Fast and Robust Mathematical Modeling of NMR Assignment Problems

by

Richard Jang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2012

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

NMR spectroscopy is not only for protein structure determination, but also for drug screening and studies of dynamics and interactions. In both cases, one of the main bottleneck steps is backbone assignment. When a homologous structure is available, it can accelerate assignment. Such structure-based methods are the focus of this thesis. This thesis aims for fast and robust methods for NMR assignment problems; in particular, structure-based backbone assignment and chemical shift mapping. For speed, we identified situations where the number of ^{15}N-labeled experiments for structure-based assignment can be reduced; in particular, when a homologous assignment or chemical shift mapping information is available. For robustness, we modeled and directly addressed the errors. Binary integer linear programming, a well-studied method in operations research, was used to model the problems and provide practically efficient solutions with optimality guarantees.

Our approach improved on the most robust method for structure-based backbone assignment on ^{15}N-labeled data by improving the accuracy by 10% on average on 9 proteins, and then by handling typing errors, which had previously been ignored. We showed that such errors can have a large impact on the accuracy; decreasing the accuracy from 95% or greater to between 40% and 75%. On automatically picked peaks, which is much noisier than manually picked peaks, we achieved an accuracy of 97% on ubiquitin.

In chemical shift mapping, the peak tracking is often done manually because the problem is inherently visual. We developed a computer vision approach for tracking the peak movements with average accuracy of over 95% on three proteins with less than 1.5 residues predicted per peak. One of the proteins tested is larger than any tested by existing automated methods, and it has more titration peak lists. We then combined peak tracking with backbone assignment to take into account contact information, which resulted in an average accuracy of 94% on one-to-one assignments for these three proteins. Finally, we applied peak tracking and backbone assignment to protein-ligand docking to illustrate the potential for fast 3D complex determination.

Acknowledgements

I would like to thank my parents and grandparents for their patience and support; my brother for keeping me grounded; my supervisor for not giving up on me; all the friends I've made at UW; my friends back home in Vancouver; and a special thanks to the WDSC members GK, HC, XQ, IK, and JC. I would like to thank Babak Alipanahi, Thorsten Dieckmann, Yay Duangkham, Xin Gao, Guy Guillemette, and Mike Piazza for thoughtful discussions; and Xiong et al. for providing me with their program and the test data for 5 proteins. This work is partially supported by NSERC Grant OGP0046506, China's MOST 863 Grant 2008AA02Z313, Canada Research Chair program, MITACS, an NSERC Collaborative Grant, Premier's Discovery Award, SHARCNET, the Cheriton Scholarship, and a grant from King Adbullah University of Science and Technology. The eNMR project (European FP7 e-Infrastructure grant, contract no. 213010, www.enmr.eu), supported by the national GRID Initiatives of Italy, Germany and the Dutch BiG Grid project (Netherlands Organization for Scientific Research), is acknowledged for the use of web portals, computing and storage facilities.

Dedication

To my loving parents whose sacrifices have given me the best opportunities.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The majority of the data analysis in nuclear magnetic resonance (NMR) spectroscopy is connecting the NMR signals to their underlying chemical structures. This is the topic of this thesis. We begin by motivating the use of NMR for studying proteins.

1.1 X-ray vs. NMR

X-ray crystallography and NMR are the predominant methods for protein 3D structure determination. Both provide an atomic view of the protein, but X-ray has been more productive in high-throughput structural genomics. As of June 1, 2011, there are over 8.6 times more x-ray structures than NMR in the protein databank (PDB). Structure determination with x-ray is rather routine and automated with robots and software [21]. This is due to generally accepted standards in experimental and quality control protocols. The SouthEast Collaboratory for Structural Genomics was able to get 85% of the 171 residue structure with PDB:1NNQ within 4.5 hours starting from mounted crystal [2]. However, the time to grow protein crystals is the rate determining step, and this can take months or even fail. Due to failed experiments, making well-diffracting crystals is more expensive than diamonds either by weight or volume [38].

In contrast, it is difficult to automate NMR with robots due to lack of standards. There exists a wide array of experiments and different combinations of such experiments. A list of the different experiment sets used by the Northeast Structural Genomics consortium can be found on their wiki site [86]. In addition, the data analysis that follows can take significantly longer than the experiments themselves due to manual analysis [27]. Although there exist various software to automate the analysis, manual analysis is still the most reliable method. Even when automated tools are used, manual verification is almost always done due to issues with trusting automation to handle the wide array of possible errors.

Unlike x-ray, NMR is limited by protein size, but this limit is constantly increasing [125]. Larger proteins typically result in more missing and overlapping signals; that is, poor sensitivity and resolution. 97% of NMR structures in the PDB are under 200 residues. One of the largest NMR structures is malate synthase G (PDB:1Y8B) at 82 kDa and 731 residues. In contrast, x-ray structures of the ribosome can be over 2000 kDa. Nevertheless, for proteins that do not form well-diffracting crystals, NMR is the only alternative for atomic resolution structures. For example, flexible regions are often absent in x-ray structures, but present for those determined by NMR. To account for variability, NMR PDB files typically contain multiple possible models consistent with the experimental data. Membrane proteins are a challenge for both methods [20, 79].

Although x-ray has produced more structures, there remains much more to investigate besides structure [19], just like there is much more to investigate once a genome has been sequenced. Protein function is an active area of research, and it will likely become more active given that a plateau in the discovery of unique folds has been reached (Figure 1.1), while the number of structures is still increasing. The holy grail is to obtain a video at atomic resolution of the protein folding pathway starting from translation in the ribosome, and then following the changes to the protein as it interacts with its environment. This video would provide information on the protein's allostery, dynamics, interactions, pathways, and stability under the conditions studied, such as a disease state or the presence of drug molecules. So far, this video can be obtained only through molecular dynamics simulations, which is only a simulation of reality. Molecular dynamics requires knowledge of the protein structure, and it is feasible for only small systems and short time scales due to the prohibitive computational costs. Protein folding occurs at a time scale that molecular

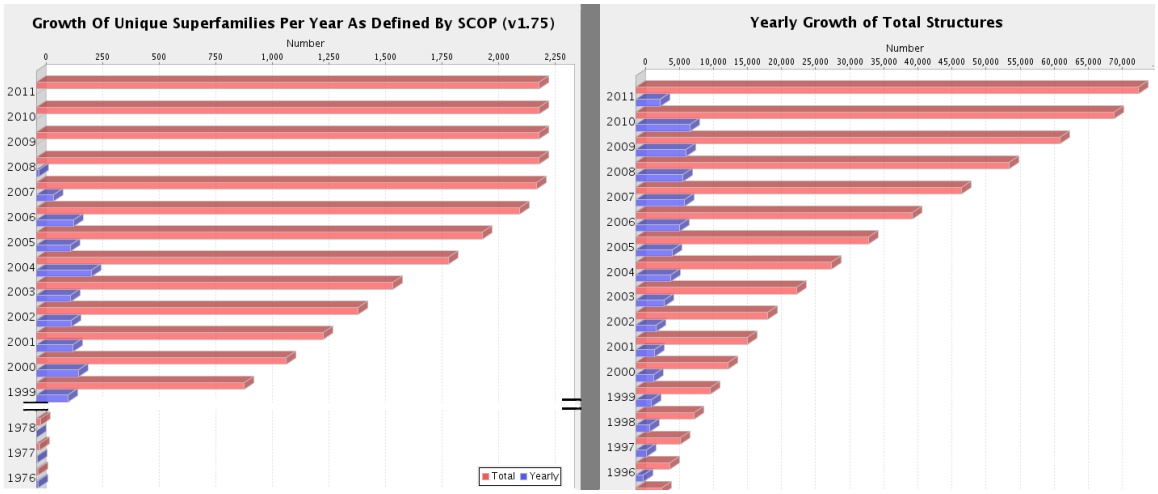

Figure 1.1: Left: Growth of new unique folds in the PDB according to the SCOP hierarchical classification database [83]. The number of folds in the superfamily group is displayed, which appears to have stopped increasing by 2009. The top level of the SCOP hierarchy is the fold group consisting of proteins with the same secondary structure and arrangement of structures, but not necessarily a common evolutionary origin. The next level is the superfamily group (shown) consisting of proteins with a common evolutionary origin. Proteins with low sequence identity, but common structural and function features are classified into the same superfamily. Right: Growth of structures in the PDB. Source: Based on `http://www.pdb.org/pdb/static.do?p=general_information/pdb_statistics/index.html`. Accessed May 2011.

dynamics currently cannot handle in general. While NMR also cannot provide this folding video, it can provide snapshots at near physiological conditions. The main advantage of NMR over x-ray is that with NMR, proteins can be studied in conditions resembling the protein's native environment. Although time-resolved x-ray crystallography can be used to study dynamics, it cannot do so in physiological conditions. There are two types of NMR: solution/liquid-state and solid-state, depending if the proteins is studied in liquids and solids, respectively. The former is used to study water-soluble proteins, while both can be used to study membrane proteins. In this work, we focus on solution methods, and we ignore membrane proteins because of the challenges they present, which is beyond the scope of this paper. In special circumstances, it is even possible to do protein structure determination of proteins inside living cells rather than using purified proteins in solution [101].

1.2 Beyond Structure Determination

Figure 1.2 gives examples of protein dynamics over various time scales and the NMR measurements for characterizing them. This thesis will focus on chemical shifts. NMR has been used to study protein-protein interactions [130], protein dynamics [76], folding pathways [57], and drug interactions [90]. Among the more successful NMR methods for drug design and screening, fragment-based methods, such as SAR (structure-activity relationship) by NMR [42, 106], have found their way in pharmaceutical companies and have resulted in discoveries that are currently undergoing clinical trials[43, 92]. In SAR by NMR, chemical shift mapping is used to identify a pair of molecules that bind to different, but nearby sites (Figure 1.3). Although each molecule may bind weakly, by linking the molecules together, a higher affinity molecule is created. Another NMR fragment-based approach is SAR by ILOE (interligand nuclear Overhauser effect) that does not require labeling of the protein target [97]. This thesis applies structure-based assignment to chemical shift mapping, which can be used to identify residues in the binding interface. By combining methods from protein-ligand docking, we demonstrate the potential for fast protein-ligand 3D complex determination.

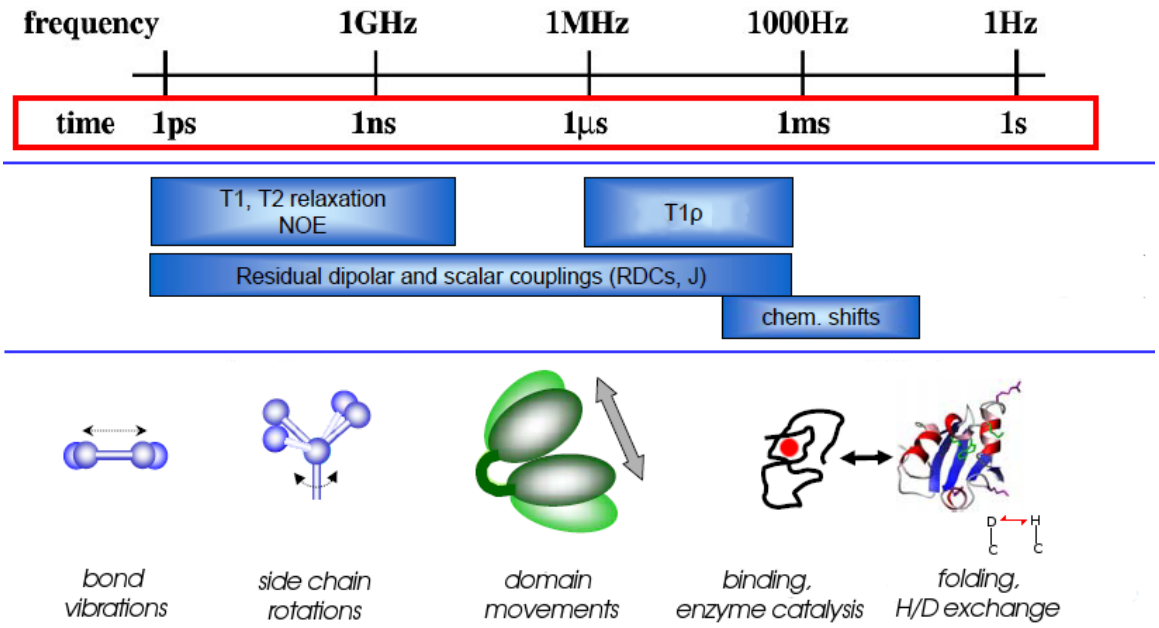

Figure 1.2: Characterizing protein dynamics over various time scales with NMR. The top row gives the time scale starting from fast motions (left) to slower motions (right). The second row gives examples of NMR measurements and experiments across the time scales. The third row illustrates various dynamic processes. Adapted from `http://www.embl-grenoble.fr/embo2008/files/THU_m_sattler.pdf`

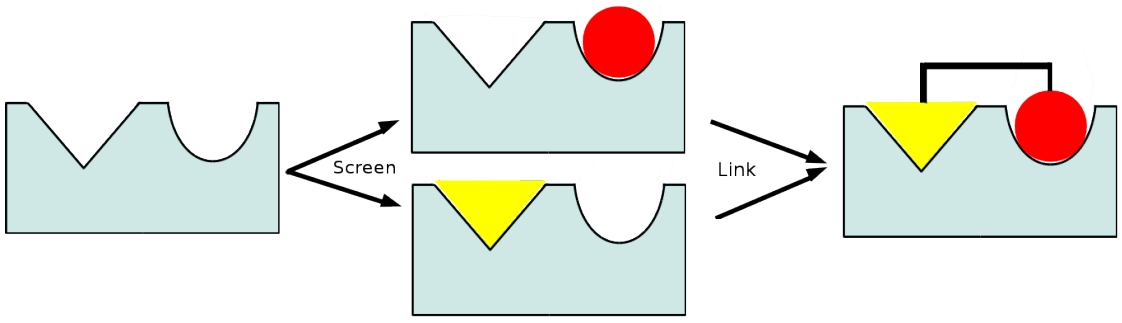

Figure 1.3: SAR by NMR. Ligands are screened for binding activity at adjacent sites and then linked to improve the affinity.

Chapter 2

Background

A review of the relevant basic NMR principles, experiments, and problems is presented for the computer scientist. It is by no means a complete introduction to NMR spectroscopy.

2.1 Physical Principles

Atoms with an odd number of nucleons, which is the number of protons plus neutrons, have a non-zero quantum mechanical property called spin. Atoms with spin $\frac{1}{2}$ are of particular interest to protein NMR spectroscopists. Examples include ^{1}H, ^{13}C and ^{15}N. To facilitate NMR spectroscopy, proteins are usually isotope labelled with ^{13}C and ^{15}N. NMR active nuclei can be visualized as magnets spinning about the axis of their magnetic moment, much like the earth (Figure 2.1). In the presence of an applied magnetic field, nuclei will be in one of two energy states. Those with their axis aligned with the field are in the low energy state, while nuclei with the axis aligned against are in the high. Regardless of the state, the applied field causes the spinning nuclei to precess/wobble, analogous to a spinning toy top. The frequency of precession is proportional to the magnetic field "experienced" by the nuclei, which depends on the local chemical environment of the atom, which in turn depends on the 3D structure of the protein. The field experienced is affected by the surrounding electrons, which, according to the laws of electromagnetism,

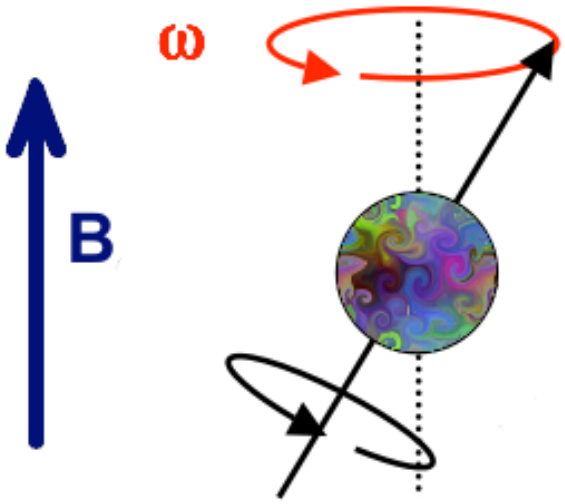

Figure 2.1: An NMR-active nuclei spinning about its magnetic moment axis (black) while undergoing precession at frequency ω (orange) in an applied magnetic field B.

will produce a magnetic field opposing the applied field. This shielding effect is less prominent for protons near electronegative atoms because the electronegative atoms have a deshielding effect of drawing electrons away from the proton. Deshielded protons experience the applied field more, and therefore have a higher frequency of precession. The precession frequency is known as *chemical shift*, and sometimes referred to as *spin* or *resonance*, and the value is relative to the frequency of standard compounds - DSS for ^{1}H and liquid ammonia for ^{15}N [119]. Equation 2.1 gives the chemical shift in standard form, where δ_a represents the value for atom a, ω represents its precession frequency in Hz, and ω_0 the frequency of a standard compound. The result is multiplied by a million to give a number in units of parts per million (ppm). Chemical shift is measured by applying energy of frequencies in the radio wave region of the electromagnetic spectrum to a protein sample in a magnetic field. If nuclei with frequency ω at the low energy state absorbs energy of the same frequency, it will transition to the high energy state, and then release the energy and relax back to the low energy state. The energy released gets detected. The energy applied produces a resonance condition, analogous to pushing a child on a swing at the correct frequency that makes the child go higher and higher.

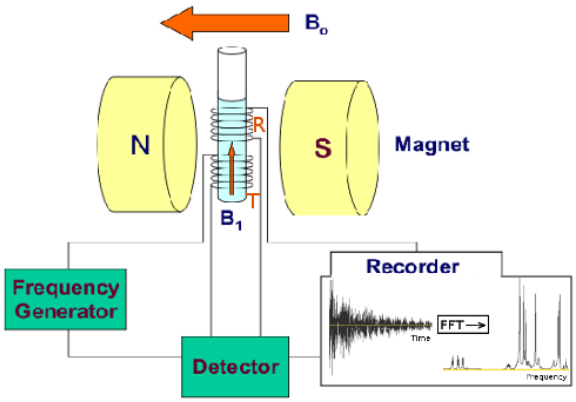

Figure 2.2: Simplified spectrometer operation. A spinning tube containing the protein sample is placed in a magnetic field B_0. A frequency generator generates alternating current on transmitter coil T, which induces magnetic field B_1, which in turn, perturbs the sample. The sample is then allowed to return to an equilibrium state from a mechanism known as relaxation. This cycle of perturbation and relaxation generates a fluctuating magnetic field in the sample, which induces an alternating current on a receiver coil R. The resonance signals get recorded by a recorder/computer, which performs a fast Fourier transform to convert the overlapping signals in the time domain to the frequency domain. Source: Adapted from `http://tonga.usip.edu/gmoyna/NMR_lectures/lecture2.pdf`

$$\delta_a = \frac{\omega - \omega_o}{\omega_0} \times 10^6 \tag{2.1}$$

Figure 2.2 gives a simplified model of the operation of a spectrometer. It is not exactly how modern spectrometers work, but it is sufficient for our purposes.

A lot of the data analysis in NMR is detective work, where the NMR signals serve as the observed evidence of some unknown structure. The goal is to link the evidence to its correct interpretation. For the evidence, we use chemical shifts. Each chemical shift value is associated with some atom, and a group of chemical shifts is associated with groups of atoms that are related to each other in 3D space or some other manner related to the structure. The difficulty is that the evidence is often incomplete and noisy. Given enough information, one can reconstruct the 3D structure, or in terms of our detective analogy, the crime. This thesis is focused on the linking part prior to the structure reconstruction. Note that the problems are not independent since knowledge of one part can help the other. We focus on using

structure information, such as from a homologous 3D structure, to simplify the linking step. For example, if experiment X indicates that chemical shift A is associated with B, which in turn is associated with C, and if we know that residue 4 is in contact with 5, and 5 with 8, then there is evidence that A is 4, B is 5, and C is 8. Of course there are other possibilities, but the number of possible combinations of assignments is reduced using this *a priori* information. This is analogous to using a detective's experience with certain types of crime scenes to better interpret the evidence. Although it might seem awkward for structure determination methods to use an input structure for structure output, this is not the case. Structure determination methods are typically iterative, so intermediate structures are used as input to produce structures that fit better the NMR data. In addition, for drug screening and NMR studies on dynamics, the output is not necessarily a structure. For example, it may suffice to know that binding is occurring, which residues are involved, and the binding affinity. An input structure can accelerate these studies if the structural changes are expected to be small upon binding, or the bound structure is similar to other known bound structures.

2.2 Experiments

NMR experiments can yield multidimensional spectra. Figure 2.3, gives an example of a 2D spectra, visualized as a contour diagram. It consists of peaks, which are local maxima whose multidimensional coordinates give the chemical shift values of a set of atoms and whose height is the intensity of the signal. Typically, peaks are selected or "picked" through manual inspection or with an automated peak picking tool and then verified manually. The set of peaks from a given spectrum shall be referred to as its peak list. We focus on experiments that do not use ^{13}C because they are more expensive due to proteins having more C than N. In addition, experiments using ^{13}C typically require double labeling that includes both ^{13}C and ^{15}N. In the online catalog of VLI Research Inc., `http://www.vli-research.com/Order_Proteins/Catalog/`, for 10 mg of the protein Ubiquitin, the ^{15}N-labeled costs $2250, the double labeled ^{13}C and ^{15}N costs $5750, and the unlabeled costs $2100.

Each peak in a 2D ^{1}H-^{15}N-HSQC spectrum consists of the backbone amide N and H^N chemical shifts

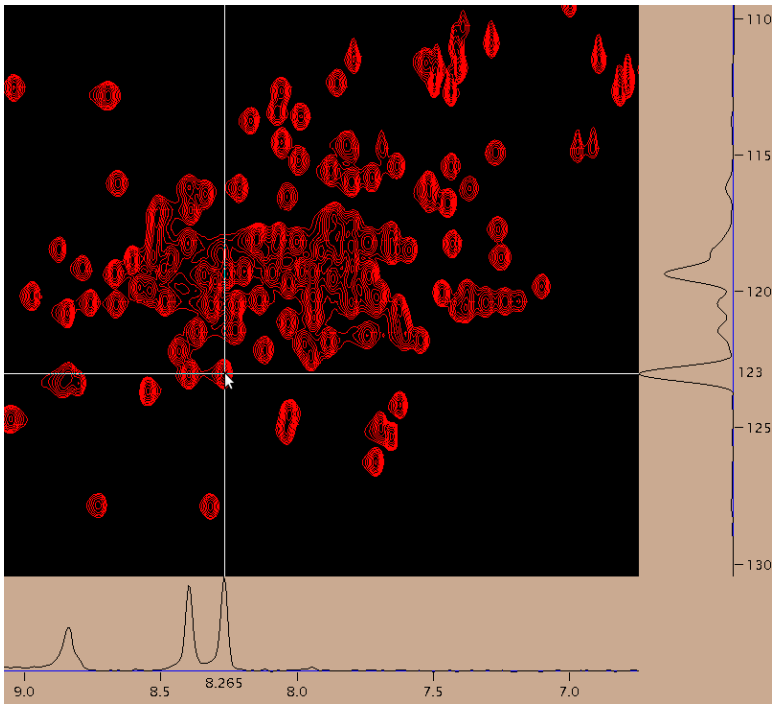

Figure 2.3: A section of a 2D ^{1}H-^{15}N-HSQC spectrum. The horizontal dimension gives the H^{N} chemical shift and the vertical the N chemical shift; both in units of ppm. The slices give the peak intensity in each dimension. The figure is a screenshot of SPARKY [37].

of some residue. The chemical shifts of the side chain N, H^N group of ASN, GLN, and TRP also appear (those for ARG and LYS are usually not visible), but they can be filtered out using expected patterns in this spectrum (Figure 2.4) and others (discussed below). As an abstraction, each HSQC peak can be treated as a 2D point in chemical shift space plus an intensity value. The δ_N and δ_{H^N} serves as an anchor for matching peaks from other spectra with dimension two or greater. δ_N is typically in the range of 100-135 ppm, and δ_{H^N} 6-11 ppm. Proline residues do not have a backbone H^N, so it has no HSQC peak.

Each peak in a 3D ^{15}N-TOCSY spectrum consists of backbone δ_N , δ_{H^N} of some residue, and the chemical shift of a proton in the side chain of that residue. Peaks with the same δ_N, δ_{H^N} can be grouped together to give the shifts of the residue's side chain protons. In the literature, a grouping of related chemical shifts, anchored by some HSQC peak, is called a *spin system*. As an example, for the spin system (121.14, 8.29, 4.19, 1.98, 0.83, 0.80 ppm), there would be TOCSY peaks for the 8.29, 4.19, 1.98, 0.83, and 0.80 ppm with δ_N , δ_{H^N} similar to 121.14, 8.29 ppm. There would also be an HSQC peak for 121.14, 8.29 ppm. Typically, a spin system consists of the chemical shifts of a single residue, and perhaps, for linking purposes, those of an adjacent residue. Note that methyl protons have only one chemical shift value visible in the TOCSY. For instance, if the above example is valine, the 0.83 ppm might represent the 3 HG1's, and 0.80 ppm the 3 HG2's. If the 6 HG's have very similar values, then there might only be one chemical shift value representing all of them. For the side chain N, H^N of GLN, there will be spin systems containing (NE2, HE21, HE22) and (NE2, HE22, HE21), where the HE's have chemical shifts around 7 ppm; similarly for ASN for ND2. For TRP, there may be a spin system consisting of the shifts for NE1, HE1, and one or more H's with a value around 7 ppm.

Grouping peaks by chemical shift, which we shall call spin system compilation, requires a definition of chemical shift distance. In this work, we use the following definitions

$$\begin{aligned} \Delta\delta_N(h,\, h') &= |\delta_N(h) - \delta_N(h')| \\ \Delta\delta_{H^N}(h,\, h') &= |\delta_{H^N}(h) - \delta_{H^N}(h')| \\ \Delta\delta_{NH}(h,\, h') &= \Delta\delta_N(h,\, h') + 10 \times \Delta\delta_{H^N}(h,\, h') \end{aligned} \tag{2.2}$$

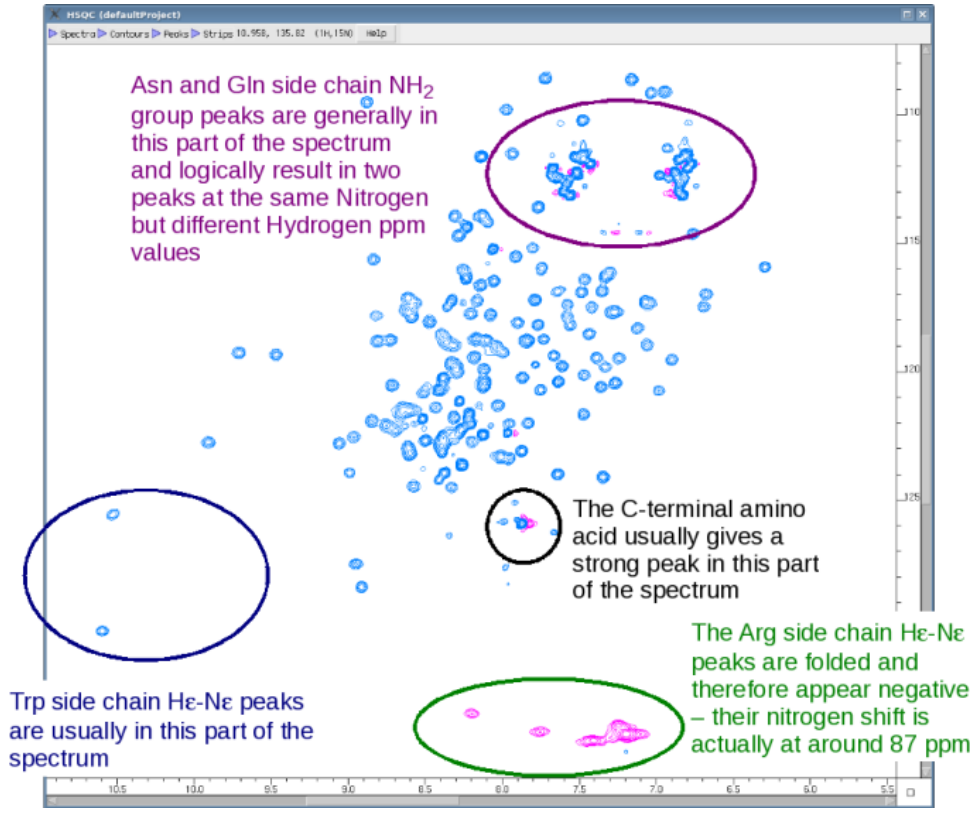

Figure 2.4: ^{15}N-HSQC spectrum showing the typical locations of side chain N, H^N chemical shifts. Folding refers to an artifact from NMR spectrum processing when signals are outside the spectrum boundaries. These artifacts can often be detected easily. Source: `http://www.protein-nmr.org.uk/spectra.html`

where $\delta_N(h)$ gives the N chemical shift of peak h, $\delta_{H^N}(h)$ the H^N chemical shift of h, and the 10 comes from the gyromagnetic ratio of ^{1}H and ^{15}N. Euclidean distance and various types of weightings can also be used [102]. *Chemical shift degeneracy*, defined as chemical shifts of very similar values but belonging to different atoms, can cause ambiguities when using chemical shift to match peaks. For instance, HSQC peaks 120.0, 7.85 ppm and 120.2, 7.81 ppm, corresponding to residues A and B, both match TOCSY peak 120.1, 7.83, 4.15 ppm that corresponds to proton C. C can belong to only one residue, unless it turns out that both residues have a proton with chemical shift at 4.15 ppm.

Each peak in a 3D ^{15}N-NOESY spectrum consists of the δ_N and δ_{H^N} of some residue, and δ_H of a nearby proton that is within 5Å of the H^N. δ_H can represent a proton in the same or different residue. Each peak also has an intensity value that is inversely proportional to the sixth power of the distance between the protons. This distance-dependent phenomenon is known as the Nuclear Overhauser Effect (NOE). However, for various reasons, such as spin diffusion and mobility, the intensity is often not an accurate reflection of distance [118]. Therefore, in structure calculation, rather than using exact distances, distance bins are used. NOEs can be weaker than expected by distance alone, so the lower distance bound of each bin is the same, but the upper bound varies depending on the intensity. For instance, NOEs are typically classified as strong, medium, and weak with distance bins 2-3Å, 2-4Å, and 2-5Å, respectively.

Given NOE (δ_N,δ_{H^N}, δ_H), to determine whether or not a pair of spins systems/residues are in contact, we match δ_N, δ_{H^N} to the amide chemical shifts of one spin system based on some distance tolerance, and δ_H to any proton chemical shift (including H^N) of another spin system. For example, given spin systems (121.1, 8.29, 4.19, 1.98, 0.83 ppm) and (106.5, 7.45, 3.79, 3.63 ppm), the NOE (121.2, 8.32, 3.82) ppm matches the amide chemical shift of the first spin system based on a 0.5, 0.05 ppm threshold, and the 3.79 ppm proton chemical shift of the second spin system based on a 0.05 ppm threshold. Note that there can be many pairs of spin systems that match a given NOE. Typically, each backbone H^N is near some proton in an adjacent residue (often another H^N or H^α), so embedded in the NOESY spectrum is sequential connectivity information. Like ^{15}N-HSQC and ^{15}N-NOESY, combinations of double-labeled ^{13}C , ^{15}N experiments can also give sequential connectivity information by correlating pairs of atoms separated by a limited number of bonds. These experiments are known as through-bond experiments, which will not

be discuss further. NOESY is a through-space experiment, which correlates atoms by distance. For the side chain N, H^N of GLN, there may be NOEs for (NE2, HE21, HE21), (NE2, HE21, HE22), (NE2, HE22, HE21), and (NE2, HE22, HE22); analogously for ASN. For TRP, there may be an NOE for (NE1, HE1, HE1) and perhaps NOEs for one or more (NE1, HE1, H), where H has a chemical shift around 7 ppm.

The magnetic fields of a pair of nearby NMR-active nuclei also interact via dipolar coupling, which depends on their distance and the angle between their displacement vector and the vector representing the direction of the external magnetic field in a global molecular coordinate frame. If the distance between the atoms is known (e.g. they are connected by a bond) and the molecule is rigid, then the distance can be treated as a constant, so we are left with the orientation of the vector in the molecular frame. This information is global, and it supplements local distance information from NOEs. In solution, dipole couplings average to zero, so alignment media, such as bicelles, are typically added to re-establish the coupling. A residual dipolar coupling (RDC) represents the orientation information, but the angle cannot be obtained directly from the RDC value due to ambiguities. Typically more than one media needs to be used, and the magnetic field axis in the molecular frame, which is in the form of an alignment tensor (a 3x3 matrix), needs to be estimated. To estimate this, at least five (in practice > 15) unique RDC assignments are needed [112]. ^{1}H-^{15}N-HSQC experiments can yield backbone N-H^N RDCs (D_{NH^N}). ^{13}C experiments can yield RDCs involving carbon and the protons attached to it.

2.3 Backbone Resonance Assignment

Resonance assignment is simply assigning each spin to its underlying atom. Depending on the context, in this paper, we shall use "assignment" to mean a single spin-atom assignment, or all the spin-atom assignments of a protein, or all the spin-atom assignments for a single residue. It will be clear from the context what is meant. The assignment problem is one of the main bottleneck steps in NMR because it can take weeks and even months for large proteins using semi-automated and manual methods. The focus here is on the assignment of backbone N, H^N chemical shifts to the underlying residues, which is equivalent to assigning each ^{15}N-HSQC peak to each amino acid residue. Proline residues cannot be assigned. The

N-terminus residue and regions of high mobility, such as long unstructured loops, also cannot be assigned due to poor NMR signals and undefined contacts.

Side chain proton assignment will not be considered except for H^{α}. In general, side chain assignment is non-trivial because it is often not possible to unambiguously assign side chain proton chemical shifts to the correct proton, especially stereospecifically. It might be worth investigating whether a pseudo atom approach of representing the other side chain protons will work. Methods for assigning side chain chemical shifts typically require a backbone assignment as input. Assigning each NOESY peak to the underlying pair of protons in contact is known as NOE assignment. It is also normally done after the backbone assignment step. The combination of side chain and NOE assignment is often needed to acquire a sufficient number of distance constraints for structure determination. Double-labeled experiments are needed in this case. The work in this thesis only performs backbone resonance and backbone NOE assignment because the focus is not on structure determination.

Traditional sequence-based, backbone assignment methods depend on connectivity information from double-labeled triple resonance experiments to connect chains of adjacent peaks together and to identify possible amino acid types of each peak for anchoring the chain onto the protein sequence (reviewed in [41, 18, 11], and illustrated in Figure 2.5). In contrast, structure-based methods, which is the focus of this thesis, use information from a homologous 3D structure to speed up assignment. This is analogous to molecular replacement in x-ray crystallography, where a homologous structure is used to help solve the phase problem. The homologous structure is used as a guide or template for matching the experimental data. This can accelerate the assignment step because the ambiguity is reduced. We shall call the homologous structure the *template* structure and the structure from which the NMR data is derived as the *target* structure. Sequence-based methods that also use NOEs and RDCs can also use a template structure. Given the recent slow growth of unique folds in the PDB (Figure 1.1) in contrast to the growth in the number of structures, it is likely that a homologous structure exists for the protein studied. So far, there is no standard approach for assignment either in terms of experiments or algorithms. Table A1 in [41] lists 44 programs for the assignment problem in general.

To see how assignment fits into structure determination, see Figure 2.6, which gives a generic structure

	N,	HN,	CA,	CB,	CA-1,	CB-1
MET 7:	121.7,	7.97,	**55.6,**	**32.3**,	55.0,	42.6
ARG 8:	121.7,	8.31,	55.7,	31.4,	**55.7,**	**32.5**

Figure 2.5: Traditional backbone assignment from double-labeled triple resonance experiments. Double-labeled NMR experiments give the chemical shifts for the atom types shown, which include the carbon chemical shifts of the current and previous residue. In this example, the 3D peaks of chemical shifts are grouped together using the amide chemical shifts. Two spin systems are shown along with their residue assignment. The carbon chemical shifts in bold are connected together between adjacent amino acids to give the backbone assignment for the polypeptide chain. This connection problem is known as the Hamiltonian Path problem.

determination pipeline. It is a simplification of the one used by the NorthEast Structural Genomics Consortium (NESG), which is one of the large scale US National Institutes of Health-funded structural genomics centers of the Protein Structure Initiative. Although it might seem awkward to use structure-based assignment for structure determination, this is not the case. The initial structure obtained from the structure calculation step is typically of poor quality because of assignment errors and incorrect distance restraints. Therefore, the pipeline is iterative, so that errors can be corrected. Intermediate structures are used to identify NOEs supported or violated by the contacts in the structures. Gradually, the quality of the structures improve. In general, it is not possible to satisify all distance restraints due to spurious NOEs. By using structure-based assignment, we can jump into the middle of the iteration. The assignment score can be used to measure how well the contacts in the structures match the NOE data. The ultimate goal is to bootstrap this pipeline with computational structure prediction methods, which is especially useful if there is no homolog, but there is a threading template. Achieving this goal requires the integration of backbone and sidechain assignment with structure determination techniques, which are all non-trival problems. For this thesis, we focus on only backbone assignment, which is the first step towards this goal.

