 without affecting the performance significantly. We used the median value of 60. The subsequently generated child paths are appended to OL if their mapping scores are high and there is room in OL, or if their mapping scores are higher than that of some existing path in OL. Another list, *Complete List* (CL), is kept in CISA to save those paths that can not be expanded further. At the time OL becomes empty, the high quality paths with their mapping positions are extracted out of CL where the conflicts are resolved in a greedy fashion. CISA chooses the most reliable string out of the remaining connectivity graph in each iteration and the corresponding path is removed from the graph. Our algorithm terminates when the connectivity graph becomes empty and returns the constructed strings with their mapping positions on the target protein.

5.4 Experiments

We have designed two experiments to test our algorithm, CISA, and to compare its performance with that of PACES.

The first experiment used 22 proteins that were tested by PACES and one real dataset `Zdom` that we obtained from AutoAssign [78]. However, we did not obtain for each protein the exact instance as tested by PACES in [22]. Therefore, we simulated them from the corresponding protein entries in BioMagResBank according to the simulation procedure described in [22]. Note that there are some more proteins that were tested by PACES and did not require simulations. These proteins subsequently were excluded from our datasets. This experiment was designed to compare the performance between CISA and PACES.

The second experiment was designed to show the computational speed of CISA and its overall assignment accuracy, for which all eligible protein entries deposited in BioMagResBank were simulated and tested. The performance of CISA on individual proteins and the average assignment accuracy were collected. Since it was possible to run PACES on all these proteins within a reasonable amount of time, we chose to run CISA only.

In the first experiment, PACES was run on every dataset for 1 iteration only because we did not manually analyze the assignment to prepare for the second iteration. In this sense, all three programs are automated without any manual adjustment. As a result, the performance of PACES reported in the following might be a little worse than that reported in [22], where PACES was usually run a few iterations on a dataset with manual adjustments in order to improve the assignment accuracy.

5.4.1 Experiment 1

In the first experiment, we used the datasets tested in [22] and followed the same simulation procedure, which used three inter-residue chemical shifts, C^{α}, C^{β}, and carbonyl C, for connectivity graph construction (tolerance thresholds were $\delta_{\alpha} =$ 0.2ppm, $\delta_{\beta} = 0.4$ppm, and $\delta = 0.15$ppm). The reason we did our own simulation in this experiment is the unavailability of the original datasets from [22]. Our simulated datasets were very close to the corresponding datasets in [22] in terms of the number

of missing spin systems (and the performance of PACES). Overall, in these datasets, the percentage of missing spin systems ranged from 3% to 39%. We find that the existence of missing spin systems challenged the robustness of our CISA in many ways, especially in its assignment accuracy. A real instance `Zdom` was also included in this experiment, which we indirectly obtained from AutoAssign [78] and did not need simulation. The performances of PACES and CISA on these 23 instances are collected in Table 5.1. Their assignment accuracies are also plotted in Figure 5.1. In summary, CISA outperformed PACES in all instances except `bmr4402` where PACES performed a little bit better than CISA (assignment accuracies 0.873 vs 0.860). The tendency of the assignment accuracies shows that their performance gap becomes larger as the instances become harder.

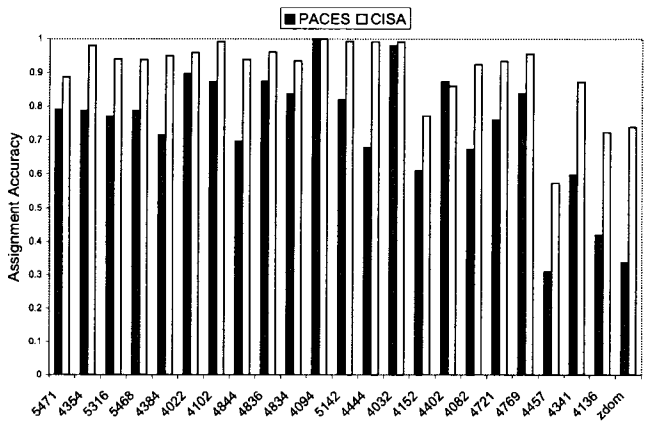


Figure 5.1: Plots of assignment accuracies for PACES and CISA on the simulated datasets for proteins from [22], using the exact dataset generation method as described, and a real dataset `Zdom` indirectly obtained from AutoAssign [78].

5.4.2 Experiment 2

The second experiment was designed to show the computational speed of CISA and its overall assignment accuracy. To this purpose, we simulated all eligible protein entries deposited in BioMagResBank using the default tolerance thresholds. We chose

Length	InstanceID	#SpinSystems	PACES	CISA
731	bmr5471	654	0.791	0.887
370	bmr4354	330	0.788	0.979
288	bmr5316	265	0.770	0.940
266	bmr5468	240	0.788	0.938
262	bmr4384	221	0.715	0.950
260	bmr4022	242	0.897	0.959
232	bmr4102	212	0.873	0.991
221	bmr4844	198	0.697	0.939
217	bmr4836	206	0.874	0.961
189	bmr4834	166	0.837	0.934
133	bmr4094	129	1.000	1.000
130	bmr5142	127	0.819	0.992
128	bmr4444	106	0.679	0.991
124	bmr4032	119	0.980	0.990
Group 1 Avg.			**0.822**	**0.961**
214	bmr4152	197	0.610	0.772
105	bmr4402 (126–230)	93	0.873	0.860
139	bmr4082	132	0.674	0.924
81	bmr4721	74	0.760	0.933
68	bmr4769	67	0.838	0.956
Group 2 Avg.			**0.751**	**0.889**
227	bmr4457	162	0.310*	0.575
192	bmr4341	117	0.598	0.872
110	bmr4136	105	0.419	0.724
71	Zdom*	65	0.338	0.738
Group 3 Avg.			**0.416**	**0.727**
Overall Avg.			**0.736**	**0.905**

Table 5.1: Assignment accuracies of PACES and CISA on simulated datasets for proteins from [22], using the exact dataset generation method as described, and a real dataset `Zdom` indirectly obtained from AutoAssign [78]. Tolerance thresholds are $\delta_\alpha = 0.2$ppm, $\delta_\beta = 0.4$ppm, and $\delta = 0.15$ppm. #SpinSystems records the number of available spin systems for one instance. The datasets are partitioned into three groups. In the first group, datasets all have carbon alpha C^α, carbon beta C^β, and carbonyl C chemical shifts of high quality; In the second group, datasets all have carbon alpha C^α, carbon beta C^β, and carbonyl C chemical shifts, but of low quality; In the third group, datasets have only carbon alpha C^α and carbon beta C^β chemical shifts of various quality. *PACES performance on this dataset was obtained by reducing tolerance thresholds to $\delta_\alpha = 0.15$ppm and $\delta_\beta = 0.3$ppm to ensure an assignment in 8 hours.

to use the chemical shift combination (H, N, C^{α}, C^{β}), and consequently the eligible proteins are those that contain all these four types of chemical shifts (though they might be obtained from different spectra). The default tolerance thresholds for C^{α} and C^{β} are 0.2ppm and 0.4ppm, respectively. To screen out some highly degenerate protein entries, we set up a 5-minute time limit for CISA on each protein. That is, if CISA could not terminate the assignment for one protein in 5 minutes, then the protein entry was discarded. We remark that 5 minutes was long enough since for most of the proteins on which CISA terminated, it terminated within seconds. One interesting discovery is that we found some proteins have significant resolution differences within their spectral profiles, for example, `bmr4402` (cf. Experiment 3) has one half of high resolution but the other half of very low resolution. Through setting up the time limit, CISA was able to detect the low resolution proteins about 20kDa in size.

In summary, CISA was able to finish the assignments for 360 proteins in total. The length of these proteins ranges from 58 to 198, and the assignment accuracy from 0.62 to 1.00. The average assignment accuracy is 0.903, which is consistent with the results in Experiment 1. The assignment accuracy versus the length of the protein is plotted in Figure 5.2, where each cross represents an instance. From the plot, we see that CISA appears insensitive to the size of proteins.

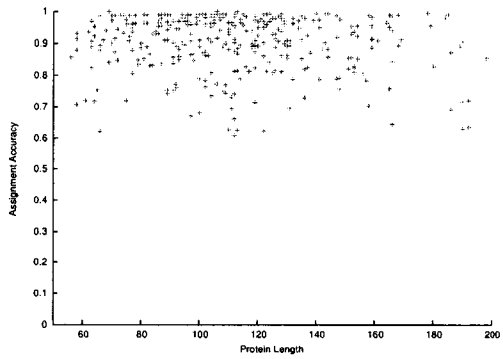


Figure 5.2: Plots of assignment accuracies for CISA on the simulated datasets for 360 proteins from BioMagResBank, where each cross represents one instance using its length.

5.5 Discussions and Conclusions

On a normal desktop with a 1.6GHz AMD-2000 processor and a 1Gb RAM, for the instances in the first two experiments, the overall running time of CISA ranged from a few seconds to 4 hours (and most of them were done in less than 20 minutes). For the instances in the third experiment, the overall running time of CISA never exceeded 30 minutes. PACES failed to finish the assignments in 8 hours for a number of instances because their connectivity graphs were too complicated and the enumeration of paths became infeasible (1Gb memory ran out). For this reason, we manually adjusted the tolerance thresholds to reduce the graph complexity in order for PACES to output some assignments. However, we admit that doing this might bring down the assignment accuracy a bit since true edges could be removed from the connectivity graph.

Through CISA, we have successfully combined the spin system signature information into the path growing in the connectivity graph, which prunes the search space more effectively compared to PACES (which failed on a number of complex instances in the first two experiments). However, in the current version of CISA the weights of edges are used only to order the child paths. We believe that some better usage of edge weights in the mapping score evaluation for a growing path would help more effectively quantifying the quality of the growing path. We have tried some simple linear functions on the edge weights and the mapping scores of paths, which turned out not to serve satisfactorily. We are currently investigating more combinations. Across all the experiments, we found that CISA spent a large portion (about 50%) of time in finding the first string. We also observed that for all instances, after 3 to 4 iterations, CISA found the best string in a straightforward way. In other words, CISA running time was mostly consumed in its first 3–4 iterations. One possible way to speed up CISA in the first string finding could be to use only high probability edges in the connectivity graph. This method is still under investigation.