2.3.1 Literature Survey

Among the older methods, GARANT [15, 14] can optionally use a homologous structure and/or chemical shift assignments from a homologous protein to improve accuracy. Their method is based on matching experimental peaks with expected peaks using a genetic algorithm, so presumably any spectra can be used

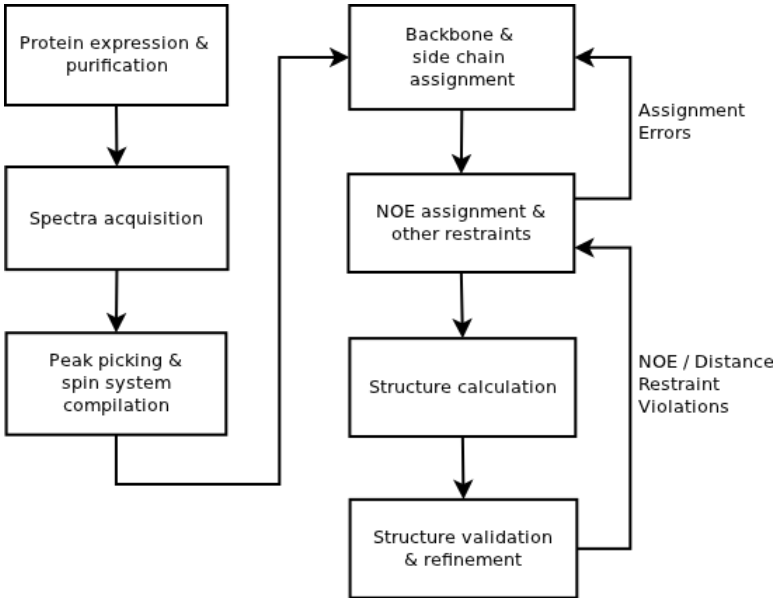

Figure 2.6: Generic solution NMR structure determination pipeline. Note that this is an extreme simplification of the process described in [80]. The pipeline is iterative as the results of subsequent steps are used to correct errors in previous steps. The step labeled "other constraints" consists of constraints other than distance constraints from NOE assignment. Examples include torsion angle and bond vector orientation.

including those for assigning side-chains. They tested their method on automatically picked peaks from the spectra of a 74-residue, 68-residue, and a 165-residue protein. For the two smaller proteins, they used 2D ^{1}H COSY (each peak gives a pair of protons that are at most 3 bonds apart), 2D ^{1}H TOCSY, and 2D ^{1}H NOESY. For the larger protein, they used 3D ^{13}C- and^{15}N-NOESY, and 3D CBCA(CO)NH (each peak gives N, H^N of residue i, and C^α or C^β of i-1). Exact accuracy numbers were not given as their results were reported as graphs, but their main conclusion is that using information from a homologous protein with automatically picked peaks is equivalent to using manually picked peaks without homology information. This indicates that fully automated methods must compensate for the poor quality output of automatic peak picking methods with homology information in order to compete with the accuracy of manual methods.

MARS [52, 53] is a sequence-based method that can use an input 3D structure in conjunction with RDCs. The structure is used to backcompute RDCs that get compared to the experimental values in the scoring function. The scoring function also matches predicted chemical shifts with experimental shifts. MARS requires as input spin systems, which they call pseudoresidues, rather than the spectra or peak lists. It assumes spin system compilation is done correctly, even though it is not a trivial problem due to chemical shift degeneracy. MARS accepts chemical shifts from N_i, H_i^N, C_i^α, $C_i{}^\beta$, C_{i-1}^α, C_{i-1}^β, and C_{i-1}^α. The assignment algorithm exhaustively tries to place pseudoresidue fragments of length 5 onto all possible 5 residue segments of the sequence. The algorithm then generates different assignment possibilities by adding noise to predicted chemical shifts. Conserved assignments are then fixed, and the process iterates using fragments smaller than 5. For the two-domain 370-residue maltose-binding protein, using the chemical shifts for N_i, H_i^N, C_i^α, $C_i{}^\beta$, C_{i-1}^α, C_{i-1}^β, and C_{i-1}^α, and RDCs D_{NH^N} and $D_{C^\alpha C'}$, and a 20% missing pseudoresidue rate, they assigned correctly 95% of the non-missing pseudoresidues.

For structure-based assignment, the Nuclear Vector Replacement (NVR) approach [66, 65] uses chemical shifts from ^{15}N-HSQC, D_{NH^N} RDCs in 2 media, sparse H^N-H^N NOEs from 3D ^{15}N-NOESY, and amide exchange rates. The problem was cast as a maximum bipartite matching problem, which they solved in polynomial time. Using close structural templates, they achieved an accuracy of over 99%. Their work was extended to handle more distant templates using normal mode analysis to obtain an ensemble of template

structures [7]. Unlike NOEs, RDC experiments are not as commonly used for backbone assignment. It is used more often for structure validation and refinement.

The contact replacement (CR) method uses the jigsaw approach [9], consisting of only ^{15}N-labeled data: 2D ^{15}N-HSQC, 3D ^{15}N-TOCSY, 3D ^{15}N-NOESY, and $^{3}J_{HNH\alpha}$ coupling constants derived from 3D HNHA. The problem was cast as a subgraph matching problem, where one graph consists of the contacts in the known protein structure, and the other consists of NOEs that connect spin system pairs. In general, the mapping of NOESY peaks to specific contacts is ambiguous due to experimental errors, missing peaks, and false peaks. Although the graph problem solved is NP-hard, the authors proved that under their noise model, the problem could be solved in polynomial time with high probability. They developed a branch-and-bound algorithm [123], which they later improved to a randomized algorithm [124]. The CR method was demonstrated to tolerate 1-2Å structural variation, 2.5 to 6x noise, and 10-40% missing NOE edges. Although they mentioned that there exists methods with close to 90% average accuracy for predicting a spin system's amino acid type class (types were grouped into 10 classes), such prediction errors were not tested. The method achieved an assignment accuracy of above 80% in α-helices, 70% in β-sheets, and 60% in loops. To our knowledge, it is the most error-tolerant structure-based method in terms of the noise level.

In NOEnet [110], the problem was also cast as a subgraph matching problem. Unlike the CR method, NOEnet generates an ensemble of assignments containing all possible assignments compatible with the NMR data, and it requires only H^N-H^N NOEs. However, it requires unambiguous NOEs, such as those from 4D NOESY experiments, so the noise is less than that handled by the CR method. NOEnet was updated to handle RDCs and chemical shifts from ^{15}N and ^{13}C-labeled proteins [112]. Filters were used to prune assignments that contain significant deviations from predicted RDCs and predicted chemical shifts. For the 259-residue EIN protein, they obtained an accuracy of 100% using N and H^N chemical shifts, H^N-H^N NOEs, and H^N-^{15}N RDCs. Using H^N-H^N NOEs and N, H^N, ${C_i}^{\alpha}$, ${C_{i-1}}^{\alpha}$ chemical shifts, they obtained an accuracy of 99.2%. The accuracy here allows for peaks to have multiple residue possibilities, whereas the accuracies reported for the other methods require one-to-one assignments.

Besides sequence-based and structure-based methods, there are structure determination methods that

do assignment indirectly. CLOUD-based methods use NOE information, antidistance constraints, and van der Waals repulsions to generate putative positions for protons [40] or residue fragments [67] by molecular dynamics or simulated annealing. The "clouds" generated are analogous to the electron density maps in x-ray crystallography. The protons or fragments are then connected and assigned to the sequence using a Monte Carlo search. Structures were generated for proteins with up to 137 residues [18].

Methods based on using Rosetta are typically used to produce or refine a structure rather than to perform backbone assignment [95, 94, 104]. Here, assigned backbone chemical shifts, but unassigned NOEs and unassigned side chain assignments, were used as inputs. There is an older version of Rosetta [74] that did backbone assignment using a Monte Carlo algorithm from N, H, and C chemical shifts, RDCs, NOEs, and structures predicted by Rosetta. The predicted structures were used to obtain predicted chemical shifts and predicted RDCs. Structures for proteins up to 140 residues were obtained from backbone assignments of up to 70% accuracy.

In general, NMR spectra are examined by visual inspection, where the peaks are picked by inspection, or by automatic methods with manual checking to remove noise. The peaks get accumulated in a peak list, and this list can change during the study as errors and inconsistencies are discovered during the assignment step. Therefore, the peak picking and assignment steps are usually done together. We aim to build a system that automates this process for ^{15}N-labeled data when a homologous structure is available. The difficulty is that automatically picked peak lists are noisier. To our knowledge, current structure-based methods are still semi-automated due to automated peak lists being of lower quality. To start, we first build upon the work of the CR method to address the limitation of ignoring type prediction errors. We show that if these errors are ignored, the assignment accuracy can be poor. Our system corrects for such errors with some success. In the absence of such errors, our system has comparable accuracy. The core of our system is a binary integer linear programming model. For the automatically picked peaks problem, as proof of concept, we solve it for human ubiquitin from a publicly available data set [44]. Instead of using HNHA, which was not available for ubiquitin, and not commonly available in the BMRB, we used chemical shifts from a homologous assignment (yeast ubiquitin), which is available in the BMRB. This work on backbone assignment is described in our paper [49].

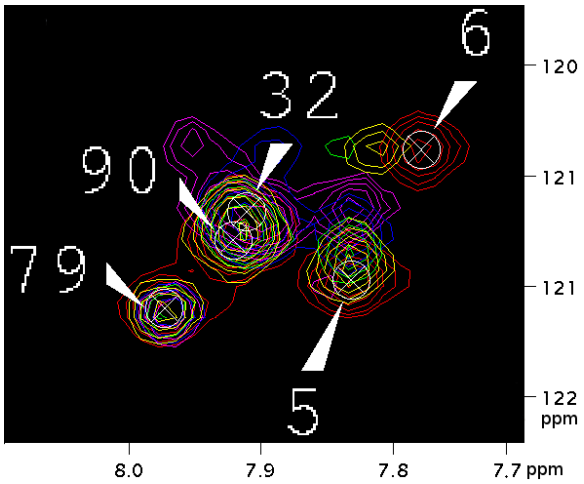

Figure 2.7: An overlay of five ^{15}N-HSQC spectra to illustrate peak tracking in fast exchange chemical shift mapping. The spectrum of the unbound protein is in red and labeled with each peak's residue number. The colors of the other spectra in order of increasing ligand concentration are yellow, green, blue, and purple.

2.4 Chemical Shift Mapping

This problem is similar to the backbone assignment problem in that the goal is to assign chemical shifts to the underlying atoms. The difference is that in chemical shift mapping, we are interested in tracking chemical shift changes when the chemical environment of the protein changes. Our focus is on identifying residues involved in ligand binding, but the methods are applicable to other changes, such as temperature and pH. Candidate residues are those whose atoms have chemical shift changes that exceed some threshold. Note that chemical shift changes might also be attributed to allosteric changes. Therefore, additional analysis is required to verify binding.

Figure 2.7 gives an overlay of five ^{15}N-HSQC spectra for a protein under increasing ligand concentrations, where the assignment is known for the unbound condition. The first spectrum containing the unbound protein shall be referred to as the *reference* spectrum, and its peaks as reference peaks. The last spectrum containing the fully-saturated protein shall be referred to as the *target* spectrum, and its peaks target peaks. The spectra excluding the reference shall be referred to as the perturbed spectra. If the assignments are known for the peaks in the reference, the path of chemical shift changes can be tracked

from reference peak to target peak via the perturbed peaks. The paths give the backbone assignment for the protein in each environment. From the figure, it appears that residue 6 has moved, although it is missing blue and purple peaks, and the paths of residues 32 and 90 are uncertain. The tracking problem is analogous to the computer vision problem of tracking points through frames of a video.

Ligand binding is an example of *chemical exchange*, which is a kinetic process characterized by a rate constant that measures the transition between the different states, such as between the free/unbound state and the bound. The situation depicted in Figure 2.7 is known as *fast exchange*. In fast exchange, the chemical shift is equal to the average of the free and bound state values weighted by their relative concentration. That is, if δ_{free} is the value for the free state, δ_{bound} is the value for the bound, then $\delta = \delta_{free}\frac{[free]}{[free]+[bound]} + \delta_{bound}\frac{[bound]}{[free]+[bound]}$, where $[free] + [bound]$ is the total protein concentration consisting of those in the free and those in the bound states. This thesis will focus on fast exchange because the SAR method is based on it, but there is another type of exchange. In *slow exchange*, both the chemical shifts for the free and bound state may appear in the spectra at the same time, with the intensity of the peaks proportional to the concentration of each state. The tracking is more difficult in this case. If the protein in Figure 2.7 undergoes slow exchange, only the red and purple peaks would be present. Ligands in fast exchange tend to bind weakly, and those in slow exchange bind tightly. Figure 2.8 summarizes the fast and slow exchange cases. In *intermediate exchange*, the peaks become exchange broadened (flat and buried under noise), so the spectra are difficult to analyze. One can try to shift the exchange to the fast or slow case by changing the temperature or external magnetic field.

Once putative binding residues have been identified, they can be used as constraints by protein-ligand docking methods to obtain the structure of the complex (Figure 2.9). In addition, binding constants for measuring the tightness of binding can be estimated from the peak paths (Section 2.4.3). For complex determination, it is useful to isolate the inter-molecular NOEs between the binding partners and the intra-molecular NOEs within. If one molecule is isotope-labeled (^{13}C, ^{15}N, or both) and the other is not, isotope-edited and filtered NOESY experiments can be used to show and hide, respectively, the signal from protons attached to the labeled isotope, and therefore, provide the inter- and intra-molecular NOEs. In general, such experiments are amenable to only the slow exchange case because of line broadening and lack

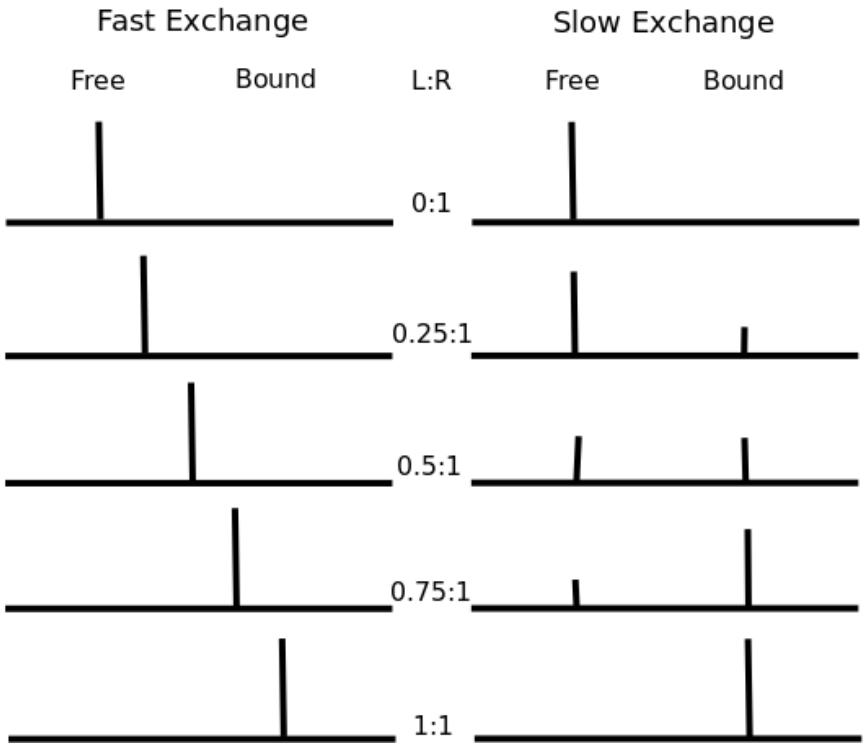

Figure 2.8: Fast and slow chemical exchange. The horizontal axis represents chemical shift, and the vertical axis represents peak intensity. L:R is the concentration ratio between the ligand and receptor. Only the chemical shift for the receptor protein is represented. In fast exchange, the chemical shift is an average weighted by the concentration of the free and bound state. For example, at 0.5:1, half the receptors are bound, so the chemical shift lies in the middle of free and bound. In slow exchange, chemical shifts for both the free and bound state may appear, where the peak intensity is proportional to the relative concentration of each state.

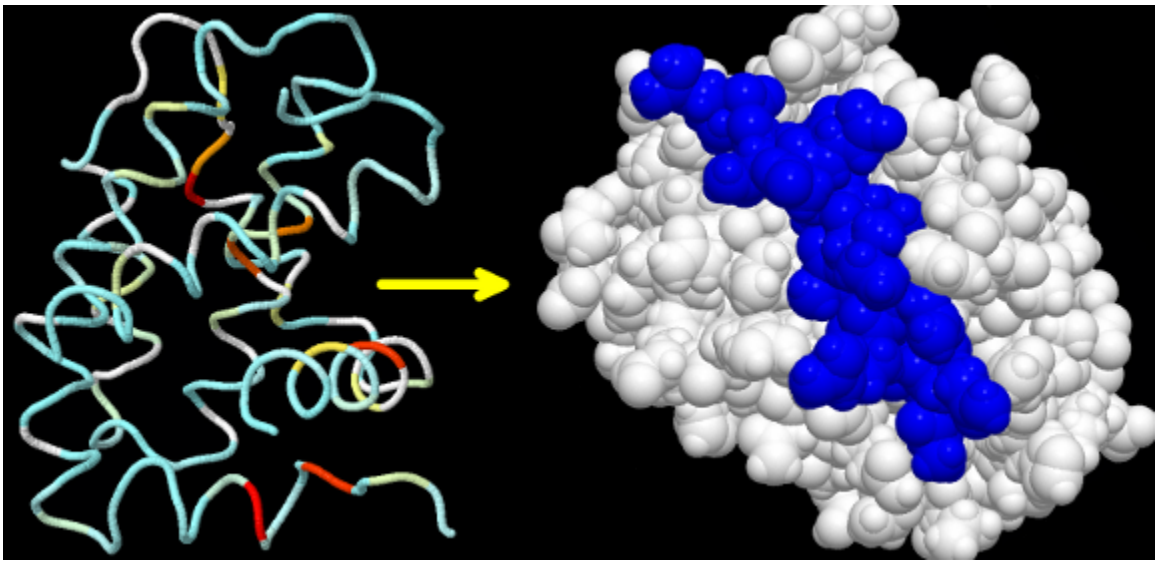

Figure 2.9: Left: Backbone diagram of $hBcl_{XL}$ with the residues highlighted by the magnitude of their chemical shift changes. In order of increasing chemical shift change, the colors are blue, yellow, orange, red. Residues in white are unassigned. Right: Spacefill diagram of the protein bound to Bak. To obtain this complex, the putative binding residues were passed in as constraints to a protein-protein docking program. Figure produced using Jmol [1].

of sensitivity due to rapid interconversion between the states in fast exchange. There are, however, cases where such experiments worked in the fast exchange case [23, 32, 87, 115, 98]. Instead, in fast exchange, transferred NOEs [34] are used to obtain the intramolecular NOEs of the bound ligand, and saturation transfer difference [72] is used to give the binding residues of the ligand. These methods are ligand-based in that they do not give information about the binding site on the larger receptor protein, which is provided by chemical shift mapping. They are also more applicable to the fast exchange case than the slow.

2.4.1 Literature Survey

Typically chemical shift mapping is done manually due to various errors and artifacts such as noise peaks, overlapping peaks, and missing peaks. Because of this, there are only a handful of automated methods for fast exchange.

Felix-Autoscreen [91] formulates the assignment of peaks in the reference spectrum to peaks in a

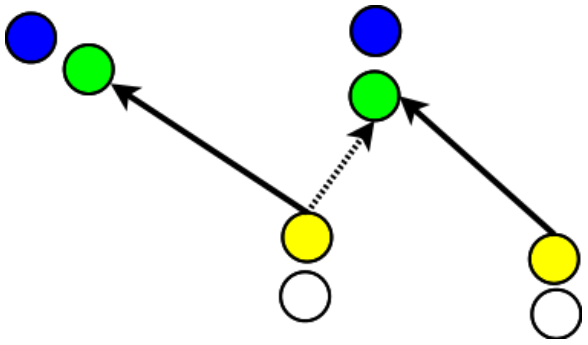

Figure 2.10: Peak walking by taking into account the path of a neighboring peak. The peak colors in increasing ligand concentration are white, yellow, green, and blue. The shortest path for the left white peak (dashed arrow) gives the wrong mapping because it results in a much longer path for the peak on the right.

perturbed spectrum as a bipartite graph matching problem, such that the sum of the chemical shift and peak shape differences is minimized. Optimizing the sum of the distances is better than choosing the peak nearest to each reference peak because the local greedy approach disregards the mappings of other peaks nearby, which results in errors (Figure 2.10). Dummy peaks were used to handle missing data, and peaks were picked on the fly during the execution of their algorithm. To handle more than one set of perturbed spectra, the bipartite matching algorithm was repeated successively, where the current perturbed spectrum becomes the new reference spectrum once it has been mapped. They tested it on a 74-residue protein domain in 8 different ligand concentrations, and obtained results similar to their manual efforts. The successive approach, however, is still a local greedy approach because it does not consider all the spectra simultaneously, so information about potential peak movements in later perturbed spectra are disregarded. Successive bipartite matching is an approximation of the k-dimensional matching problem (Section 4.1).

NvMap [35] also uses a greedy algorithm to successively match perturbed spectra. However, unlike FELIX-Autoscreen, the sum of the distances was not used. Instead, the pair of reference and perturbed peaks with the shortest distance was chosen and removed from consideration, and then the process was repeated for the next shortest. They tested their method on 97 residues of the SUMO protein on 2 different ligands, each at 6 different ligand concentrations. They obtained an average accuracy of 95%. The main

source of error was overlapping peaks within a spectrum, where only one of the peaks was picked and added to the peak list.

Of note is an older method, MUNIN [28], which is an automated method that does not identify the peak paths, but rather spectra similarity. Given a set of spectra, where each spectrum contains the protein with a different ligand, MUNIN identifies the peaks that are present in some spectra and absent in others using a 3-way decomposition method on the set without peak picking. If one or more spectra contains a ligand known to bind or not to bind, then it is often possible to identify whether binding has occurred in the other spectra. MUNIN was tested on a small region of the spectra of a 74-residue protein domain, where it identified the only spectrum with binding. MUNIN, however, does not identify the binding residues.

For large proteins, ambiguous mappings are inevitable. Rather than finding the unique mapping between peaks in the target to peaks in the reference, we find a set of plausible reference peaks for each target peak, where plausibility is determined by a scoring function. If the assignment for the reference is known, then the mappings give a set of possible residues for each target peak; e.g., ILE 3, LEU 27, LEU 78. We want this set to be small, but yet contain the correct residue. In this thesis, we present a novel peak walking model that describes the movements that peaks can make, and an approach that generates high scoring mappings by enumerating high scoring paths based on this model. Unlike previous methods, errors are modeled explicitly without using dummy peaks. We call our method PeakWalker. We tested it on 3 proteins with publicly available peak lists: UbcH5B titrated with Not4 [114]; hBcl$_{\mathrm{XL}}$ with BH3I-1 [62, 78]; and histone H1 at 2 different temperatures [109]. At 218 residues minus a removed flexible loop region R45 to A84, which was removed from the DNA sequence prior to NMR, hBcl$_{\mathrm{XL}}$ is much larger than the proteins tested by other automated methods. The average accuracy on the test set is over 95%, with an average of less than 1.5 amino acids predicted per target peak. We compared PeakWalker to a greedy approach similar to that used by NvMap, but modified to return multiple mappings. We also tested PeakWalker by varying the number of noise peaks.

We then describe an updated version of our structure-based backbone assignment method, called PeakAssigner, which takes the output of PeakWalker as input, and then resolves the mapping ambiguities using 3D ^{15}N-NOESY and the 3D structure of a homologous protein. In chemical shift perturbation

studies, a 3D structure is often available, such as from the Protein Data Bank (PDB)[17]. It is often the case that the bound structure of the protein is similar across different ligands that can bind to it, so that one bound structure can be used for studying different ligands. For drug screening, we often look for bound structures that are similar to the complex with a known drug. Here, we want to find drugs that bind similarly, but have fewer side effects. It is unlikely that a drug with a significantly different bound structure will have the desired effect. Therefore, structure-based assignment methods are ideal for disambiguating the possible mappings. On $hBcl_{XL}$, UbcH5B, and histone H1, PeakAssigner achieves an average accuracy of over 94%. The updated assignment method no longer uses 3D ^{15}N-TOCSY and a homologous resonance assignment to avoid their limitations. For large proteins, the TOCSY becomes crowded, and the homologous assignment can vary from the true assignment depending on the similarity of the proteins and the experimental conditions. Currently, there are no automated backbone resonance assignment methods that use only a series of ^{15}N-HSQC spectra and ambiguous NOEs from ^{15}N-NOESY spectra, and do backbone NOE assignment and H^{α} assignment, simultaneously. Although NOE and H^{α} assignment is not the main output of our algorithm, we show that by performing them, there is an improvement in backbone assignment accuracy, on average. This is demonstrated with simulated NOESY peaks from the PDB structures 1KA5, 1EGO, 1G6J, 1SGO, and 1YYC. PeakWalker and PeakAssigner are described in our paper [50].

2.4.2 Binding Models

Early binding models are based on shape complementarity, where one molecule acts as a "lock", while the other acts as the "key". Then it was observed that some molecules can still bind even though the shape of the key does not fit the lock. When the lock initially does not fit the key, the induced-fit model explains that molecules can undergo a conformational change upon binding to adapt to the shapes of each other. This model was later extended to one based on conformational selection and population shift [69]. The idea is that copies of the same protein are distributed in different conformational states, including one that is close to the bound state. Upon changes to the environment, such as the introduction of a binding partner, the distribution becomes shifted towards the bound state.

2.4.3 Binding Tightness

For the sake of completeness, we briefly describe how binding constants can be estimated once the peak tracking is known. Consider the following chemical equilibrium

$$R + L \underset{k_{-1}}{\overset{k_{+1}}{\leftrightarrow}} RL$$

where R represents the free receptor, L the ligand, RL the bound receptor, k_{+1} the rate coefficient of the forward reaction, and k_{-1} the rate coefficient of the reverse reaction. The dissociation constant

$$K_d = \frac{k_{-1}}{k_{+1}} = \frac{[R] + [L]}{[RL]}$$

measures the tightness of binding, where $[\cdot]$ represents concentration. The smaller the value of K_d, the tighter the binding. K_d is estimated for each atom that is potentially part of the binding site, and then the average is taken. To estimate K_d, one can use least squares data fitting between the experimental and predicted chemical shifts. The predicted values depend on a binding model. For a fast exchange, single binding site model with no allostery, the predicted value is given by $\delta = \frac{\delta_R + \delta_{RL} K_d^{-1}[L]}{1 + K_d^{-1}[L]}$ [62] where δ_R are δ_{RL} are the experimental chemical shifts of the atom in the free state and bound states, respectively. If the model is unknown, different models can be tested to find the best fit.

The different exchange cases are defined by the rate coefficients. Defining the exchange rate as $K_{ex} = k_{+1} + k_{-1}$, a site is in fast exchange if $K_{ex} \gg \Delta\nu$, slow exchange if $K_{ex} \ll \Delta\nu$, and intermediate exchange if $K_{ex} \approx \Delta\nu$, where $\Delta\nu$ is the resonance frequency difference in Hz between the atom in the free and bound state [100].

2.5 Binary Integer Linear Programming

Our approaches use binary integer linear programming (BILP). A *linear program* (LP) in standard form is expressed as

$$\begin{aligned} \max \quad & c^T x \\ \text{subject to} \quad & Ax \leq b \\ & x \geq 0 \end{aligned}$$

where x is the $n \times 1$ vector of real variables to be solved, c is an $n \times 1$ coefficient vector, A is an $m \times n$ constraint matrix, and the right-hand side b is an $m \times 1$ vector of bounds. $c^T x$ is called the objective function to be optimized, subject to the constraints $Ax \leq b$ and $x \geq 0$. This problem is known to be solvable in polynomial time, but if x is constrained to be integers, the problem becomes an *integer linear program* (ILP), which is NP-hard. ILPs are typically solved using branch-and-bound with linear programming relaxations to obtain the bounds. If x is constrained to be binary, taking on the boolean values of 0 or 1, then the problem becomes a *binary integer linear program* (BILP), which is still NP-hard, but there exists specialized branch-and-bound methods on boolean variables, such as the Balas Additive Algorithm [10]. For BILPs where A, b, and c are all integers, the problem becomes a *pseudo-boolean program* (PBP). PBPs can exploit methods [3, 127] from both the operations research community, such as those for BILP, and also methods from the computer science constraint programming community, such as constraint propagation. A special case of PBP is the *boolean satisfiability (SAT) problem*, where is $c = 0$, so the problem becomes a feasibility problem rather than optimization. Our problems are similar to pseudo-boolean programs except that c does not consist of all integers. We did not consider rounding the values to integers to get a PBP because the integer programming solver CPLEX, which we used, still outperforms PBP solvers on PBPs according to the Pseudo-Boolean Competition 2010 (See OPT-SMALLINT-LIN, under which our problems fall, at `http://www.cril.univ-artois.fr/PB10/results/ranking.php?idev=36`).

ILP solvers can return a solution with score guaranteed to be at most $N\%$ away from the optimal score, where N is a user-specified parameter called the gap. The smaller the gap, the longer the run time.

At 0%, it results in brute-force search. Unless stated otherwise, we used a gap of 1%. For generating multiple solutions, the sequential algorithm, introduced by Greisdorfer et al. [39] and generalized to more than two solutions [29], can be used to generate solutions that are guaranteed to be within a certain percentage of the optimal solution and have maximum diversity as measured by a diversity measure, such as average pairwise hamming distance. The one tree algorithm can also be used [29] with the same optimality guarantees without the diversity maximization, but with better efficiency. We used the one tree algorithm with CPLEX 12.2 and the sequential algorithm with CPLEX 9.13, which did not have the one tree.

Chapter 3

Backbone Resonance Assignment

The goal is robust methods that can tolerate sparse and noisy data. In terms of missing data and noise for ^{15}N-labeled data, the CR method is the most robust, so we improve upon their ideas to get better error handling, optimality guarantees, and modeling flexibility. Modeling flexibility and efficiency often do not go together. In practice however, our BILP modeling approach runs faster than the CR method because we can incorporate constraints directly into the model to limit the search space. Incorporating such constraints is more difficult to do with the CR method.

We decided to avoid ^{13}C-labeling for not only reasons of cost, but also because the vast majority of available methods use such data, so there is little room for improvement. However, if data from ^{13}C-labeling is available, our models are capable of accepting such data with minor modifications, which should only improve the accuracy of our results. Our approach provides an alternative perspective to the problem. We also do not use RDC data, although our models can accommodate it, since it is not often used for backbone assignment. Since this thesis builds from the CR method, we first review graph matching and summarize their graph data structures, which is one of the key contributions of their work, before presenting our methods.

The input data consists of amide chemical shifts from ^{15}N-HSQC, side chain proton chemical shifts

from ^{15}N-TOCSY-HSQC, J-coupling constants from HNHA, and H^N-H NOEs from ^{15}N-NOESY-HSQC. Each spin system looks like the following example: 123.0, 8.94, (5.27, 2.9, 6.8 FYWH), J=4 (helix), [8.6, 8.94, 4.72, 4.81]. The first two numbers give the amide chemical shifts. The next set of numbers gives the side chain proton chemical shifts, and possible amino acid types, which can be predicted from those shifts (Section 3.7.3). The next number gives $^3J_{HNH\alpha}$, which is used to predict the secondary structure type (Section 3.6). The final set of numbers in square parenthesis gives the chemical shifts of the protons that are in contact with H^N, which has chemical shift 8.94. The predicted amino acid and secondary structure types are used to limit the possible residues for this spin system; e.g., F3, Y45, Y82, W50, H8, H65.

3.1 Graph Matching

The field of pattern recognition using the representation of objects as graphs is a well studied problem in computer science [25]. Graph matching is also commonly used in bioinformatics and chemoinformatics [96]. One approach to comparing two graphs is finding the largest subgraph common to both. This *maximum common subgraph* (MCS) approach, also known as *subgraph isomorphism*, has different variations depending on the constraints on the common subgraph; e.g., whether or not a vertex common to both graphs can have an extra edge in only one of the graphs and still be allowed in the common subgraph. Figure 3.1 illustrates the maximum common node-induced subgraph (MCIS) and maximum common edge subgraph (MCES) of a pair of graphs. Both versions are NP-hard [36]. The problem is not only NP-hard, but also APX-hard [26], which means it is difficult to even approximate.

A MCIS requires a one-to-one correspondence; that is, if vertices X and Y in the MCIS has an edge between them, then in both graphs to be compared, there must be an edge between X and Y. If vertices X and Y in the MCIS does not have an edge between them, then in both graphs, there must not be an edge between them. In MCES, both graphs can have extra edges. The problem of finding the MCS can be reduced to finding the maximum clique in a modular product/association/compatibility graph [13]. A clique is a subset of vertices where every pair of vertices is connected by an edge.

The MCS problem can also be viewed as a graph editing problem. It is analogous to the sequence

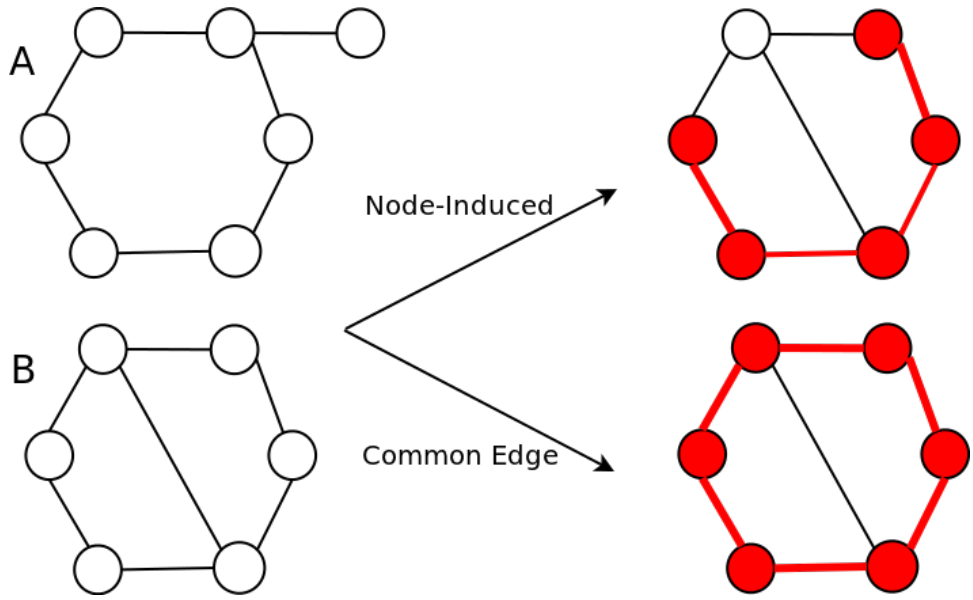

Figure 3.1: Different types of maximum common subgraphs (highlighted in red) for graphs A and B. Note that in the node-induced case, only one vertex incident to the diagonal edge can be in the common subgraph because if both vertices are in the common subgraph, then both graphs must have this diagonal edge, which is not the case. Also note that the maximum common subgraph is not unique because we could have included the other vertex instead.

alignment problem, where the goal is to "edit" one string using the fewest character additions, deletions, and substitutions to transform it to the other string. Unlike strings, minimizing the graph edit distance is more difficult because graphs in general do not have a linear structure that can be exploited.

MCS is not the only graph-based approach for assignment. If both graphs consists of only vertices and no edges, a new graph can be created from the two using the same vertices. Edges are added between vertices from one graph to vertices in the other, where the weight of the edge measures the degree that the vertex pair matches according to some criteria. To illustrate, in the NVR approach, one set of vertices represents H^N-N internuclear bond vectors, the other represents D_{NH^N} RDCs, and the edge weights represent the similarity between RDCs predicted from the bond vectors and the experimental values. The goal is to find a one-to-one mapping or matching from one set of vertices to the other, such that the total weight of the edges is maximized. A matching is simply a set of edges, where the edges do not share any vertices. This is known as the maximum bipartite graph matching problem, and it is solvable in polynomial time; for example, by the Hungarian algorithm [64].

3.2 Backbone Assignment as Graph Matching

Many backbone assignment methods cast the problem as a graph matching problem. NOE-based methods are typically based on MCS, where one graph represents the contacts and the other the NOE connectivities. In this section, we describe the graph representation of the CR method and any changes that were made to the representation.

3.2.1 Contact Graph

The template structure is represented by a *contact graph* (CG). Each vertex represents a proton; e.g., H^N of ALA 25. Only the 3D coordinates of H^N and H^α protons are considered. Vertices are also labeled with the amino acid type and secondary structure type of the residue containing the proton. Each edge is labeled by the H^N - H^α or H^N - H^N contact that it represents, where a contact is defined by a distance threshold.

H^N - H^α or H^N - H^N are the only interaction types considered. Among all the types of protons, NOEs for contacts involving amide protons, and alpha protons are more reliable because they are observed more frequently whenever such contacts are expected [31]. The edges are directed such that the tail represents the H^N of some residue and the head represents the H^α or H^N of another residue. An H^N - H^N contact will have two edges, one for each direction because in each NOE (δ_N, δ_{H^N}, δ_H), the δ_N, δ_{H^N} represent only one of the amide groups; there should exist another NOE whose δ_N, δ_{H^N} matches that of the other amide group.

3.2.2 Interaction Graph

Proton chemical shift correlations in the ^{15}N-NOESY are represented by an NMR *interaction graph* (IG). Each vertex represents the chemical shift of a proton of unknown identity from an associated spin system; e.g. 8.94 ppm from (123.1, 8.94, 5.26, 2.22, 1.86) ppm. Vertices are labeled by their predicted amino acid type and secondary structure type. Amino acid type predictions were obtained from the RESCUE software, version 1 [93] using the chemical shifts as input. RESCUE classifies each spin system into one of ten possible amino acid classes. In our work, we used all classes with positive reliability score rather than the highest scoring class to compensate for errors made by RESCUE. Secondary structure type predictions can be obtained from $^3J_{HNH\alpha}$ coupling constants [122].

Each edge represents an NOE match; e.g., NOE (121.2, 8.32, 3.82) ppm matches spin system (121.1, 8.29, 4.19, 1.98, 0.83 ppm) and spin system (106.5, 7.45, 3.79, 3.63 ppm) based on a $\delta_N, \delta_{H^N}, \delta_H$ threshold of 0.5, 0.05, 0.05 ppm (the matching chemical shifts are underlined). Each edge is directed from the amide proton chemical shift to the other proton chemical shift. We consider only proton chemical shifts for H^N obtained from the HSQC peaks, and those for H^α in the range 2-6 ppm obtained from TOCSY peaks. Like the CR method, contacts for other protons, such as to H^γ, were not considered because assigning the non-alpha protons unambiguously is often not possible due to overlap. However, it is worth investigating if leaving such assignments as ambiguous would improve or hinder overall assignment.

For each interaction, a match score is associated, which measures the match of the NOESY peak to the pair of spin systems. Given NOESY peak (δ_N, δ_{H^N}, δ_H) that matches δ_N^A, $\delta_{H^N}^A$ of spin system A and

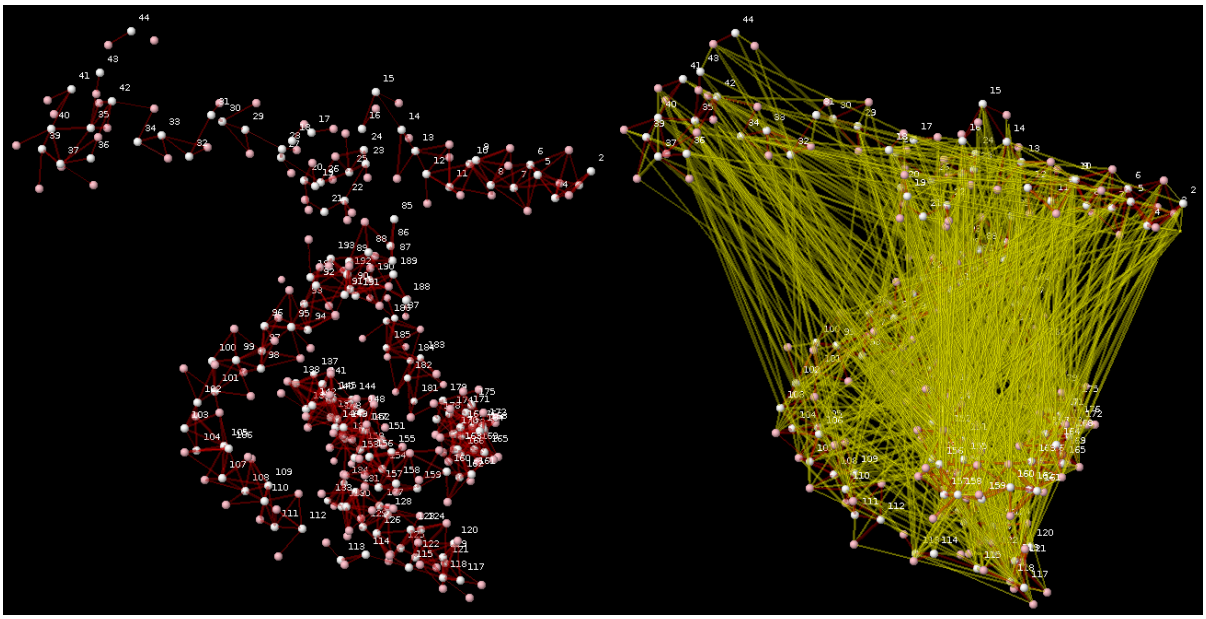

Figure 3.2: Left: Contact graph. H^N in white and H^α in pink. H^N is labeled by its residue number. Right: Interaction graph. The maximum common subgraph is highlighted in red. The H^Ns are labeled by their residue number in the IG, but in general, the labeling is not known.

some δ_H^B (H^Nor H^α) of spin system B, the match score is defined as $erfc(\frac{|\delta_H - \delta_H^B|}{0.02 \times \sqrt{2}})$ as used in [124], where $erfc$ is the complementary error function. The score ranges from 0 to 1, where if $|\delta_H - \delta_H^B|$ is small, i.e. the match is close, then the score is closer to 1. The score does not use δ_N, δ_{H^N} of the NOE.

An NOE can match the chemical shifts of many possible pairs of spin systems. In the IG, there are edges for every possible match even though there is only one correct match. In general, IGs have many more edges than the CG. The *noise ratio or level* is defined as the number of IG edges over the number of CG edges. Hidden within the IG is some form of the CG, perhaps with extra and missing edges. The goal is to find it. Figure 3.2 illustrates the graphs, and shows the matching subgraph embedded in the IG. For illustrative purposes, edges were removed from the IG, so that the matching subgraph could be visible.

3.2.3 NMR Graph Matching Approach

A CG vertex matches an IG vertex if the CG vertex's amino acid and secondary structure type is present in the IG vertex's set of predicted amino acid and secondary structure types. A CG edge matches an IG edge if the end point vertices match and the interaction type matches.

For simplification, protons belonging to the same residue in the CG were grouped together into a single vertex, and proton chemical shifts in the IG belonging to the same spin system were grouped together into a single vertex. Contacts between a pair of residues were grouped together into a single undirected edge in the CG, but with the directions of each interaction preserved. NOESY peak matches between a pair of spin systems were grouped together into a single undirected edge in the IG. A pair of CG and IG edges match if their list of types have at least one type in common, and both pairs of end point vertices match. For each edge match, we associate an edge score by taking the sum of the match scores for all interaction types in common.

To find the best match, we used the MCES to maximizes the sum of the edge scores, subject to the constraint that the vertices and edges match. MCIS was not chosen because NMR data, as represented by the IG, is typically noisy and contains missing data. Since a polypeptide chain is sequential in nature, the IG has an embedded Hamiltonian path. In graph theory, a Hamiltonian path is a path that visits each vertex exactly once. Due to proline residues not having a backbone amide group, the Hamiltonian path is fragmented. Finding this path and using it to guide the assignment was the approach taken by the CR method. Since contacts outside the Hamiltonian path might be missed, we chose to solve the subgraph isomorphism problem directly. We modeled the problem as a quadratic assignment problem and cast it as a BILP. Our BILP is similar to the quadratic assignment linearization by Adams and Johnson used for the maximum clique problem [22, 128]. In the quadratic assignment problem, the goal is to place N facilities in N locations, such that the cost is minimized, where the cost depends on the distance between the locations and the flow (e.g. weight of the load transferred) between the facilities. Formally, the cost is $\min_{\pi} \sum_{i \neq j} f_{ij} \cdot d(\pi(i), \pi(j))$, where π is the assignment of facilities to locations, the sum is over all pairs of facilities, f_{ij} is the flow between facilities i and j, and $d(\cdot, \cdot)$ is the distance between the locations. The problem is APX-hard, both the minimization and maximization versions, although currently the

minimization version is more difficult to approximate [84]. Our NMR problem is not as difficult since it is a special case. The problem was proven to be solvable in polynomial time with high probability [124]. In practice, however, the algorithm based on that theory runs extremely slowly because the number of possibilities to ensure the high probability can still be large. Using a BILP has the advantage of using constraints to prune the possibilities.

BILPs are not new to the MCS problem [54]. They have even been used in NMR. In IPASS [5], a BILP was used for sequence-based assignment that relies on sequential connectivity from double-labeled triple resonance experiments. We use a BILP for structure-based, ^{15}N-labeled-only NOESY-based backbone assignment. A BILP was also used for assignment using RDCs by NVR-BIP [6]. Our BILP is different from both. IPASS is based on Hamiltonian Path, and NVR-BIP is based on maximum bipartite matching. Ours is based on maximum common subgraph. We chose a BILP because it models the problem naturally as we will show. The commercial optimization package ILOG CPLEX® version 9.13 was used as the solver.

3.3 Backbone Assignment BILP Model

Define V_c, V_i to be the set of vertices in the CG and IG, respectively. Define E_c, E_i to be the set of edges in the CG and IG, respectively. For notation purposes, (a, b) shall denote a graph edge between vertices a and b, and $\{a, b\}$ shall denote the assignment of residue a to spin system b.

3.3.1 Binary Variables

$X_{a,s,b,t}$ Equals to 1 if spin system s is assigned to amino acid residue a, and spin system t is assigned to amino acid b; and 0 otherwise. This variable represents the assignment of edge $(a, b) \in E_c$ to edge $(s, t) \in E_i$, where vertex a is assigned to vertex s, and b to t. $X_{a,s,b,t}$ is equivalent to $X_{b,t,a,s}$ (only one such variable exists), but $X_{a,t,b,s}$ is different from $X_{a,s,b,t}$ because a is assigned to t in the former, and a to s in the latter.

$X_{a,\,s}$ Equals to 1 if spin system s is assigned to amino acid residue a; and 0 otherwise. This variable represents the assignment of vertex $a \in V_c$ to vertex $s \in V_i$. The set of all $X_{a,\,s}$ variables set to 1 gives the backbone assignment.

3.3.2 Objective Function Coefficients

$W(X_{a,\,s,\,b,\,t})$ The score of assigning the contacts between residues a and b to the NOESY peak matches between spin systems s and t. The score is equal to the sum of the match scores for all interaction types in common. The score is non-negative. It represents the edge match score.

$W(X_{a,\,s})$ The score of assigning amino acid a to spin system s. Currently it is not used and set to 0. This represents the vertex match score.

3.3.3 Objective Function

$$\max_{X} \left(\sum_{\{a,\,s\} \in A} W(X_{a,\,s}) \cdot X_{a,\,s} + \sum_{(a,\,b) \in E_c} \sum_{(s,\,t) \in E_i(a,\,b)} W(X_{a,\,s,\,b,\,t}) \cdot X_{a,\,s,\,b,\,t} \right) \tag{3.1}$$

where

A The set of all $\{a,\, s\}$ that 1) match, where $a \in V_c$ and $s \in V_i$, and 2) involved in at least one edge match; i.e., there exists $(a,\, b) \in E_c$ and $(s,\, t) \in E_i$ such that $(s,\, t)$ matches $(a,\, b)$.

$E_i(a,\, b)$ The set of all edges in the IG that match the edge $(a,\, b) \in E_c$. An edge $(s,\, t) \in E_i$ matches edge $(a,\, b)$ if one: the edges have interactions in common, and two: if either the types of a matches the types of s and the types of b matches that of t, or a with t and b with s.

The objective function expresses the total edge and vertex match score of the assignment. The first summation is over all vertices that match and that are involved in at least one edge match. The second summation is over all edges that match.

3.3.4 Constraints

1. Each amino acid a is assigned to at most one spin system. $\sum_s X_{a,\,s} \leq 1$
2. Each spin system s is assigned to at most one amino acid. $\sum_a X_{a,\,s} \leq 1$
3. For edges that match, the pairs of end point vertices must also match. That is, $\forall\{a,\,s\} \in A,\ \forall(a,\,b) \in E_c$, if $X_{a,\,s,\,b,\,t} = 1$ then $X_{a,\,s} = 1$ and $X_{b,\,t} = 1$ and vice versa. This is $\sum_{t \text{ such that } (s,\,t) \in E_i(a,\,b)} X_{a,\,s,\,b,\,t} \leq X_{a,\,s}$.

3.3.5 Discussion

The size of the model is dependent on the number of possible residues for each spin system; e.g., GLU 5, GLU 35, ASP 79. This is known as the *spin system typing*. The spin system typing can also be defined in the other direction as the possible spin systems for each residue. This typing can be obtained from the amino acid type and secondary structure type predictions, which are used to limit the possible residues. The number of variables is proportional to the number of pairs of edges that match. The number of pairs of edges that match depend on the end point vertices matching, which in turn depends on the spin system typing. In the worst case, all $|E_c| \times |E_i|$ pairs of edges match.

Constraint 3 accounts for the majority of the constraints. The number of such constraints is $\sum_r |Typing(r)| \times Degree(r)$, where the summation is over all residues, $|Typing(r)|$ is the number of possible spin systems for the given residue, and $Degree(r)$ is the number of edges incident to the vertex that represents residue r in the CG. If the typing can be bounded by a constant, then the above summation will be bounded by a constant factor on the number of edges in the contact graph. Theoretically, an NOE corresponds to only one contact, but for efficiency reasons, like the CR method, the BILP presented here does not enforce the constraint that each NOE is assigned to at most one contact, and therefore, does not do backbone NOE assignment. However, both can be modified to do it. Section 4.3 describes a BILP that also does NOE assignment, but with more variables and constraints. Results for this new BILP will be presented in the next chapter.

Since we look for the MCES instead of the MCIS, the objective function contains variables for possible edge matches only rather than non-matches. Extra edges can get unmatched, especially since the IG tends to have more edges than the CG. Since all scores are non-negative, non-matches implicitly have a score of 0. Missing edges also have a score of 0 because it means an unmatched edge in the other graph. We also do not assign vertices that are isolated, unless the vertices can be unambiguously assigned, such as being the only ones with a particular amino acid and secondary structure type combination. Constraints 1 and 2 allow extra amino acids or spin systems to be unassigned because they are inequalities rather than equalities. Note that the above formulation does not enforce that the common subgraph be connected, so contacts in different domains of the protein can get matched, while the parts in-between are not matched.

A proof of correctness for Constraint 3 is given in Section 4.3.5, where an improved version of the BILP is given. The idea is as follows: if $X_{a,\,s} = 1$ and $X_{b,\,t} = 1$, the left hand side of Constraint 3 can be zero, so missing edges are allowed. However, edge match scores are always non-negative and we are maximizing the score, so if an edge match exists between $(a,\, b)$ and $(s,\, t)$, we are guaranteed that one of the edge match variables on the left is set to 1. Also, an edge in one graph cannot be assigned to more than one edge in the other graph; e.g., $X_{a,\,s,\,b,\,t} = 1$ and $X_{a,\,u,\,b,\,v} = 1$, or $X_{a,\,s,\,b,\,t} = 1$ and $X_{i,\,s,\,j,\,t} = 1$ cannot occur because one of Constraints 1 and 2 would be violated.

The advantage of using a BILP is that the scores in the objective function coefficients only need to be computed once when searching through the space of all possible common subgraphs. In addition, by relaxing the constraint that all variables are 0 or 1 by allowing them to be in the interval between 0 and 1, linear programming can be used to obtain an upper bound on the score, which is used to prune suboptimal solutions. Note that if all the $X_{a,\,s,\,b,\,t}$ variables are 0, the problem becomes a maximum bipartite matching problem. In this case, we can relax the constraint that the $X_{a,\,s}$ variables are integers because the constraint matrix is totally unimodular [24], so linear programming, which is not NP-hard, is guaranteed to give an integer optimal solution.

3.4 Model Generalizations

The BILP model can be adapted to accommodate different situations by setting, adding, or removing variables, modifying their coefficients, and adding or removing constraints.

3.4.1 Different Types of Data

Although we considered only chemical shift data in the scoring function, the objective function can easily be modified to account for different types of data. This can be done as long as the data can be modeled as matching features to residues and pairs of features to pairs of residues. Different weights on $W(X_{a,\,s,\,b,\,t})$ can be used to emphasize matches to specific types of contacts, such as long range β-sheet contacts and local H^{α} and H^{N} contacts in α-helices. H^N-H^N NOEs, as used in NOEnet, can be encoded in the $W(X_{a,\,s,\,b,\,t})$ terms.

For ^{13}C-labeling, if there is carbon connectivity evidence that supports that spin systems s and t is associated with adjacent amino acids a_i and a_{i+1}, the value of $W(X_{i,\,s,\,i+1,\,t})$ can be increased. The variable $X_{i,\,s,\,i+1,\,t}$ can also be removed if there is insufficient connectivity and contact information. The CR method focused on finding common Hamiltonian path fragments. Similar to carbon connectivity, the score for each edge match for adjacent residues can be scaled up to emphasize the Hamiltonian path, so that the objective function contains a weighted version of the Hamiltonian path length.

The BILP model is also suitable for the iterative expectation/maximization framework of the NVR method. For RDC data, once an alignment tensor has been estimated, back-computed RDCs can be computed and compared with the experimental values to yield a value for each $W(X_{a,\,s})$. After running the BILP, the assignment information can be used to update the alignment tensor, which in turn, is used to update the $W(X_{a,\,s})$ terms.

3.4.2 A Priori Assignment

BILP solvers can start from an initial solution to improve performance. This initial solution can even be a partial assignment. If specific spin system-amino acid assignments are known, the corresponding variables can be fixed to 1. The ability to fix specific assignments and to start from an existing assignment allows for a semi-automated approach, where the returned assignment is examined and corrected manually. The BILP can then be rerun using the new information rather than starting from scratch.

3.4.3 Multiple Solutions

The maximum common subgraph is not necessarily unique, so there may be multiple best scoring assignments. The methods mentioned in Section 2.5 can be used to generate multiple solutions. Examining the variability of each amino acid's possible assignments among a set of near optimal assignments allows one to assess the reliability. Assignments that do not change among the set of top assignments can be trusted more than those that vary. The set of assignments can be used in consensus methods. For instance, the above BILP can be used to generate a consensus assignment by setting each $W(X_{a,\,s})$ to the number of times amino acid a is assigned to spin system s.

3.4.4 Approximate Matching

The input 3D structure may not match exactly the structure from NMR. To account for small differences, extra edges can be added to the CG. For instance, to account for some alpha helices containing 3_{10}-helix turns, edge match variables for H_i^{α}-H_{i+2}^{N} contacts can be added to the BILP, perhaps with score lower than that for H_i^{α}-H_{i+3}^{N} alpha helix contacts. Off by one contacts for beta sheets can also be added; e.g., $H_i^{\alpha} - H_{j+1}^{N}$, $H_i^{\alpha} - H_{j-1}^{N}$, $H_{i+1}^{\alpha} - H_j^{N}$, $H_{i-1}^{\alpha} - H_j^{N}$. Larger conformational changes; however, are more difficult to handle because they require the prediction of non-local contacts.

3.5 Type Prediction Errors

The above graph matching requires an edge match to have the corresponding end point vertices match in amino acid and secondary structure type. If there are type prediction errors, there will be assignment errors. We build from the CR approach by handling such errors, which were ignored. To try to recover from them, we identify assignments that are believed to be reliable, set their variables to 1, and then rerun the algorithm by relaxing the type matching requirement for edge matches. The idea is similar to manual and automated iterative NMR methods, such as MARS and NVR, that identify reliable assignments and then redo the assignment with those assignments removed. Figure 3.4 summarizes our approach, and Figure 3.3 gives the intuition behind our approach. Figures like Figure 3.3 can be produced by our method to gauge the reliability of each assignment.

To determine whether or not a particular residue-spin system assignment should be fixed, we examined the percentage of contacts matched involving that residue given the assignments of the other residues. Due to erroneous assignments, an overly tight criteria for identifying fixed assignments, which we shall refer to as *anchors*, may exclude many correct anchors and result in a large problem size. An overly loose criteria, can result in incorrect anchors that may lead to further incorrect assignments when the algorithm is rerun. We chose a 50% cutoff, and then used progressively tighter criteria. We chose 50% because the majority of the missing edge percentages in our data are below 50% (Table 3.2). The approach that we used for identifying the confident assignments may appear rather weak, but in practice, a very large problem size would have resulted if we had used a tigher criteria. To facilitate manual analysis, the percentage of contacts matched and the number of sequential neighbors connected can be outputted for each residue.

After the BILP is solved with the anchors, some of the anchors may no longer satisfy the anchor criteria because the number of contacts matched might have changed if the assignments for the unfixed residues also changed. If the new set of anchors is different from the previous, the BILP is rerun again with the new set. This is done until the set stabilizes. From our tests, the number of iterations did not exceed 5. Although there is a finite number of possible anchors, it is possible that the loop does not terminate if the same sequence of anchor sets appear in a cycle. This can be detected by remembering previous sets.

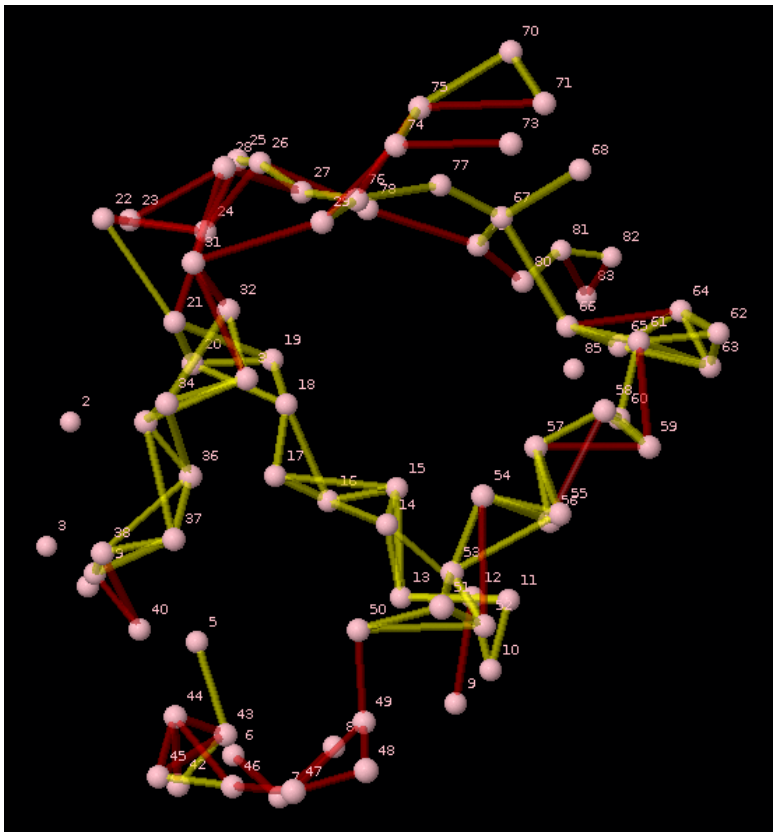

Figure 3.3: Intuition behind fixing assignments. Each pink sphere represents a backbone H^N atom labeled by its residue number. For simplicity, only the H^N-H^N contacts are shown. Contacts in yellow have an NOE assignment; while those in red do not. The amide chemical shift assignment for residues 16, which has all of its contacts assigned, is more likely correct than compared to residue 48, which has no contacts assigned. For the amide chemical shifts assigned to 48, we assume that a type prediction error is a possibility, so we resolve the BILP while allowing these chemical shifts to be assigned to residues of any type. For 16, its backbone assignment would be fixed during the resolving.

For tightening the criteria, we considered sequential neighboring contacts, nonlocal contacts between β-sheet residues, and local helix contacts ($i \pm 5$). We first required only one sequential neighbor to be matched in addition to the 50% criteria. Then we required both neighbors; assignments for residues at the end points will never be fixed. Finally, we additionally required that β-sheet residues have at least one β-sheet contact match, and that α-helix residues have at least one local contact match to a residue before the α-helix residue and one local contact match after. We did not attempt to optimize the set of criteria for fixing assignments because we do not have enough training data for the optimization.

The solution with the optimal score is not necessarily the correct assignment due to noisy edges in the IG, increased ambiguity from relaxing the type matching requirement, incorrect fixed assignments, and inaccuracies in the scoring function in modeling the problem. To account for this, multiple solutions of the initial BILP were generated and each one entered the loop. At the end, we take the best scoring assignment. Alternatively, one can take the consensus assignment as described in Section 3.4.3. Fortunately, from our tests, we found that the score of the correct assignment is near the best scoring, and that the assignments did not differ significantly. To explore possible solutions for the unfixed residues, multiple solutions can be generated on the BILP with variables corresponding to the anchors set to 1.

The following anchor tests illustrate the behavior of anchors under various criteria. These results are based on the new BILP described in Section 4.3, but the general patterns should be similar. Table 3.1 shows the anchor accuracy and the number of anchors for different groups of assignments with different average accuracies per group. A 50% contact match criteria was used to identify the anchors. As the assignment accuracy decreases, the anchor accuracy does not drop significantly, but the number of anchors decreases noticeably. This is expected because fewer assignments will satisfy the anchor criteria when the assignment accuracy is low, and it is unlikely that wrong assignments will satisfy the criteria. All the assignments are for 1KA5. The pattern is similar for other proteins.

Figure 3.5 shows the relationship between the number of anchors and accuracy of anchors for different fractions of contacts matched. Assignments for 1KA5 with accuracies between 35% to 100% were generated. In general, as the required fraction of contacts increases, the number of anchors decreases and the number of incorrect anchors decreases. The decrease in the number of anchors appears more pronounced. Figure

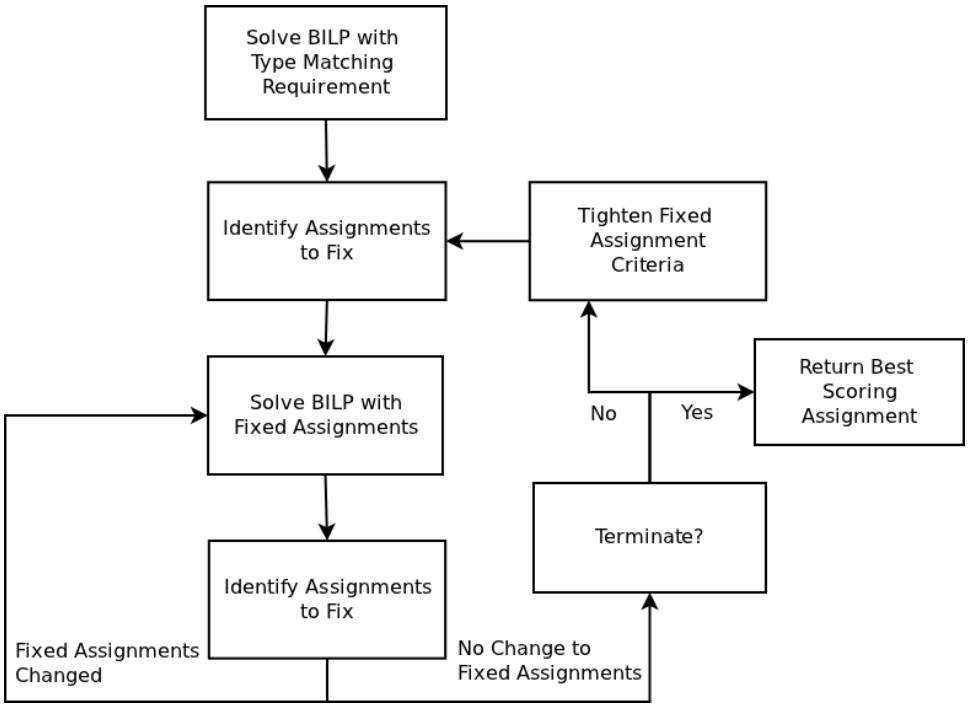

Figure 3.4: Iterative BILP with fixed assignments.

Table 3.1: Number of anchors as a percentage of the total number of spin systems and anchor accuracy for groups of assignments with different average accuracies. A 50% contact match criteria was used to identify the anchors.

Avg. Assignment Accuracy (%)	Anchor Accuracy (%)	Num. Anchors (% of Spin Systems)
100	100	52.2
95.3	100	41.3
90.5	99.8	33.9
84.9	100	25.1
79.6	99.8	19.6
74.7	99.2	15.7
70.9	100	11.6
64.9	99.4	10.4
59.9	100	6.3
54.9	97.5	5.9

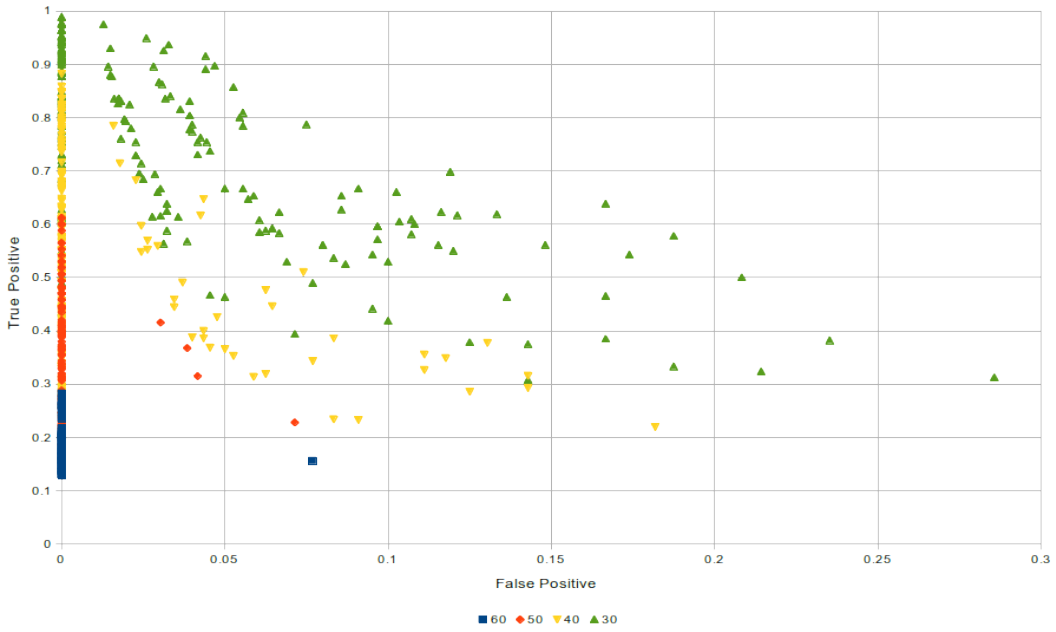

Figure 3.5: True and false positive rates for anchors based on the fraction of contacts matched. The fraction ranges from 30% to 60%. The true positive rate is the fraction of correct assignments identified as anchors, and the false positive rate is the fraction of anchors that are incorrect.

3.6 shows the relationship for the number of sequential neighbors with a contact match. In all cases, the fraction of all contacts matched was set to 50%. Unlike the previous figure, there does not appear to be any significant difference in the number of anchors and accuracy here. When the β-sheet and α-helix contact requirements as described above were added, the differences were also not significant (results not shown). It appears that the 50% criteria is already stringent enough that adding additional requirements does not change the anchors substantially. The pattern is similar for other proteins.

Note that the results here are based on the new BILP that does NOE assignment directly by enforcing that each NOESY peak is associated with at most one contact. In the non-NOE assignment case, a contact match occurs if a pair of vertices in one graph is assigned to a pair of vertices in the other graph, and the edge between the pairs share at least one interaction type. This effectively makes the contact matching requirement less stringent because an NOE can satisfy more than one contact. Therefore, in this case, adding the sequential neighbors and secondary structure anchor criteria should reduce the number

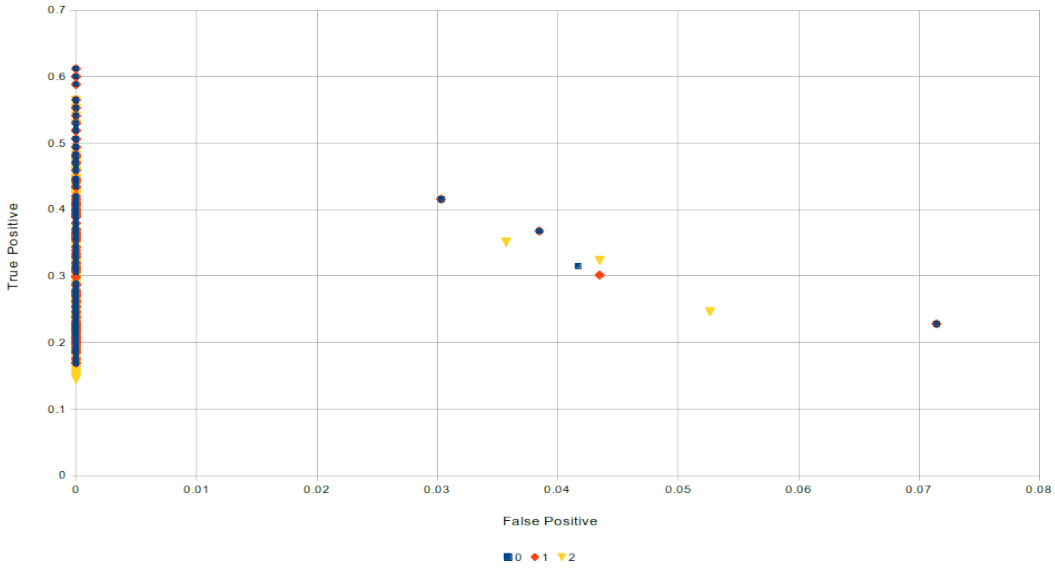

Figure 3.6: True and false positive rates for anchors based on the number of sequential neighbors that have a contact match. The required fraction of all contacts matched is 50% for all cases.

of anchors and decrease the false positive rate. For 1KA5, based on the number of anchors observed, a 50% cutoff in the first BILP is actually closer to a 25% cutoff in the new BILP. For a 25% contact match criteria for 1KA5, the false positive rate was 5.7% and the fraction of correct assignments that are anchors was 84%. When the requirement of two sequential neighbor contacts was added, the values decreased to 4.0% and 50.0%, respectively. When the requirement of secondary structure contacts was also added, the values decreased to 1.8% and 37.5%, respectively.

The above anchor behavior provides insight into our iterative BILP approach. Initially, when the assignment accuracy is low due to type prediction errors, a less stringent anchor criteria ensures that there is a tractable number of anchors. After relaxing the type matching requirement for the non-anchors, the score of the assignment tends to increase as more contacts get matched, and the assignment also tends to have higher accuracy. Because the assignment accuracy is higher, we can use stricter anchor criteria and still have a tractable number of anchors. The number of anchors still tends to decrease, which is necessary for correcting errors in previous anchors. For 1KA5 with both amino acid and secondary structure type prediction errors, the initial number of anchors was 63 out of 85 possible, and the final number was 48 for

the old BILP. Note that we cannot ignore all the type predictions because it would allow each spin system to be assigned to any residue. The resulting BILP will be too large to run. Among the results in Table 3.5, the largest BILP solved had about 150,000 variables and 25,000 constraints. This is for 1YYC for the case with 70 anchors out of 158 possible spin systems.

3.6 Results

We tested the performance of our method on the simulated data used by the CR method, which consisted of 9 proteins. We obtained from the authors 5 interaction graphs, derived from the following NMR structures from the PDB: 1KA5, 1EGO, 1G6J, 1SGO, and 1YYC. The data for 2NBT, 1RYJ, 2FB7, and 1P4W were simulated according to their simulation method described in [123], where only one of the NMR models was used to generate the NOESY peaks. Although the simulated data was derived from one of the models in the PDB file, similar to the CR experiments, we tested the data using every model in the PDB file as the template structure, where the number of models per PDB file ranged from 10 to 32. The structural noise (in RMSD) of the models within each PDB file is given in Table 3.2, which summarizes the test set. Note that the number of spin systems is less than the number of residues due to proline residues and missing HSQC data. For measuring accuracy, we used the number of correct assignments over the number of spin systems. A spin system assignment is incorrect for assignments to the incorrect residue, for assignments to a residue with missing data, and for non-assignments when a correct assignment exists. To control noise, our method automatically increases the distance cutoff at 0.25 Å increments until the noise level is under 8. This gave a small improvement over using a fixed 4 Å cutoff. The cutoffs used ranged from 4 to 4.5Å . We used the same distance cutoffs in the CR software.

For the case with no type prediction errors, Table 3.3 compares our method with the CR method, where the first row of each entry gives our results, while the row below gives the CR's. On 8 of the 9 proteins, our average accuracy on the entire protein was better. In two instances, the accuracy was better by over 20%. Our method was better able to maximize the score as shown in column 4 of the table. In many instances, the score was higher than the score of the correct assignment, which indicates inaccuracies in the scoring.

Table 3.2: Summary of the test set for backbone assignment. From left to right: template structure, number of residues in the template (total/helix/sheet/loop); number of spin systems (total/helix/sheet/loop); number of prolines; noise level (number IG edges per CG edge); percentage of contacts missing in the NMR data (total/helix/sheet/loop); average pairwise RMSD of the models in the template PDB file (total/helix/sheet/loop).