Chapter 6

GASA: A Graph-Based Automated NMR Backbone Resonance Sequential Assignment

The traditional automated assignment procedure involves three steps, namely peak grouping, connectivity determination and string assignment. This procedure is widely used by most automated assignment systems. In this chapter, we describe a novel assignment procedure to separate the assignment procedure not into physical steps, but only into virtual steps and use their outputs to cross validate each other. The novelty lies in the places where the ambiguities in the peak grouping step could be resolved by the connectivity determination and the ambiguities in the connectivity determination could be resolved by the string assignment. In such a way, all ambiguities in the whole assignment procedure would be resolved globally and optimally.

6.1 Overview

The traditional automated assignment procedure involves three separate steps, which respectively group resonance peaks from multiple spectra into spin systems, predict the connectivity among the resultant spin systems to assemble them into strings, and then to map strings to non-overlapping consecutive amino acid residues in the target protein. This is illustrated in Figure 6.1, where the scoring scheme quantifies the chemical shift signature information for each steps if necessary. Several assignment

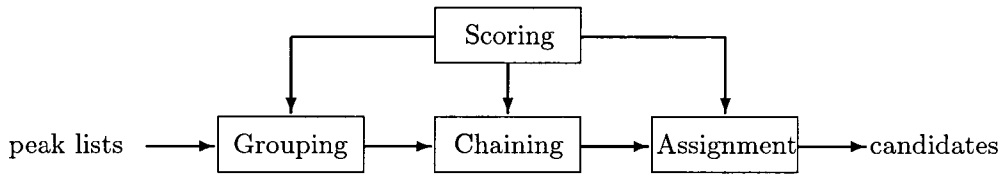


Figure 6.1: The flow chart of the peak assignment process.

methods have adopted the traditional procedure to automate the resonance assignment process. Among them, AutoAssign [78] and RIBRA [73] are two programs that fully automate the whole assignment process while most of other programs assume that the perfect spin systems are given as input, and focus on the design of computational models for connectivity determination and string assignment. Within these two programs, the peak grouping is especially addressed with a binary-decision model, which considers the HSQC peaks as base peaks and subsequently maps the

peaks from other spectra to these base peaks. The H and N chemical shift values of the mapped peaks must fall within the pre-specified H and N chemical shift tolerance thresholds of the base peaks. However, the binary-decision model in the peak grouping inevitably suffers from its sensitivity to the tolerance thresholds. From one protein dataset to another, the chemical shift tolerance thresholds vary because of the experimental condition and the structure complexity. Large tolerance thresholds could create too many ambiguities in both spin system identification and connectivity determination, which lead to a dramatic decrease of assignment accuracy. On the other hand, small tolerance thresholds would produce too few spin systems that would hardly lead to a useful assignment when the resolution of spectral data is low.

While there are a considerable number of assignment programs, the assignment accuracy remains unsatisfactory in practice. Even worse, if the given spectral data is of low resolution, most programs often fail to output a meaningful assignment. Through a thorough investigation, we first identified that the bottleneck in most automated assignment programs is the performance of the peak grouping task. Our previous work [68, 69] showed that the quality and quantity of spin systems produced in the peak grouping could have the most significant effect on the assignment. Nevertheless, the widely used binary-decision model is inefficient in producing the spin systems of high quality for spectra with typical resolution.

Second, we found that in the traditional procedure, which is the basis of most automated assignment program, each task is conducted individually . The input of each task is assumed to contain enough information to produce some meaningful output. However, for low resolution spectral data, the ambiguities that appear in one task seem very hard to be internally resolved. Though it is sometimes possible to output multiple candidates, the uncertainties might cause more ambiguities in the succeeding tasks. Consequently, the whole process would fail to produce a meaningful resonance assignment. In the previous chapter, we have shown that by incorporating the assignment verification into the connectivity determination, we can provide a better approach for resolving the ambiguities in connectivity determination. We believe that by combining the peak grouping with the connectivity determination and string assignment, we could effectively resolve the ambiguities appearing in the peak grouping stage and present a better solution to the automated

resonance assignment.

6.2 GASA Algorithm

The input data to our program consists of the protein sequence and a set of NMR peak lists. Except for the HSQC spectrum, our approach does not require any other specific NMR spectra as long as they are sufficient for the assignment purpose. For ease of exposition and fair comparison with RIBRA [73], we assume the availability of spectral peaks containing chemical shifts for carbon alpha and carbon beta, as well as the HSQC peak list. Thus the peak lists we use to conduct the experiments and comparison include HSQC, CBCA(CO)NH and HNCACB, although our approach can accept many other combinations. Given these three peak lists, RIBRA tried to find the two closest CBCA(CO)HN peaks and the four closest HNCACB peaks for each peak in the HSQC spectrum under the constraint that the H and N differences between these peaks are within the given tolerance thresholds. If more than 6 peaks are found, RIBRA generates all possible combinations to represent all legal spin systems. The true spin systems are filtered out in the later process in RIBRA. The difference between the peak grouping model applied in RIBRA and the general binary-decision approach used in AutoAssign is that the ambiguities appearing in the peak grouping could be automatically resolved to some extent in RIBRA, while in AutoAssign, additional manual work has to be conducted or more peak lists are required to provide the redundant information for resolving the ambiguities. Nonetheless, we argue that the peak grouping model in RIBRA is still susceptible to the change of pre-chosen tolerance thresholds because large tolerance thresholds could make RIBRA produce a huge number of legal spin systems while small tolerance thresholds would lead to too few spin systems to perform the assignment process.

To eliminate the sensitivity to the given tolerance thresholds in peak grouping and provide a computational model for automatically resolving ambiguities and conducting the sequential assignment, we designed a two-stage Graph-based Approach for Sequential Assignment (acronym **GASA**) that not only addresses the hard issues in the peak grouping but also presents a new model to automate the sequential assignment process. In the first stage, we propose a two-way nearest neighbor search

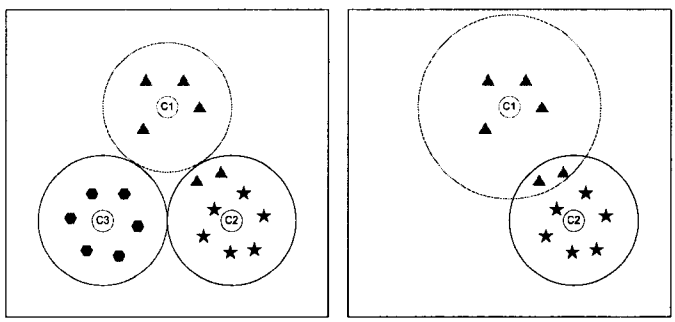
approach that eliminates the requirement of user-specified H and N chemical shift tolerance thresholds. The output of the first stage is two lists of spin systems. One list contains the perfect spin systems and the other the imperfect spin systems. In the second stage, connectivity determination is performed to resolve the ambiguities contained in the imperfect spin systems, and the string assignment would be included as a subroutine to determine the confident connectivity information. In our approach, once the ambiguities in the imperfect spin systems are resolved, connectivity determination and string assignment would be completed at the same time.

6.2.1 Filtering

The task of filtering is to find all perfect spin systems without asking for the tolerance thresholds. In all peak grouping models we have seen, the tolerance thresholds are required as the cut-offs that decide if two peaks should reside in the same spin system or not. As a result, different tolerance thresholds would clearly produce different sets of possible spin systems, and for the spectral data with the low resolution, a minor change of tolerance thresholds would lead to the a difference in the final assignment. Thus the question of how to choose the tolerance thresholds is a very challenging issue in automated resonance assignment. An intuitive solution to this issue is to use an exhaustive search that automatically tests all possible tolerance thresholds. Obviously, this solution is very time-consuming and not applicable especially for the large protein. During our investigation, we noticed that the peaks residing in the same spin system usually have closer H and N chemical shifts than those in different spin systems. Hence we could use the nearest neighbour method to differentiate peaks in different spin systems. The peaks in the HSQC spectrum would be considered as centers, and each peak in CBCA(CO)NH and HNCACB would be distributed to the closest center. Given a center $C = (H_C, N_C)$ and a peak $P = (H_P, N_P, C_P^{\alpha/\beta})$, the distance between them is defined as

$$D_\Delta = \sqrt{\left(\frac{H_P - H_C}{\sigma_H}\right)^2 + \left(\frac{N_P - N_C}{\sigma_N}\right)^2}, \tag{6.1}$$

where σ_H and σ_N are the standard deviations of H and N chemical shifts that are collected from BioMagResBank (`http://www.bmrb.wisc.edu`). In the ideal case,



(a) (b)

Figure 6.2: Problems in the peak grouping.(a) There are 3 HSQC peaks as 3 centers C_1, C_2, C_3. Each peak is associated with the closest center. Only C_3 forms a perfect spin system with 6 associated peaks. (b) C_1 finds the top 6 closest peaks to form a perfect spin system and meanwhile C_2 forms a perfect spin system with rest of peaks

each center should have 6 peaks distributed to it in total. However, due to the chemical shift degeneracy, some centers may have less than 6 or even 0 peaks because the peaks belonging to them might be closer to other centers, which makes those other centers have more than 6 peaks. Figure 6.2 illustrates a simple example with 3 HSQC peaks as centers for this situation. In the ideal case, each center should have 6 peaks in total. However, only one perfect spin system with center C_3 is formed because the two peaks belonging to center C_1 are closer to center C_2, which creates ambiguities in both spin systems. Nevertheless, if we place the focus on center C_1, we may find that its two peaks residing in the wrong spin system are still in its top 6 closest peaks. If the spin system with center C_1 is formed by adding these two peaks (see Figure 6.2(b)), the spin system with center C_2 also becomes a perfect spin system. We designed a bidirectional nearest neighbour model in Filtering, which consists of two steps: Residing and Inviting. In the Residing step, we associated each peak in the CBCA(CO)NH and HNCACB spectrum with the closest HSQC peak. If the HSQC peak with its associated peaks in the CBCA(CO)NH and HNCACB spectrum form a perfect spin system, the resultant spin system is inserted into the list of perfect spin systems and the contained peaks are removed from the nearest neighbour model. In the Inviting step, each peak in the HSQC spectrum looks

for the top k closest peaks in the CBCA(CO)NH and HNCACB spectrum, and if a perfect spin system can be formed within these k peaks, the similar procedure would be conducted for the formed perfect spin system and its contained peaks. The parameter k is related to the number of peaks contained in a perfect spin system. It is usually specified as 1.5 times the number of peaks in a perfect spin system. It can be automatically computed in the program with respect to the input peak lists. The aforementioned two steps would be continually executed until no perfect spin system can be found and two lists of spin systems are produced. One list contains the perfect spin system and the other list saves the imperfect spin systems. The user could specify the maximal H and N tolerance thresholds to speed up the process, but a minor differences in the maximal tolerance thresholds would not affect the performance of this model. The pseudocode of Filtering is in the following;

Phase 1: ***Filtering***

Input: HSQC, CBCA(CO)NH, HNCACB peak lists.