Template	No. Residues	No. Spin Sys	No. PRO	Noise (x)	Missing (%)	RMSD (Å)
1KA5	88/40/23/25	85/39/23/23	1	5.5/5.6/5.9/5.3	21/20/21/22	0.2/0.2/0.1/0.2
1EGO	85/40/19/26	81/40/19/22	3	5.6/5.4/5.8/6.3	22/22/26/19	1.6/1.4/0.9/2.3
1G6J	76/18/22/36	72/18/22/32	3	4.4/3.5/5.1/4.8	33/31/32/35	1.1/0.6/0.4/1.5
1SGO	139/46/28/65	136/46/28/61	3	5.5/4.7/4.0/7.4	41/38/49/40	10.9/7.3/5.5/14.1
1YYC	174/36/72/66	158/36/70/52	10	6.6/5.2/7.5/7.3	38/35/38/40	4.0/2.5/1.6/6.0
2NBT	66/-/16/50	60/-/16/48	5	3.4/-/3.6/3.3	36/-/22/40	3.4/-/1.7/3.8
1RYJ	70/9/27/34	67/9/27/31	2	3.1/2.0/3.1/3.8	28/33/29/25	1.5/1.0/0.9/1.9
2FB7	80/-/32/48	73/-/32/41	7	3.1/-/3.0/3.2	34/-/30/36	5.4/-/2.0/6.8
1P4W	87/66/-/21	82/65/-/17	3	5.5/5.3/-/6.7	31/28/-/40	1.1/0.7/-/1.9

In particular, the contribution of an NOE can get counted twice (see Figure 4.5). For 2NBT, where 40% of loop contacts were missing, we did slightly worse, but the scores were greater than the score of the correct assignment; similarly for the helix residues in 1RYJ. Since amino acids in helices tend to have local contacts with nearby amino acids, in many of our tests, we observed that missing NOE edges and typing errors produced local errors in helices. For 1RYJ, the accuracy for helices using a $(i \pm 2)$ window, *i.e.*, allowing a spin system to be assigned within two residues away from the correct residue, was 100%.

For recovering from type prediction errors, we first tested with only amino acid type prediction errors, and then we tested with both amino acid and secondary structure typing errors. For 1KA5, 1EGO, 1G6J, 1SGO, and 1YYC, we ran RESCUE Version 1 [93] on the experimental proton chemical shifts from the protein's entry in the BMRB. Table 3.4 gives the results for amino acid type prediction errors. For comparison, we included the result for when type prediction errors were ignored. Such errors were ignored when type matching was strictly enforced for edge matching and the iterative algorithm was not used to correct for the errors. In general, error correction resulted in large improvements. For 1G6J, the amino acid typing accuracy was high, so the improvement was minimal. For 1YYC, the improvement was significant even though the typing accuracy was low. The accuracy, however, varied substantially depending on the model used as the template. Nevertheless, the template with the best score yielded an accuracy of 89.9%,

Table 3.3: Comparison between the BILP model and the CR method for correct amino acid and secondary structure typing. For each protein, the first row gives our results, while the second row gives the CR's. From left to right: template structure; average accuracy over all the models (total/helix/sheet/loop); accuracy ranges (total/helix/sheet/loop); number of times the assignment score was greater than, less than, or equal to the score of the correct assignment.

Template	Avg Acc. (%)	Acc. Range (%)	Times Score >,<, = Ref
1KA5	100/100/100/100	100/100/100	0, 0, 16
	94/100/76/100	98-93/91-74/100	0, 16, 0
1EGO	98/100/100/93	100-97/100/100/100-90	15, 0, 5
	96/96/100/93	100-92/100-90/100/100-79	4, 12, 4
1G6J	97/100/100/94	100-95/100/100/100-90	25, 2, 5
	91/100/ 87/88	97-89/100/100-86/100-85	0, 32, 0
1SGO	96/97/100/94	100-86/100-95/100/100-70	13, 3, 4
	80/95/95/62	88-71/100-87/100-86/76-45	0, 20, 0
1YYC	97/99/96/98	100-93/100/100-91/100-92	17, 0, 3
	72/92/62/72	76-67/100-89/69-53/79-64	0, 20, 0
2NBT	91/-/98/88	96-85/-/100-93/95-79	10, 0 , 0
	92/-/95/90	100-88/-/100-88/96-82	1, 9, 0
1RYJ	97/98/96/96	97-94/100-88/96/96-93	20, 0, 0
	82/100/70/86	82-75/100/70/88-72	0, 20, 0
2FB7	96/-/97/96	100-91/-/100-93/100-90	7, 0, 3
	92/-/94/90	95-88/-/100-94/95-83	0, 10, 0
1P4W	99/100/-/97	100-97/100/-/100-88	4, 0, 16
	77/77/-/77	91-63/91-63/-/90-58	0, 20, 0

Table 3.4: Assignment accuracy for amino acid typing errors and correct secondary structure typing. From left to right: template structure; average accuracy for strict type matching enforcement; average accuracy for iterative error correction over all the models (total/helix/sheet/loop); accuracy ranges for iterative error correction (total/helix/sheet/loop); amino acid typing accuracy; number of times the assignment score was greater than, less than, or equal to the score of the correct assignment. Values in parenthesis give the accuracy within an $i \pm 2$ window.

Template	Avg Acc Strict (%)	Avg Acc Iter (%)	Range Acc Iter (%)	A.A. Typing Acc (%)	Times Score >, <, = Ref
1KA5	86	100/100/100/100	100/100/100/100	89	0, 0, 16
1EGO	86	94/92(99)/100/94	100-91/100-87/100/100-90	90	15, 3, 2
1G6J	92	94/100/93/91	97-87/100/100-90/100-78	96	7, 25, 0
1SGO	82	92/90(100)/95/93	96-87/100-84/100-82/96-83	92	7, 13, 0
1YYC	59	77/86 (92)/81/66	94-68/100-58/100-52/90-50	79	0, 20 , 0

which increased to 94.1% if an ($i \pm 2$) window was considered. This indicates that using multiple templates, such as those generated by normal mode analysis [7], can improve accuracy. In these tests, for the tightest anchor criteria, we required both sequential neighbors to have contact matches, but we did not require nonlocal β-beta sheet and local α-helix contact matches.

The standard method for predicting secondary structures from ${}^3J_{HNH\alpha}$ coupling constants [122] is similar to the following: if the coupling value is between 2.5 and 5.5, the spin system is predicted as helix. If the value is between 8 and 11.5, the spin system is predicted as β-sheet; otherwise, it is predicted as loop. From a test set of the following BMRB entries with accession numbers 4267, 4071, 2151, 4458, 4376, 4136, 4784, 4347, 4163, 4297, plus ubiquitin experimental values from the literature [116], we obtained an average typing accuracy of 60% with a range of 50-69%. This will likely be too low for resonance assignment, so we classified coupling constants into classes consisting of two secondary structure types, which dramatically increased the average accuracy at the cost of increased problem size and increased ambiguity. For values less than 6.5, we classify it as helix and loop; otherwise we classify it as β-sheet and loop. With this, we obtained an average accuracy of 92% with a range of 82-100%.

Table 3.5 gives the results for both amino acid and secondary structure typing errors. The secondary structure class prediction errors were introduced at random to yield accuracies below 92%. Unlike the previous tests, for the tightest anchor criteria here, we required nonlocal β-beta sheet and local α-helix contact matches as described in Section 3.5. Handling both typing errors resulted in a significantly larger

Table 3.5: Assignment accuracy for both amino acid and secondary structure typing errors. From left to right: template structure; accuracy for strict type matching; accuracy of the best scoring model for iterative error correction (total/helix/sheet/loop); amino acid typing accuracy; secondary structure typing accuracy; percentage difference in score of the best scoring assignment compared to the correct assignment (+ means score of our assignment was higher). Values in parenthesis gives the accuracy in a ($i \pm 2$) window.

Template	Acc Strict (%)	Acc Best Score Iter (%)	A.A. Typing Acc (%)	S.S. Typing Acc (%)	Diff Ref Score (%)
1KA5	72	100/100/100/100	89	91	0
1EGO	65	97/95(100)/100/100	90	85	-1.5%
1SGO	63	88/82(91)/96/88	92	87	-3.0%
1G6J	75	91/100/86/90	96	90	+0.5%
1YYC	40	70/91/71/53	79	91	-3.1%

search space. For instance, for 1YYC, it took over 40 hours versus less than 9 minutes if there are only amino acid typing errors. For the convenience of time, we tested each target using only the first model in the template. Column 2 of Table 3.5 shows that low assignment accuracies can result if spin system type prediction errors are ignored, even if the type prediction accuracy is high. Relative to Table 3.3, where the accuracies are 95% or better, the accuracies here for the no error handling case are between 40-75%. Error handling improved this to 70-100%. For 1KA5, the assignment accuracy did not change from the previous test. For 1EGO, the accuracy actually improved because of the tighter criteria for fixing assignments. The larger 1SGO struggled to maximize the score, but the accuracy is still much higher than without the iterative algorithm. For 1YYC, its large size combined with its low amino acid typing accuracy, produced poor quality anchors, but there is still a large improvement over the case without the iterative algorithm.

Ubiquitin, a commonly used protein to test resonance assignment methods, was used to test the method on peak lists from experimentally derived spectra as opposed to simulated data. We obtained ^{15}N HSQC, ^{15}N TOCSY, and ^{15}N NOESY data from Richard Harris's The Ubiquitin Resource Page [44]. Peaks were picked manually by inspecting the spectra with SPARKY [37]. Results for peaks picked automatically by the automated peak picking tool PICKY [4] are presented in Section 3.7.4. Ubiquitin has 76 residues and 3 prolines. The reference solution has 70 assigned residues. The noise level is 4.6 at a 4 Å cutoff, and the missing edge percentage is 28.3%. HSQC peaks without an H^{α} chemical shift were correctly filtered out as noise. For amino acid type prediction, RESCUE performed poorly, giving an accuracy of 68.6%. The errors appear to be due to missing peaks that are hidden by peak overlap. Using a higher resolution

TOCSY spectrum might improve accuracy. We performed the typing manually using each type's expected number of proton chemical shifts and their expected range of values. Manual typing gave an accuracy of 90%, where the average number of possible amino acid types per spin system is 3.3 with a range of 1 to 8 (Appendix 5). We used the results of manual typing for assignment. RESCUE version 2 [70], yielded an accuracy of 90% as well, but only after the peaks were manually assigned to their spin systems. Amino acid type prediction can be performed simultaneously with manual peak picking and spin system compilation, so we chose to do the type prediction manually to save time versus using RESCUE after the spin systems have been compiled. Section 3.7.3 describes our modification of the RESCUE method for simultaneous amino acid type prediction and spin system compilation.

Experimental ${}^3J_{HNH\alpha}$ coupling constants were obtained from the literature [116]. Eight spin systems did not have J-coupling values, so their predicted class included all three secondary structure types. The accuracy of secondary structure type prediction was 91%, yielding a combined typing accuracy of 83%. Model 1 from PDB 1D3Z was used as the template structure. The template structure was not derived from the NMR data. The best scoring assignment has accuracy 87.1%, with 64.3% on α-helix (85.7% with $i \pm 2$ window), 95.7% on β-sheet, and 90.0% on loops. When type prediction errors were ignored, the overall accuracy dropped to 59%.

Although the accuracy for helix residues is low, many of the errors are due to a ± 1 assignment position error due to the HSQC peak of a nearby amino acid not being present in the NMR data. We also obtained a consensus assignment by generating 10 solutions from the best scoring assignment with anchors based on the tightest anchor criteria described in Section 3.5. Consensus gave an accuracy of 91% (62 out of 68 predictions) with 78% for helices (92% $i \pm 2$) and the other types unchanged. This result on non-simulated data is comparable to the result on the 1G6J simulated data, which is also for ubiquitin, except that this test is slightly more difficult. This data set is missing HSQC peaks for 2 residues in addition to the prolines. The simulated data is missing only prolines and the initial methionine. The local assignment errors in helices show the limitations of using only backbone proton contact information. Since our BILP model can accommodate different sources of information, it is of interest to test the relative contribution of each source to assignment.

3.7 Assignment From Automatically Picked Peaks

We tested our method on ubiquitin on automatically picked peaks. Using such peak lists is more challenging because of increased noise and decreased sensitivity. This will be evident later when we compare the peak lists of the manual and automated cases. The software PICKY was used to automatically pick peaks from the same ubiquitin spectra as the manual case. Unlike the manual case, we did not use J-coupling constants from 3D HNHA for two reasons. One, we did not have the spectra, and more importantly, obtaining a high yield of J-coupling constants from HNHA becomes more difficult as the size of the protein increases due to spectral overlap. In one study on calcium-free calmodulin [63], only 98 coupling constants were obtained out of 132 possible. The increased noise from automatically picked peaks would exacerbate this problem. J-coupling values do not seem to be commonly available in the BMRB, especially for large proteins. Using G-matrix Fourier transform NMR experiments might alleviate these limitations [12], but more studies are needed. Using HNHA to predict secondary structure is not as popular as chemical shift-based methods that use ^{13}C-labeled data, such as TALOS [103]. The reason is that chemical shift-based methods are faster, more accurate, and less sensitive to protein size problems [16]. However, these methods require carbon chemical shifts.

Since we did not have secondary structure information to limit the number of possible residues for the chemical shifts, we used homologous assignment information from a homologous protein, which is typically available in protein mutant studies. By using homology information, the number of possible residues is reduced, which speeds up the assignment process. Figure 3.7 summarizes our approach. After peak picking, TOCSY peaks with similar amide chemical shifts were grouped together to give the side chain proton chemical shifts of some spin system (Section 3.7.1). NOESY peaks with similar amide chemical shifts were grouped together to give the interactions from a given spin system. After grouping, the groups of TOCSY and NOESY peaks were calibrated, so that their amide chemical shifts match the corresponding HSQC peak (Section 3.7.2). The grouped TOCSY peaks were then used to identify possible amino acid types and to identify the H^{α} chemical shifts (Section 3.7.3).

The NOESY interactions, predicted amino acid types, H^{α} chemical shift assignments, and homologous

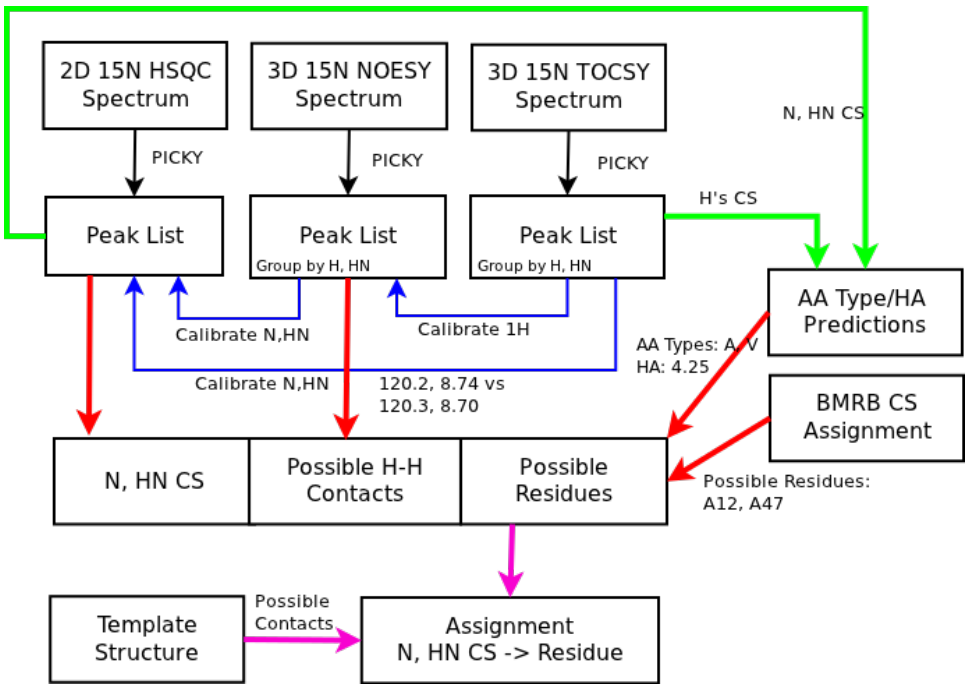

Figure 3.7: Automated backbone assignment with homologous assignment and structure. CS denotes chemical shift. Arrows denote the flow of information, starting from NMR spectra. Different colored arrows represent different steps. The possible residues and the possible contacts from the NOESY peak list are used to build the IG. The contacts in the template structure are used to build the CG. The backbone assignment step is iterative (not shown). Fixed assignments are used to reduce the set of possible residues for the other spin systems in the IG.

assignment information were used to created the IG. The homologous assignment was obtained from the BMRB (described below). The IG and CG were passed as input to the assignment step, which uses the BILP described earlier. In the first iteration, only the homologous assignment information, which is assumed to be more reliable than the amino acid type predictions, was used for identifying possible residues for each spin system. The scoring function was augmented to take into account the amide chemical shift difference between the spin systems and BMRB chemical shifts. It is similar to the match score described earlier, but for $X_{a,s}$. In subsequent iterations, when anchors are used, both the amino acid type predictions and the BMRB chemical shifts were used to identify the possible residues for the unfixed residues.

To identify which BMRB entry best matched the HSQC peaks, we used an amide chemical shift distance scoring function together with the Hungarian algorithm [64] to score the match of the chemical shifts in the BMRB entry to the HSQC peaks. This is simply a problem of matching a pair of 2D point sets. If the number of amide chemical shifts were different, dummy nodes were added. Assignments to dummy nodes were given the worst scores, so they could be ignored. Among 6 ubiquitin candidates, BMRB entry 15410 (human ubiquitin) had the best score, and yielded an assignment accuracy of 100%, with 70 correctly assigned residues by the Hungarian algorithm alone. Therefore, we used the next best scoring entry, 4769 (yeast ubiquitin), which had 54 correct assignments. Any residue with δ_N , δ_{H^N} that matched an HSQC peak within a 0.75 ppm, 0.1 ppm tolerance was a possible residue for the spin system. If a particular spin system still had no possible residues, the thresholds were doubled for this spin system. This yielded an average of 2 residues per spin system with a range of 1 to 6. 62 spin systems had the correct residue in its list of possible residues. However, because of noise peaks and erroneous H^{α} assignments, the assignment accuracy can be less.

Chemical shifts are extremely sensitive to the environmental conditions and to conformation changes from mutations. If we had used BMRB entry 5387, which is also human ubiquitin, but in a different solvent, the Hungarian assignment would have given 41 correct. The RMSD between the two human ubiquitins is only 1.8 Å, but about 40% of the chemical shifts are different. Therefore, use of the BMRB is limited to similar proteins in similar conditions. This situation occurs often in chemical shift mapping, where the chemical shifts of the unbound form of the protein is used to study different drug molecules.

Such chemical shift assignment information is invaluable in such studies.

The list of possible residues for each spin system, or just one residue for fixed spin systems was used to filter edges in the IG that lack support from the contacts. For the edge between a pair of spin systems with possible residue sets R_1 and R_2, if there did not exist at least one interaction in this edge that matched any contact from the set of contacts in $R_1 \times R_2$, then the edge was deleted. If a NOESY peak did not match any IG edge, it was deleted. For filtering, residues involving alpha helices and beta sheets in the input protein structure were considered in contact if they were 6.5Å apart. We used a cutoff larger than 4Å to account for structural variability because the input structure was not derived from the NMR data. For contacts involving loops, we used a 12Å cutoff because loop regions are, in general, structurally more variable. Prior to pruning, the noise ratio was 7.8. Afterwards, it was 2.7. Without noise filtering, the noise ratio was over 12. Pruning resulted in 287 NOESY peaks, compared to 1552 peaks in the unfiltered peak list. Fortunately, edge pruning did not remove any correct NOESY peaks and edges. Noisy HSQC peaks were identified by sorting the peaks by intensity, and then using a 4 standard deviation cutoff to remove low intensity peaks. After removal, the number of HSQC peaks was 84 versus 90 in the initial list. No peaks corresponding to the backbone amides were removed.

In the other direction, the IG edges were used to prune the list of possible residues that lack NOE support. Using all 4Å non-loop contacts from residue R, we tested spin system S from R's possible spin systems for NOE support. If there did not exist any matching IG edge among all possible pairs of spin systems involving S and the possible spin systems of the residues at the other end of the contacts, then R was removed from S's list of possible residues. This pruning step resulted in one correct residue removed.

By using automatically picked NOESY peaks, the missing edge percentage increased to 36-45%, compared to 28% for manual. The percentage is a range because our iterative assignment process rebuilds the graphs using the assignment results from the previous iteration. The manual NOESY peak list had 376 peaks versus 287 here. From manual inspection of the automatic peak list, it appeared that some low intensity peaks corresponding to some H^N- H^N contacts were not picked. Perhaps an approach of integrating peak picking directly in the assignment step would help. For example, when deciding whether or not a pair of spin systems have an NOE that matches some contact, the peak picker could adjust the

noise threshold according to the distance of the contact. In general, peak picking is a non-trivial problem and an urgent target for improvement [118].

The consensus assignment after the first assignment step yielded 61 correct assignments out of 62 assignments made. The ILP was run for 4 iterations; after which, the best score did not change. The final consensus assignment yielded 67 correct assignments out of 69, for an accuracy of 97%. Five spin systems with incorrect BMRB residue matches were assigned correctly due to correct amino acid type prediction. One residue with incorrect BMRB matches and incorrect amino acid type predictions was correctly assigned due to relaxing the type matching requirement for the unfixed residues. One error was due to an assignment for a residue not in the reference solution, but which was clearly incorrect. The other error was for the residue adjacent to that residue.

3.7.1 Peak Grouping

We took a naive approach for grouping peaks by mimicking the manual way. Using the amide chemical shifts, 3D peaks matching at least one HSQC peak within a 0.5, 0.05 ppm tolerance were projected onto a 1D line representing the proton chemical shift. Peaks on the same line/strip were treated as a group if they were within 0.5, 0.05 ppm of each other. A peak can occur in more than one strip due to peak overlap. Ambiguities were resolved using the approach in Section 3.7.3. Peaks not in any strips were removed. For TOCSY peaks only, an additional processing step was done. Peaks within a strip were split into at most 4 possible groups defined by the maximum, minimum N and maximum, minimum H^N chemical shift values of the peaks in the strip (Figure 3.8). Peaks were then assigned to the closest corner. If there were groups of size at least 2 (since we want each strip to have at least $H^N$and H^α), and the difference in the assigned corners of the groups was larger than 0.25 ppm for N or 0.025 ppm for H^N, the strip was split. Using statistics from the BMRB, strips with no H^N and putative H^α atoms were removed. H^αs typically have chemical shift values between 2 and 6 ppm. Initially there were 2019 TOCSY peaks. After processing, there were 311. Strips include other proton chemical shifts besides those for H^N and H^α because they are needed for amino acid type prediction.

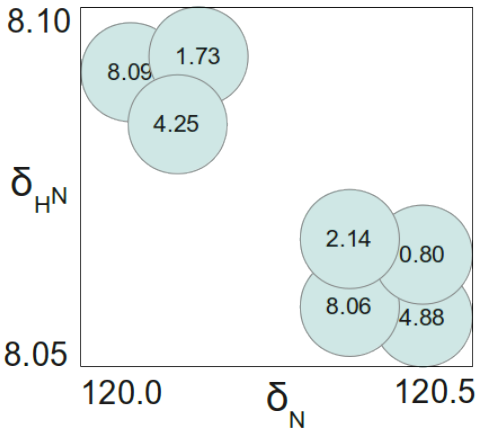

Figure 3.8: Splitting the peaks in a TOCSY strip using the maximum, minimum N and maximum, minimum H^N chemical shift values of the peaks in the strip. Here, there appears to be 2 different groups. One with backbone amide chemical shifts near 120.5, 8.05 ppm, and the other with shifts near 120.0, 8.10 ppm. The proton chemical shift of each 3D peak is given inside the circle.

3.7.2 Peak List Calibration

The amide chemical shifts δ_N , δ_{H^N} were used to group NOESY and TOCSY peaks together, and then assign each group to its corresponding HSQC peak (by representing each group with a 2D peak). The amide chemical shifts of each group do not necessarily correspond exactly to the values of their HSQC peak. Finding offsets to add to each peak to make their amide chemical shifts match their HSQC peak is called *calibration*. If calibration can be done with a common global offset added to every peak in the NOESY or TOCSY, one can simply look for an unambiguous HSQC peak associated with some group, assuming the chemical shift difference is not too large. This can be done manually quite easily using computer-aided assignment software, such as CARA [58]. For our case, a global offset approach was sufficient. In general, a global offset is sufficient unless, when measuring the different spectra, the experimental conditions change that results in a change in structure.

If there is local noise about this global offset, and the noise is small relative to this offset, one can take the average offset using the HSQC peaks that can be unambiguously associated with some group. If desired, manual fine tuning can then be performed to get exact matches. As the size of the protein increases, finding unambiguous matches becomes more difficult. As an alternate method, we propose a

maximum clique approach with rectangles. Each HSQC peak is associated with a rectangle with width equal to the N chemical shift match tolerance and height the H^N tolerance. We used 0.5 ppm and 0.05 ppm, respectively. For each group of NOESY/TOCSY peaks with a common amide chemical shift, we associate a rectangle with width equal to some user-specified maximum δ_N calibration offset and height equal to some δ_{H^N} maximum offset. We shall refer to these limits as search tolerances. Figure 3.9 describes the method, which is based on the maximum clique of the intersection graph of rectangles. This problem can be solved in O($n$log$n$) time [48]. An analogous method can be used to calibrate the side chain proton chemical shifts, δ_H, of NOESY peaks to the side chain protons of TOCSY peaks. Each NOESY δ_H is represented by an interval centered about its chemical shift value and with some user-specified maximum width for the offset. Each TOCSY peak is represented by an interval with width equal to the H chemical shift match tolerance, which is 0.05 ppm in our case. Instead of a maximum clique problem on the intersection graph of rectangles, we have a maximum clique problem on the intersection graph of 1D line intervals. This can be solved in linear time [77].

Table 3.6 compares the average unambiguous match method to the rectangle clique. Using 143 amide chemical shifts from BMRB entry 15624, we generated for each peak a corresponding NOESY group with a common global noise offset and a local noise offset that differed for each peak. This was done a 1000 times using different local noise generated at random. $\Delta\delta_{NH}$ from Equation 2.2 was used to measure the calibration error. In all tests, we used 1.0, 0.1 ppm for the search tolerance. If the local noise is small (uniformly distributed within $\pm$0.1, 0.01 ppm), the unambiguous method has smaller error, but when the local noise increased to 0.15, 0.015 ppm, the method resulted in some peaks no longer matching its corresponding HSQC peak according to the 0.5, 0.05 ppm threshold. The clique method did not have this problem. We also tried different combinations of positive and negative global offsets, but the results were similar. Generating missing and noise peaks in the NOESY group peaks (10-20% missing, 10-20% noise) did not increase the error significantly in either case.

For ubiquitin, the local noise was negligible, so both methods worked fine. To remove the influence of noise peaks, we used only the top 50% most intense NOESY and TOCSY peaks to determine the calibration offsets. If calibration was not performed for ubiquitin, the missing edge percentage increased

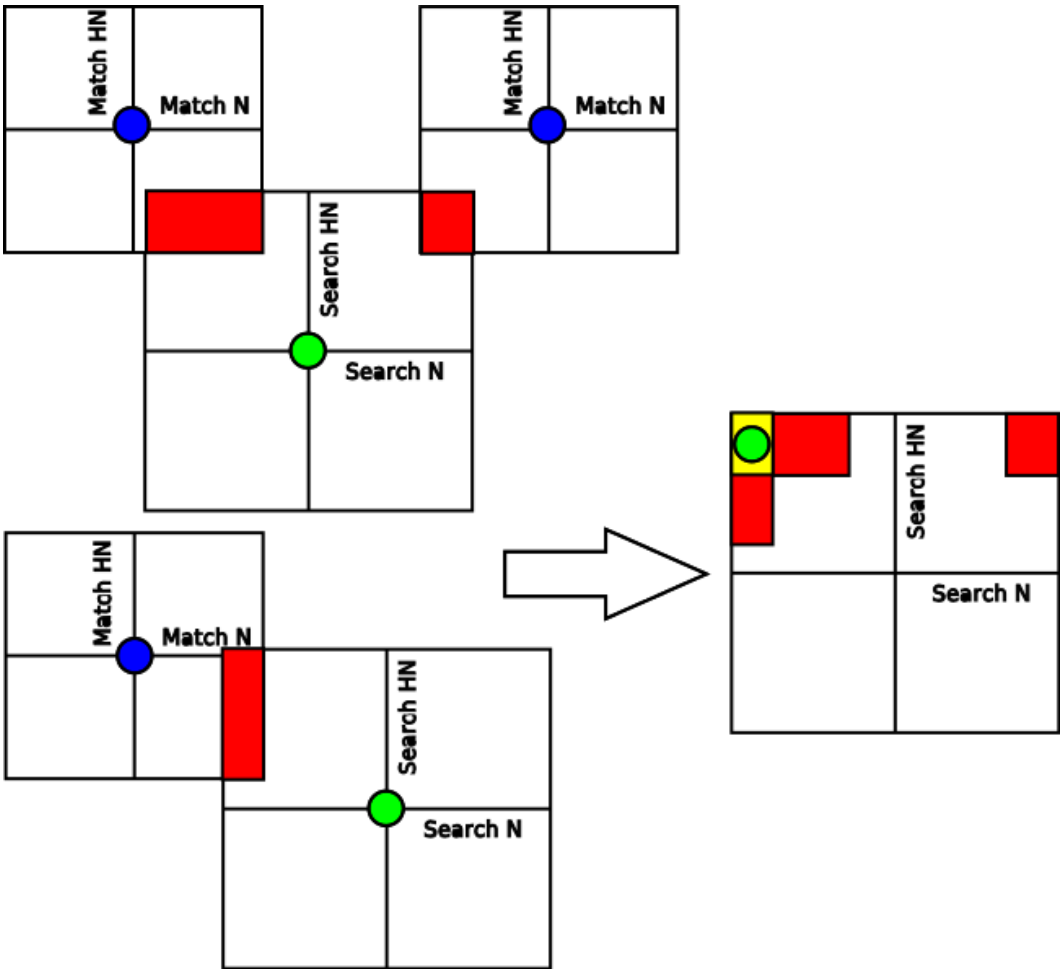

Figure 3.9: Rectangle intersection approach for calibrating the green peaks to the blue peaks. Left of the arrow: For each green peak rectangle, we find all blue peak rectangles that intersect it. This is done by first subtracting the centroid of the green peak to all intersecting peaks, so that the green peak is at the origin. Search N, Search H^N give the allowable amide chemical shift offsets to add to the green peaks. Blue peaks match the green peak if the green peak has amide chemical shift within Match N, Match H^N of the blue. The red region gives the intersection. We illustrate this for 2 green peaks. Right of the arrow: All the red regions are then intersected with a rectangle with dimensions Search N, Search H^N and centered at the origin. The region with the most overlap is in yellow. This region gives the global offset to add to every green peak. If all green peaks are moved to the yellow, 2 blue peaks get matched as opposed to 1 if the green is moved to the red region on the right.

Table 3.6: Comparison between the rectangle clique and average unambiguous match calibration methods. Local noise was generated randomly in the interval defined by column 2; allowing for positive and negative values. For each known peak to peak association, the average $\Delta\delta_{NH}$ was computed. Avg Missed is the number of peaks out of 143, on average out of the 1000 trials, whose calibrated chemical shifts no longer match those of their associated peak based on the match tolerances of 0.5, 0.05 ppm for N, H^N.

		Clique		Unambiguous	
Global Offset N, H^N	Local Noise (+/-) N, H^N	Avg Dist	Avg Missed	Avg Dist	Avg Missed
0.9, 0.09	0, 0	0.44	0	0	0
0.9, 0.09	0.1, 0.01	0.44	0	0	0
0.9, 0.09	0.15, 0.015	0.44	0	0.54	44.2
0.9, 0.09	0.2, 0.02	0.45	0	0.81	65.9

by 12%. Since the removal of peaks in the peak grouping step may change calibration, calibration and grouping were repeated until no additional peaks were removed. Peak list processing is known to have a significant impact on assignment accuracy [88]. Another calibration approach that is worth investigating is the Hausdorff distance from the computer vision community [47].

3.7.3 Amino Acid Type Prediction and H^α Assignment

Our BILP for backbone assignment requires a list of possible residues for each spin system, and the chemical shifts representing the H^α atoms. One way to identify the possible residues is to predict the possible amino acid types; e.g. GLY, ALA. RESCUE 2 mimics the manual typing approach of comparing the chemical shifts to their expected values in the BMRB. However, it assumes that the spin systems have already been formed. In our case, forming spin systems requires grouping all TOCSY peaks with a common amide chemical shift to get all the proton chemical shifts of the residue. An automated approach was given in Section 3.7.1, but it allows a peak to occur in more than one strip due to peak overlap. Therefore, we extended the Bayesian scoring model in RESCUE 2 to perform not only amino acid type prediction, but also unambiguous TOCSY peak grouping, and H^α assignment using the ambiguous strips of TOCSY peaks as input. Unlike the side chain assignment problem, we do not have an *a priori* backbone assignment, so we can obtain only a distribution of possible assignments. For our problem, we considered only the distribution of H^α chemical shifts for each spin system. To our knowledge, we are not aware of any

other automated approach that can perform amino acid type prediction and H^{α} prediction simultaneously without an *a priori* backbone assignment.

To obtain the amino acid type predictions, we sampled the posterior probability distribution of $\vec{X}$, which is the set consisting of predictions (i, R), where the residue represented by the amide chemical shift of HSQC peak i is assigned to amino acid type R. Letting $\vec{H}$ represent the set of all HSQC peaks and $\vec{T}$ represent the TOCSY peaks, and applying Bayes' rule, we have

$$\begin{aligned} P(\vec{X} \mid \vec{H}, \vec{T}) &= \sum_{\vec{Z}} P(\vec{Z}, \vec{X} \mid \vec{H}, \vec{T}) \\ &= \frac{1}{P(\vec{H}, \vec{T})} \sum_{\vec{Z}} P(\vec{Z}, \vec{X}, \vec{H}, \vec{T}) \\ &= \alpha \sum_{\vec{Z}} \big(P(\vec{T} \mid \vec{H}, \vec{Z}, \vec{X}) \times P(\vec{H} \mid \vec{Z}, \vec{X}) \times P(\vec{Z} \mid \vec{X}) \times P(\vec{X}) \big) \end{aligned} \tag{3.2}$$

where α is a normalization constant to denote $P(\vec{H}, \vec{T})$, which is related to the input peak lists, and $\vec{Z}$ is the set of all possible assignment cases, defined by the following sets $\vec{A}$, $\vec{M}$, and $\vec{N}$. $(i, j, R_k) \in \vec{A}$ denotes an assignment of TOCSY peak j to proton R_k (e.g. H^{α}, H^{β}) of some residue represented by HSQC peak i and with amino acid type R. $(i, R_k) \in \vec{M}$ denotes proton R_k assigned to HSQC peak i and missing its TOCSY peak. $j \in \vec{N}$ denotes TOCSY peak j is noise. Denote δ_j to collectively represent the chemical shift values of some peak, either TOCSY or HSQC. The summation in Equation 3.2 can be expanded to give

$$\sum_{(\vec{A},\vec{M},\vec{N})} \Bigg(\Big(P(\vec{T}_A \,|\, \vec{A},\, \vec{X}_A) \times P(\vec{A} \,|\, \vec{X}_A) \Big) \times P(\vec{M} \,|\, \vec{X}_M) \times \Big(P(\vec{T}_N \,|\, \vec{N}) \times P(\vec{N}) \Big) \times \Big(P(\vec{H} \,|\, \vec{X}) \times P(\vec{X}) \Big) \Bigg) =$$

$$\sum_{(\vec{A},\vec{M},\vec{N})} \Bigg(\prod_{i,\, j,\, R_k \in \vec{A}} \Big(P(\delta_j \,|\, \overline{R_k \text{ missing}},\, \overline{j \text{ noise}}) \times P(\overline{R_k \text{ missing}},\, \overline{j \text{ noise}} \,|\, i,\, R) \Big) \times \prod_{(i,\, R_k) \in \vec{M}} \Big(P(R_k \text{ missing} \,|\, i,\, R) \Big) \prod_{j \in \vec{N}} \Big(P(\delta_j \,|\, j \text{ noise}) \times P(j \text{ noise}) \Big) \times \prod_{(i,\, R) \in \vec{X},\, i \in \vec{H}} \Big(P(\delta_i \,|\, i,\, R) \times P(i,\, R) \Big) \Bigg) \tag{3.3}$$

where the top bars $^-$ denote negation. $\vec{T}_A$ denotes the TOCSY peaks in $\vec{A}$, and $\vec{X}_A$ the $(i,\, R)$ amino acid type assignments in $\vec{A}$; similarly for $\vec{M}$ and $\vec{N}$. Since the space of $\vec{Z}$ is large, we considered only the top scoring assignments rather than all. That is, we first maximize the product inside the summation with respect to $\vec{X}$ and $\vec{Z}$, and then sample again to obtain the next largest assignment, and so on. For each HSQC peak, the number of times it is assigned to each amino acid type is recorded. The number of times each TOCSY peak is assigned to H^{α} of each amino acid type that is assigned to an HSQC peak is also recorded. Since multiplying many fractions may result in a product that gets rounded to 0, instead of maximizing, we take the negative logarithm and minimize. This replaces the product with a sum of terms; i.e., $-\log \prod_j P_j = -\sum_j \log P_j$. By using binary variables to denote whether or not a particular assignment is selected, we can use a BILP to sample the top scoring assignments; e.g., $-\sum_i x_i \log P_i$, where x_i equals to 1 if an assignment, denoted by i and consisting of a specific TOCSY peak, HSQC peak, proton type, and amino acid type, is selected. The $-\log P_i$ terms serve as objective function coefficients. These terms are non-negative, where higher probabilities is associated with smaller values. This sum is minimized when all the variables are zero, which is not what we want, so the coefficients are shifted by subtracting each coefficient by the largest coefficient value, so that all coefficients are negative. RESCUE 2, however, does not need such a global optimization approach because when given the peak groupings, each spin system can be amino acid typed independently of each other. Compared to Markov chain Monte Carlo approaches, our sampling approach is deterministic. Both approaches have issues with determining the required number of samples. In our case, the BILP gap parameter determines the number of samples.

Typing BILP

The decision variables and their objective function coefficients ($-log$) are

- $X_{i,\,R}$ is set to 1 if amino acid of type R is assigned to HSQC peak i. Only peaks whose chemical shifts are within 3.5 standard deviations of the mean values of R are considered (3.5 for 99.9% confidence interval).

- $C(X_{i,\,R}) = P(\delta_i \,|\, i,\, R) \times P(i,\, R) = G(\delta_i^N, \mu_R^N, \sigma_R^N) \times G(\delta_i^{HN}, \mu_R^{HN}, \sigma_R^{HN}) \times \frac{count(R)}{len}$, where G is the Gaussian density function with mean μ and standard deviation σ obtained from BMRB statistics for type R (as used in RESCUE 2). $count(R)$ is the number of residues of type R in the protein sequence, and len is the length of the sequence.

- $X_{\delta_j,\,R_k,\,i}$ is set to 1 if TOCSY peak j is assigned to proton R_k, and HSQC peak i is assigned to type R. Only TOCSY peaks whose δ_j^N and δ_j^{HN} is within 0.5, and 0.05 ppm of those of i , and whose δ_j^H is within 3.5σ of R_k are considered.

- $C(X_{\delta_j,\,R_k,\,i}) = P(\delta_j \,|\, \overline{R_k \,\text{missing}},\, \overline{j \,\text{noise}}) \times P(\overline{R_k \,\text{missing}},\, \overline{j \,\text{noise}} \,|\, i,\, R) = G(\delta_j^N, \mu_R^N, \sigma_R^N) \times G(\delta_j^{HN}, \mu_R^{HN}, \sigma_R^{HN}) \times G(\delta_j^H, \mu_R^H, \sigma_R^H) \times countBMRB(\overline{R_k \,\text{missing}}) \times (\frac{I_j}{max_j})$, where $countBMRB(\overline{R_k \,\text{missing}})$ is the number of chemical shifts for proton R_k in the BMRB statistics divided by the maximum number of proton chemical shifts in the BMRB for type R. max_j is the largest intensity of the TOCSY peaks nearby j according to a chemical shift threshold. I_j is the intensity of peak j.

- $X_{R_k,\,i}$ is set to 1 if proton R_k is missing its peak, and R is assigned to HSQC peak i.

- $C(X_{R_k,\,i}) = P(R_k \,\text{missing} \,|\, i,\, R) = 1 - countBMRB(\overline{R_k \,\text{missing}})$.

- X_j is set to 1 if TOCSY peak j is assigned as a noise peak.

- $C(X_j) = P(\delta_j \,|\, j \,\text{noise}) \times P(j \,\text{noise}) = \frac{1}{K} \times P(j \,\text{noise}) = \frac{1}{K} \times (1 - \frac{I_j}{max_j})$, where K is the product of the N, H^N, and H chemical shift ranges. We used $33 \times 4 \times 11$.

The constraints are

$$X_{i,\,R} = \sum_{j} X_{\delta_j,\,R_k,\,i} + X_{R_k,\,i} \quad \forall i \in \vec{H},\, \forall R_k \in R \tag{3.4}$$

$$\sum_{R_k,\,i} X_{\delta_j,\,R_k,\,i} + X_j \;= 1 \quad \forall j \in \vec{T} \tag{3.5}$$

$$\sum_{R} X_{i,\,R} \;\leq\; 1 \quad \forall i \in \vec{H} \tag{3.6}$$

$$\sum_{i} X_{i,\,R} \;\leq\; count(R) \quad \forall R \in \text{AATypes} \tag{3.7}$$

$$X_{i,\,R} \;\leq\; \sum_{j} X_{\delta_j,\,R_{H\alpha},\,i} \quad \forall i \in \vec{H}, \forall R \tag{3.8}$$

If HSQC peak i is assigned type R, Constraint 3.4 ensures that each proton type is assigned either to a TOCSY peak or is missing its TOCSY peak. Constraint 3.5 ensures that each TOCSY peak is assigned to either a proton or as noise. Constraint 3.6 ensures that each HSQC peak is assigned to at most one amino acid type. We allow HSQC peaks to have no amino acid type assignment. Constraint 3.7 ensures that the number spin systems assigned to a specific amino acid type does not exceed the number of such residues in the protein sequence. Constraint 3.8 ensures that if an HSQC peak is assigned a type, it will have an H^{α} assigned to a TOCSY peak. Alternatively, assignments to H^{α} atoms can be encouraged by scaling the scores rather than using a constraint, so that missing H^{α} TOCSY peaks are allowed, but we did not do this. Other constraints are possible, such as bounds on the number of noise peaks, which can be estimated by the number of TOCSY peaks and the expected number of TOCSY peaks based on the amino acid sequence.

We used the sequential algorithm to generate multiple solutions. For ubiquitin, the number of solutions was set to the number of residues. Even with this number, all the scores of the solutions were still within 1% of the optimal. To increase the sample space, all amino acid types in the class of the type predicted were included with the prediction. We used the same classes as RESCUE 1 except that we grouped S and T together, and V and A together because these residues have similar BMRB statistics. To ensure that not all previous predictions get regenerated including the types from the same class, after obtaining each solution, we added the following set of constraints, where boolean variable b_i is equal to 1 whenever one of the previously generated type predictions for HSQC peak i gets selected again.

$$\begin{aligned} \sum_{R \in Type(i)} X_{i,\,R} &= b_i \quad \forall i \in \vec{H} \\ \sum_i b_i &\leq (\sum_{i,\,R} X_{i,\,R}) - 1 \end{aligned} \tag{3.9}$$

$Type(i)$ consists of the previously generated type predictions for HSQC peak i. The first constraint ensures that b_i is 1 if and only if one of the previously generated type predictions is predicted for i. The second expression ensures that a new type prediction for at least one HSQC peak will be returned. For each HSQC peak/spin system, all amino acid types with nonzero count were kept as possible types. For each spin system, the TOCSY peak with the highest count that was assigned to an H^{α} was kept. For predictions containing GLY, which has two H^{α}s, the TOCSY peak with the second highest count was also kept.

3.7.4 Results

We obtained 64 correct type predictions out of 70 predictions for ubiquitin. 14 HSQC peaks did not have any type prediction. It turned out that all of these were noise peaks. The accuracy is better than our manual prediction result; however, on average, each spin system had 4.4 amino acid types, versus 3.3 for manual. The range is 2 to 13. The range is large because one amino acid class had 7 residue types (FYWHDNC). Better methods are needed to differentiate among the residues in this class. Due

to the large number of amino acid types predicted per spin system and the lack of secondary structure type prediction information, the search space for assignment will be large. Therefore, except for the H^{α} predictions, the amino acid type predictions were not used in the first iteration of assignment. They were used after anchors have been determined. Only the homologous BMRB assignments were used for typing in the first iteration.

Type prediction errors were due to overlapped peaks, unidentified noise peaks, missing peaks, and peaks matching the mean proton chemical shift values of the incorrect residue better than the correct. If peak shape information was available, one of the noise peaks could have been removed because its shape was distorted. For H^{α} assignment, 65 out of 70 spin systems had at least one H^{α} correct. Assignment errors were due to amino acid type prediction errors, H^{β} of serine or threonine being picked as the H^{α}, and noise peaks that cannot be pruned based on intensity.

The assignment accuracy with the automatically picked peaks (67/69) is better than that with manual (62/68). This is likely attributed to the use of homologous assignment information, which helped to reduce both the noise level and the search space despite a larger initial noise level, lack of secondary structure prediction information, and more missing edges. Although this was only one test, it highlights additional challenges besides the backbone assignment step: peak picking, peak grouping/spin system formation, peak list calibration, spin system type prediction, and H^{α} side chain assignment without an *a priori* backbone assignment. To our knowledge, there are no fully automated methods for structure-based backbone assignment using only ^{15}N-labeled data.

3.8 Conclusion

So far, we have only considered structure-based assignment with a homologous structure. For non-fully automatic assignment for 1EGO, we tried using the more distant structure 1HQO as the template, which has 12% sequence identity with 1EGO and a structure alignment of 2.4 Å on 76 aligned residues using the structure alignment software CE [105]. For no amino acid or secondary structure type prediction errors, the accuracy was 80%. Since the secondary structure boundaries of the protein structures do not align

exactly, for graph vertex matching, we allowed residues along the secondary structure boundaries (within 2 residues) to also take on the secondary structure type of the adjacent structure. By doing this the accuracy increased to 84%. For amino acid typing errors only, the accuracy dropped significantly to 65% for inexact boundary matching (44% for the loops), but the drop was even more for exact boundary matching at 54%. Inexact boundary matching helped, but there is definitely room for improvement.

When we tried fully automatic assignment on two other proteins, we discovered some practical limitations. In the BMRB's time-domain library, the data we needed was only available for 2 proteins: the 112 residue At3g17210 from Arabidopsis thaliana and the 120 residue ubiquitin-like domain of tubulin-folding cofactor B. Finding proteins with ^{15}N-HSQC and 3D ^{15}N-NOESY was not the issue. It was finding the 3D ^{15}N-TOCSY that was challenging. It turns out that there is a practical reason why 3D ^{15}N-TOCSY is not popular, especially for large proteins. For large proteins, the spectra becomes very crowded with many overlapping peaks. For the two proteins, which are much larger than ubiquitin, we had difficulty performing manual amino acid type prediction because we could not determine which peaks should be grouped together. Automated amino acid type prediction did no better. The set of predicted amino acid types for each spin system was too large to be useful. For the two proteins, their data sets actually consisted of both ^{15}N- and ^{13}C-labeled data likely because using only ^{15}N was not sufficient. For very large proteins, the protein is often protonated, so amino acid type prediction from the TOCSY proton chemical shifts will not be possible for most residues due to suppressed signals.

Using the BMRB for identifying the possible residues also has its limitations. Chemical shifts are extremely sensitive to environmental conditions and conformational changes, such as from mutations. When we tried using human ubiquitin in a different solvent, BMRB 5387, the result from applying the Hungarian algorithm gave 41 correct assignments, compared to 70 for human ubiquitin from BMRB 15410. Instead of using TOCSY and the BMRB, perhaps other sources of data, such as RDCs, could be used.

In the next chapter, we used chemical shift mapping to identify the possible residues for each spin system. Chemical shift mapping is often used for monitoring chemical shift perturbations due to the binding of a ligand, which is important in NMR drug discovery [90]. In this case, homologous structure information is often available, so rapid assignment from only ^{15}N-labeled data is desirable as a starting

point for further analysis. Manual analysis and additional experiments can then be performed to obtain the assignments for the unassigned residues, and for the residues with few contact-NOE matches. The next chapter will also present the updated BILP for backbone assignment that also does backbone NOE assignment. It is used in the chemical shift mapping algorithm.

Chapter 4

Chemical Shift Mapping

In this chapter, we present a computer vision approach to fast-exchange mapping, followed by an attempt at slow-exchange mapping. Chapter 3 described a graph-based approach to backbone assignment. In this chapter, we provide a non-graph-based perspective.

4.1 Peak Walking Problem

In the fast-exchange situation as depicted in Figure 2.7, there is a noticeable peak "walking" pattern. This problem is similar to the computer vision problem of tracking objects through the frames in a video [126]. In our case, the objects are peaks, represented as 2D points.

PeakWalker is based on k-dimensional maximum matching (Figure 4.1), which is NP-Complete and APX-complete for $k > 2$ [55, 129]. For $k = 2$, the problem is maximum bipartite matching, which is solvable in polynomial time [64].

Definition 1. *In* maximum k-dimensional matching*, we are given disjoint sets of vertices* $\{T_i \mid i \in [0, 1, \ldots, k-1]\}$*, where vertices in one set have edges to vertices in only the adjacent sets. The goal*

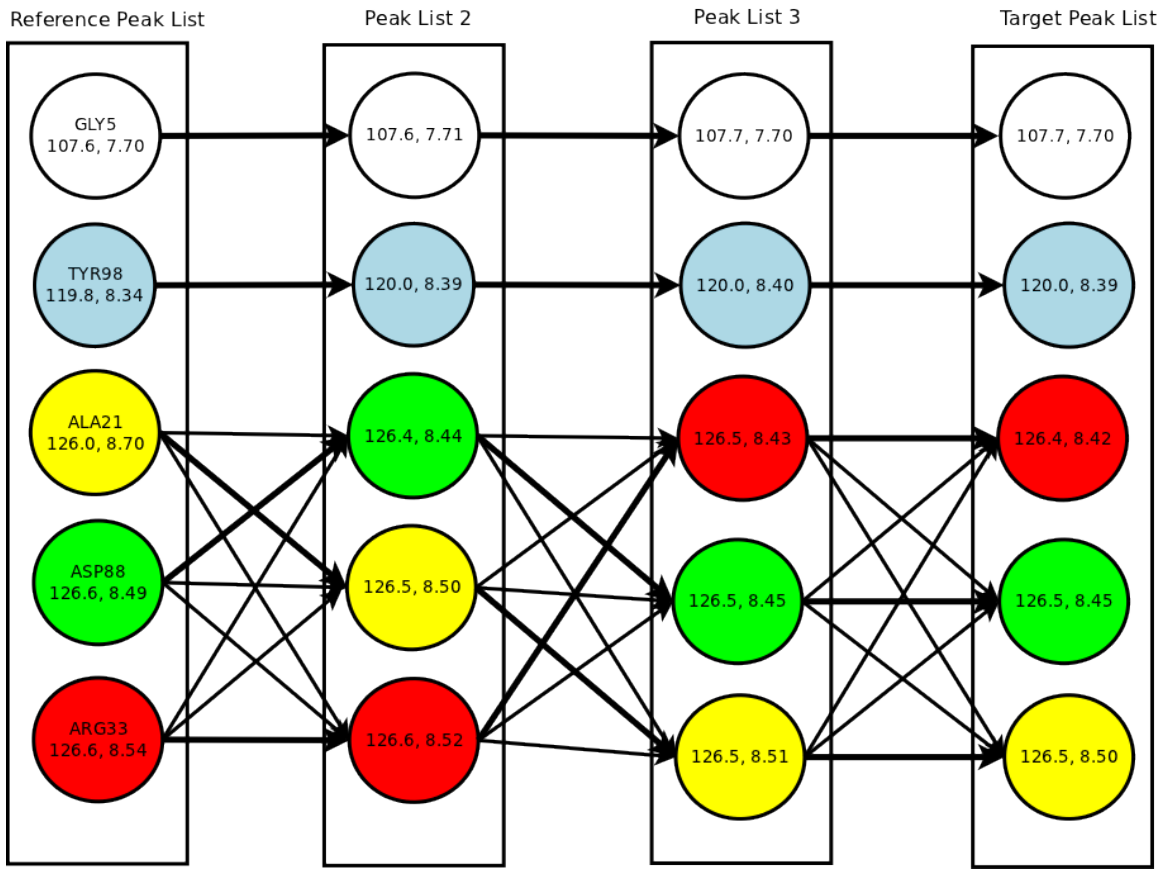

Figure 4.1: Fast-exchange peak walking as k-dimensional matching. Peaks belonging to the same residue have the same color. Peaks in adjacent peak lists with N, H^N chemical shift values within 1.0, 0.2 ppm are connected by an edge.

is to find the largest set of paths $M \subseteq T_0 \times T_1 \times ... \times T_{k-1}$ *, such that the paths do not share vertices; i.e., they are independent paths.*

The fast-exchange case can be viewed as k-dimensional matching if we let $\{T_i \,|\, i \in [0, 1, \ldots, k-1]\}$ denote the peak lists in increasing ligand concentrations, where T_0 denotes the reference peaks, whose residue assignments are known, and T_{k-1} denotes the target peaks, whose assignments are unknown. Each peak is represented by a vertex. An edge is drawn between vertices in adjacent peak lists if the chemical shift change is within some distance threshold. That is, for peaks $h \in T_i$ and $h' \in T_{i+1}$, an edge is drawn

between them if $\Delta\delta_N(h,\, h') \leq t_N$ and $\Delta\delta_{H^N}(h,\, h') \leq t_{H^N}$, where t_N and t_{H^N} are user-specified thresholds and the $\triangle\delta$'s are defined in Equation 2.2. Euclidean distance and various types of weights can also be used to measure chemical shift change [102]. For UbcH5B and histone H1, 1.0 ppm and 0.2 ppm were used for t_N and t_{H^N}, respectively. This is comparable to the thresholds used by FELIX-Autoscreen [91]. Smaller thresholds of 0.75 ppm and 0.125 ppm were used for hBcl_{XL} because it has more perturbed spectra, so the chemical shift changes are expected to be more gradual.

Our problem is a constrained version of maximum weighted k-dimensional matching, where we find the set of paths M that maximizes a scoring function under the constraint that the paths are limited by the peak movements defined by our peak walking model. In our model, we handle various cases as illustrated in Figure 4.2. Figure 4.3 shows the allowable transitions from the perspective of a single peak. A peak in T_i can transition to nearby peaks in T_{i+1} within t_N and t_{H^N}. These transitions shall be referred to as consecutive transitions. A peak can also disappear permanently, or disappear in T_{i+1}, but then reappear in T_{i+2}. The former shall be referred to as a disappearing/missing transition, and the latter a jump. One explanation for disappearing peaks is that the underlying atom is undergoing intermediate exchange. Only jumps of length 2 are explicitly modeled. Finally, a peak in T_i can correspond to a residue with no peaks in T_j , $\forall j < i$. These shall be referred to as new peaks. Transitions correspond to directed edges in the graph. New peaks have no predecessor peak, and disappearing peaks have no successor. Both of these peaks result in subpaths. Peaks that have almost identical chemical shifts can have only one peak present in the peak list due to peak overlap. To handle this, we define two peak states: ambiguous and unambiguous. A peak can be in only one state. An ambiguous or overlapped peak allows multiple transitions, while an unambiguous peak allows only one in- and one out-transition. Ambiguous peaks allow paths to share peaks subject to a penalty. The number of in- and out-transitions for peaks in any state are equal because peaks can only be created or destroyed in the ways allowed by our model. To limit the number of possible paths, only consecutive transitions are allowed for ambiguous peaks. A peak that corresponds to noise is modeled implicitly. Noise peaks are those not assigned to any path.

Definition 2. *The* mappings *for peak $h_i \in T_{k-1}$ is the set of its possible residues $R(h_i)$. This set is obtained by first finding M, the maximum weighted k-dimensional matching on the graph defined by the*

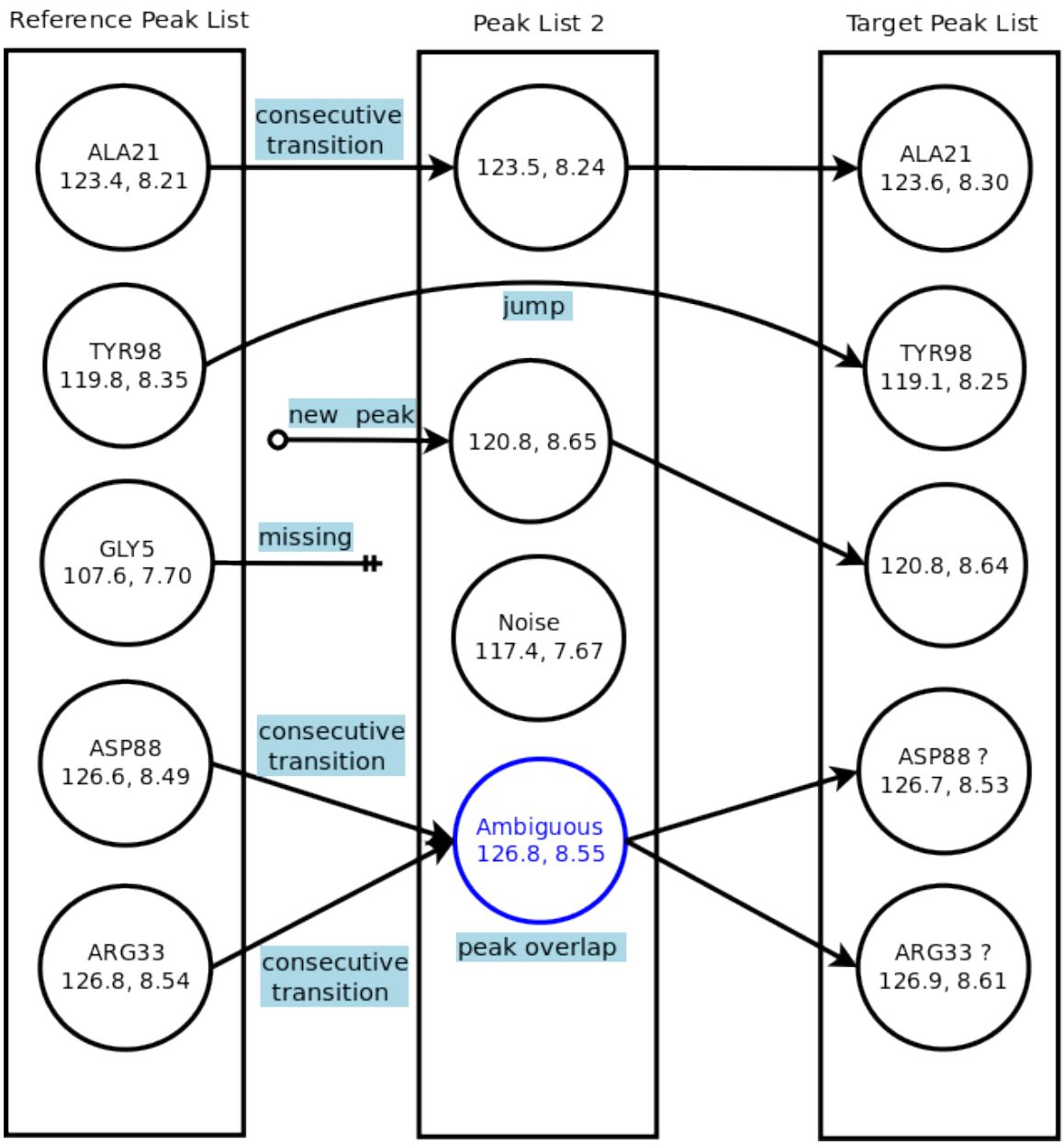

Figure 4.2: Examples of the transitions and errors in our peak walking model for fast exchange.

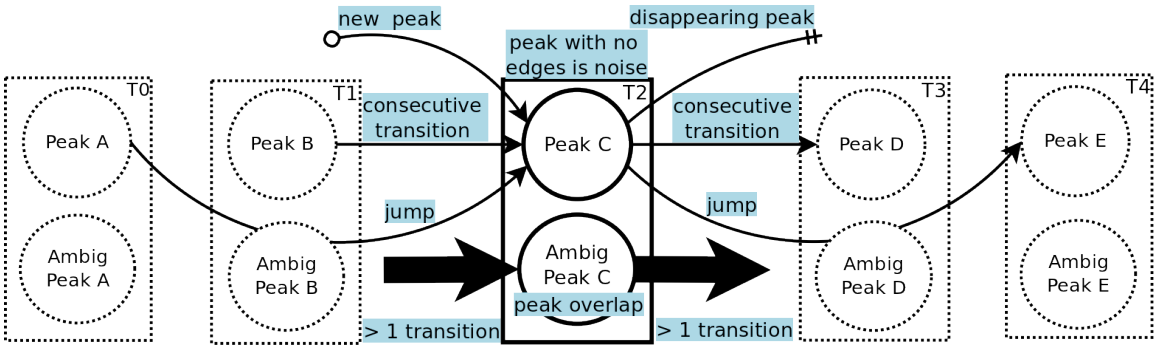

Figure 4.3: Peak walking model for fast exchange from the perspective of a single peak labeled Peak C. The allowable transitions include new peak, consecutive, jump, and disappearing. A peak is either ambiguous or unambiguous.

peak walking model and with weight on the transitions and peak states. Let S be the amino acid sequence of the protein, and one-to-one function $f_0 : T_0 \to S$ be the known reference assignment. For paths in M with end points $h_j \in T_0$ and $h_i \in T_{k-1}$, add $f_0(h_j)$ to $R(h_i)$. If $|R(h_i)| > 1$, or if $|R(h_i)| = 1$ and $R(h_i) \cap R(h_l) \neq \emptyset$ for $h_l \neq h_i$, then $R(h_i)$ is ambiguous.

The optimal and near optimal sets of paths were generated to obtain different mappings per peak. This was done by modeling the problem as a BILP and using the one-tree algorithm [29] to generate multiple solutions. We used CPLEX® version 12.2 as the solver. Integer linear programming has been used successfully in the computer vision problem of tracking multiple objects [51, 81].

4.2 BILP Model for Fast-Exchange

Before presenting the model, we present some background on reified/indicator and logical constraints which give BILPs additional expressiveness. CPLEX supports such constraints by translating them into equivalent linear constraints on binary variables. We found that the automatic translations resulted in models that run very slowly (likely because of Big M values - a very large number used to ensure that all inequalities hold), so we performed the translation manually. How the automatic translation is done is undocumented, so in the following sections, we provide our manual translations.

4.2.1 Reified Constraints

Reified constraints allow the truth value of a logical condition to be stored; i.e. $Y = 1 \Leftrightarrow C = 1$, which is equivalent to $Y = C$. This means whenever condition C is true (equal to 1), Y is 1; and when C is false, Y is 0. Note that $Y = 1 \Rightarrow X = 1 \equiv Y \leq X$. We shall use $\equiv$ to denote equivalent.

1. $Y = 1 \Leftrightarrow \sum_i X_i \geq K$, for some positive integer K, $\equiv$ (both (a) and (b) are added to the BILP)

   (a) $Y = 1 \Rightarrow \sum_i X_i \geq K \equiv KY \leq \sum_i X_i$. To see this, whenever $Y = 1$, the summation is greater than or equal to K; whenever $Y = 0$, the summation is always true.

   (b) $Y = 0 \Rightarrow \sum_i X_i \leq K - 1 \equiv \sum_i X_i \leq MY + K - 1$. Whenever $Y = 0$, the constraint is enforced, and whenever $Y = 1$, the constraint is always true.

4.2.2 Logical Constraints

Logical conditions can be combined and then reified. We shall show this for AND, OR, and the absolute value. For notation purposes, we shall use $X \Leftrightarrow Y$ to denote $X = 1 \Leftrightarrow Y = 1$.

1. AND ($\wedge$)

   (a) $A \wedge B \equiv A + B \geq 2$

   (b) $X \Leftrightarrow A \wedge B \equiv$

      i. $X \Rightarrow A \wedge B \equiv 2X \leq A + B$

      ii. $\bar{X} \Rightarrow \bar{A} \vee \bar{B}$, where $\bar{X} = 1 - X$, $\equiv A + B \leq X + 1$

2. OR ($\vee$)

   (a) $A \vee B \equiv A + B \geq 1$

   (b) $X \Leftrightarrow A \vee B \equiv$

      i. $X \Rightarrow A \vee B \equiv X \leq A + B$

ii. $\bar{X} \Rightarrow \bar{A} \wedge \bar{B} \equiv A + B \leq 2X$

3. For $|X| \geq K$ and $A \Leftrightarrow |X| \geq K$, where X is an integer variable, K is a positive integer, and A is a binary variable, we refer the reader to Appendix 5. We ended up not using the result of this part, but it is left here for completeness and to illustrate the expressiveness of logical constraints since we could not find this in any textbook.

We now present the BILP for fast exchange.

4.2.3 Binary Variables

The variables indicate the transitions and peak states.

- $X_{hih'}$ Equals to 1 if peak $h \in T_i$ transitions to $h' \in T_{i+1}$. This variable represents a consecutive transition.
- X_{hi} Equals to 1 if $h \in T_i$ is a single unambiguous peak. Equals to 0 if it is an ambiguous peak. This variable represents the peak state.
- D_{hi} Equals to 1 if $h \in T_i$ is missing its peaks in T_j, $\forall j > i$. This represents a peak that disappears and no longer reappears.
- $J_{hih'}$ Equals to 1 if $h \in T_i$ is missing in T_{i+1}, but transitions to $h' \in T_{i+2}$. This represents a jump.
- N_{hi} Equals to 1 if $h \in T_i$ has no associated peaks in T_j, $\forall j < i$. This represents a new peak.

4.2.4 Objective Function Coefficients

The objective function is maximized. The coefficients score the transitions and peak states, so the sum of the coefficients multiplied by their corresponding variables gives the score of the paths. Ideally, if a database of peak lists and chemical shift mappings are available, these coefficients could be obtained through training with machine learning techniques, so that the manual mapping process can be modeled.

Unfortunately this database does not exist, so we used our best judgment to scale the scores relative to each other.

- $C(X_{hih'}) = \Phi(\Delta\delta_N(h',\, h),\, 0,\, \text{tolN}) + \Phi(\Delta\delta_{H^N}(h',\, h),\, 0,\, \text{tolHN})$. This is the score of a consecutive transition, where $\Phi(x,\, \mu,\, \sigma) = 2 \times (1 - cdf(x,\, \mu,\, \sigma))$. *cdf* is the cumulative distribution function of a normally distributed variable with mean μ and standard deviation σ. tolN and tolHN were set to values, such that t_N and t_{H^N} (Section 4.1), respectively, correspond to 2 standard deviations from a mean value of 0. The score is a number between 0 and 1, with small chemical shift differences being closer to 1 (because x is positive, so *cdf* returns a value of at least 0.5).

- $C(\overline{X_{hi}}) = -2 \times (k - i - 1) \times (\Phi(\frac{3t_N}{4},\, 0,\, \text{tolN}) + \Phi(\frac{3t_{H^N}}{4},\, 0,\, \text{tolHN}))$, where $\overline{X_{hi}} = 1 - X_{hi}$. The score is negative to penalize ambiguous peaks. We require ambiguous peaks to have at least 2 paths worth of transitions from i to $k-1$ to compensate for the penalty. The penalty decreases with increasing i because there are fewer transitions remaining in the path.

- $C(D_{hi}) = \Phi(t_N,\, 0,\, \text{tolN}) + \Phi(t_{H^N},\, 0,\, \text{tolHN})$. This is the score for disappearing peaks. We give such peaks a positive score similar to a consecutive transition with a chemical shift difference of t_N and t_{H^N}.

- $C(J_{hih'}) = 0.75 \times (\Phi(\Delta\delta_N(h',\, h),\, 0,\, \text{tolN}) + \Phi(\Delta\delta_{H^N}(h',\, h),\, 0,\, \text{tolHN}))$. This is the score for jumps. The 0.75 encourages consecutive transitions over jumps of the same chemical shift difference.

- $C(N_{hi}) = -(k - i - 1) \times (\Phi(\frac{3t_N}{4},\, 0,\, \text{tolN}) + \Phi(\frac{3t_{H^N}}{4},\, 0,\, \text{tolHN}))$. This is the score for new peaks. The score is negative to ensure that there must exist compensating transitions from i to $k-1$.

- Peaks corresponding to noise have no transitions, and they get set to unambiguous because we are maximizing and unambiguous peaks are not penalized.

4.2.5 Constraints

1. For each peak (ambiguous or unambiguous), the number of in-edges is equal to the number of out-edges. Even if a peak disappears permanently (an out-edge), the peak must have come from a

previous transition or be a new peak, which is considered an in-transition. From Figure 4.3, we can see that this constraint is $\forall i \in [1,\, k-2],\ \forall h \in T_i, \quad \sum_{h'} X_{h'(i-1)h} + \sum_{h'} J_{h'(i-2)h} + N_{hi} = \sum_{h'} X_{hih'} + D_{hi} + \sum_{h'} J_{hih'}$.

2. Ambiguous peaks are limited to only consecutive transitions. To get rid of jumps, define the reified constraint $J_{hi} = 1 \leftrightarrow \sum_{h'} J_{h'(i-2)h} \geq 1$, $\forall i \in [2,\, k-1]$, $\forall h \in T_i$, where J_{hi} is a binary variable. Then jumps are removed with $J_{hi} \leq X_{hi}$ since if $X_{hi} = 0$ (ambiguous), then $J_{hi} = 0$ and $\sum_{h'} J_{h'(i-2)h} = 0$. Disappearing and new peaks are handled similarly.

3. For each unambiguous peak, the number of in-transitions is bounded above by 1; similarly for out-transitions. Define the reified constraints $I_{hi} = 1 \leftrightarrow \sum_{h'} X_{h'(i-1)h} + \sum_{h'} J_{h'(i-2)h} + N_{hi} \leq 1$, and $O_{hi} = 1 \leftrightarrow \sum_{h'} X_{hih'} + D_{hi} + \sum_{h'} J_{hih'} \leq 1$. Then the constraint is expressed as $I_{hi} = X_{hi}$ and $O_{hi} = X_{hi}$. This, combined with Constraint 2, also handles, for ambiguous peaks, the constraint that the number of consecutive in-transitions is greater than 1 and the number of consecutive out-transitions is greater than 1.

4. Consecutive transitions generally do not zig-zag. That is, peaks typically do not take a large step in one direction and then take a large step in the reverse direction. To enforce this, let $h \in T_i$, $h' \in T_{i+1}$, $h'' \in T_{i+2}$. If $0.5 \leq \Delta\delta_N(h,\, h') \leq t_N$, $0.05 \leq \Delta\delta_{H^N}(h,\, h') \leq t_{H^N}$, $0.5 \leq \Delta\delta_N(h',\, h'') \leq t_N$, $0.05 \leq \Delta\delta_{H^N}(h',\, h'') \leq t_{H^N}$, then consider the following vectors: $V_{hh'} = (\delta_N(h') - \delta_N(h),\, 10 \times (\delta_{H^N}(h') - \delta_{H^N}(h)))$ and $V_{h'h''} = (\delta_N(h'') - \delta_N(h'),\, 10 \times (\delta_{H^N}(h'') - \delta_{H^N}(h')))$. The consecutive transitions h to h' to h'' zig-zag if the angle between $V_{hh'}$ and $V_{h'h''}$, $\theta_{hh'h''}$, is between 105 and 180 degrees. When h transitions to h', transitions from h' to h'' that result in zig-zag are prevented by adding the constraint $X_{hih'} \leq Z_{h'(i+1)}$, where we have the reified constraint $Z_{h'(i+1)} = 1 \longleftrightarrow (\sum_{h''|\, \theta_{hh'h''} \in [105,180]} X_{h'(i+1)h''} = 0)$. Thus, if $X_{hih'} = 1$, then all consecutive transitions from h' to h'' that cause zig-zag are prevented because the sum is forced to 0.

5. For ambiguous peaks, the number of consecutive transitions entering must equal to the number leaving. We want $\sum_{h'} X_{h'(i-1)h} = \sum_{h'} X_{hih'}$ whenever $X_{hi} = 0$. To implement this constraint, let $D_{hi} = \sum_{h'} X_{h'(i-1)h} - \sum_{h'} X_{hih'}$ and define the reified constraint $W_{hi} \leftrightarrow |D_{hi}| \geq 1$, where W_{hi} is

Protein	Num Peak Lists	Range of Num Peaks in Lists	Num Residues
$hBcl_{XL}$	11	120-136	178
Ubch5b	5	119-127	147
Histone H1	2	86-97	92

Table 4.1: PeakWalker test set.

a binary variable. The reified constraint can be implemented using the results in Appendix 5. We want $W_{hi} \leq X_{hi}$ because when $X_{hi} = 0$, W_{hi} is forced to 0, which in turn, forces D_{hi} to 0. Note that constraints 1 and 2 already imply this constraint, so this is not needed.