Residing: For each peak in CBCA(CO)NH and HNCACB, find the closest peak in HSQC. Remove those peaks that form perfect spin systems.

Inviting: For each peak in HSQC, find top k peaks in CBCA(CO)NH and HNCACB. Remove those peaks that form perfect spin systems.

Stop if no perfect spin system is found.

6.2.2 Resolving

The goal of the Resolving step is to identify the true peaks contained in the imperfect spin system and then to conduct the connectivity determination and string assignment. Nevertheless, it is very hard to distinguish between true peaks and false peaks when each imperfect spin system is individually checked. During our development, we found that in most cases the spin systems containing true peaks could produce more confident connectivity information than those containing false peaks. Hence we believed that we could extract the true peaks from the imperfect spin systems

through the search of high confident connectivity information contained among the perfect spin systems and imperfect spin systems, and those peaks used in building confident connections would have a high probability of being true peaks.

The relationships between spin systems are formulated into the *connectivity graph* similar to CISA as we discussed in the previous chapter. Given two perfect spin systems $v_i = (\mathrm{H}_i, \mathrm{N}_i, \mathrm{C}_i^{\alpha}, \mathrm{C}_i^{\beta}, \mathrm{C}_{i-1}^{\alpha}, \mathrm{C}_{i-1}^{\beta})$ and $v_j = (\mathrm{H}_j, \mathrm{N}_j, \mathrm{C}_j^{\alpha}, \mathrm{C}_j^{\beta}, \mathrm{C}_{j-1}^{\alpha}, \mathrm{C}_{j-1}^{\beta})$ if both $|\mathrm{C}_i^{\alpha} - \mathrm{C}_{j-1}^{\alpha}| \leq \delta_{\alpha}$ and $|\mathrm{C}_i^{\beta} - \mathrm{C}_{j-1}^{\beta}| \leq \delta_{\beta}$ hold, then there is an edge from v_i to v_j with its weight calculated as

$$\frac{1}{2}\left(\frac{|\mathrm{C}_i^{\alpha} - \mathrm{C}_{j-1}^{\alpha}|}{\delta_{\alpha}} + \frac{|\mathrm{C}_i^{\beta} - \mathrm{C}_{j-1}^{\beta}|}{\delta_{\beta}}\right);$$

Here both δ_{α} and δ_{β} are pre-determined tolerance thresholds, which are typically set to 0.2ppm and 0.4ppm, though minor adjustments are sometimes necessary to ensure a sufficient amount of connectivities. Given one perfect spin system $v_i = (\mathrm{H}_i, \mathrm{N}_i, \mathrm{C}_i^{\alpha}, \mathrm{C}_i^{\beta}, \mathrm{C}_{i-1}^{\alpha}, \mathrm{C}_{i-1}^{\beta})$ and one imperfect spin system $v_j = (\mathrm{H}_j, \mathrm{N}_j, \mathrm{C}_{j1}^{\alpha}, \mathrm{C}_{j2}^{\alpha}, \cdots, \mathrm{C}_{jm}^{\alpha}, \mathrm{C}_{j1}^{\beta}, \mathrm{C}_{j2}^{\beta}, \cdots, \mathrm{C}_{jn}^{\beta})$, we check each legal combination $v_j' = (\mathrm{H}_j, \mathrm{N}_j, \mathrm{C}_{jl}^{\alpha}, \mathrm{C}_{jk}^{\beta}, \mathrm{C}_{jp}^{\alpha}, \mathrm{C}_{jq}^{\beta})$ where $l, k \in [1, m]$ and $p, q \in [1, n]$. The carbon chemical shifts with subscripts l, k represent the intra chemical shift and those with subscripts p, q representing the inter chemical shifts. If both $|\mathrm{C}_i^{\alpha} - \mathrm{C}_{jp}^{\alpha}| \leq \delta_{\alpha}$ and $|\mathrm{C}_i^{\beta} - \mathrm{C}_{jq}^{\beta}| \leq \delta_{\beta}$ hold, then there is an edge from v_i to v'_j with its weight calculated as

$$\frac{1}{2}\left(\frac{|\mathrm{C}_i^{\alpha} - \mathrm{C}_{jp}^{\alpha}|}{\delta_{\alpha}} + \frac{|\mathrm{C}_i^{\beta} - \mathrm{C}_{jq}^{\beta}|}{\delta_{\beta}}\right); \tag{6.2}$$

If both $|\mathrm{C}_{jl}^{\alpha} - \mathrm{C}_i^{\alpha}| \leq \delta_{\alpha}$ and $|\mathrm{C}_{jk}^{\beta} - \mathrm{C}_i^{\beta}| \leq \delta_{\beta}$ hold, then there is an edge from v'_j to v_i with its weight calculated as

$$\frac{1}{2}\left(\frac{|\mathrm{C}_{jl}^{\alpha} - \mathrm{C}_i^{\alpha}|}{\delta_{\alpha}} + \frac{|\mathrm{C}_{jk}^{\beta} - \mathrm{C}_i^{\beta}|}{\delta_{\beta}}\right);$$

It is possible that there are multiple connections between one perfect spin system and one imperfect spin system but at most one connection could be true. Given two imperfect spin systems, no connection is allowed.

With the connectivity graph constructed, we use essentially the same heuristic search algorithm in CISA [67], in which the search is guided by the quality of the generated path mapping to the target protein. Given a path, its quality or mapping

score is measured by the average likelihood of the best mapping position for the path on the target protein. The edge weights are used to order the edges coming out of the ending spin system in the current path to provide the candidate spin systems for the current path to grow to. It has been observed that a sufficiently long path itself is able to detect the succeeding spin system by taking advantage of the discerning power of the scoring scheme ([68]). In each iteration, GASA starts with an **Open List** (OL) of paths and seeks to expand the one with the best mapping score. Another list, **Complete List**(CL), is used in the algorithm to save those completed paths. In the following, we briefly describe the algorithm for finding connectivity determination and resolving the ambiguities in imperfect spin systems.

Phase 2: ***GASA***

OL Initialization: Let G denote the constructed connectivity graph. We first search for all unambiguous edges in G. We expand those edges into simple paths with a pre-defined length L by both tracing their head vertices backward and their tail vertices forward. The tracing would stop if either of the following conditions is satisfied. (1) The new traced vertices are sitting in the paths. (2) The length of the path is L. The paths stored in OL are sorted in the non-increasing order of their mapping scores. The size of OL is shrunk to a fixed size S and only the first S paths in OL are kept for the trade-off between computing time and accuracy.

Path Growing: In this step, the algorithm tries to bidirectionally expand the top ranked path stored in OL. Denote this path as P, the first vertex in P as h and the last vertex in P as t. All directed edges incident to h and incident from t are considered to generate potential child paths. For every potential child path, the algorithm finds its best mapping position in the target protein and calculates the best mapping score. If its mapping score is higher than that of some path already stored in OL, then the child path is added into OL (and the path with least mapping score is removed from OL). If none of the potential child paths of P is added into OL or P is not expandable in either direction, path P is added into CL. The algorithm proceeds to consider the top ranked path in OL iteratively and the growing process is done when OL

becomes empty.

CL Finalizing: Let P denote the path of the highest mapping score (tie is broken to the longest path) in CL. Other paths in CL with both length and score less than 90% of the length and score of path P are discarded from further consideration. The remaining paths are considered to contain reliable connectivities and would be examined further.

Connectivity Filtering: Only those edges occurring in at least 90% of the paths in CL are chosen as reliable connectivities and the other edges are removed from further consideration. Subsequently, paths with edges removed are broken down into shorter pieces.

Ambiguities Resolving: At this stage, the paths in CL are considered to contain only reliable connectivities. The longest one of them is the target in this iteration. Denote this path as P. The spin systems on P are removed from the connectivity graph G, as well as the edges incident to/from them. For the imperfect spin systems in P, the peaks used to build the connections in P could be considered as true peaks. If the remaining connectivity graph is still non-empty, the algorithm proceeds to the next iteration. Otherwise, it terminates and reports the assignment, i.e., the strings it found and their mapping positions on the target protein.

6.3 Experiments

Four experiments are designed to evaluate the value of our work by comparing the performance of GASA with recently developed methods.

In the previous chapter, we compared CISA, which is a subcomponent of GASA, with PACES [22] on the PACES datasets. In the first two experiments, we use our simulated dataset to make a full comparison with more published methods on connectivity determination. The test results further demonstrate the performance of combining connectivity determination with string assignment by comparing GASA with RANDOM [47], PACES, and MARS [45]. Our simulated dataset contains 12 proteins from [76], which do not have solved structures and thus would not bias the chemical shift signature information.

The purpose of the third and the fourth experiments is to to demonstrate the advantage of merging all peak grouping, connectivity determination, and string assignment together into a single iterated process by comparing GASA with RIBRA. In the third experiment, we re-examine the 5 released datasets by RIBRA, which are simulated from the real protein NMR data deposited in BioMagResBank. In this experiment, GASA performed basically as good as RIBRA. In the fourth experiment, we sought out another simulation which we thought was much closer to the reality to determine whether the results for RIBRA data are representative.