4.2.6 Results

PeakWalker was tested on $hBcl_{XL}$, UbcH5B, and histone H1. Table 4.1 gives the characteristics of the test set. Note that $hBcl_{XL}$ is much larger than ubiquitin. $hBcl_{XL}$ consisted of 11 peak lists. The reference peak list contained 148 peaks, while the target contained 142. UbcH5B consisted of 5 peak lists. The reference contained 127 peaks, while the target also contained 127. Histone H1 consisted of 2 peak lists. The reference contained 97 peaks, while the target contained 86. Unlike the other proteins, the assignment for Histone H1 was unknown, so we had to perform the chemical shift mapping manually to obtain a reference solution. Due to ambiguities inherent with chemical shift mapping, especially using only 2 peak lists, we produced both an ambiguous mapping, and for testing purposes, our best guess unambiguous mapping.

The peak lists of $hBcl_{XL}$, UbcH5B, and histone H1 were edited to introduce additional errors. To obtain errors due to overlapped peaks, peaks within the same peak list that have $\Delta\delta_N \leq 0.1$ ppm and $\Delta\delta_{H^N} \leq 0.01$ ppm were merged into a single peak. Such peaks would likely appear as a single peak when viewing the spectra. Multiple peaks could be merged into a single peak. In $hBcl_{XL}$, 5 residues had identical chemical shifts in the target list. After merging, $hBcl_{XL}$ had 136 peaks in the reference list and 122 in the target. UbcH5B had 127 in the reference and 123 in the target. There were no changes to the Histone H1 lists. To simulate noise peaks, in each peak list, we introduced noise peaks in the range of the N and H^N chemical shifts, 99-133 ppm and 6.25-10.75 ppm, respectively. Unless stated otherwise, the number of

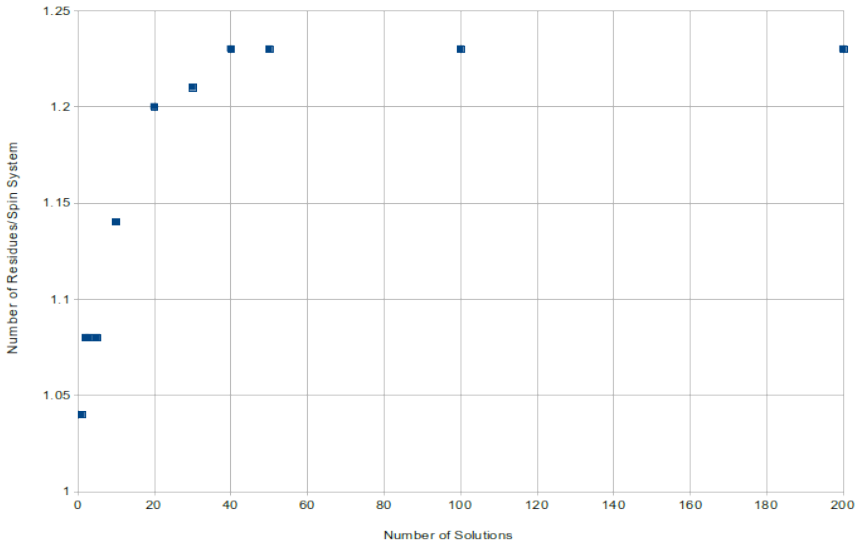

Figure 4.4: The number of residues per target peak/spin system as a function of the number of solutions for histone H1.

noise peaks added to each peak list is equal to 10% of its size prior to the addition.

Multiple solutions were generated to obtain multiple possible paths. The number of solutions generated is dependent on the gap tolerance provided to CPLEX. Unless specified otherwise, a gap of 1% was used. To determine the number of solutions that should be generated, various numbers were tested to determine their effect on the average number of residues predicted per peak. We observed that as the number of solutions increased, the average number of residues plateaus, so we used the value at the start of the plateau as the number of solutions (Figure 4.4). Likely, no new mappings were generated because paths containing these mappings caused a violation of the gap optimality criteria.

For comparison purposes, we implemented the greedy approach in NvMap, but also added no zig-zagging as described above, and jump handling of arbitrary length by allowing unmatched peaks in T_i to be matched to peaks in T_j for any $j > i$. The same chemical shift thresholds as those used by PeakWalker were used. None of the existing approaches deal directly with ambiguous mappings. To generate these without generating many mappings per peak, we used a naive approach. For $h_i \in T_{k-1}$, where h_i is matched to $h_j \in T_0$, in increasing order of $\Delta\delta_{NH}(h_j, h_b)$ for any $h_b \neq h_i$ in T_{k-1}, add $f_0(h_j)$ to $R(h_b)$ (f_o

Protein	Method	Num Correct	Num Correct Range	Acc (%)	Acc Range (%)	Avg Num Res/Peak
$hBcl_{XL}$	Greedy	110.9	110-111	95.7	94-96.5	1
	Greedy	111.1	111-112	90.7	89.5-91.7	1.7
	PeakW	116.3	116-117	96.8	95.9-97.5	1.4
Ubch5b	Greedy	114.6	113-115	94.2	91.5-96.7	1
	Greedy	116.9	116-118	94.4	93.5-95.2	1.2
	Greedy	120.8	120-122	97.2	96-98.4	1.5
	PeakW	120.4	119-123	98.1	96.0-99.2	1.2
Histone $H1^U$	Greedy	78.1	76-83	91.4	89.4-93.3	1
	Greedy	83.0	83-83	95.5	94.3-96.5	1.5
	PeakW	85.1	85-86	99.3	97.7-100	1.3
Histone $H1^A$	Greedy	72.0	72-72	82.8	80.9-83.7	2.0
	PeakW	76.0	76-76	88.8	87.4-89.4	1.3

Table 4.2: Comparison between Greedy and PeakWalker. Results for the best guess unambiguous mapping (U) and the ambiguous mapping (A) are given for Histone H1. The mappings for a peak is correct in the ambiguous case if it contains all possible residues.

and R defined in Section 4.1) until a maximum number of additional mappings have been added. Various values for the maximum were tested. We did not test the consecutive bipartite matching approach because it is an approximation of k-dimensional matching, which we solve directly.

Table 4.2 compares the accuracy between the greedy algorithm and PeakWalker. Different values for the maximum number of candidate residues were tested with greedy. Accuracy is defined as the number of target peaks whose possible mappings contain the correct residue divided by the number of peaks with mappings predicted, including noise peaks. Since one could predict mappings for only a few peaks and still have high accuracy, we have also included the number of peaks whose mappings contain the correct residue. The numbers are averages over 10 trials, where each trial used different noise peaks. The average number of residues predicted per peak varied by at most 0.1 in the trials (not shown). For Histone H1, the accuracy for the ambiguous case is defined as the number of target peaks whose mappings include all the possible residues divided by the number of peaks with mappings. In general, PeakWalker has comparable or better accuracy, and comparable or more correct predictions with fewer candidate residues per peak. Greedy does not perform as well because it does not consider complete alternative paths; it focuses on only alternative neighboring peaks. Returning multiple paths generally improved the results for both the number of correct and the accuracy. For $hBcl_{XL}$, if only one path is returned for each peak, then the

number of correct assignments dropped to 110.2 with an accuracy of 94.1%. For Ubch5b, the result was 115.2 and 95.6%., and for histone H1, the result for the unambiguous mapping case was 81.1 and 95.1%.

The peak lists of $hBcl_{XL}$ contained the most errors among the proteins. Out of 136 peaks in the reference, only 114 had a complete path without any missing peaks between the reference and target. 12 residues did not have any peak in the reference list, but had peaks in the other lists. There was one residue with a jump of length 2, and 3 residues with a jump of length 3. There were no jumps longer than 3. Despite not explicitly modeling jumps of length 3, on average PeakWalker got 2.4 of those mappings correct because neighboring peaks provided alternative paths. For UbcH5B, all the target peaks had corresponding peaks in the reference peak list. There were 2 jumps of length 2, and 4 jumps of length 3. On average, PeakWalker got 3.2 out of those 4 correct. There were no jumps in histone H1.

We also tested $hBcl_{XL}$ using only 6 peak lists instead of 11 by taking every other list. This corresponds to performing fewer NMR experiments. The accuracy decreased slightly to 95.7% with 114.9 correct predictions. $hBcl_{XL}$ was also tested with no overlapped peaks merged in the input. This corresponds to the result if all overlapped peaks could be predicted. For this test, at a cost of optimality, the gap was set to 4% to keep the run time to less than 5 mins per trial on an Intel Core 2 Duo T9300 laptop with 3 GB RAM. Nevertheless, the accuracy was 98.7% with an average number of correct mappings of 138.0 (an increase of over 21), at an average of 1.7 residues per peak. This indicates that peak overlap can hide many peak mappings, which can be a problem if these residues are involved in binding. However, binding residues tend to have chemical shift changes upon binding, so to completely hide such a residue, every time it moves there must exist at least another peak with similar chemical shift to overlap it. In the case of $hBcl_{XL}$, peak overlap masked only the target peak of one known binding residue with significant shift change, but the residue's other peaks were not masked, so the movement of these peaks were detected.

Table 4.3 displays the results of a noise test on $hBcl_{XL}$. The results are averages over 10 trials. The number of noise peaks added ranged from 0 to 50% of the number of peaks prior to addition. All the tests in Table 4.2 had 10% noise. The accuracy at 10% is actually slightly larger than the accuracy at 0% because by chance, some noise peaks provided alternative paths from the target peak to its correct reference. Accuracy depends on the location of the noise peaks relative to non-noise peaks. In general,

Noise (%)	Num Correct	Num Correct Range	Acc (%)	Acc Range (%)
0	116	116-116	96.7	96.7-96.7
10	116.3	116-117	96.8	95.9-97.5
20	115.8	115-117	95.8	95-97.5
30	115.5	114-116	94.9	91.9-97.5
40	115.2	114-116	95.3	93.4-96.6
50	115.2	113-117	93.5	91.1-95.1

Table 4.3: Results for PeakWalker on $hBcl_{XL}$ with various noise levels.

the number of correct predictions and the accuracy decreases with increasing noise, but the decrease is relatively graceful for randomly distributed noise.

4.3 Backbone Assignment Version 2.0

The previous chapter presented a BILP for backbone resonance assignment, which was used for assignment from automatically picked peaks for a small protein. However, using 3D ^{15}N-TOCSY and a homologous assignment limited the practicality of the method. In large proteins, the TOCSY can have many overlapped peaks. Using a homologous assignment to find possible residues for each target peak is similar to chemical shift mapping using only 2 peak lists. There is usually more ambiguity when fewer peak lists are used. In addition, the experimental conditions for the protein in the homologous assignment can be significantly different from that of the current study.

In this chapter, we present a BILP that no longer uses TOCSY. The TOCSY was previously used to identify possible amino acid types for each target peak, and this was used to reduce the number of possible mappings. To reduce the possibilities without TOCSY, a series of perturbed spectra could be used. The TOCSY was also used to obtain the chemical shifts of the H^{α} atoms for matching against NOESY peaks. Such H^{α} chemical shifts are available in the NOESY spectrum, but in a more noisy form. We have also added a further improvement. The constraint that each NOESY peak is assigned to at most one contact was not enforced in our previous algorithm. In adding this constraint, our new algorithm not only performs resonance assignment, but also backbone NOE assignment and H^{α} assignment, simultaneously. The two

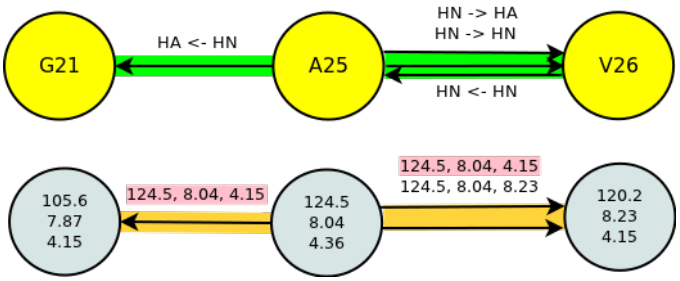

Figure 4.5: Difference between Assignment 1.0 and 2.0. The top line gives the contacts of residue A25 to two residues. The line below gives the spin systems assigned to those residues, and the NOEs between the central spin system and two other spin systems. One NOE, highlighted in pink, has two directed edges. In 1.0, an undirected edge match occurs (green with orange) when the edges share at least one interaction type in common, which is the case for the two edges in this example. However, the NOE in pink indirectly gets assigned to two contacts (to the two HN-HA contacts). The score for this NOE gets counted twice. In 2.0, constraints are added so that an NOE can only be assigned to at most one contact.

BILPs are identical except for additional constraints, and how we determine the possible residues for each amide chemical shift - amino acid and secondary structure type versus peak walking paths. Figure 4.5 illustrates the difference between the two assignment methods. Figure 4.6, illustrates how PeakWalker is combined with PeakAssigner to obtain one-to-one assignments. Although NOE and H^{α} assignment is not the main output of our algorithm, we show that by performing them, there is an improvement in backbone assignment accuracy, on average. This is demonstrated with simulated NOESY peaks from the protein structures 1KA5, 1EGO, 1G6J, 1SGO, and 1YYC. On $hBcl_{XL}$, UbcH5B, and histone H1, we obtain an average accuracy of over 94%.

Unlike the old BILP, PeakAssigner does not do iterative typing error correction because it was not necessary due to the high accuracy of the input from PeakWalker and the small number of possible residues per spin system. Iterative typing error correction can take hours depending on the size of the protein and the number of errors, so using peak tracking input can potentially save a lot of time. The new BILP will be presented formally as an "assignment problem" in the combinatorial optimization context rather than as a maximum common subgraph problem, although one can be converted to the other. A few definitions are required before we can formally define the problem. Note that PeakWalker can be run independently of PeakAssigner. PeakAssigner is simply used to incorporate contact information from a

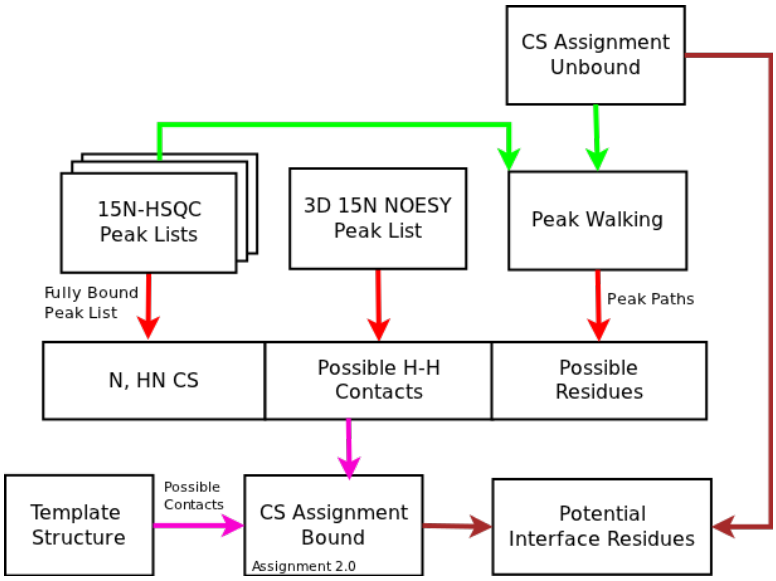

Figure 4.6: Combining peak walking with backbone assignment. CS denotes chemical shift. Arrows with the same color are in the same step. To determine the potential binding interface residues, the amide chemical shifts of the protein in the fully bound state is assigned and then compared to the chemical shifts from the unbound state.

homologous structure to resolve ambiguities.

4.3.1 Definitions

Definition 3. *The* mappings *or* typings *for backbone amide chemical shifts* δ_N, δ_{H^N} *of HSQC peak* $h_i \in T_{k-1}$, *is the set of its possible residues* $R(h_i)$*; e.g., ALA 3, LEU 57.*

In chemical shift mapping studies, possible residues for each peak can be obtained by tracing paths from peaks of unknown assignment to peaks with known assignment. It is often the case that peak tracing results in a many-to-one mapping between residues and backbone amide chemical shifts. The goal of backbone assignment is to find a one-to-one mapping; that is, to pick at most one residue in $R(h_i)$ for each h_i using additional evidence to determine the most likely mapping. In our case, we use contact information from the NOESY.

Definition 4. *A* contact *c consists of two protons and an interaction type: c*[0] =$H_a{}^N$*, which is the amide proton of some amino acid denoted by a, and c*[1]=$H_b{}^N$ *or* $H_b{}^\alpha$*, the amide or alpha proton of another amino acid denoted by b. For* H^α*, it is possible that* $a = b$*. Let* $P(c)$ *be the proton type (*H^N *or* H^α*) of c*[1].

In NOE assignment, contacts are mapped to NOESY peaks. A NOESY peak can represent an intra-residue or inter-residue contact. For intra-residue contacts, usually only H^α - H^N is visible among contacts to side chain protons. The other protons are usually too far away from H^N. A spin system can be extracted from an intra-residue peak because it gives the chemical shifts of the protons within the same residue.

Definition 5. *A NOESY peak p* ($\delta_N(p)$, $\delta_{H^N}(p)$, $\delta_H(p)$) induces *an* H^α *chemical shift for HSQC peak h* ($\delta_N(h)$, $\delta_{H^N}(h)$) *if the amide chemical shifts match and the proton chemical shift of the NOESY peak is similar to a typical Hα chemical shift. Formally, all of this means* $\Delta\delta_N(p, h) \leq \sigma_N$, $\Delta\delta_{H^N}(p, h) \leq \sigma_{H^N}$, *and* $\delta_H(p)$ *matches within 3 standard deviations of the mean value of* $\delta_{H^\alpha}(T(a))$ *of at least one amino acid* $a \in R(h)$*, where* $T(a)$ *is the amino acid type of a. The mean and standard deviations of each amino acid type were obtained from the Biological Magnetic Resonance Data Bank (BMRB) [113].* σ_N, σ_{H^N} *are match tolerances.*

For match tolerances, we used 0.5, 0.05 ppm. Since the intensity of NOESY peaks is inversely proportional to the distance of the underlying protons in contact, and intra-residue H^N, H^α's are relatively close, we can expect the intensity of intra-residue H^N-H^α NOESY peaks to be large. Among the 8 closest (by $\Delta\delta_{NH}(p, h)$) NOESY-induced H^α peaks of HSQC peak h , we took the 4 most intense peaks as a possible induced $\delta_{H^\alpha}(h)$.

It is possible that the N, H^N chemical shifts of a NOESY peak matches the chemical shifts of more than one HSQC peak, and the H chemical shift matches the H^N shift of more than one HSQC peak or more than one induced H^α shift. For NOE assignment, we define the concept of a NOESY peak match to account for this ambiguity. If there is only one possible NOESY peak match for a given NOESY peak, then that peak is *unambiguous*.

Definition 6. *A* NOESY peak match *n consists of a pair of HSQC peaks that match the chemical shifts of the contacting protons as represented by the NOESY peak, the NOESY peak itself, plus an interaction*

type. Formally, n consists of $n[0] = (\delta_N(s), \delta_{H^N}(s))$ of HSQC peak s, $n[1] = \delta_{H^N}(t)$ or an induced $\delta_{H^\alpha}(t)$ of HSQC peak t, $n[2] = (\delta_N(p), \delta_{H^N}(p), \delta_H(p))$ of some NOESY peak p; where, $\Delta\delta_N(s, p) \leq \sigma_N$, $\Delta\delta_{H^N}(s, p) \leq \sigma_{H^N}$, and $\Delta\delta_H(t, p) \leq \sigma_H$. We used 0.05 ppm for σ_H. For H^α, it is possible that $s = t$. Define $P(n)$ to be the proton type of $n[1]$.

Assigning NOESY peaks to contacts is equivalent to assigning NOESY peak matches to contacts if we constrain NOESY peaks to be assigned to at most one NOESY peak match.

To limit the number of assignment possibilities, we consider only the residues and contacts that *match* the experimental peaks.

Definition 7. *Amino acid a* matches *HSQC peak h if $a \in R(h)$.*

Definition 8. *Contact c* matches *NOESY peak match n if the end points match and the interaction types match. Formally, this is i) $a \in R(s)$, where amino acid $a = c[0]$ and peak $s = n[0]$ (a is a possible amino acid type for HSQC peak s); ii) $b \in R(t)$, where $b = c[1]$ and $t = n[1]$ (b is a possible amino acid type of HSQC peak t), and iii) $P(c) = P(n)$.*

4.3.2 Problem Statement

Given the above definitions, we define the backbone structure-based assignment problem.

Definition 9. *Let S be the amino acid sequence, H the set of all backbone amide chemical shifts, C the set of all contacts from the 3D structure of a homologous protein, P the set of all NOESY peaks, and N the set of all NOESY peak matches. The* backbone resonance assignment *is a one-to-one function $f : H \to S$, where $h \in H$ matches $f(h)$, and the* backbone NOE assignment *is a one-to-one function $g : N \to C$, where $n \in N$, $n[2] \in P$, and n matches the contact $g(n)$. The best resonance and NOE assignment is one that maximizes the scoring function*

$$\left(\sum_{h \in H} \sum_{f(h) \in S} w_f(h, f(h))\right) + \left(\sum_{n \in N} \sum_{g(n) \in C} w_g(n, g(n))\right)$$

The scoring functions $w_f : H \times S \to \mathbb{R}$ and $w_g : N \times C \to \mathbb{R}$ weigh each individual resonance and NOE assignment, respectively. Since N contains peaks in H, the two functions are dependent.

Unlike other definitions, this one takes into account both backbone resonance and NOE assignment simultaneously. Bipartite graph and linear assignment versions of the problem only look for f, while subgraph isomorphism and quadratic assignment versions look for both f and g. The former problem is solvable in polynomial time, while the latter is NP-hard.

Our BILP for backbone assignment is based on the following input: The possible residues $R(h_i)$ for each 2D HSQC peak h_i, a 3D ^{15}N-NOESY peak list, and a 3D structure from a homologous protein. The main assumption is that the homologous protein has similar structure. For binding, if the structure is of the unbound form, then the assumption is that the structure does not change significantly upon binding. For B-cell lymphoma-extra large (Bcl-xL), this is the case. For a protein like calmodulin (CaM), this is not the case. However, if a structure is known for the CaM bound form with some other ligand, then it is likely that other molecules that bind to CaM have similar bound structures. Therefore, the structure of one bound form can be used for assignment for other bound molecules. This has applications for drug screening on a library of molecules.

4.3.3 Binary Variables

The variables indicate individual resonance (vertex) and NOE (edge) assignments. Note that contacts are assigned to NOESY peak matches rather than NOESY peaks, but assigning contacts to NOESY peak matches is equivalent to assigning contacts to NOESY peaks because constraints are used to ensure that each NOESY peak is assigned to at most one NOESY peak match and vice versa.

- $X_{a,h}$ Equal to 1 if amino acid a is assigned to HSQC peak h, where a matches h.
- $X_{c,n}$ Equal to 1 if contact c is assigned to NOESY peak match n, where c matches n.

4.3.4 Objective Function Coefficients

A linear objective function is maximized. The coefficients are the weights, and they are non-negative.

- $w_f(X_{a,h}) = 3 \times (1 - \frac{\Delta\delta_{NH}(f_0^{-1}(a),h) - min(h)}{max(h) - min(h)})$ This is the score of assigning amino acid a with reference peak $f_0^{-1}(a)$ (defined in Section 4.1) to target peak h. Shorter peak paths have higher score because we do not expect many chemical shift changes except for the residues in the binding interface. $min(h)$ and $max(h)$ is the smallest and largest, respectively, $\Delta\delta_{NH}$ between h and the reference peaks of the amino acids in $R(h)$. This normalizes the score based on the lengths of the alternative paths for h. It is possible to use a chemical shift prediction program to predict the chemical shift of a residue, and then measure its distance to the target peak. We tried the program ShiftX [85], but it did not improve the accuracy of the results.

- $w_g(X_{c,n}) = \Phi(\Delta\delta_N(p,\, s),\, 0,\, \frac{\sigma_N}{2}) + \Phi(\Delta\delta_{H^N}(p,\, s),\, 0,\, \frac{\sigma_{HN}}{2}) + \Phi(\Delta\delta_H(p,\, t),\, 0,\, \frac{\sigma_H}{2}) + F(c)$, where $\Phi(x,\, \mu,\, \sigma) = 2 \times (1 - cdf(x,\, \mu,\, \sigma))$. $cdf(x,\, \mu,\, \sigma))$ is the cumulative distribution function of a normally distributed variable with mean μ and standard deviation σ. We used a mean of 0, and standard deviation values such that x equal to 0.5 and 0.05 ppm for δ_N and δ_{H^N}, respectively, corresponds to 2 standard deviations. This weight is the score of assigning contact c to NOESY peak match n, where HSQC peak $s = n[0]$, HSQC peak $t = n[1]$, and NOESY peak $p = n[2]$. NOESY peak matches, where the HSQC peaks have very similar chemical shifts to the NOESY peak chemical shifts, have higher score. $F(c)$ is a weight on the type of contact. In the absence of missing NOESY peaks, contacts involving adjacent amino acids should have a NOESY peak match, so it is natural for adjacent amino acid contacts to have higher weight than nonadjacent.

4.3.5 Constraints

1. Each amino acid a is assigned to at most one HSQC peak. This is $\sum_h X_{a,h} \leq 1$.
2. Each HSQC peak h is assigned to at most one amino acid. This is $\sum_a X_{a,h} \leq 1$.
3. Each contact c is assigned to at most one NOESY peak match. This is $\sum_n X_{c,n} \leq 1$.

4. Each NOESY peak $p = n[2]$ of NOESY peak match n is assigned to at most one contact. This is $\sum_{c,\, n[0],\, n[1]} X_{c,n} \leq 1$.

5. Each pair of HSQC peaks $n[0]$, $n[1]$ of NOESY peak match n has at most one NOESY peak. This is $\sum_{c,\, n[2]} X_{c,n} \leq 1$.

6. Contact c is assigned to NOESY peak match n if and only if amino acid $c[0]$ is assigned to HSQC peak $n[0]$, and $c[1]$ is assigned to $n[1]$.

   (a) $\forall c,\, \forall h,\, \sum_{n \,|\, h=n[0]} X_{c,n} \leq X_{c[0],h}$

   (b) $\forall c,\, \forall h',\, \sum_{n \,|\, h'=n[1]} X_{c,n} \leq X_{c[1],h'}$

   Proof. Suppose contact c is assigned to NOESY peak match n. Then $X_{c,n} = 1$ and the summation on the left-hand side is 1, and at most 1 based on constraints 3-5. This forces $X_{c[0],h} = 1$ and $X_{c[0],h'} = 1$, so $c[0]$ is assigned to $h = n[0]$ and $c[1]$ is assigned to $h' = n[1]$. Suppose $X_{c[0],h} = 1$ and $X_{c[1],h'} = 1$, then since we are maximizing the score, the summations on the left-hand sides must be set to one if a matching NOESY peak exists. Note that we cannot have $X_{c,n} = 1$, where $h = n[0]$ and $X_{c,n'} = 1$, where $h' = n'[1]$ and $n \neq n'$ because it violates constraint 3. □

7. Each H^{α} proton, z_a of an amino acid a, is assigned to at most one induced H^{α} peak, y_h of HSQC peak h. Let $b_{z_a,y_h} = 1 \leftrightarrow \sum_{c,\, n \,|\, z_a \in c[1],\, y_h = n[1]} X_{c,n} \geq 1$ be a reified constraint, where $b_{z_a,y_h} = 1$ when z_a is assigned to y_h. The summation is over all $X_{c,n}$ that contain z_a and y_h. The constraint is then $\forall z_a,\, \sum_{y_h} b_{z_a,y_h} \leq 1$.

8. Each induced H^{α} peak, y_h of HSQC peak h, is assigned to at most one H^{α} proton. This constraint is then $\forall y_h,\, \sum_{z_a} b_{z_a,y_h} \leq 1$.

4.3.6 Multiple Assignment Possibilities

To identify robust assignments, we find additional near-optimal solutions to yield different assignment possibilities for each peak. The idea is that an assignment in the optimal solution is more reliable if it

PDB ID	1KA5	1EGO	1G6J	1SGO	1YYC
Noise (X)	5.6	5.3	3.8	8.2	7.6
Acc v2.0 (%)	100	95.6	93.5	87.9	95.2
Range v2.0 (%)	100	89.9-100	93.1-94.4	81.8-99.3	90-99.4
Acc v1.0 (%)	100	92.8	94.4	86.7	89.4
Range v1.0 (%)	100	89.9-97.5	94.4-94.4	76.7-96.3	75.5-96.8
NOE Acc (%)	92.6	89.0	94.2	86.6	89.8
Range NOE (%)	91.0-94.0	79.4-94.7	92.3-96.2	81.8-93.5	84.9-94.2

Table 4.4: Comparison between the two BILP assignment methods. The new version is denoted as v2.0 and the old v1.0. The accuracy is the number of correct one-to-one mappings divided by the number of mappings. NOE assignment accuracy is the number of correct NOESY peak to contact assignments divided by the number of assignments. NOE assignment accuracy is only for v2.0 because v1.0 does not do it.

is also in the near-optimal solutions. We used CPLEX 12.2 and the one-tree algorithm. A consensus assignment can be obtained from the set of solutions, by setting the objective function terms $w_f(X_{a,h})$ to the number of times amino acid a was assigned to peak h, and $w_g(X_{c,n})$ to the number of times contact c was assigned to NOESY peak match n.

4.4 Results

To compare the combined NOE and resonance assignment approach of PeakAssigner with the old BILP, we ran both on data simulated from the structures 1KA5, 1EGO, 1G6J, 1SGO, and 1YYC, which were part of the test set in our previous work. Rather than using the simulated data provided by the authors of the CR method, which was done in the previous tests, we simulated the data ourselves so that we could trace the results back to the data. In these tests, the possible mappings for each peak contained the correct residue, and there were no H^{α} assignment errors. Later tests will introduce such errors. NOESY peaks were simulated using chemical shift data from the protein's BMRB entry: 2030 for 1KA5, 491 for 1EGO, 5387 for 1G6J, 6052 for 1SGO, and 6515 for 1YYC.

The results are given in Table 4.4. Each PDB file contained multiple 3D models. The table shows the average result from using every pair of structures, where one was the template structure and the other was the target. The noise level is defined as the number of NOESY peak matches divided by the number of

Accuracy	Target	Typing Size	Type Error Probability	Missing H^{α} Probability
151/151 151/151	3FDL	4	0	0
151/151 145/151	3FDL	5	0	0
150/150 149/150	3FDL	6	0	0
149/150 -	3FDL	7	0	0
131/144 112/148	3FDL	5	0.1	0
125/141 102/148	3FDL	5	0.15	0
135/144 -	3FDL *	5	0.1	0.05
114/144 101/147	LOMETS *	4	0.1	0.05
127/142 115/144	LOMETS *	3	0.1	0.05

Table 4.5: Comparison between the two BILP assignment methods under various parameters. For each row, the top line gives the result for v2.0 and the line below for v1.0. 3FDL was used as the template for all cases. Results with a '-' did not complete within an hour. The target proteins for NOE simulation include 3FDL and a structure obtained from the protein threading server LOMETS [121]. A distance cutoff of 4.5Å was used for simulation. The superposition of 3FDL and the LOMETS model is 13.6 Å, and 2.3 Å over 110 residues in the structure alignment result from CE [105]. Typing size is the number of possible residues per target peak. The possible residues excluded the correct residue according to Type Error Probability. Missing H^{α} chemical shifts were introduced with probability given in the last column. (*) In the last 3 cases only, missing NOEs were introduced with probabilities 0, 0.05, 0.21, 0.41, 0.51 for contacts with distances 1.0, 2.0, 3.0, 4.0, 4.5Å, respectively.

Protein	UbcH5B	Histone H1	$hBcl_{XL}$	$hBcl_{XL}$*
Num Correct	119.5	66.2	101.9	99.8
Num Correct Range	119-120	65-67	101-103	99-101
Acc (%)	98.0	94.7	95.6	94.5
Acc Range (%)	97.5-98.4	92.9-95.7	94.4-97.2	93.5-95.2
Num H^N-H^N Correct	157	114	116.1	118.3
Acc H^N-H^N (%)	92.7	90.9	90.0	86.3
Num H^N-H^α Correct	168.2	104.9	128.6	0
Acc H^N-H^α (%)	75.6	65.4	63.1	0

Table 4.6: One-to-one backbone assignment results from PeakWalker input. The input many-to-one mappings for $hBcl_{XL}$ had a 96.7% accuracy with 116 correct and 1.3 residues per peak on average. The input for UbcH5B had values of 98.3%, 120, and 1.2, respectively. The input for histone H1 had values of 98.8%, 85, and 1.3. The NOE assignment accuracy for each contact type is defined as the percentage of the number of contacts of the given type that were assigned to the correct NOESY peak. * The last column gives the results of using only H^N-H^N contacts for $hBcl_{XL}$.

contacts. With the exception of 1G6J, which has a low noise level, our new method was better, especially when the noise level increased. We also tested 1SGO with different noise levels by using different values for the match tolerance. For a noise level of 4.6, the old method was 0.5% more accurate, but for noise levels from 5.5 to 10.3, the new method did 0.2 to 4.2% better. Larger proteins typically have higher noise levels due to increased peak overlap.

Table 4.5 compares the approaches on 3FDL as the template structure. The accuracy for the old BILP for a typing size of 5 residues per peak versus 6 residues is surprisingly worst despite less ambiguity. This is likely because the residues in the typings, which were generated randomly, affected the possible NOESY peak matches, so the typing composition influenced the accuracy. The new BILP performed better in general. Surprisingly, when the typing size/level of ambiguity increased, the new BILP ran faster than the old even though the number of variables in some cases was twice as many. In two cases, the old BILP failed to complete within 1 hour. When there are errors, the search space tends to increase because of increased ambiguity. The additional constraints in the new BILP likely pruned the search space better. For a typing size of 5 on the LOMETS target with type error probability of 0.1, missing H^α probability 0.05, and missing NOEs, both methods did not complete within an hour (result not shown in the table). For the test cases in Table 4.4, the old BILP ran faster likely because the search space was small due to no typing errors.

Table 4.6 shows the assignment results for hBcl$_{XL}$, UbcH5B, and histone H1 with the new BILP. The values are averages over 10 trials, where each trial is a different NOESY peak simulation. Section 4.4.1 describes the NOESY peak simulation in detail, and Section 4.4.2 describes the input template structures. Peak mappings were obtained from PeakWalker, and the unambiguous reference mapping was used to measure the accuracy on histone H1. As expected, the resonance assignment accuracies were slightly less than those for the input many-to-one mappings. However, the number of correct assignments for hBcl$_{XL}$ and histone H1 was less than expected when comparing to Table 4.2. This is likely due to differences between the contacts in the template and target structures. Their superpositions were greater than that for UbcH5B, and the templates had fewer residues than the target. When we used the target as the template structure for resonance assignment, the number of correct assignments increased by 8.9 for hBcl$_{XL}$ and 6.1 for histone H1. Other types of errors, such as missing NOESY peaks, had only a small influence on the number of correct assignments. Another factor is that our accuracy definition did not take into account peaks that were assigned to the wrong residue, but have almost identical chemical shift to the correct peak of the wrong residue. When this is taken into account, the number of correct assignments increased by about 2.6 for hBcl$_{XL}$. There was no change for histone H1 because its peak lists had no overlapped peaks.

When comparing PeakWalker one-to-one mappings, which did not take into account contact information, to PeakWalker-PeakAssigner one-to-one mappings, the accuracies were similar, but PeakWalker alone had more number of correct except for UbcH5B, where the combination had over 4 more correct on average. The reason for this result is that PeakAssigner currently does not produce assignments for residues not in the template structure. For hBcl$_{XL}$, the template structure had 24 fewer residues compared to the number of target peaks, and for histone H1, 7 fewer residues. For UbcH5B, the number of residues is the same as the number of peaks. It is possible to modify PeakAssigner to go with PeakWalker's results if a residue does not have an assignment due to not being in the template structure. In terms of getting the binding residues correct, the combination of PeakWalker and PeakAssigner did slightly better. For hBcl$_{XL}$, both approaches got the same binding residues correct, but for UbcH5B, the combination got one extra binding residue correct due to an ambiguous mapping. Histone H1 was not a binding situation, but

Num Correct Input	111/123	111/123	111/115	111/115
Avg Num Res/Peak	2.3	3.3	2	3
Num Correct Avg	92.2	86.3	95.1	94.3
Num Correct Range	90-95	80-93	93-97	93-96
Acc (%)	91.8	86.5	94.6	94.3
Acc Range (%)	89.1-95.0	79.2-93.0	92.1-96.9	92.1-96.0

Table 4.7: One-to-one assignment results for $hBcl_{XL}$ with different input many-to-one mappings. The results are averages over 10 trials, where each trial is a different NOESY simulation.

rather the protein at two different temperatures. For this protein, PeakWalker alone made 2 incorrect assignments that the combination got correct. There was no instance where the combination made an incorrect assignment, but PeakWalker alone made the correct one-to-one mapping. The main issue with running PeakWalker alone is that some noise peaks received one-to-one mappings, whereas the combination did not make any assignment for them because they were not supported by contact information.

Despite using ambiguous induced H^{α} chemical shift assignments, the H^{N}-H^{α} contact assignments were over 60%, even with a 5% H^{α} missing rate. Nevertheless, the comparable results for $hBcl_{XL}$ that used only H^{N}-H^{N} contacts indicate that the accuracy was not impacted significantly when H^{α} was not used.

Table 4.7 shows the results for $hBcl_{XL}$ with different many-to-one input mappings. When the number of candidate residues per peak increased, the accuracy and the number of correct assignments decreased. However, the decrease was much more pronounced for the input with poorer accuracy. The decrease in the other case was minimal, so erring on producing a few extra possible mappings is less detrimental if it can be done accurately.

Once backbone assignment is done, one can compute the chemical shift change between each target peak and its assigned reference peak. Residues with large changes might indicate their involvement in binding. The binding affinity can then be determined by computing the dissociation constant, which can be obtained from model fitting using the peak paths and the predicted paths according to some model of binding [62]. Figure 4.7 shows the chemical shift changes of $hBcl_{XL}$. For this protein, residues with large changes are involved in binding or near binding residues, but this is not always the case for all proteins because changes can also be attributed to allosteric changes. Except for 2 residues involved in binding, the reference solution and PeakAssigner agree. Residue 196 was not in the input structure for assignment, and

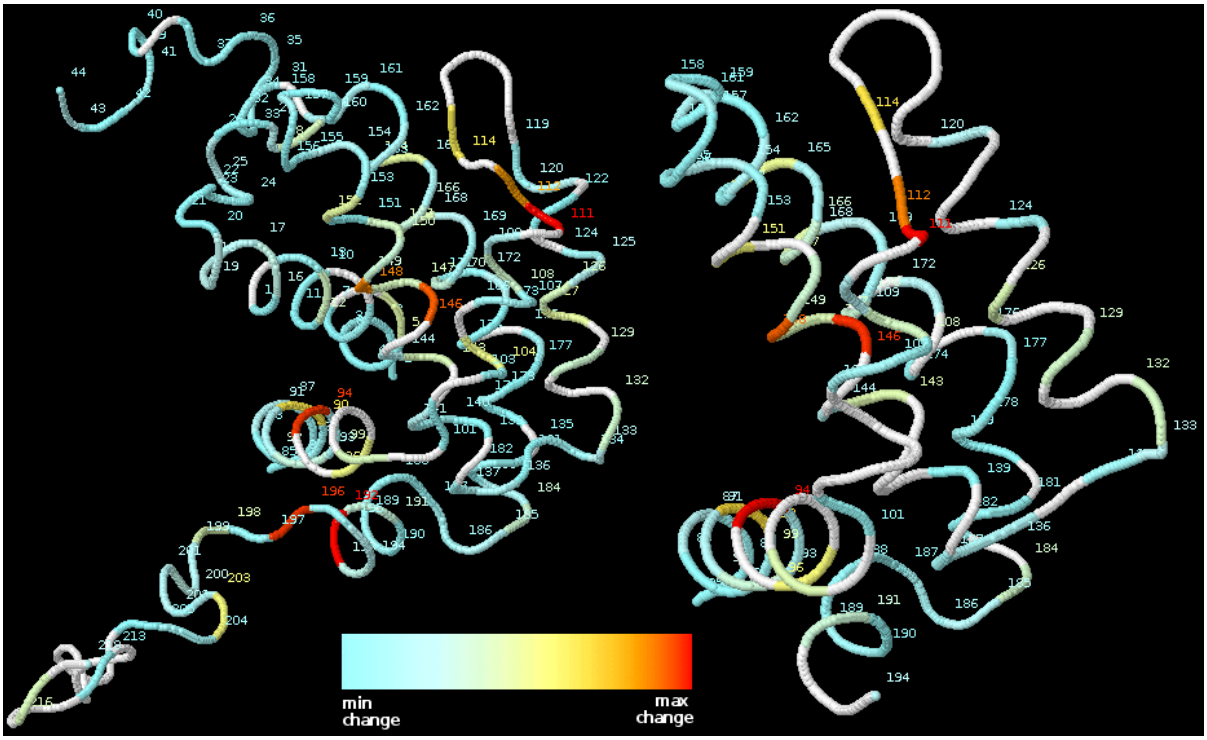

Figure 4.7: The chemical shift changes for the residues of hBcl$_{XL}$ upon binding. (LEFT) gives the known changes on the correct NMR structure. (RIGHT) gives the changes based on the backbone assignment on the template structure 3FDL. Residues are labeled by their residue number. Unassigned residues are unlabeled and colored white.

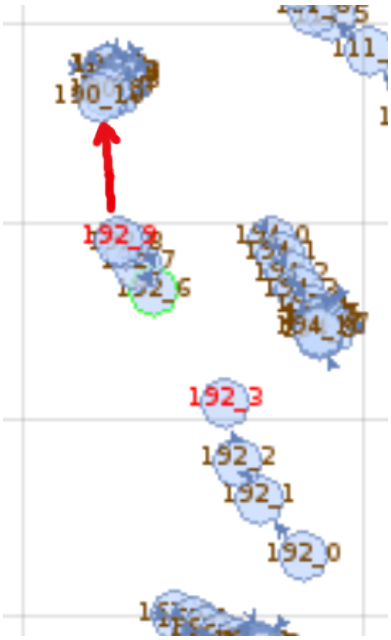

Figure 4.8: One possible peak walking path for residue 192 of hBcl$_{XL}$. The vertical spacing is δ_{H^N} at 0.1 ppm units, and the horizontal spacing is δ_N at 1.0 ppm units. Peaks are denoted by circles with the residue number followed by the peak list number after the underscore. Peak list 0 is the free protein. The red arrow points to the true final peak position, which was concealed due to peak overlap with residue 190. Red labels indicate peaks missing their next transition and the green circle indicates a peak predicated as missing all its previous peaks. The peak for 192 is not present in lists 4 and 5. The grey circle is a noise peak.

the target peak for 192 was concealed due to peak overlap with residue 190. However, 192 was correctly predicted as missing its peak by PeakWalker, and correctly predicted as having a large shift change using its peaks in the other peak lists (Figure 4.8). Figure 4.9 shows the result of docking the Bak peptide from 1BXL to the homology model for hBcl$_{XL}$ using the putative binding residues 90, 94, 111, 112, 114, 146, 148, and 192 as constraints.

Figure 4.10 highlights the residues with chemical shift changes in UbcH5B. The residues with the largest changes: 4, 60, 94, 96 match or are close to the actual residues in the interface: 3, 5, 58, 59, 63, 94, 96, 98. Overlapped peaks resulted in residue 58 not being assigned. 2 peak lists had 58's peak overlapped with that of residue 132, so it ended up being predicted as being missing by PeakWalker. Missing transitions can be flagged for manual analysis. FELIX-Autoscreen [91] uses peak shape information to identify overlapped

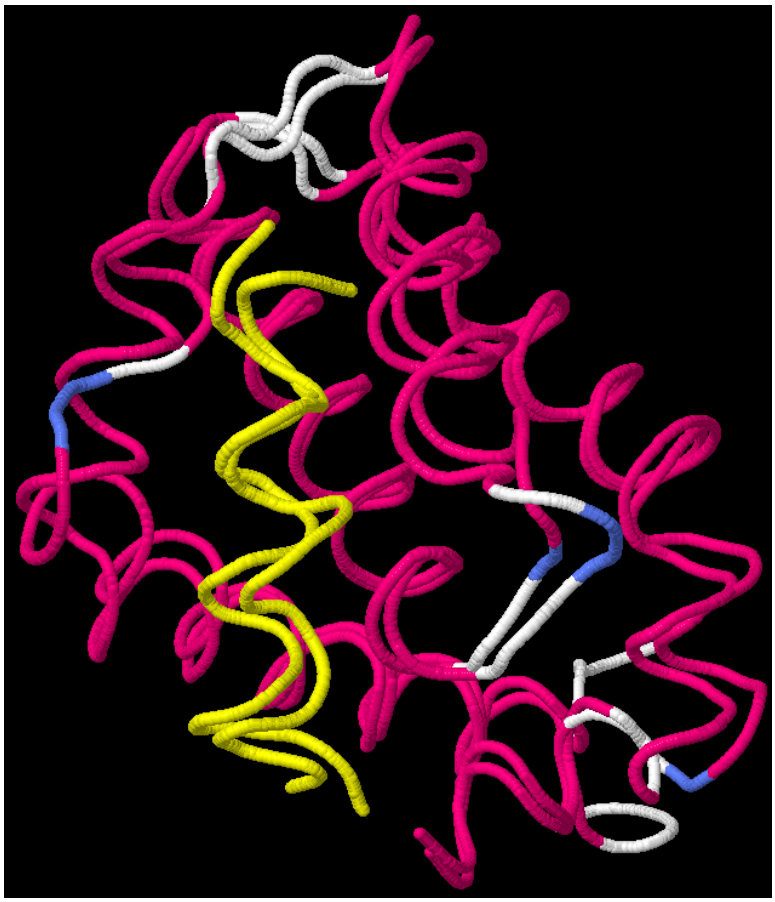

Figure 4.9: Structure alignment of $hBcl_{XL}$-Bak protein-protein complex with 1BXL. The complex was obtained by docking the Bak peptide (yellow) of 1BXL to the homology model for $hBcl_{XL}$ using putative binding residues 90, 94, 111, 112, 114, 146, 148, and 192 as constraints. ClusPro [61, 60] was used for protein-protein docking, where the lowest energy structure from the largest cluster is shown. MM-align [82] was used for the structure alignment.

Figure 4.10: The chemical shift changes for the residues in UbcH5B. (LEFT) shows the changes based on the backbone assignment on the template. In increasing order of shift change, the colors are blue, yellow, orange, and red. Residues are labeled by their residue number. Unassigned residues are unlabeled and colored white. (RIGHT) gives the actual structure of the complex with the non-blue and non-white residues from the left highlighted. The ligand, Not4, is highlighted in yellow.

peaks. Peak shape can be incorporated into our model by scoring peaks as more likely ambiguous if they have shapes resembling the sum of two or more peaks. However, shape information was not available in the peak lists. They would have to be extracted from the raw spectra.

Residues 10, 11, and 12 do not appear to be part of the binding interface, but they have chemical shift changes likely attributed to dynamics. We passed the 12 residue with the largest chemical shift changes (4, 8, 10, 11, 12, 16, 60, 63, 94, 96, 98, 99) to ClusPro [61, 60] for docking, but unfortunately, it was unable to find the correct docking conformation. Compared to Bak for $hBcl_{XL}$, which is a small helix, Not4 is a larger ligand, so it is likely chemical shift mapping information is also needed for the ligand for complex

determination. The ligand should be easier to study since it is often smaller than the receptor. We provided the docking program HADDOCK [30], with the above residues as active residues, minus residues 10, 12, 60, 99, which did not meet the $> 40\%$ relative solvent accessibility criteria of HADDOCK for being active. Solvent accessibility was computed using NACCESS [46] on the template structure. We also provided the known active residues of the ligand: 16, 17, 18, 19, 40, 48, 49, 50, 52, 55, 56, 57. HADDOCK generates ambiguous interaction restraints from the active residues to support ambiguously assigned intermolecular NOEs. Figure 4.11 shows the best docking result, which was the third best scoring docking model among the top 5 solutions returned. The first and second best scoring models had the ligand near the correct area, but in the wrong orientation. Note that the restraints are ambiguous, so the docking program's scoring function was relied upon to determine which residues were supposed to be in contact. Using chemical shift information in the docking scoring function [111], and interface information from transferred NOEs and saturation transfer difference NMR should improve the results. We also tried ClusPro with the above residues, but none of the top 10 solutions returned were correct.

As an alternative to ligand chemical shift information, differential chemical shifts can be used to orient the ligand [73]. This method is useful when the comparison between the chemical shifts of the free and bound protein fails due to most chemical shifts having changed. Instead, chemical shift mapping is done on the target with different mutants of the ligand, assuming they bind. The mappings are examined for differences. It is assumed that large chemical shift differences between the cases is due to the mutations, and that the mutations do not change the docking significantly. Since the mutations are known, we know which part of the ligand is associated with which residues of the target.

Histone H1, which was studied under two different temperatures, was not a binding test case, so we decided not to pursue it any further.

4.4.1 NOESY Peak Simulation

NOESY peaks were simulated using the contacts in the 3D structure (within 4.5Å), N and H^N chemical shifts from the target peak list, and H^α chemical shifts from either ShiftX predictions [85] or from the BMRB depending on availability. For $hBcl_{XL}$, we used the protein threading server LOMETS [121] to

Figure 4.11: Structure alignment of the predicted and actual UbcH5B-Not4 complexes. Not4 is in blue and yellow. The native complex is in yellow and red. HADDOCK was used for docking. MM-align was used for the alignment.

obtain the target structure. The structure chosen among the possibilities returned by LOMETS was the one that used 1LXL as the threading template. It consisted of 178 residues after the flexible loop region was removed. ShiftX was used to obtain the H^{α} chemical shift values. For Ubch5b, we used the structure named "ubch5b-not4_1.pdb" that was provided with the peak lists, and ShiftX for the H^{α} chemical shifts. The structure consisted of 147 residues. For histone H1, we used 1UST for the structure and BMRB entry 6161 for the H^{α} chemical shifts. It consisted of 92 residues.

A global offset to calibrate the N, H^N chemical shifts of the NOESY against the HSQC is assumed to have already been obtained from a calibration step, so we simulated only local calibration errors. Local calibration noise, randomly distributed between 0 and 0.15 ppm for N, 0 and 0.015 ppm for H^N, were introduced to NOESY peaks. Global calibration can be performed manually relatively quickly compared to backbone assignment. Missing inter-residue contacts were introduced with the following probabilities (0, 0.05, 0.21, 0.41, 0.51) for contacts within the following distances (1.0, 2.0, 3.0, 4.0, 4.5)Å, respectively. Missing intra-residue H^N-H^{α} contacts were introduced with probability 0.05. With size 10% of the number of NOESY peaks, NOESY peaks corresponding to noise were added in the range 99-133 ppm for N, 6.25-10.75 ppm for H^N, and 2-6 ppm for H^{α}.

4.4.2 Template Structures

The homology-modeling server SWISS-MODEL [8, 59, 89] was used to obtain the templates as input to assignment. Reduce [120] was used to add the coordinates of hydrogen atoms to the templates. As input to SWISS-MODEL, the template used for $hBcl_{XL}$ was 3FDL. It consisted of 154 residues. Residues 27 to 82 were not present in the file. The 3D superposition between the target and template was 13.6Å. However, the structure alignment using residues 85-194 was 2.3Å according to the program CE [105]. The template for Ubch5b was 2ESK, which consisted of 147 residues. The superposition was 2.4Å, where all residues aligned. The template for histone H1 was 1YQA, which consisted of 85 residues. The superposition was 4.9Å, but the structure alignment was 2.0Å using residues 9-82.

4.5 Discussion

Even with NOE information, a one-to-one mapping for all residues is not always possible. Our approach, however, facilitates an iterative semi-automated approach. Once assignments and paths have been verified, perhaps using additional information, the corresponding variables can be removed from the BILP, and then the model resolved. Multiple near-optimal solutions can be returned to account for ambiguity. NOEs matching multiple contacts can also be outputted to facilitate manual analysis.

For single site binding, peaks trajectories are usually relatively straight. Our peak walking model assumes peaks follow a non-zig-zagging trajectory, so curves are allowed as long as they do not zig-zag. Curves can occur when there are multiple binding sites or a conformational change upon binding. Complex determination when there is conformational change is still a challenge for flexible docking methods. This is especially the case when there is a large change, such as in CaM (Figure 4.12). In the documentation of RosettaDock, it says "docking will not capture the flexibility of a molecule like calmodulin" (`http://rosettadock.graylab.jhu.edu/documentation`). There does not appear to be many docking methods that can handle large backbone changes. SnugDock [107] can model the flexibility of light and heavy chains and antibody CDR loops in antibody-antigen docking. HADDOCK can model backbone flexibility about a hinge residue [56] if it is given the hinge residue and the correct interface contacts, but as ambiguous interaction restraints. It can determine the structure by cutting the protein at the hinge and then performing multibody docking with connectivity and flexibility constraints between the boundary of the two cut domains. CaM was not tested, so we tried it on HADDOCK. This version of HADDOCK is currently only available with guru-level access on their webserver, which we obtained from the authors.

We started from the free-form structure with calcium bound (middle of Figure 4.12). The HADDOCK paper provided guidelines on predicting hinge residues, so we tested them by using HingeProt [33] to predict the hinge residue. It predicted residues 17, 64, 78, and 90. 17 and 90 are near the ends of helices, 64 is in a beta sheet, and 78 is near the middle of a long helix, so according to their guidelines, 78 is the most likely hinge residue. It turns out 78 is closest to the residue with the largest chemical shift change, which is 76. Specifying residue 78 as the hinge, residues 73-83 as fully flexible, and the correct active

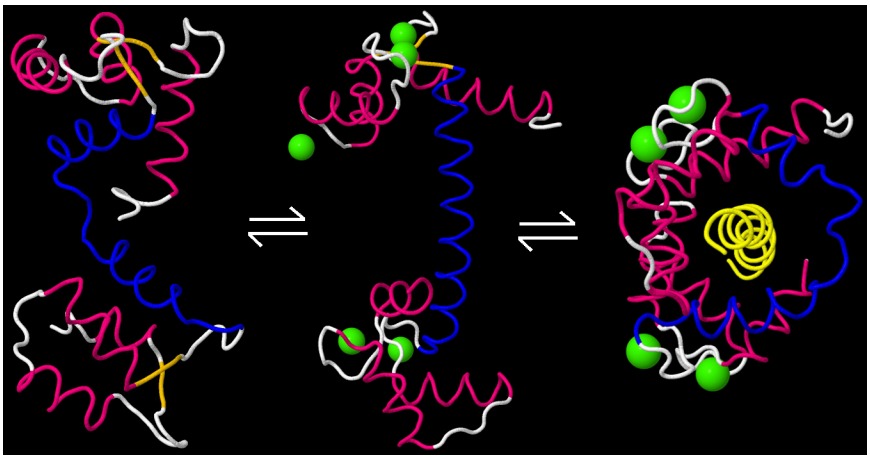

Figure 4.12: The different conformations of calmodulin. Upon binding of calcium (green spheres), the central helices of free calmodulin (LEFT) join into a single helix (MIDDLE). Calcium-bound calmodulin wraps around the peptide upon binding (RIGHT).

residues of both binding partners, but not which pairs of residues are in contact, HADDOCK gave all incorrect conformations.

Next we tried providing it with some contacts in an ambiguous form, but which are all correct contacts. We provided it with the following, where chain A is the protein and residues without a chain designation are in the ligand: 19:A-{12, 13, 14, 15, or 22}, 32:A-{4 or 24}, 36:A-1, 48:A-2, 51:A-{2 or 6}, 54:A-{6 or 7}, 55:A-6, 68:A-{8 or 10}, 71:A-{10, 11, or 12}, 72:A-{11, 12, or 13}, 77:A-{7 or 8}, 84:A-{16 or 18}, 88:A-{19 or 20}, 92:A-20, 109:A-24, 141:A-19, 145:A-{20, 22, or 24}. It was able to find a solution with RMSD of 3.9Å in the MM-align alignment among the top 3 solutions returned, so it seems that intermolecular contacts are needed in this case (Figure 4.13). We then tried providing HADDOCK with about half the number of contacts (from only residues 36, 48, 54, 68, 72, 88, 109, and 149 of chain A), and the result was almost as good, except that the helix peptide was off by one helix turn. We also tried providing about a quarter of the number of contacts (from only residues 36, 54, 88, and 109), but this time docking was not successful. Given that NOE data is typically sparse and ambiguous, the question remains how much data loss and ambiguity can be tolerated in general. This is a topic for future work.

Our peak walking approach is applicable to the SAR by NMR method, which uses fast exchange

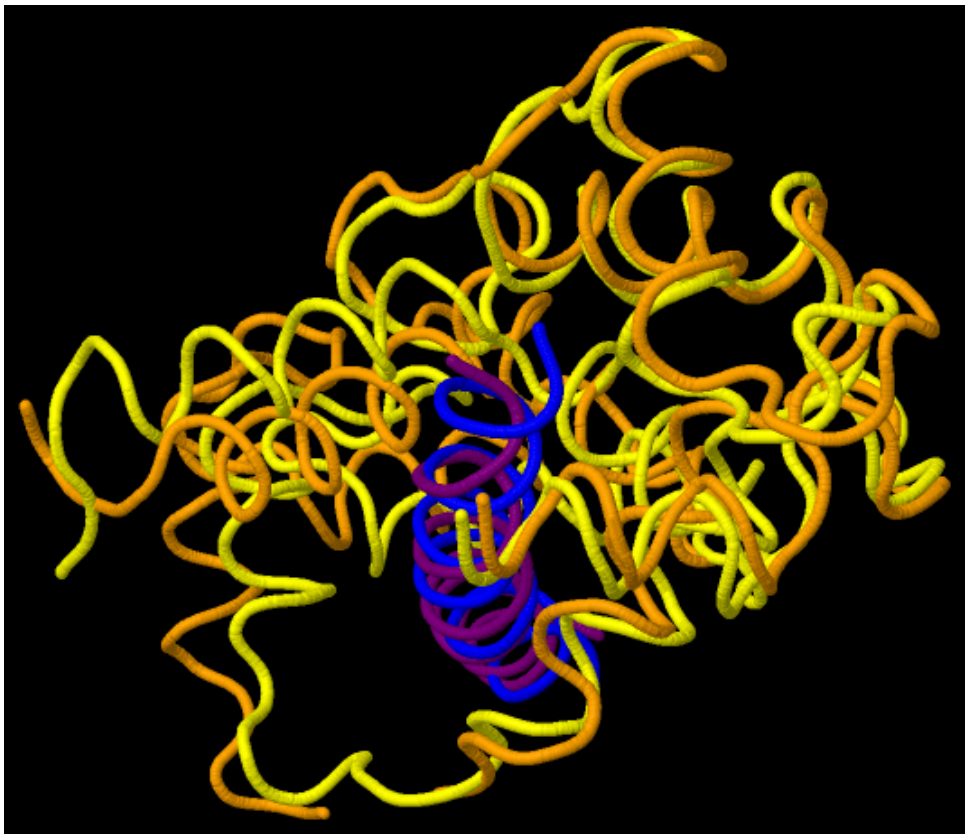

Figure 4.13: Calmodulin complexed with the calmodulin binding domain of calcineurin predicted using HADDOCK with the ligand from PDB 2JZI and the free form from 1EXR. The structure alignment with the bound form 2JZI is shown. The break at the hinge is visible at the bottom.

chemical shift mapping. However, for high-throughput screening, the SAR by NMR method is better suited for validating hits rather than for generating hits from a large compound library. Nevertheless, advances are making it more practical [108], and the number of hits to validate can still be large. According to a senior scientist at Genentech, a pharmaceutical company, "the true strength and potential (of NMR) are found in the follow-up of initial hits. NMR offers a unique set of atomic-level observables for the hit-to-lead advancement not found in other techniques." [71]. Ligand-based methods, such as transferred NOE and saturation transfer difference, are more applicable for high-throughput screening because they require less experimental time, less protein, and are less limited by the size of the target protein (reviewed in [68]). However, ligand-based methods give information only about the ligand, such as whether or not binding is possibly occurring. Chemical shift mapping is then used to obtain target binding site information, which can be used by protein-ligand docking methods to produce the structure of the complex [108]. For large proteins, selective labelling of individual amino acid types, segmental labelling of specific segments of the chain, and deuteration are used to simplify the spectrum [117]. In these cases, peak tracking without contact information may be sufficient for identifying the binding residues. SAR by NMR is based on identifying weakly binding ligands and then linking them to produce a stronger binder. In most cases, these ligands are in fast-exchange, which has limitations in obtaining intermolecular NOEs as discussed in Section 2.4. Our approach, however, does not depend on intermolecular NOEs.

In the next section, we present a model for slow exchange peak tracking and preliminary results for CaM. The results prompted us to build a combined BILP model that combines the peak walking BILP and backbone assignment BILP. This combined model allows peak path information to be incorporated directly into the backbone assignment algorithm. For SAR by NMR, the slow exchange model is more applicable for validating the linked ligand rather than screening for weak binders.

4.6 Slow Exchange BILP

We develop a peak walking model for the slow exchange case, and present preliminary results for calmodulin (CaM). Currently, we are not aware of any automated methods for slow exchange apart from those that

use the same approach as fast exchange. In general, peak tracking is more difficult here because there are no intermediate peaks to track peak movements in increments, and the number of peaks in the spectra can be almost double the number in fast exchange. Similar to the fast exchange case, we model the problem as a k-dimensional matching problem. The difference is that we allow vertices in the graph to represent not only one peak, but also two. In addition, in the scoring function we consider for a pair of peaks their intensities relative to the concentration ratio of the protein and ligand.

We define 3 types of vertices based on 3 different peak/residue states. A *free* vertex represents a peak corresponding to a residue in the free form. A *freebound* vertex represents a pair of peaks corresponding to the same residue in both the free and bound forms. A *bound* vertex represents a peak corresponding to a residue in the bound form only. Figure 4.14 illustrates the possible transitions from each state. From the free state, a residue can transition to any of the 3 states. Note that a transition does not necessarily mean that the chemical shifts have changed. From the freebound state, a residue can remain in this state or transition to the bound state. A residue in the freebound state cannot transition back to the free state. Once in the bound state, a residue must remain there. All peaks in the initial peak list are in the free state. In the final peak list, we assume the protein is fully saturated with the ligand, so no residues are in the freebound state; they must be in one of the other two states. We also allow a residue to transition to a missing state, where its peaks disappear in all subsequent peak lists. A missing transition from the freebound state means that both peaks are missing. Similar to the fast exchange case, the problem is modeled with a BILP.

4.6.1 Binary Variables

The variables represent the transitions/edges between vertices, where each vertex represents a peak or a pair of peaks in some state and from some peak list.

- $X_{hish's'}$ Equals to 1 if peak $h \in T_i$ in state $s \in \{\text{free, bound}\}$, or a pair of peaks h_a, h_b in state $s = \text{freebound}$, transitions to peak $h' \in T_{i+1}$ in state $s' \in \{\text{free, bound}\}$ or a pair of peaks h'_a, h'_b in $s' = \text{freebound}$. For $s' = \text{missing}$, h' is empty.

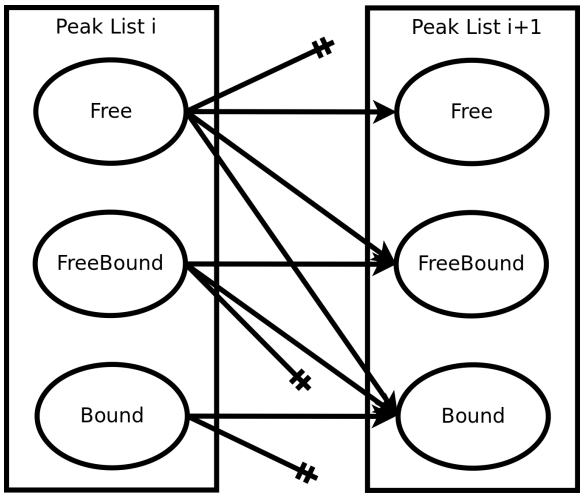

Figure 4.14: Slow exchange peak tracking model. The arrows describe the possible transitions from each state. Transitions with no arrows at the end represent missing transitions, which correspond to the peaks being missing in subsequent peak lists.

4.6.2 Objective Function Coefficients

The objective function is maximized. The score depends on the length of the assigned paths, where shorter paths have higher scores, and for freebound transitions, how well the intensity ratio matches the expected ratio.

- $C(X_{hi[\text{free}]h'[\text{free}]}) = \Phi(\Delta\delta_N(h',\, h),\, 0,\, 0.25) + \Phi(\Delta\delta_{H^N}(h',\, h),\, 0,\, 0.025)$, where Φ is defined in Section 4.3.4.

- $C(X_{hi[\text{free}]h'_a h'_b[\text{freebound}]}) = \Phi(\Delta\delta_N(h'_a,\, h),\, 0,\, 0.25) + \Phi(\Delta\delta_{H^N}(h'_a,\, h),\, 0,\, 0.025) +$
  $\Phi(|\frac{I(h'_a)}{I(h'_a)+I(h'_b)} - R_i|,\, 0,\, 0.15)$, where $I(\cdot)$ gives the intensity of the given peak, R_i is the expected intensity ratio based on the concentration ratio of ligand to protein, and h'_a is closer to h than h'_b is to h based on $\Delta\delta_{NH}$.

- $C(X_{hi[\text{free}]h'[\text{bound}]}) = 0.001$. Since the chemical shift of h' for a given residue can be very different from its h, we set this score to be a constant to give no preference for the different h's in this transition.

- $C(X_{h_a h_b[\text{freebound}]h'_a h'_b[\text{freebound}]}) = \Phi(\Delta\delta_N(h'_a, h_a), 0, 0.25) + \Phi(\Delta\delta_{H^N}(h'_a, h_a), 0, 0.025) + \Phi(\Delta\delta_N(h'_b, h_b), 0, 0.25) + \Phi(\Delta\delta_{H^N}(h'_b, h_b), 0, 0.025) + \Phi(|\frac{I(h'_a)}{I(h'_a)+I(h'_b)} - R_i|, 0, 0.15)$, where h'_a is closer to h_a than to h_b.

- $C(X_{h_a h_b[\text{freebound}]h'_b[\text{bound}]}) = \Phi(\Delta\delta_N(h'_b, h_b), 0, 0.25) + \Phi(\Delta\delta_{H^N}(h'_b, h_b), 0, 0.025)$, where h'_b is closer to h_b than to h_a.

- $C(X_{h_b[\text{bound}]h'_b[\text{bound}]}) = \Phi(\Delta\delta_N(h'_b, h_b), 0, 0.25) + \Phi(\Delta\delta_{H^N}(h'_b, h_b), 0, 0.025)$

- Transitions corresponding to missing peaks have a score of -0.001.

4.6.3 Constraints

- Define the following auxiliary variables for each vertex. $O_{his} = \sum_{h's'} X_{hish's'}$, which represents the sum of the variables corresponding to the out-edges from vertices that contain peak $h \in T_i$ in state s. $I_{his} = \sum_{h's'} X_{h'[i-1]s'hs}$, which represents the sum of the variables corresponding to the in-edges into vertices that contain peak $h \in T_i$ in state s.

- The number of in-edges, and the number of out-edges is bounded by one to prevent path overlap. This is $I_{his} \leq 1$ and $O_{his} \leq 1$, respectively.

- Analogous to the fast-exchange case, we have the number of in-edges equal to the number of out-edges. This is $O_{his} = I_{his}$.

- Define the following auxiliary variables for each peak. $O_{hi} = \sum_{sh's'} X_{hish's'}$, which represents the sum of the variables corresponding to the out-edges from vertices that contain peak $h \in T_i$ in any state. $I_{hi} = \sum_{h's's} X_{h'[i-1]s'hs}$, which represents the sum of the variables corresponding to the in-edges into vertices that contain peak $h \in T_i$ in any state.

- Since a vertex can contain more than one peak, to ensure that each peak gets assigned to at most one state and path, we have $I_{hi} \leq 1$, $O_{hi} \leq 1$, and $O_{hi} = I_{hi}$.

4.6.4 Combined BILP

The BILP for chemical shift mapping can be combined with the BILP for backbone assignment assuming that the number of possible paths per peak is not too large. A combined BILP enables the constraints to work together; e.g. if residue A is supposed to be assigned to peak S, and residue B has a path to peak T, and A and B are in contact, but S and T have no NOEs, then there is evidence against B being assigned to T. Accounting for contact information can help eliminate incorrect paths. The variables and constraints of both the above slow exchange BILP and the backbone assignment BILP are copied to a new BILP. To this new BILP, we add the following.

- Let P_{as} be a path between residue a (reference peak) and spin system s (target peak). Note that there can be more than one path between them. Define a new variable $Y_{P_{as}}$ associated with this path. This variable is equal to one when all the variables corresponding to the transitions in the path are set. That is, the constraint is $Y_{P_{as}} = 1 \longleftrightarrow \sum_{t \in P_{as}} X_t = k$, where k is the path length, which is equal to the number of peak lists-1, and the summation is over all variables corresponding to the transitions (Section 4.6.1). This is equivalent to $Y_{P_{as}} = 1 \rightarrow \sum_{t \in P_{as}} X_t = k$ and $Y_{P_{as}} = 0 \rightarrow \sum_{t \in P_{as}} X_t \leq k-1$, which in turn, is equivalent to $-\sum_{t \in P_{as}} X_t + kY_{P_{as}} \leq 0$ and $\sum_{t \in P_{as}} X_t - Y_{P_{as}} \leq k-1$.
- Since there can be more than one path between a and s, we enforce the constraint $\sum_{P'_{as}} Y_{P'_{as}} \leq 1$, where the summation is over all such paths between a and s.
- Each residue a can be in at most one path. This is $\sum_s \sum_{P_{as}} Y_{P_{as}} \leq 1$
- Each spin system s can be in at most one path. This is $\sum_a \sum_{P_{as}} Y_{P_{as}} \leq 1$
- The path variables from a to s are linked to their corresponding assignment variables $X_{a,s}$ from Section 4.3.3. This is $X_{as} = 1 \leftrightarrow \sum_{P'_{as}} Y_{P'_{as}} = 1$. Since $\sum_{P'_{as}} Y_{P'_{as}} \leq 1$, the constraint can be written as $X_{as} = \sum_{P'_{as}} Y_{P'_{as}}$.

So far, this has been implemented for only slow exchange. A similar BILP applies for fast exchange.

4.6.5 Preliminary Results

For testing CaM, we generated peak lists using the chemical shifts in BMRB 6541 (free form) and BMRB 15624 (bound form). Four peak lists with saturation levels 0:1, 1:4, 3:4, and 1:1 were generated. The number of peaks in each list is 146, 238, 238, and 143, respectively. Residues in 6541 with backbone δ_N, δ_{H^N} within 0.5, 0.05 ppm of their corresponding values in 15624 were assumed to have stayed in the free state. The other residues were assumed to be in the freebound state in the intermediate peak lists. Peak intensities were generated based on the saturation levels. For the case with no noise peaks, no missing NOEs, and no errors, except for 3 residues present in 6541, but not in 15624, fast exchange PeakWalker performed poorly with less than 100 residues correct because it does not use the intensity values. Cutoffs of 2.0 ppm for the N chemical shifts and 0.4 ppm for H^N were used.

We first tested PeakWalkerSlow without backbone assignment; that is, without the combined BILP. No noise was added to the intensities to vary the values away from their expected values, but the level of ambiguity was increased by using an intensity window of 30%, that is an intensity cutoff of $\pm$15% difference from the expected intensity ratio was used to determine whether or not a pair of peaks corresponds to the freebound state. PeakWalkerSlow got 132 correct at 5.7 peaks/residue. All 11 incorrect had chemical shift changes outside the 2.0, 0.4 ppm cutoffs. Unfortunately, CaM undergoes a large conformational change upon binding (hinge motion in a long helix), and those 11 residues are important for either binding or hinge motion. A 4.0, 0.8 ppm cutoff would be needed to cover the chemical shift changes of all residues, but this will result in a prohibitive number of possible peaks per residue. One possible solution to this problem is an iterative approach described below.

If noise is added to the intensity values, but still resulting in the values within $\pm$15% difference of the expected, the accuracy dropped significantly to 107 correct at 7 peaks/residue due to increased ambiguity. We then tried the combined BILP, which used the free form structure 1EXR as the template. NOEs were generated from the bound form 2JZI. The one-to-one assignment had 125 correct, which is better than the solution that allows for 7 possible peaks per residue. There is a computational cost with using a combined BILP, as it had over 82,000 variables and almost 35,000 constraints. An iterative approach that fixes paths, identifies reliable assignments, identifies commonly occurring paths in multiple solutions, and uses

intermolecular NOEs might be able to correct for some of the errors. For peaks matching intermolecular NOEs, perhaps the 2.0, 0.4 ppm cutoffs could be increased to help identify the residues with significant chemical shift changes. This investigation will be future work. So far, the results are for the case with no errors. For the case with errors and missing data, which is the norm, additional data will likely be needed to reduce the ambiguity. If the hinge region and binding site location can be narrowed down, such as with segmental isotope labeling, then the search problem becomes much easier. This comes with an added financial cost, but the time savings from avoiding additional experiments might be worth the cost.

Chapter 5

Conclusion and Future Work

Structure-based assignment offers the possibility to bridge experimental structure determination with computational prediction methods. If there is no homolog, structure prediction methods, such as threading, can be used to obtain a template. Structure-based assignment methods can then be used to measure the agreement between the NOEs and the contacts in the predicted structure. In CASD-NMR [99], which is a protein structure determination by NMR competition, participants are provided with backbone and side chain chemical shift assignment, and unassigned NOEs. Given this information, CS-DP-Rosetta [94], uses the DP-score [45], to measure this agreement. The score is used as a filter to select the final structure model from a large set of candidate models. The score is similar to our assignment 1.0 method in that NOEs are not directly assigned to contacts. The difference is that the backbone assignment is provided as the input. It is of interest whether an assignment 2.0 approach would produce better structures. Conformation sampling methods would then be needed to sample contacts not in the template structure, but in the unknown NMR structure and hidden in the set of NOEs.

Note that backbone NOEs normally provide only local contact information; e.g., local contacts corresponding to adjacent residues and alpha helices. Non-local beta sheet contacts are also visible, but in general backbone NOEs do not provide a sufficient number of non-local contacts for structure determination because structures with different tertiary structure can have the same local contact patterns.