All three programs in the first two experiments, RANDOM, PACES, and MARS, reported the same statistic (they may use different terms), which we denoted as *accuracy*. The definition of *accuracy* is

$$accuracy = \frac{number\ of\ correctly\ assigned\ spin\ systems}{number\ of\ available\ spin\ systems}.$$

To make fair comparison, we also provide the same statistic in the first two experiments. This also helps us justify our simulation and tests by comparing our results with those reported in the original publications of these three programs.

RIBRA, however, defines two different criteria, namely *precision* and *recall*, to measure the performance. In particular,

$$precision = \frac{number\ of\ correctly\ assigned\ amino\ acids}{number\ of\ assigned\ amino\ acids},$$

$$recall = \frac{number\ of\ correctly\ assigned\ amino\ acids}{number\ of\ amino\ acids\ with\ known\ answers}.$$

We use the same criteria in the third and fourth experiments to facilitate the comparison.

6.3.1 Dataset Generation

In the literature, the simulation procedure of peak lists or spin systems from data entries deposited in BioMagResBank is basically the same in all simulations. The difference is what type of errors should be simulated and how to simulate them. In [76], 14 proteins were carefully chosen to form datasets for simulation studies on the proposed constrained bipartite matching model for sequential assignments. These proteins do not have solved atomic structures and were not used to derive the scoring scheme adopted in our experiments. Among these proteins, `bmr4309`

and bmr4393 data entries in BioMagResBank do not contain carbon beta chemical shifts and thus cannot be used for our simulation purposes. As a result, only 12 of them were included in our datasets, whose lengths range from 66 to 215.

We first introduce how we simulate the spin systems for the first two experiments. The following is the simulation procedure for generating the spin system containing H, N, C^{α}, and C^{β} chemical shifts. Other types of chemical shifts can be added in the same way. For each of these 12 proteins, we extracted its data entry from BioMagResBank to obtain all the chemical shift values for the amide proton H, the directly attached nitrogen N, the carbon alpha C^{α}, and the carbon beta C^{β}. For each amino acid residue, except proline and glycine, the four chemical shifts together with carbon alpha C^{α} and carbon beta C^{β} chemical shifts from the preceding residue formed the initial spin system. In the case of proline residues, we excluded them from the simulation because in the real NMR data, there would not be spin systems for prolines since there would not be HSQC peaks for them. Next, for each initial spin system, chemical shifts for intra-residue C^{α} and C^{β} were perturbed by adding to them randomized errors that follow independent normal distributions with 0 mean and constant standard deviations.

Next, we describe how we simulate the peak lists for the third and fourth experiments. The following is the simulation procedure for the *Perfect* HSQC, CBCA(CO)NH and HNCACB peak lists, which is also applied for generating other spectral peak lists. Given one data entry in BioMagResBank, we extracted all the chemical shift values for the amide proton H, the directly attached nitrogen N, the carbon alpha C^{α}, and the carbon beta C^{β}. For each amino acid residue, except proline, its H and N chemical shifts form a peak in HSQC peak list, its H and N chemical shifts with C^{α} and C^{β} chemical shifts from the preceding residue form two inter peaks respectively in CBCA(CO)NH peak list, and its H and N chemical shifts with its own C^{α} and C^{β} chemical shifts and with C^{α} and C^{β} chemical shifts from the preceding residue form two intra-residue peaks and two inter-residue peaks respectively in HNCACB peak list. For glycine, it has at most two inter-residue peaks and one intra-residue peak in the HNCACB spectrum since it does not have the C^{β} chemical shift. If the preceding residue is glycine, then only one inter-residue peak in the CBCA(CO)NH spectrum and at most two intra-residue peaks and one

inter-residue peak in the HNCACB spectrum are simulated. Many types of errors can be added into the perfect peak lists to make the simulated data closer to the reality. For example, we can simulate missing peaks by removing some peaks from peak lists, generate some false peaks in the peak lists as noise and simulate chemical shift divergence by including some measuring errors into each peak.

6.3.2 Experiment 1

In the first experiment, we applied the aforementioned procedure for spin system generation with the widely accepted tolerance thresholds for C^{α} and C^{β} chemical shifts, which were $\delta_{\alpha} = 0.2$ppm and $\delta_{\beta} = 0.4$ppm, respectively [78, 22, 6, 45]. Subsequently, the standard deviations of the normal distributions were set to $0.2/2.5 = 0.08$ppm and $0.4/2.5 = 0.16$ppm, respectively. These 12 instances, with suffix 1, are summarized in Table 6.1. In order to test the robustness of all three programs, we generated another set of 12 instances through doubling the tolerance thresholds (that is, $\delta_{\alpha} = 0.4$ppm and $\delta_{\beta} = 0.8$ppm). They, having suffix 2, are also summarized in Table 6.1. We use #CE to denote the number of correct edges (i.e. true edges) in the connectivity graph and #WE to denote the number of wrong edges. Both these two quantities tell to some extent how good the tolerance thresholds are. For every vertex in the graph, the number of edges coming out is called its *out-degree*. The average out-degree of the graph is defined to be the sum of the out-degrees over all the vertices (or equivalently, the number of edges in the graph) divided by the number of vertices. Such a notion of average out-degree (denoted as Avg.OD) captures the complexity (or the density) of the connectivity graph. Obviously, Table 6.1 tells that instances in the second set are much harder than the corresponding ones in the first set, where the complexity of an instance could be measured by the average out-degree of the vertices in the connectivity graph.

All four programs — RANDOM, PACES, MARS, and GASA — were called to run on both sets of instances. The performance results of RANDOM, PACES, MARS, and GASA on both sets of instances are collected in Table 6.2. Their assignment accuracies on two sets are also plotted in Figure 6.3.

In summary, RANDOM achieved on average 50% assignment accuracy (We followed the exact way of determining accuracy as described in [6], where 1000 itera-

Length	$\delta_\alpha = 0.2$ppm, $\delta_\beta = 0.4$ppm				$\delta_\alpha = 0.4$ppm, $\delta_\beta = 0.8$ppm			
	InstanceID	#CE	#WE	Avg.OD	InstanceID	#CE	#WE	Avg.OD
66	bmr4391.1	63	20	1.30	bmr4391.2	63	46	1.72
68	bmr4752.1	65	43	1.64	bmr4752.2	65	120	2.80
78	bmr4144.1	71	20	1.26	bmr4144.2	71	77	2.06
86	bmr4579.1	82	81	1.96	bmr4579.2	82	219	3.58
89	bmr4316.1	84	118	2.61	bmr4316.2	84	309	4.62
105	bmr4288.1	93	25	1.26	bmr4288.2	93	89	1.94
112	bmr4670.1	101	24	1.12	bmr4670.2	101	100	1.79
114	bmr4929.1	109	34	1.30	bmr4929.2	109	117	2.05
115	bmr4302.1	107	18	1.16	bmr4302.2	107	87	1.80
116	bmr4353.1	97	30	1.30	bmr4353.2	97	106	2.07
158	bmr4027.1	147	71	1.48	bmr4027.2	147	252	2.70
215	bmr4318.1	190	157	1.82	bmr4318.2	190	553	3.90

Table 6.1: 24 instances for the first experiment: 'Length' denotes the length of a protein, measured by the number of amino acid residues therein; '#CE' records the number of Correct Edges in the connectivity graph, which ideally should be equal to the number of available spin systems minus 1, and '#WE' records the number of Wrong Edges, respectively; 'Avg.OD' records the average Out-Degree of the connectivity graph.

tions for each instance have been run), which is roughly the same as that claimed in the original paper [6]. PACES performed better than RANDOM, but it failed on seven instances where the connectivity graphs were too complex (computer memory ran out, see Discussion for more information). The collected results for PACES on these seven instances were obtained through manually reducing the tolerance thresholds to remove a significant portion of edges from the connectivity graph. We implemented a scheme that if PACES did not finish an instance in 8 hours, then the tolerance thresholds would be reduced by 25%, for example, from $\delta_\alpha = 0.2$ppm to $\delta_\alpha = 0.15$ppm. Both GASA and MARS outperformed PACES and RANDOM in all instances, and even more significantly on the second set of more difficult instances, while GASA performs slightly better than MARS on two datasets.

Length	$\delta_\alpha = 0.2$ppm, $\delta_\beta = 0.4$ppm				
	InstanceID	RANDOM	PACES	MARS	GASA
66	bmr4391.1	0.63	0.72	0.87	0.97
68	bmr4752.1	0.35	0.79	0.97	0.94
78	bmr4144.1	0.33	0.53	0.97	0.99
86	bmr4579.1	0.51	0.62*	0.91	0.98
89	bmr4316.1	0.36	0.40*	0.96	0.99
105	bmr4288.1	0.55	0.71	0.95	0.98
112	bmr4670.1	0.62	0.77	0.88	0.95
114	bmr4929.1	0.63	0.86	0.97	0.91
115	bmr4302.1	0.64	0.73	0.92	0.95
116	bmr4353.1	0.43	0.79	0.85	0.95
158	bmr4027.1	0.32	0.82	0.93	0.99
215	bmr4318.1	0.38	0.54*	0.81	0.84
Avg.		0.48	0.69	0.90	0.95
Length	$\delta_\alpha = 0.4$ppm, $\delta_\beta = 0.8$ppm				
	InstanceID	RANDOM	PACES	MARS	GASA
66	bmr4391.2	0.55	0.69	0.85	0.91
68	bmr4752.2	0.30	0.74‡	0.90	0.88
78	bmr4144.2	0.31	0.38	0.97	0.99
86	bmr4579.2	0.32	0.43‡	0.75	0.80
89	bmr4316.2	0.30	0.18†	0.92	0.83
105	bmr4288.2	0.38	0.53	0.93	0.91
112	bmr4670.2	0.39	0.57	0.81	0.87
114	bmr4929.2	0.43	0.77	0.97	0.94
115	bmr4302.2	0.45	0.49	0.80	0.91
116	bmr4353.2	0.43	0.61	0.80	0.90
158	bmr4027.2	0.30	0.32	0.81	0.85
215	bmr4318.2	0.22	0.45‡	0.75	0.70
Avg.		0.37	0.51	0.85	0.87

Table 6.2: Assignment accuracies of RANDOM, PACES, MARS, and GASA in the first experiment. *PACES performance on these 3 datasets were obtained by reducing tolerance thresholds to $\delta_\alpha = 0.15$ppm and $\delta_\beta = 0.3$ppm (75%). †PACES performance on this dataset was obtained by reducing tolerance thresholds to $\delta_\alpha = 0.3$ppm and $\delta_\beta = 0.6$ppm (75%). ‡PACES performance on these 3 datasets were obtained by reducing tolerance thresholds to $\delta_\alpha = 0.2$ppm and $\delta_\beta = 0.4$ppm (50%).

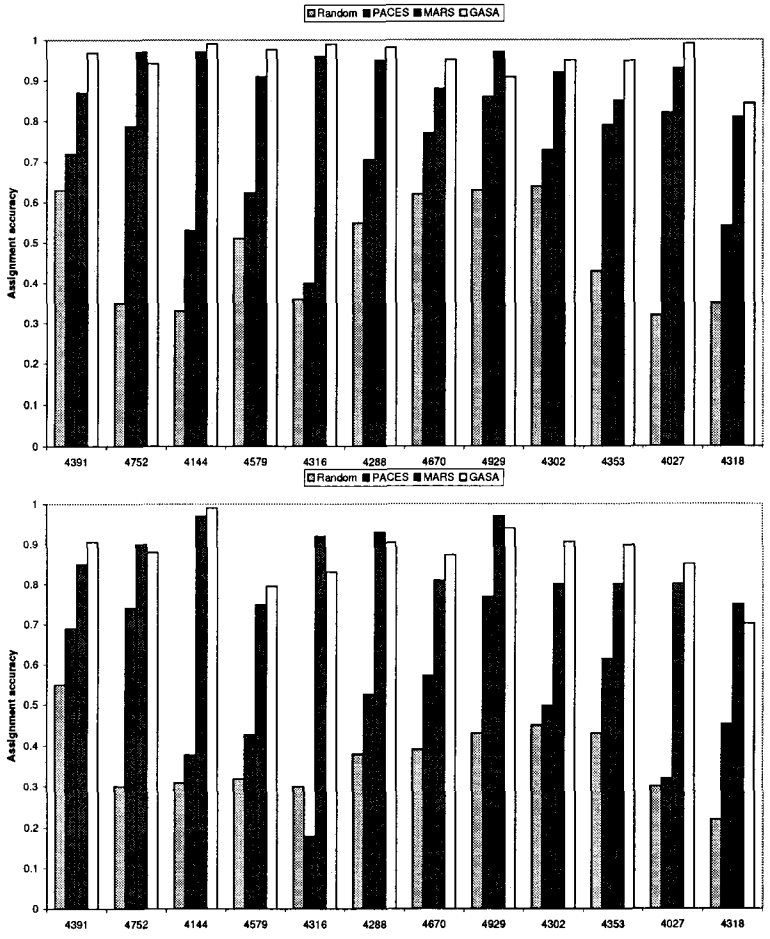


Figure 6.3: Plots of assignment accuracies for RANDOM, PACES, MARS, and GASA on two sets of instances with different tolerance thresholds, using C^{α} and C^{β} chemical shifts for connectivity inference.

6.3.3 Experiment 2

The instances used in the second experiment are for the same set of proteins used in the first experiment, excluding `bmr4391` and `bmr4316` because their data entries do not have carbonyl C chemical shifts. The experiment was designed to compare the performance among PACES, MARS, and GASA. Five chemical shifts, H, N, C^{α}, C^{β}, and carbonyl C, were included. The RANDOM is excluded in this experiment because it only processes C^{α} and C^{β} chemical shifts. As with the dataset generation in the first experiment, a spin system here included additionally the chemical shifts for the intra-residue carbonyl C and for the carbonyl C in the preceding residue. C^{α}, C^{β}, and carbonyl C chemical shift values were used to infer the connections. The tolerance threshold for carbonyl C chemical shift was set at $\delta = 0.15$ppm, and subsequently the standard deviation in the error distribution was set at $0.15/2.5 = 0.06$ppm. For the same reason as in the first experiment, we also generated another set of more difficult instances to test the robustness of both programs through doubling the tolerance thresholds. These two sets of 20 instances are summarized in Table 6.3.

Length	$\delta_{\alpha} = 0.2$ppm, $\delta_{\beta} = 0.4$ppm, $\delta = 0.15$ppm				$\delta_{\alpha} = 0.4$ppm, $\delta_{\beta} = 0.8$ppm, $\delta = 0.30$ppm			
	InstanceID	#CE	#WE	Avg.OD	InstanceID	#CE	#WE	Avg.OD
68	bmr4752.1	65	15	1.21	bmr4752.2	65	95	2.60
78	bmr4144.1	71	3	1.03	bmr4144.2	71	45	1.61
86	bmr4579.1	82	53	1.63	bmr4579.2	82	188	3.25
105	bmr4288.1	93	1	1.01	bmr4288.2	93	32	1.33
112	bmr4670.1	101	7	1.06	bmr4670.2	101	39	1.37
114	bmr4929.1	109	8	1.06	bmr4929.2	109	60	1.54
115	bmr4302.1	107	4	1.03	bmr4302.2	107	46	1.54
116	bmr4353.1	97	10	1.09	bmr4353.2	97	37	1.38
158	bmr4027.1	157	11	1.06	bmr4027.2	157	91	1.57
215	bmr4318.1	190	25	1.13	bmr4318.2	190	214	2.12

Table 6.3: 20 instances for the second experiment. For the meanings of the notations, refer to the caption for Table 6.1.

The performances of PACES, MARS, and GASA on both sets of instances are collected in Table 6.4. Their assignment accuracies on two sets are also plotted in Figure 6.4. In summary, GASA and MARS outperformed PACES significantly on both test sets.

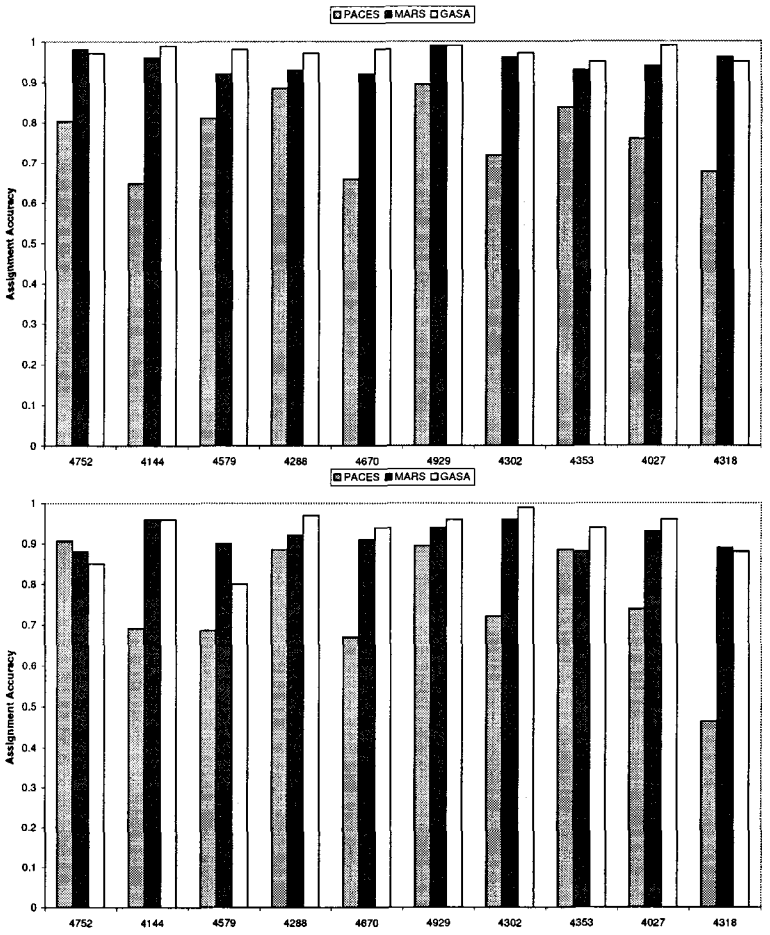


Figure 6.4: Plots of assignment accuracies for PACES, MARS and GASA on two sets of instances with different tolerance thresholds, using C^{α}, C^{β}, and carbonyl C chemical shifts for connectivity inference.

Length	$\delta_\alpha = 0.2$ppm, $\delta_\beta = 0.4$ppm, $\delta = 0.15$ppm				$\delta_\alpha = 0.4$ppm, $\delta_\beta = 0.8$ppm, $\delta = 0.30$ppm			
	InstanceID	PACES	MARS	GASA	InstanceID	PACES	MARS	GASA
68	bmr4752.1	0.78	0.98	0.97	bmr4752.2	0.88	0.88	0.85
78	bmr4144.1	0.60	0.96	0.99	bmr4144.2	0.64	0.96	0.96
89	bmr4579.1	0.78	0.92	0.98	bmr4579.2	0.66*	0.90	0.80
105	bmr4288.1	0.79	0.93	0.97	bmr4288.2	0.79	0.92	0.97
112	bmr4670.1	0.60	0.92	0.98	bmr4670.2	0.61	0.91	0.94
114	bmr4929.1	0.86	0.99	0.99	bmr4929.2	0.86	0.94	0.96
115	bmr4302.1	0.68	0.96	0.97	bmr4302.2	0.68	0.96	0.99
116	bmr4353.1	0.71	0.93	0.95	bmr4353.2	0.75	0.88	0.94
158	bmr4027.1	0.71	0.94	0.99	bmr4027.2	0.69	0.93	0.96
215	bmr4318.1	0.60	0.96	0.95	bmr4318.2	0.41*	0.89	0.87
Avg.		0.71	0.95	0.98		0.70	0.92	0.93

Table 6.4: Assignment accuracies of PACES, MARS and GASA in the second experiment. *PACES performance on these 2 datasets were obtained by reducing tolerance thresholds to $\delta_\alpha = 0.3$ppm, $\delta_\beta = 0.6$ppm, and $\delta = 0.225$ppm (75%).

6.3.4 Experiment 3

In RIBRA, 5 datasets were simulated from the data entries deposited in BioMagResBank. Among them, one is a **Perfect** dataset generated by using almost the same aforementioned simulation procedure, and the other four datasets reflect four different types of errors respectively. The first dataset, called **False positive**, is generated by respectively adding 5% fake carbon peaks into perfect CBCA(CO)NH and HNCACB peak lists. The second one, called **False negative**, is generated by randomly removing a small portion of inter-residue carbon peaks from perfect CBCA(CO)NH and HNCACB peak lists. The third one, called **Grouping error**, is generated by adding H, N, C^α and C^β measuring errors into inter-residue peaks in the perfect CBCA(CO)NH peak list. The fourth one, called **Linking error**, is generated by adding C^α and C^β measuring errors into inter-residue peaks in the perfect HNCACB peak list.

Table 6.5 presents the performances of RIBRA and GASA on these 5 datasets. As shown, there is no significant difference among the performances on the **Perfect**, **False positive** and **Link error** datasets. GASA shows more robustness on the **False negative** dataset with missing data while RIBRA performs better on the **Grouping error** dataset. Through a detailed investigation, we found that these 5 datasets contain the C^β inter-residue and intra-residue peaks with 0 C^β chemical

Dataset	RIBRA		GASA	
	Precision	Recall	Precision	Recall
Perfect	0.98	0.92	0.98	0.93
False positive	0.98	0.92	0.97	0.92
False negative	0.96	0.77	0.96	0.89
Grouping error	0.98	0.89	0.91	0.81
Linking error	0.96	0.89	0.96	0.90
Average	0.97	0.88	0.96	0.89

Table 6.5: Comparison results for RIBRA and GASA in experiment 2.

shifts for glycine, indicating that in the RIBRA simulation, glycine would have two inter-residue peaks and two intra-residue peaks in the HNCACB spectrum and the amino acid residues with the preceding glycine would have two inter-residue peaks in the CBCA(CO)NH spectrum. However, this is not true in real NMR spectral data. In fact, a large number of ambiguities in the sequential assignment result from glycine because it produces various legal combinations in the peak grouping thus making the identification of perfect spin systems even harder. For example, the spin systems containing 3,4 and 5 peaks have the same chance to be perfect spin systems as those containing 6 peaks and meanwhile they could be considered as the spin systems with missing peaks. Thus the peak grouping is much easier on the dataset with the simulated C^{β} peaks for glycine. Since the GASA algorithm is designed to deal with the real spectral data, we deleted the peaks with 0 carbon chemical shifts. This is why our performance on the **Grouping error** dataset is not as good as RIBRA. To verify our hypothesis, we randomly selected 14 proteins with length ranging from 69 to 186 in the **Grouping error** dataset, and removed all the peaks with 0 C^{β} chemical shift. Both RIBRA and GASA were tested on them. RIBRA achieved 0.88 precision and 0.73 recall, and GASA achieved 0.89 precision and 0.79 recall (See Table 6.6). One could argue that the C^{β} peaks with 0 chemical shifts for glycine can be artificially simulated in real NMR spectral data by using glycine's expected H and N chemical shifts, since the primary protein sequence is known. However, the large ranges of H and N chemical shifts for glycine would make the simulated C^{β} peaks be processed as fake peaks in many cases. Therefore, we think the C^{β} peak for glycine should not be generated in these simulations. Another weakness in the RIBRA simulation is that in the construction of **Grouping error**

BMRB Entry	Len	Missing	RIBRA		GASA	
			Precision	Recall	Precision	recall
4579	86	4	0.83	0.65	0.90	0.82
4688	111	9	0.71	0.45	0.89	0.77
4790	118	28	0.78	0.63	0.96	0.74
4898	86	4	0.86	0.68	0.92	0.82
4938	132	4	0.87	0.71	0.85	0.77
4954	97	3	0.97	0.82	0.86	0.82
4984	151	7	0.85	0.65	0.85	0.75
5003	112	7	0.96	0.66	0.88	0.79
5107	101	6	0.86	0.83	0.81	0.72
5130	130	3	0.98	0.88	0.91	0.85
5148	98	22	0.92	0.91	0.96	0.85
5168	69	3	0.93	0.82	0.91	0.89
5272	186	39	0.81	0.64	0.85	0.74
5337	111	21	0.95	0.84	0.91	0.80
Avg			0.88	0.73	0.89	0.79

Table 6.6: Comparison results for RIBRA and GASA on 14 proteins without C^{β} peaks for glycine.

datasets, RIBRA kept the perfect HSQC and HNCACB peak lists untouched and only added some measuring errors into the inter-residue peaks in the CBCA(CO)NH peak list. This simulation looks a bit far from the reality because the chemical shifts deposited in BioMagResBank have been manually adjusted. Even though the HSQC spectrum is a very reliable experiment, the deposited H and N chemical shifts in BioMagResBank are still slightly different from the measured values in a real HSQC spectrum. We believe that to simulate a real NMR spectral dataset, perturbing chemical shifts in all perfect peak lists is necessary. In Experiment 4, we present our simulation and the corresponding comparison results with RIBRA.

6.3.5 Experiment 4

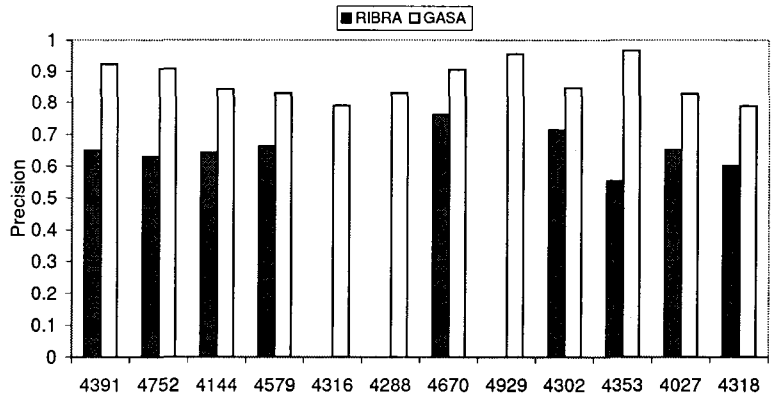
The purpose of Experiment 4 is to provide more convincing results based on a better simulation. Again, we still used the same dataset in the first two experiments to conduct the simulation. To make fair comparison with RIBRA, we simulated only three peak lists, HSQC, CBCA(CO)NH and HNCACB from each protein in the dataset, although our program can deal with many other combinations. For each of

these 12 proteins, we first build the perfect HSQC, CBCA(CO)NH and HNCACB by using the general simulation procedure mentioned above. There is no C^{β} peak for glycine in the CBCA(CO)NH and HNCACB spectrum. For each peak in the HSQC, CBCA(CO)NH and HNCACB spectrums, the contained H, N, C^{α} or C^{β} chemical shifts were perturbed by adding to them randomized errors that follow independent normal distributions with 0 means and constant standard deviations. We chose the same tolerance thresholds as those used in RIBRA, which were $\delta_{\mathrm{H}} =$ 0.06ppm for H, $\delta_{\mathrm{N}} = 0.8$ppm for N, $\delta_{\alpha} = 0.2$ppm for C^{α} and $\delta_{\beta} = 0.4$ppm for C^{β}, respectively. Subsequently, the standard deviations of the normal distributions were set to $0.06/2.5 = 0.0024$ppm, $0.8/2.5 = 0.32$ppm $0.2/2.5 = 0.08$ppm and $0.4/2.5 = 0.16$ppm, respectively.

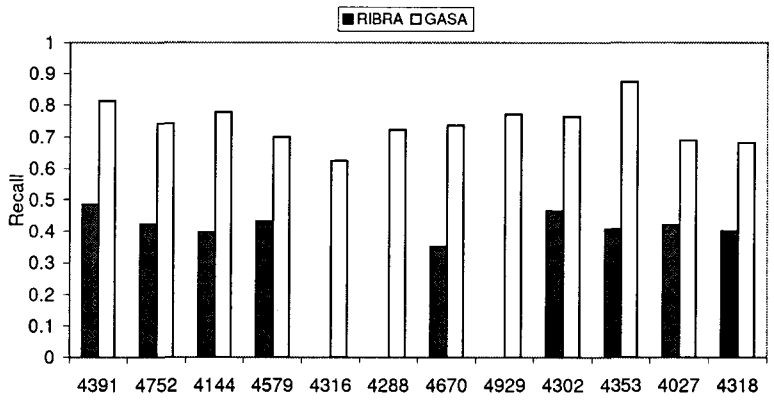
The comparison results of the second experiment on these 12 proteins are summarized in Table 6.7. The precision and recall are also plotted in Figure 6.5. In summary, GASA outperformed RIBRA in all instances while RIBRA failed to solve three instances, which are `bmr4316`, `bmr4288` and `bmr4929`. As shown in Table 6.7, RIBRA only achieved 0.65 precision and 0.42 recall on average, which are noticeably worse than what it claimed in [73], while GASA achieved 0.87 precision and 0.74 recall. The possible explanations could be (1) the simulation procedure in this experiment did not generate the C^{β} peaks with 0 carbon chemical shift for glycines, which causes more ambiguities in the peak grouping. (2) In our simulated dataset of Experiment 4, the chemical shifts in all perfect peak lists were perturbed with random measuring errors, which generated more uncertainties in all operations in the sequential assignment.

6.4 Summary

Peak grouping is one of the three stages in the automated procedure of the NMR sequential assignment. Though the quality and quantity of spin systems produced in the peak grouping have the most significant effect on the assignment, there has been surprisingly little work done to improve the precision of peak grouping. This chapter addresses the hard issues in the peak grouping, such as how to eliminate the sensitivity to the pre-specified tolerance thresholds. We developed a novel two-stage graph-based algorithm, called **GASA** and evaluated its performance in four ex-



(a)



(b)

Figure 6.5: Plots of precision (a) and recall (b) for RIBRA and GASA in Experiment 4.

BMRB Entry	Len	Missing	RIBRA			GASA		
			Grouped	Precision	Recall	Grouped	Precision	recall
4391	66	7	44	0.65	0.49	52	0.92	0.81
4752	68	2	44	0.63	0.42	54	0.91	0.74
4144	78	10	42	0.64	0.40	63	0.84	0.78
4579	86	3	54	0.66	0.43	70	0.83	0.70
4316	89	4	N/A	N/A	N/A	67	0.79	0.62
4288	105	9	N/A	N/A	N/A	84	0.83	0.72
4670	112	10	47	0.76	0.35	83	0.90	0.74
4929	114	4	N/A	N/A	N/A	89	0.96	0.77
4302	115	8	70	0.71	0.47	97	0.85	0.77
4353	116	18	72	0.55	0.41	89	0.97	0.87
4027	158	10	96	0.65	0.42	123	0.83	0.69
4318	215	24	127	0.60	0.40	165	0.79	0.68
Avg				0.65	0.42		0.87	0.74

Table 6.7: Comparison results for RIBRA and GASA in Experiment 4.

periments. In the first two experiments, GASA outperformed RANDOM, PACES, and MARS, which indicates that combining the chaining and assignment together would efficiently resolve the ambiguities and then make a better assignment. The third experiment was conducted on the datasets released by RIBRA. Our program performed as well as RIBRA on the Perfect, False positive and Link error datasets. GASA showed more robustness on the False negative dataset with missing data, while RIBRA was good at handling the Grouping error dataset. To provide more convincing results, we provided a better simulation in the fourth experiment, which was much closer to the reality. We found strong improvements in all instances compared to RIBRA. The performance comparisons with RANDOM, PACES, MARS, and RIBRA demonstrated the fact that GASA might be more promising for practical use.

Chapter 7

Conclusions and Future Work

This thesis describes our research on automated sequential resonance assignment for NMR protein structure determination. We believe several possible improvements could be possible and may form the basis for future research.

7.1 Conclusions

It is well known that NMR sequential resonance assignment is a critical process in protein NMR structure determination. The precision of resonance assignment has a significant effect on the accuracy of protein structure calculation. In this thesis, we have reviewed the literature in NMR resonance assignment and conducted a thorough analysis on computational issues not fully resolved in NMR sequential resonance assignment. We have also developed some generic models to tackle these issues respectively, which are listed below.

Peak Grouping

Peak grouping takes as input the peak lists extracted from multi-dimensional NMR spectra, and outputs the spin systems that contain the chemical shifts for atoms from the common residues. Many existing methods neglected this process in which the only available computational model is based on a binary-decision process. However, in reality, the quality of the peak lists is not sufficient to make the peak grouping a trivial task, and the simple binary-decision model is ineffective in producing the spin systems of high quality for most cases except for high resolution NMR spectra. We reported that the quality and quantity of spin systems produced in the peak grouping have the most significant effect on the sequential assignment, and the peak grouping is the most important stage throughout the whole process that is worthy of more attention. We have developed a novel two-stage graph-based algorithm, called GASA, which outperformed the latest work RIBRA. The performance comparisons with RIBRA demonstrated that GASA could be more promising for practical use.

Connectivity Determination

Given a set of spin systems, the task of connectivity determination is to extract the pair-wise relationships between spin systems, which constrain that some

pairs of spin systems should be assigned to consecutive residues in the target protein. We have designed a best-first search algorithm, called CISA, based on the novel heuristics to perform the connectivity determination. This algorithm improves the assignment accuracy significantly compared to two most recently proposed sequential resonance assignment programs, RANDOM and PACES.

String Assignment

The assignment process of identified spin systems with connectivity information has been formulated as a constrained weighted bipartite matching problem between strings of spin systems and a sequence of amino acids with predicted secondary structures. This problem is NP-hard. We have developed an integer programming solver without the sacrifice of time efficiency, which can compute the highly confident assignment within seconds.

Scoring Scheme

Accurately quantifying the signature information of chemical shifts provides a foundation for accurate and complete sequential resonance assignment in protein NMR spectroscopy. Most studies assume that the chemical shift follows a normal distribution and use the normal density functions to derive the likelihood that weighs the mapping of a spin system to an amino acid, which is not necessarily true based on our experimental results. We have designed a statistics based scoring scheme by using Bayesian learning. Extensive simulation studies have been conducted to validate the different scoring schemes, and the one with the best performance has been implemented on a public web server.

The experimental results revealed that our models outperform existing methods on a number of simulated datasets and have the potential to automate NMR sequential resonance assignment.

7.2 Future Work

We believe our research to date has made a significant contribution to the field of NMR protein structure determination. The models and algorithms that we developed have been proved to outperform most recent methods, although it has not

fully satisfied the expectation of NMR researchers. I foresee a greater improvement will be made if we extract more knowledge from NMR data. Future attempts are outlined below.

Many uncertainties should be taken into account in the scoring scheme, such as the accuracy of predicted secondary structure. Combining evidence from other approaches properly will provide a more accurate estimation. Using advanced learning models, such as Bayesian networks, may improve our current scoring scheme.

For connectivity determination, our heuristic algorithm CISA has successfully combined the spin system signature information into the path growing process in the connectivity graph, which prunes the search space more effectively than PACES [22]. However, in the current version of CISA the weights of edges are used only to order the child paths. Taking the idea from RANDOM [47] that uses edge weights as edge selection probabilities, we believe that some better usage of edge weights into the mapping score evaluation for a growing path would help to quantify the quality of the growing path more effectively. We have tried some simple linear functions on the edge weights and the mapping scores of paths that turned out not to serve satisfactorily. We are currently investigating more combinations.

A possible disadvantage of the current version of GASA is that wrong edges included during the OL initialization might continue to stay in and thus would lead to erroneous final assignments. Although this is very unlikely to happen according to our extensive simulation studies, we feel that some mechanism might need to be set up to shuffle low mapping score paths that would be considered once every a few iterations during the path growing step.

The last but not the least, we will extend our work by including structure calculation, since the protein structure is the final target of NMR sequential resonance assignment. We have realized the importance of the inseparability of NMR sequential resonance assignment in this dissertation. We believe that incorporating the structure calculation into NMR sequential resonance assign-

ment could be a worthwhile approach to protein structure determination via NMR.

Bibliography

[1] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[2] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering Points To Identify the Clustering Structure. In *Proceedings of ACM SIGMOD'99 International Conference on Management of Data*, pages 49–60, 1999.

[3] C. Antz, K. P. Neidig, and H. R. Kalbitzer. A general Bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis. *Journal of Biomolecular NMR*, 5:287–296, 1995.

[4] V. I. Arnold. *Mathematical methods of classical mechanics.* Springer, 1978.

[5] H. S. Atreya, S. C. Sahu, K. V. R. Chary, and G. Govil. A tracked approach for automated NMR assignments in proteins (TATAPRO). *Journal of Biomolecular NMR*, 17:125–136, 2000.

[6] C. Bailey-Kellogg, S. Chainraj, and G. Pandurangan. A random graph approach to NMR sequential assignment. In *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2004)*, pages 58–67, 2004.

[7] C. Bailey-Kellogg, A. Widge, J. J. K. III, M. J. Berardi, J. H. Bushweller, and B. R. Donald. The NOESY Jigsaw: automated protein secondary structure and main-main assignment from sparse, unassigned NMR data. *Journal of Computational Biology*, 7:537–558, 2000.

[8] M. C. Baran, Y. J. Huang, H. N. Moseley, and G. T. Montelione. Automated analysis of protein NMR assignments and structures. *Chemical Reviews*, 104:3541–3556, 2004.

[9] C. Bartels, M. Billeter, P. Güntert, and K. Wüthrich. Automated sequence-specific assignment of homologous proteins using the program GARANT. *Journal of Biomolecular NMR*, 7:207–213, 1996.

[10] C. Bartels, P. Güntert, M. Billeter, and K. Wüthrich. GARANT – A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *Journal of Computational Chemistry*, 18:139–149, 1997.

[11] E. D. Becker. *High Resolution NMR: Theory and Chemical Applications.* Academic Press, 2000.

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

[13] BioMagResBank. `http://www.bmrb.wisc.edu`. University of Wisconsin. Madison, Wisconsin.

[14] R. Boppana and M.M. Halldorsson. Approximating Maximum Independent Sets by Excluding Subgraphs. *BIT.*, 32:180-196, 1992.

[15] W. Braun, C. Bosch, L. R. Brown, N. Go, and K. Wüthrich. Combined use of proton-proton overhauser enhancements and a distance geometry algorithm for determination of polypeptide conformations. *Biochimica et Biophysics Acta*, 667:377–396, 1981.

[16] W. Braun and N. Go. Calculation of protein conformations by proton-proton distance constraints. *Journal of Molecular Biology*, 186:611–626, 1985.

[17] A. T. Brunger. *X-PLOR, Version 3.1. A system for X-ray Crystallography and NMR.* Yale University Press, 1992.

[18] N. E. G. Buchler, E. P. R. Zuiderweg, H. Wang, and R. A. Goldstein. Protein heteronuclear NMR assignments using mean-field simulated annealing. *Journal of Magnetic Resonance*, 125:34–42, 1997.

[19] E. A. Carrara, F. Pagliari, and C. Nicolini. Neural networks for the peak-picking of nuclear magnetic resonance spectra. *Neural Networks*, 6:1023–1032, 1993.

[20] Z.-Z. Chen, T. Jiang, G.-H. Lin, J. J. Wen, D. Xu, and Y. Xu. Improved approximation algorithms for NMR spectral peak assignment. In *Proceedings of the 2nd Workshop on Algorithms in Bioinformatics (WABI 2002)*, volume 2452 of *Lecture Notes in Computer Science*, pages 82–96. Springer, 2002.

[21] Z.-Z. Chen, T. Jiang, G.-H. Lin, J. J. Wen, D. Xu, J. Xu, and Y. Xu. Approximation algorithms for NMR spectral peak assignment. *Theoretical Computer Science*, 299:211–229, 2003.

[22] B. E. Coggins and P. Zhou. PACES: Protein sequential assignment by computer-assisted exhaustive search. *Journal of Biomolecular NMR*, 26:93–111, 2003.

[23] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. Introduction to Algorithms McGraw Hill, 2001

[24] S. A. Corne and P. Johnson. An artificial neural network of classifying cross peaks in two-dimensional NMR spectra. *Journal of Magnetic Resonance*, 100:256–266, 1992.

[25] G. Cornilescu, F. Delaglio, and A. Bax. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *Journal of Biomolecular NMR*, 13:289–302, 1999.

[26] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[27] J. L. Devore. *Probability and Statistics for Engineering and the Science.* Duxbury Press, December 1999. Fifth Edition.

[28] A. E. Ferentz and G. Wagner. NMR spectroscopy: a multifaceted approach to macromolecular structure. *Quarterly Review Biophysics*, 33:29–65, 2000.

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

[30] A. V. Goldberg and R. Kennedy. An efficient cost scaling algorithm for the assignment problem. *Mathematical Programming*, 71:153–178, 1995.

[31] W. Gronwald and H. R. Kalbitzer. Automated structure determination of proteins by NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 44:33–96, 2004.

[32] P. Güntert. Structure calculation of biological macromolecules from NMR data. *Quarterly Reviews of Biophysics*, 31:145–237, 1998.

[33] W. F. van Gunsteren and H. J. C. Berendsen. Computer simulation of molecular dynamics: methodology, applications and perspectives in chemistry. *Angewandte Chemie International edition*, 29:992–1023, 1990.

[34] P. Güntert. Automated NMR protein structure calculation. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 43:105–125, 2003.

[35] P. Güntert, W. Braun, and K. Wüthrich. Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *Journal of Molecular Biology*, 217:517–530, 1991.

[36] P. Güntert, C. Mumenthaler, and K. Wüthrich. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *Journal of Molecular Biology*, 273:283–298, 1997.

[37] P. Güntert, Y. Q. Qian, G. Otting, M. Muller, W. Gehring, and K. Wüthrich. Structure determination of the Antp ($C_{39} \rightarrow S$) homeodomain from nuclear magnetic resonance data in solution using a novel strategy for the structure calculation with the programs DIANA, CALIBA, HABAS and GLOMSA. *Journal of Molecular Biology*, 217:531–540, 1991.

[38] P. Güntert, M. Salzmann, D. Braun, and K. Wüthrich. Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. *Journal of Biomolecular NMR*, 18:129–137, 2000.

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

[40] T. F. Havel and K. Wüthrich. A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular ^{1}H-^{1}H proximities in solution. *Bulletin of Mathematical Biology*, 46:673–698, 1984.

[41] T. Herrmann, P. Güntert, and K. Wüthrich. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *Journal of Molecular Biology*, 319:209–227, 2002.

[42] T. K. Hitchens, J. A. Lukin, Y. Zhan, S. A. McCallum, and G. S. Rule. MONTE: An automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. *Journal of Biomolecular NMR*, 25:1–9, 2003.

[43] B. Johnson and R.A. Blevins. NMRView: A computer program for the visualization and analysis of NMR data. *J. Biomol. NMR* , 4: 603614, 1994.

[44] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292:195–202, 1999.

[45] Y.-S. Jung and M. Zweckstetter. Mars – robust automatic backbone assignment of proteins. *Journal of Biomolecular NMR*, 30:11–23, 2004.

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

[47] C. B. Kellogg, S. Chainraj, and G. Pandurangan. A random graph approach to NMR sequential assignment. In *RECOMB'04*, pages 58–67, 2004.

[48] G. J. Kleywegt, R. Boelens, and R. Kaptein. A versatile approach toward the partially automatic recognition of cross peaks in 2D ^{1}H NMR spectra. *Journal of Magnetic Resonance*, 88:601–608, 1990.

[49] R. Koradi, M. Billeter, M. Engeli, P. Güntert, and K. Wüthrich. Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *Journal of Magnetic Resonance*, 135:288–297, 1998.

[50] M. Leutner, R. M. Gschwind, J. Liermann, C. Schwarz, G. Gemmecker, and H. Kessler. Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *Journal of Biomolecular NMR*, 11:31–43, 1998.

[51] K. B. Li and B. C. Sanctuary. Automated resonance assignment of protein using heteronuclear 3D NMR. 1. Backbone spin systems extraction and creation of polypeptides. *Journal of Chemical Information and Computational Science*, 37:359–366, 1997.

[52] G.-H. Lin, D. Xu, Z. Z. Chen, T. Jiang, J. J. Wen, and Y. Xu. An efficient branch-and-bound algorithm for the assignment of protein backbone NMR peaks. In *Proceedings of the First IEEE Computer Society Bioinformatics Conference (CSB 2002)*, pages 165–174, 2002.

[53] G.-H. Lin, J. Wagner, and X. Wan. Score: A web server for scoring spin systems in protein NMR spectroscopy. *Bioinformatics*, 2005. Submitted.

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

[55] D. Malmodin, C. H. Papavoine, and M. Billeter. Fully automated sequence-specific resonance assignments of hetero-nuclear protein spectra. *Journal of Biomolecular NMR*, 27:69–79, 2003.

[56] J. L. Markley, A. Bax, Y. Arata, C. W. Hilbers, R. Kaptein, B. D. Sykes, P. E. Wright, and K. Wüthrich. Recommendations for the presentation of NMR structures of proteins and nucleic acids. *Journal of Molecular Biology*, 280:933–952, 1998.

[57] L. J. McGuffin, K. Bryson, and D. T. Jones. The PSIPRED protein structure prediction server. *Bioinformatics*, 16:404–405, 2000.

[58] H. N. B. Moseley and G. T. Montelione. Automated analysis of NMR assignments and structures for proteins. *Current Opinion in Structural Biology*, 9:635–642, 1999.

[59] M. Nilges, G. M. Clore, and A. M. Gronenborn. Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Letters*, 229:317–324, 1988.

[60] *Pilot Projects for the Protein Structure Initiative (Structural Genomics).* National Institute of General Medical Sciences, Washington, D.C., 1999. `http://www.nih.gov/grants/guide/rfa-files/RFA-GM-99-009.html`.

[61] A. Rouh, A. Louis-Joseph, and J. Y. Lallemand. Bayesian signal extraction from noisy FT NMR spectra. *Journal of Biomolecular NMR*, 4:505–518, 1994.

[62] C. M. Slupsky, R. F. Boyko, V. K. Booth, and B. D. Sykes. SMARTNOTEBOOK: a semi-automated approach to protein sequential NMR resonance assignments. *Journal of Biomolecular NMR*, 27:313–321, 2003.

[63] *Virtual Textbook of Organic Chemistry* . William Reusch, Michigan State University, 1999 `http://www.cem.msu.edu/~reusch/VirtTxtJml/Spectrpy/nmr/nmr1.htm`.

[64] T. A. Tatusova and T. L. Madden. Blast 2 sequences - a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters*, 174:247–250, 1999.

[65] T. Tegos, Z.-Z. Chen, and G.-H. Lin. Heuristic Search in Constrained Bipartite Matching with Applications to Protein NMR Peak Assignment. Journal of Bioinformatics and Computational Biology, 2005. In Press.

[66] N. Tjandra, J. G. Omichinski, A. M. Gronenborn, G. M. Clore, and A. Bax. Use of dipolar ^{1}H-^{15}N and ^{1}H-^{13}C couplings in the structure determination of magnetically oriented macromolecules in solution. *Nature Structural Biology*, 4:732–738, 1997.

[67] X. Wan and G.-H. Lin. CISA: Combined NMR Resonance Connectivity Information Determination and Sequential Assignment. *Manuscript* Preparing for Submission.

[68] X. Wan, T. Tegos, and G.-H. Lin. Histogram-based scoring schemes for protein NMR resonance assignment. *Journal of Bioinformatics and Computational Biology*, 2(4):747-764, 2004.

[69] X. Wan, D. Xu, C.M. Slupsky, and G.-H. Lin. Automated protein NMR resonance assignments. In *Proceedings of the Second IEEE Computer Society Bioinformatics Conference (CSB 2003)*, pages 197–208, 2003.

[70] M.P. Williamson, T.F. Havel, and K. Wüthrich. Solution conformation and proteinase inhibitor IIA from bull seminal plasma by proton NMR and distance geometry. *Journal of Molecular Biology*, 182:295–315, 1985.

[71] D. S. Wishart and B. D. Sykes. The ^{13}C Chemical-Shift Index: A simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *Journal of Biomolecular NMR*, 4:171–180, 1994.

[72] D. S. Wishart, B. D. Sykes, and F. M. Richards. The Chemical Shift Index: A fast and simple method of the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, 31:1647–1651, 1992.

[73] K.-P. Wu, J.-M. Chang, J.-B. Chen, C.-F. Chang, W.-J. Wu, T.-H. Huang, T.-Y. Sung, and W.-L. Hsu. RIBRA – an error-tolerant algorithm for the NMR backbone assignment problem. In *Proceedings of the 9th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2005)*, pages 103–117, 2005.

[74] K. Wüthrich. *NMR of Proteins and Nucleic Acids.* Wiley, John & Sons, New York, 1986.

[75] K. Wüthrich, M. Billeter, and W. Braun. Polypeptide secondary structure determination by nuclear magnetic resonance observation of short proton-proton distances. *Journal of Molecular Biology*, 180:715–740, 1984.

[76] Y. Xu, D. Xu, D. Kim, V. Olman, J. Razumovskaya, and T. Jiang. Automated assignment of backbone NMR peaks using constrained bipartite matching. *IEEE Computing in Science & Engineering*, 4:50–62, 2002.

[77] H. Y. Zhang , S. Neal and D. S. Wishart, RefDB: A database of uniformly referenced protein chemical shifts. *Journal of Biomolecular NMR*, 25:173–195, 2003.

[78] D. E. Zimmerman, C. A. Kulikowski, Y. Huang, W. F. M. Tashiro, S. Shimotakahara, C. Chien, R. Powers, and G. T. Montelione. Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of Molecular Biology*, 269:592–610, 1997.