Non-local (long range) contacts, such as side chain information from doubled-labeled NMR experiments, or global orientation information, such as RDCs, is necessary for structure determination. In the case of large proteins, where the protein is often deuterated, the set of distance restraints is usually sparse. One can supplement this set with additional distance restraints from consensus contacts from an ensemble of possible templates. This assumes that the unknown structure also has these common structural fragments due to evolutionary reasons. The advantage of structure prediction methods is that they can predict such templates without a dense set of NOEs, which is required by traditional NMR structure determination methods. Our approach of using limited NMR data is a "mini-version" of large protein NMR, where the data is very sparse, so backbone-only approaches and structure-based approaches become much more important. Our research provides a glimpse of the computational challenges that should arise there.

Local contact information is often sufficient to obtain the backbone assignment if given a homologous structure because sequential connectivity and secondary structure contacts can be exploited for linking the spin systems. This backbone assignment information, however, when combined with interface information has the potential for 3D complex determination if flexible docking methods are able to use the interface information with an energy function to fold the protein as it undergoes conformational change upon binding. This is not yet a reality, but the trend is towards complementing wet labs with *in-silico* methods that enable experimental efforts to be reduced. Another difficulty is determining the contacts and hinge residue with NMR, especially in the slow exchange case. Sequence-only approaches for detecting the hinge residue [33] can be used to help differentiate between binding interface and hinge residues. NMR data is typically sparse, and assignment errors are typically present. The tolerance of docking methods to these issues would need to be investigated. Since peak tracking is easier in the fast-exchange case, and SAR drug screening is based mainly on this case, our peak walking approach has potential to speed up these studies. Ultimately, to confirm the utility of our methods, a large-scale study on a drug library will need to be performed. This will require calibration with pharmaceutical companies. To encourage collaboration, we must build an easy-to-use web service that automates the peak list to structure/complex pipeline while allowing for user intervention at each step. This is the long-term goal. The short-term goal is to tackle the conformational change problem. This will likely require using side chain NOEs or RDCs, and conformation

sampling methods from the protein structure prediction and protein-ligand docking field.

When this thesis was conceived, the original topic was fully automated assignment. Computers and NMR have been around for decades, but fully automated methods have still not caught on. The vast number of possible deviations from the norm due to noise, missing data, ambiguity, and unexpected observations make it difficult to trust full automation as a routine method; plus it is difficult for automated methods to account for every possible case. Currently, only semi-automated methods are routine. Instead of full automation, we focused on using limited NMR data, so that computer methods can still take on an important role. However, full *in-silico* structure prediction, especially for large complexes, is currently a challenge because of the size of the sample space and the poor ability of current docking energy functions to accurately capture the binding affinity [75]. It is known that local docking (when the binding site is limited to a specific location) is a much easier problem than global docking, so a combination of fast *in vitro* and *in silico* methods is the way to go for studying proteomes and interactomes.

Appendix A

Ubiquitin Spin Systems from Manually Picked Peaks

The first two columns give the reference assignment, followed by the spin system chemical shifts, and then the amino acid type predictions as single letter codes.

2 GLN 123.077 8.942 5.26 2.22 1.86 1.6 K R I L Q E

3 ILE 115.284 8.303 4.135 1.77 0.626 A I L V

4 PHE 118.681 8.606 5.622 3.02 2.85 F Y H D N

5 VAL 121.423 9.299 4.77 1.89 0.68 V I L

6 LYS 128.011 8.913 5.286 2.9 1.689 1.377 K R V I L

7 THR 115.67 8.747 4.904 1.18 A

8 LEU 121.5 9.13 4.77 4.28 1.91 1.77 1.01 I L T

9 THR 106.043 7.636 4.8 4.401 1.26 G T

10 GLY 109.377 7.823 4.319 3.583 G

11 LYS 122.085 7.262 4.323 2.89 1.67 1.389 1.21 K R I L

12 THR 120.784 8.646 5.03 3.93 1.05 T

13 ILE 127.85 9.538 4.5 1.87 0.85 I L A V

14 THR 121.865 8.738 4.95 4.02 1.11 T

15 LEU 125.305 8.733 4.735 1.35 1.2 0.7 A T V I L

16 GLU 122.656 8.121 4.87 2.21 2.08 1.86 E Q M

17 VAL 117.705 8.938 4.68 2.3 0.7 0.4 V L

18 GLU 119.465 8.653 5.051 2.3 2.165 1.57 E Q M

20 SER 103.593 7.023 4.342 4.13 3.8 G S T

21 ASP 124.065 8.046 4.657 4.17 2.92 2.54 1.62 E Q M D N

22 THR 109.184 7.888 4.882 4.34 3.6 1.26 T S

23 ILE 121.407 8.519 3.63 2.47 0.79 F Y H D N V I L

25 ASN 121.543 7.924 4.523 3.188 2.844 F Y H D N

26 VAL 122.314 8.099 3.39 2.35 0.97 0.7 V I L R K

27 LYS 119.146 8.56 4.565 2.15 1.43 V I L

28 ALA 123.66 7.978 4.14 1.61 A

29 LYS 120.402 7.859 4.182 2.14 1.8 1.577 M Q E R K

30 ILE 121.514 8.281 3.477 2.36 0.66 V I L K R

31 GLN 123.775 8.55 3.815 2.487 2.228 1.933 1.5 K R E Q

32 ASP 119.914 8.019 4.315 2.782 A D N

33 LYS 115.613 7.421 4.291 3.14 2 1.82 1.604 K R E Q M

34 GLU 114.482 8.722 4.553 2.105 1.669 V E Q M

35 GLY 109.052 8.506 4.115 3.9 G

36 ILE 120.466 6.148 4.41 1.4 0.91 V I L

39 ASP 113.808 8.526 4.4 2.71 D N Y

40 GLN 117.051 7.819 4.44 2.402 1.81 Q E M

41 GLN 118.227 7.484 4.2 2.52 1.907 1.66 Q E M R K

42 ARG 123.207 8.5 4.458 3.07 1.677 1.446 R K

43 LEU 124.586 8.832 5.356 1.527 1.146 0.76 V I L

44 ILE 122.401 9.082 4.93 1.73 0.66 V I L

45 PHE 125.362 8.845 5.118 3 2.8 F H D N

46 ALA 133.047 8.947 4.77 3.7 0.85 T

47 GLY 102.654 8.126 4.071 3.42 G S

48 LYS 122.178 7.978 4.59 3.11 1.87 1.49 K R I

49 GLN 123.153 8.636 4.53 2.22 1.96 D N Q E M

50 LEU 125.88 8.556 4.063 1.47 1 T V A

51 GLU 123.309 8.387 4.477 2.371 1.961 Q E M

52 ASP 120.535 8.162 4.36 2.543 D N

54 ARG 119.515 7.463 4.702 3.11 2.221 2 1.8 1.6 R K

55 THR 108.986 8.827 5.214 1.11 G A T

56 LEU 118.214 8.149 4.037 2.1 1.2 V Q E A

57 SER 113.696 8.478 4.227 3.773 G S T

58 ASP 124.691 7.932 4.263 2.98 2.274 E Q M D N

59 TYR 115.947 7.254 4.628 3.468 2.507 E Q M D N

60 ASN 116.179 8.152 4.327 3.287 2.77 E Q M D N H

61 ILE 119.064 7.246 3.36 1.4 1.09 0.44 -0.36 I L V

62 GLN 125.127 7.62 4.46 2.289 1.89 E Q M D N

63 LYS 120.763 8.495 3.953 2 1.9 1.48 G E Q M

64 GLU 114.733 9.311 3.316 2.44 2.22 D N H S E Q

65 SER 115.123 7.662 4.612 3.91 3.61 S

66 THR 117.579 8.737 5.26 4.08 0.91 T

67 LEU 127.837 9.42 5.07 1.612 0.65 I L V

68 HIS 119.307 9.234 5.122 3.1 2.896 D N H

69 LEU 124.223 8.291 5.157 1.62 1.32 1.1 0.86 0.75 I L K R

70 VAL 126.921 9.189 4.35 2 0.86 V A

71 LEU 123.274 8.109 5 1.65 0.926 I L V

72 ARG 123.967 8.593 4.241 3.15 1.73 1.5 K R

```
73 LEU 124.723 8.352 4.374 1.6 0.89 I L V
74 ARG 122.163 8.438 4.269 3.19 1.8 1.63 K R T
75 GLY 111.275 8.491 4.76 3.946 G S
76 GLY 115.236 7.942 3.763 3.754 G
```

The following are spin systems that were filtered out as being non-backbone or noise; i.e., those with TOCSY and NOE patterns resembling side chain N, H^N chemical shifts, and those with no putative H^α or H^N.

```
112.571 6.820
112.583 7.281
111.646 7.543 6.816
112.228 7.728 6.739 4.770
111.645 6.815 7.537
110.065 7.838 6.887
112.224 6.736 7.721
110.273 7.654 6.808
110.064 6.887 7.836
111.382 6.747 7.684
110.272 6.813 7.651
111.369 7.682 6.747
112.554 7.669 6.835 4.770
124.853 7.037
96.448 4.748
124.183 7.251
123.875 7.119
123.871 7.188
```

Appendix B

Absolute Value as Linear Constraints

Instead of binary, let X be an integer variable. Let $|X| \leq M$ for some positive integer M. To implement $|X| \geq K$, where K is a positive integer, we need 4 constraints. A $\equiv$ B shall be used to denote that A is equivalent to B. First define a binary variable Z, such that

1. $(X \geq 0 \Rightarrow Z = 1) \wedge (X < 0 \Rightarrow Z = 0) \equiv$

   (a) $X \geq 0 \Rightarrow Z = 1 \equiv X \leq (M+1)Z - 1$. To see this, whenever $X \geq 0$, Z is forced to 1 (cannot be 0) in order for the inequality to hold; but when $X < 0$, Z can be either 1 or 0 and the inequality still holds.

   (b) $X < 0 \Rightarrow Z = 0 \equiv MZ \leq X + M$. To see this, whenever, $X < 0$, Z is forced to 0 in order for the inequality to hold; but when $X \geq 0$, Z can be either 1 or 0 and the inequality still holds.

2. Now back to $|X| \geq K$. This is equivalent to $X \geq K$ when $X \geq 0$, and $-X \geq K$ when $X < 0$.

   (a) When $X \geq 0$, we have case 1(a). We want $Z = 1 \Rightarrow X \geq K \equiv (1 - Z)(-K - M) \leq X - K$. This can be seen by substituting $Z = 1$ to give the desired inequality, and $Z = 0$ to give an inequality that is always true for all values of X.

(b) When $X < 0$. we have case 1(b). We want $Z = 0 \Rightarrow -X \geq K \equiv Z(K + M) \geq X + K$. This can be seen by substituting $Z = 0$ to give the desired inequality, and $Z = 1$ to give an inequality that is always true for all values of X.

To implement $A \Leftrightarrow |X| \geq K$, consider the above 4 constraints. Part 1 is the same, but we replace Part 2 with the condition that $|X| \geq K$ holds only when $A = 1$, rather than always holding. Note that $Y = 1 \Rightarrow X = 1 \equiv \overline{Y} \vee X$. The new constraints consist of the following, where A_+ and A_- are binary variables.

1. $(A_+ = 1 \Leftrightarrow X \geq K)$

   (a) Define $C_{X \geq K} \Leftrightarrow X \geq K$. It is $-M(1 - C_{X \geq K}) + KC_{X \geq K} \leq X$ and $X \leq MC_{X \geq K} + K - 1$. This can be seen by substituting values for the binary variable $C_{X \geq K}$ as in Section 4.2.1.

   (b) Define $C_{X \leq K-1} \Leftrightarrow X \leq K - 1 \equiv (C_{X \leq k-1} = 1 \Rightarrow X \leq K - 1) \wedge (C_{X \leq k-1} = 0 \Rightarrow X \geq K)$. It is $X \leq M(1 - C_{X \leq K-1}) + K - 1$ and $X \geq K(1 - C_{X \leq K-1}) - MC_{X \leq K-1}$.

   (c) $(A_+ = 1 \Leftrightarrow X \geq K) \equiv (A_+ = 1 \Rightarrow C_{X \geq K} = 1) \wedge (A_+ = 0 \Rightarrow C_{X \leq K-1} = 1) \equiv (\bar{A_+} \vee C_{X \geq K}) \wedge (A_+ \vee C_{X \leq K-1}) \equiv C_{\bar{A_+} \vee X \geq K} \wedge C_{A_+ \vee X \leq K-1}$, where the C's are obtained by reifying the OR statements using the result in Section 4.2.2. $C_{\bar{A_+} \vee X \geq K} \wedge C_{A_+ \vee X \leq K-1} \equiv C_{(\bar{A_+} \vee C_{X \geq K}) \wedge (A_+ \vee C_{X \leq K-1})}$, which is obtained by reifying the AND statement.

2. $(A_- = 1 \Leftrightarrow -X \geq K)$

   (a) Define $C_{-X \geq K} \Leftrightarrow -X \geq K$. It is $-M(1 - C_{-X \geq K}) + KC_{-X \geq K} \leq -X$ and $-X \leq MC_{-X \geq K} + K - 1$.

   (b) Define $C_{-X \leq K-1} \Leftrightarrow -X \leq K - 1$. It is $-X \leq M(1 - C_{-X \leq K-1}) + K - 1$ and $-X \geq K(1 - C_{-X \leq K-1}) - MC_{-X \leq K-1}$.

   (c) $(A_- = 1 \Leftrightarrow -X \geq K) \equiv (A_- = 1 \Rightarrow C_{-X \geq K} = 1) \wedge (A_- = 0 \Rightarrow C_{-X \leq K-1} = 1) \equiv (\bar{A_-} \vee C_{-X \geq K}) \wedge (A_- \vee C_{-X \leq K-1}) \equiv C_{\bar{A_-} \vee (-X \geq K)} \wedge C_{A_- \vee (-X \leq K-1)} \equiv C_{(\bar{A_-} \vee C_{-X \geq K}) \wedge (A_- \vee C_{-X \leq K-1})}$, obtained by reifying the OR and then the AND statement.

3. $(A = 1 \Leftrightarrow |X| \geq K) \equiv (A = 1 \Leftrightarrow ((X \geq 0 \wedge A_+) \vee (X < 0 \wedge A_-)) \equiv (A = 1 \Leftrightarrow ((Z \wedge A_+) \vee (\overline{Z} \wedge A_-))$. Each of the two AND terms can be reified using 2 constraints each as in Section 4.2.2, and then the OR term can be reified using 2 constraints to give $A \Leftrightarrow C_{(Z \wedge A_+) \vee (\overline{Z} \wedge A_-)}$, which is $A = C_{(Z \wedge A_+) \vee (\overline{Z} \wedge A_-)}$.

References

[1] Jmol: an open-source java viewer for chemical structures in 3D. `http://www.jmol.org`.

[2] Southeast collaboratory for structural genomics. `http://www.secsg.org/methods.htm`, Accessed June 2011.

[3] Tobias Achterberg, Timo Berthold, Thorsten Koch, and Kati Wolter. Constraint integer programming: A new approach to integrate CP and MIP. In Laurent Perron and Michael A. Trick, editors, *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems, 5th International Conference, CPAIOR 2008*, volume 5015 of *Lecture Notes in Computer Science*, pages 6–20. Springer, 2008.

[4] Babak Alipanahi, Xin Gao, Emre Karakoc, Logan Donaldson, and Ming Li. PICKY: a novel SVD-based NMR spectra peak picking method. *Bioinformatics*, 25(12):268–275, Jun 2009.

[5] Babak Alipanahi, Xin Gao, Emre Karakoc, Shuai Cheng Li, Frank Balbach, Guangyu Feng, Logan Donaldson, and Ming Li. Error tolerant NMR backbone resonance assignment and automated structure generation. *J Bioinform Comput Biol*, 9(1):15–41, Feb 2011.

[6] M.S. Apaydin, B. Catay, N. Patrick, and B.R. Donald. NVR-BIP: Nuclear vector replacement using binary integer programming for NMR structure-based assignments. *The Computer Journal*, page bxp120, 2010.

[7] M.S. Apaydin, V. Conitzer, and B.R. Donald. Structure-based protein NMR assignments using native structural ensembles. *J. Biomol. NMR*, 40(4):263–276, 2008.

[8] Konstantin Arnold, Lorenza Bordoli, Jürgen Kopp, and Torsten Schwede. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201, Jan 2006.

[9] C. Bailey-Kellogg, A. Widge, J. Kelly, J. Brushweller, and B.R. Donald. The NOESY jigsaw: Automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J. Comput. Biol.*, 7:537–558, 2000.

[10] Egon Balas, Fred Glover, and Stanley Zionts. An additive algorithm for solving linear programs with zero-one variables. *Operations Research*, 13(4):517–549, July 1965.

[11] Michael C Baran, Yuanpeng J Huang, Hunter N B Moseley, and Gaetano T Montelione. Automated analysis of protein NMR assignments and structures. *Chem Rev*, 104(8):3541–3556, Aug 2004.

[12] Ravi Pratap Barnwal, Ashok K Rout, Kandala V R Chary, and Hanudatta S Atreya. Rapid measurement of 3J(H N-H alpha) and 3J(N-H beta) coupling constants in polypeptides. *J Biomol NMR*, 39(4):259–263, Dec 2007.

[13] H.G. Barrow and R.M. Burstall. Subgraph isomorphism, matching relational structures and maximal cliques. *Information Processing Letters*, 4(4):83–84, January 1976.

[14] Christian Bartels, Martin Billeter, Peter Güntert, and Kurt Wüthrich. Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. *Journal of Biomolecular NMR*, 7(3):207–213, 1996-05-01.

[15] Christian Bartels, Peter Güntert, Martin Billeter, and Kurt Wüthrich. GARANT-a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *J. Comput. Chem.*, 18(1):139–149, 1997.

[16] Mark V Berjanskii, Stephen Neal, and David S Wishart. PREDITOR: a web server for predicting protein torsion angle restraints. *Nucleic Acids Res*, 34(Web Server issue):W63–W69, Jul 2006.

[17] 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 Res*, 28(1):235–242, Jan 2000.

[18] Guillermo A Bermejo and Miguel Llinás. Structure-oriented methods for protein NMR data analysis. *Prog Nucl Magn Reson Spectrosc*, 56(4):311–328, May 2010.

[19] Michael Bieri, Ann H Kwan, Mehdi Mobli, Glenn F King, Joel P Mackay, and Paul R Gooley. Macromolecular NMR spectroscopy for the non-spectroscopist: beyond macromolecular solution structure determination. *FEBS J*, 278(5):704–715, Mar 2011.

[20] Roslyn M Bill, Peter J F Henderson, So Iwata, Edmund R S Kunji, Hartmut Michel, Richard Neutze, Simon Newstead, Bert Poolman, Christopher G Tate, and Horst Vogel. Overcoming barriers to membrane protein structure determination. *Nat Biotechnol*, 29(4):335–340, Apr 2011.

[21] Martin Billeter, Gerhard Wagner, and Kurt Wüthrich. Solution NMR structure determination of proteins revisited. *J Biomol NMR*, 42(3):155–158, Nov 2008.

[22] I.M. Bomze, M. Budinich, P.M. Pardalos, and M. Pelillo. The maximum clique problem. In *Handbook of Combinatorial Optimization*, pages 1–74. Kluwer Academic Publishers, 1999.

[23] Alexander L Breeze. Isotope-filtered NMR methods for the study of biomolecular structure and interactions. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 36(4):323–372, 2000.

[24] R. Burkard, M. Dell'Amico, and S. Martello. *Assignment Problems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2009.

[25] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 2004.

[26] Pierluigi Crescenzi and Viggo Kann. A compendium of np optimization problems. `http://www.csc.kth.se/~viggo/wwwcompendium/wwwcompendium.html`, Accessed July 2011.

[27] Yang Daiwen, Xu Yingqi, and Zheng Yu. NOESY-based strategy for assignments of backbone and side chain resonances of large proteins without deuteration. *Nature Protocol Exchange*, May 2006.

[28] Charlotta S Damberg, Vladislav Yu Orekhov, and Martin Billeter. Automated analysis of large sets of heteronuclear correlation spectra in NMR-based drug discovery. *J Med Chem*, 45(26):5649–5654, Dec 2002.

[29] E. Danna, M. Fenelon, Z. Gu, and R. Wunderling. Generating multiple solutions for mixed integer programming problems. *Integer Programming and Combinatorial Optimization*, pages 280–294, 2007.

[30] Sjoerd J de Vries, Marc van Dijk, and Alexandre M J J Bonvin. The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc*, 5(5):883–897, 2010.

[31] J. F. Doreleijers, M. L. Raves, T. Rullmann, and R. Kaptein. Completeness of NOEs in protein structure: a statistical analysis of NMR. *J Biomol NMR*, 14(2):123–132, Jun 1999.

[32] C. Eichmüller, W. Schüler, R. Konrat, and B. Kräutler. Simultaneous measurement of intra- and intermolecular NOEs in differentially labeled protein-ligand complexes. *J Biomol NMR*, 21(2):107–116, Oct 2001.

[33] Ugur Emekli, Dina Schneidman-Duhovny, Haim J Wolfson, Ruth Nussinov, and Turkan Haliloglu. Hingeprot: automated prediction of hinges in protein structures. *Proteins*, 70(4):1219–1227, Mar 2008.

[34] J. Fejzo, C. A. Lepre, J. W. Peng, G. W. Bemis, Ajay, M. A. Murcko, and J. M. Moore. The SHAPES strategy: an NMR-based approach for lead generation in drug discovery. *Chem Biol*, 6(10):755–769, Oct 1999.

[35] Lisa Fukui and Yuan Chen. NvMap: automated analysis of NMR chemical shift perturbation data. *Bioinformatics*, 23(3):378–380, Feb 2007.

[36] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman, 1979.

[37] T. D. Goddard and D. G. Kneller. Sparky 3. University of California, San Francisco.

[38] Marek Grabowski, Maksymilian Chruszcz, Matthew D Zimmerman, Olga Kirillova, and Wladek Minor. Benefits of structural genomics for drug discovery research. *Infect Disord Drug Targets*, 9(5):459–474, Nov 2009.

[39] P. Greistorfer, A. Lokketangen, S. Vob, and D. Woodruff. Experiments concerning sequential versus simultaneous maximization of objective function and distance. *Journal of Heuristics*, 14(6):613–625, 2008.

[40] Alexander Grishaev and Miguel Llinás. CLOUDS, a protocol for deriving a molecular proton density via NMR. *Proc Natl Acad Sci U S A*, 99(10):6707–6712, May 2002.

[41] Paul Guerry and Torsten Herrmann. Advances in automated NMR protein structure determination. *Q Rev Biophys*, pages 1–53, Mar 2011.

[42] Philip J Hajduk. SAR by NMR: putting the pieces together. *Mol Interv*, 6(5):266–272, Oct 2006.

[43] Philip J Hajduk and Jonathan Greer. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat Rev Drug Discov*, 6(3):211–219, Mar 2007.

[44] R. Harris. The ubiquitin NMR resource page. `http://www.biochem.ucl.ac.uk/bsm/nmr/ubq/index.html`.

[45] Yuanpeng J Huang, Robert Powers, and Gaetano T Montelione. Protein NMR recall, precision, and f-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *J Am Chem Soc*, 127(6):1665–1674, Feb 2005.

[46] S.J. Hubbard and J.M. Thornton. *'NACCESS', Computer Program*. Department of Biochemistry and Molecular Biology, University College London, 1993.

[47] D.P. Huttenlocher and E.W. Jaquith. Computing visual correspondence: incorporating the probability of a false match. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pages 515 –522, jun 1995.

[48] Hiroshi Imai and Takao Asano. Finding the connected components and a maximum clique of an intersection graph of rectangles in the plane. *Journal of Algorithms*, 4(4):310 – 323, 1983.

[49] Richard Jang, Xin Gao, and Ming Li. Towards fully automated structure-based NMR resonance assignment of 15N-labeled proteins from automatically picked peaks. *Journal of Computational Biology*, 18(3):347–363, 2011.

[50] Richard Jang, Xin Gao, and Ming Li. Combining automated peak tracking in SAR by NMR with structure-based backbone assignment from 15n-noesy. *BMC Bioinformatics*, 13(Suppl 3):S4, 2012.

[51] Hao Jiang, S. Fels, and J. J. Little. A linear programming approach for multiple object tracking. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition CVPR '07*, pages 1–8, 2007.

[52] Young-Sang Jung and Markus Zweckstetter. Backbone assignment of proteins with known structure using residual dipolar couplings. *J Biomol NMR*, 30(1):25–35, Sep 2004.

[53] Young-Sang Jung and Markus Zweckstetter. Mars – robust automatic backbone assignment of proteins. *J Biomol NMR*, 30(1):11–23, Sep 2004.

[54] D. Justice and A. Hero. A binary linear programming formulation of the graph edit distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(8):1200–1214, Aug. 2006.

[55] Viggo Kann. Maximum bounded 3-dimensional matching is MAX SNP-complete. *Inf. Process. Lett.*, 37(1):27–35, 1991.

[56] Ezgi Karaca and Alexandre M J J Bonvin. A multidomain flexible docking approach to deal with large conformational changes in the modeling of biomolecular complexes. *Structure*, 19(4):555–565, Apr 2011.

[57] S. L. Kazmirski, K. B. Wong, S. M. Freund, Y. J. Tan, A. R. Fersht, and V. Daggett. Protein folding from a highly disordered denatured state: the folding pathway of chymotrypsin inhibitor 2 at atomic resolution. *Proc Natl Acad Sci U S A*, 98(8):4349–4354, Apr 2001.

[58] Rochus Keller. *The Computer Aided Resonance Assignment Tutorial*. CANTINA Verlag, 2004.

[59] Florian Kiefer, Konstantin Arnold, Michael Künzli, Lorenza Bordoli, and Torsten Schwede. The SWISS-MODEL repository and associated resources. *Nucleic Acids Res*, 37(Database issue):D387–D392, Jan 2009.

[60] Dima Kozakov, Ryan Brenke, Stephen R Comeau, and Sandor Vajda. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*, 65(2):392–406, Nov 2006.

[61] Dima Kozakov, David R Hall, Dmitri Beglov, Ryan Brenke, Stephen R Comeau, Yang Shen, Keyong Li, Jiefu Zheng, Pirooz Vakili, Ioannis Ch Paschalidis, and Sandor Vajda. Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19. *Proteins*, 78(15):3124–3130, Nov 2010.

[62] Janarthanan Krishnamoorthy, Victor C K Yu, and Yu-Keung Mok. Auto-FACE: an NMR based binding site mapping program for fast chemical exchange protein-ligand systems. *PLoS One*, 5(2):e8943, 2010.

[63] H. Kuboniwa, S. Grzesiek, F. Delaglio, and A. Bax. Measurement of hn-h alpha J couplings in calcium-free calmodulin using new 2d and 3d water-flip-back methods. *J Biomol NMR*, 4(6):871–878, Nov 1994.

[64] Harold W. Kuhn. The hungarian method for the assignment problem, 2010.

[65] C.J. Langmead and B.R. Donald. An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments. *J. Biomol. NMR*, 29(2):111–138, 2004.

[66] C.J. Langmead, A. Yan, R. Lilien, L. Wang, and B.R. Donald. A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. *J. Comput. Biol.*, 11(2-3):277–298, 2004.

[67] Alexander Lemak, Carlos A Steren, Cheryl H Arrowsmith, and Miguel Llinás. Sequence specific resonance assignment via multicanonical monte carlo search using an ABACUS approach. *J Biomol NMR*, 41(1):29–41, May 2008.

[68] Christian Ludwig and Ulrich L Guenther. Ligand based NMR methods for drug discovery. *Front Biosci*, 14:4565–4574, 2009.

[69] B. Ma, S. Kumar, C. J. Tsai, and R. Nussinov. Folding funnels and binding mechanisms. *Protein Eng*, 12(9):713–720, Sep 1999.

[70] Antoine Marin, Thérèse E Malliavin, Pierre Nicolas, and Marc-André Delsuc. From NMR chemical shifts to amino acid types: investigation of the predictive power carried by nuclei. *J. Biomol. NMR*, 30(1):47–60, Sep 2004.

[71] Till Maurer. Advancing fragment binders to lead-like compounds using ligand and protein-based NMR spectroscopy. *Methods Enzymol*, 493:469–485, 2011.

[72] Moriz Mayer and Bernd Meyer. Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angewandte Chemie International Edition*, 38(12):1784–1788, 1999.

[73] Ales Medek, Philip J. Hajduk, Jamey Mack, and Stephen W. Fesik. The use of differential chemical shifts for determining the binding site location and orientation of protein-bound ligands. *Journal of the American Chemical Society*, 122(6):1241–1242, 2000.

[74] Jens Meiler and David Baker. Rapid protein fold determination using unassigned NMR data. *Proc Natl Acad Sci U S A*, 100(26):15404–15409, Dec 2003.

[75] Adrien S.J. Melquiond, Ezgi Karaca, Panagiotis L. Kastritis, and Alexandre M.J.J. Bonvin. Next challenges in proteinprotein docking: from proteome to interactome and beyond. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2011.

[76] Anthony Mittermaier and Lewis E Kay. New tools provide new insights in NMR studies of protein dynamics. *Science*, 312(5771):224–228, Apr 2006.

[77] Rolf Möhring. Algorithmic graph theory and perfect graphs. *Order*, 3(2):207–208, June 1986.

[78] Yu-Keung Mok. Auto-FACE download. `http://www.dbs.nus.edu.sg/staff/henry.htm`, accessed May 2010.

[79] Pierre Montaville and Nadège Jamin. Determination of membrane protein structures using solution and solid-state NMR. In Jean-Jacques Lacapère, editor, *Membrane Protein Structure Determination*, volume 654 of *Methods in Molecular Biology*, pages 261–282. Humana Press, 2010.

[80] Gaetano T Montelione and Thomas Szyperski. Advances in protein NMR provided by the NIGMS protein structure initiative: impact on drug discovery. *Curr Opin Drug Discov Devel*, 13(3):335–349, May 2010.

[81] C. Morefield. Application of 0-1 integer programming to multitarget tracking problems. *Automatic Control, IEEE Transactions on*, 22(3):302–312, 1977.

[82] Srayanta Mukherjee and Yang Zhang. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res*, 37(11):e83, Jun 2009.

[83] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, Apr 1995.

[84] Viswanath Nagarajan and Maxim Sviridenko. On the maximum quadratic assignment problem. *Math. Oper. Res.*, 34:859–868, November 2009.

[85] Stephen Neal, Alex M Nip, Haiyan Zhang, and David S Wishart. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *Journal of Biomolecular NMR*, 26(3):215–240, 2003.

[86] NESG Wiki. Common nmr experiment sets. `http://www.nmr2.buffalo.edu/nesg.wiki/Common_NMR_experiment_sets`, accessed January 2012.

[87] Daniel Nietlispach, Helen R Mott, Katherine M Stott, Peter R Nielsen, Abarna Thiru, and Ernest D Laue. Structure determination of protein complexes by NMR. *Methods Mol Biol*, 278:255–288, 2004.

[88] Norma H Pawley, Jason D Gans, and Ryszard Michalczyk. APART: automated preprocessing for NMR assignments with reduced tedium. *Bioinformatics*, 21(5):680–682, 2005.

[89] Manuel C. Peitsch. Protein modeling by E-mail. *Nat Biotech*, 13(7):658–660, July 1995.

[90] Maurizio Pellecchia, Ivano Bertini, David Cowburn, Claudio Dalvit, Ernest Giralt, Wolfgang Jahnke, Thomas L James, Steve W Homans, Horst Kessler, Claudio Luchinat, Bernd Meyer, Hartmut Oschkinat, Jeff Peng, Harald Schwalbe, and Gregg Siegal. Perspectives on NMR in drug discovery: a technique comes of age. *Nat Rev Drug Discov*, 7(9):738–745, Sep 2008.

[91] Chen Peng, Stephen W Unger, Fabian V Filipp, Michael Sattler, and Sándor Szalma. Automated evaluation of chemical shift perturbation spectra: New approaches to quantitative analysis of receptor-ligand interaction NMR spectra. *J Biomol NMR*, 29(4):491–504, Aug 2004.

[92] Andrew M Petros, Jeffrey R Huth, Thorsten Oost, Cheol-Min Park, Hong Ding, Xilu Wang, Haichao Zhang, Paul Nimmer, Renaldo Mendoza, Chaohong Sun, Jamey Mack, Karl Walter, Sarah Dorwin, Emily Gramling, Uri Ladror, Saul H Rosenberg, Steven W Elmore, Stephen W Fesik, and Philip J Hajduk. Discovery of a potent and selective bcl-2 inhibitor using SAR by NMR. *Bioorg Med Chem Lett*, 20(22):6587–6591, Nov 2010.

[93] J. L. Pons and M. A. Delsuc. RESCUE: An artificial neural network tool for the NMR spectral assignment of proteins. *J. Biomol. NMR*, 15(1):15–26, 1999.

[94] Srivatsan Raman, Yuanpeng J Huang, Binchen Mao, Paolo Rossi, James M Aramini, Gaohua Liu, Gaetano T Montelione, and David Baker. Accurate automated protein NMR structure determination using unassigned NOESY data. *J Am Chem Soc*, 132(1):202–207, Jan 2010.

[95] Srivatsan Raman, Oliver F Lange, Paolo Rossi, Michael Tyka, Xu Wang, James Aramini, Gaohua Liu, Theresa A Ramelot, Alexander Eletsky, Thomas Szyperski, Michael A Kennedy, James Prestegard, Gaetano T Montelione, and David Baker. NMR structure determination for larger proteins using backbone-only data. *Science*, 327(5968):1014–1018, Feb 2010.

[96] J.W. Raymond and P. Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.*, 16(7):521–533, 2002.

[97] Michele F Rega, Bainan Wu, Jun Wei, Ziming Zhang, Jason F Cellitti, and Maurizio Pellecchia. SAR by interligand nuclear overhauser effects (ILOEs) based discovery of acylsulfonamide compounds active against Bcl-x(L) and Mcl-1. *J Med Chem*, 54(17):6000–6013, Sep 2011.

[98] Mikhail Reibarkh, Thomas J Malia, Brian T Hopkins, and Gerhard Wagner. Identification of individual protein-ligand NOEs in the limit of intermediate exchange. *J Biomol NMR*, 36(1):1–11, Sep 2006.

[99] Antonio Rosato, James M Aramini, Cheryl Arrowsmith, Anurag Bagaria, David Baker, Andrea Cavalli, Jurgen F Doreleijers, Alexander Eletsky, Andrea Giachetti, Paul Guerry, Aleksandras Gutmanas, Peter Güntert, Yunfen He, Torsten Herrmann, Yuanpeng J Huang, Victor Jaravine, Hendrik R A Jonker, Michael A Kennedy, Oliver F Lange, Gaohua Liu, Thérèse E Malliavin, Rajeswari Mani, Binchen Mao, Gaetano T Montelione, Michael Nilges, Paolo Rossi, Gijs van der Schot, Harald Schwalbe, Thomas A Szyperski, Michele Vendruscolo, Robert Vernon, Wim F Vranken, Sjoerd de Vries, Geerten W Vuister, Bin Wu, Yunhuang Yang, and Alexandre M J J Bonvin. Blind testing of routine, fully automated determination of protein structures from NMR data. *Structure*, 20(2):227–236, Feb 2012.

[100] G.S. Rule and T.K. Hitchens. *Fundamentals of protein NMR spectrosopy.* Focus on structural biology. Springer, 2006.

[101] Daisuke Sakakibara, Atsuko Sasaki, Teppei Ikeya, Junpei Hamatsu, Tomomi Hanashima, Masaki Mishima, Masatoshi Yoshimasu, Nobuhiro Hayashi, Tsutomu Mikawa, Markus Wälchli, Brian O Smith, Masahiro Shirakawa, Peter Güntert, and Yutaka Ito. Protein structure determination in living cells by in-cell NMR spectroscopy. *Nature*, 458(7234):102–105, Mar 2009.

[102] Frank H Schumann, Hubert Riepl, Till Maurer, Wolfram Gronwald, Klaus-Peter Neidig, and Hans Robert Kalbitzer. Combined chemical shift changes and amino acid specific chemical shift mapping of protein-protein interactions. *J Biomol NMR*, 39(4):275–289, Dec 2007.

[103] Yang Shen, Frank Delaglio, Gabriel Cornilescu, and Ad Bax. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol NMR*, 44(4):213–223, Aug 2009.

[104] Yang Shen, Oliver Lange, Frank Delaglio, Paolo Rossi, James M Aramini, Gaohua Liu, Alexander Eletsky, Yibing Wu, Kiran K Singarapu, Alexander Lemak, Alexandr Ignatchenko, Cheryl H Arrowsmith, Thomas Szyperski, Gaetano T Montelione, David Baker, and Ad Bax. Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci U S A*, 105(12):4685–4690, Mar 2008.

[105] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, 11(9):739–747, Sep 1998.

[106] S. B. Shuker, P. J. Hajduk, R. P. Meadows, and S. W. Fesik. Discovering high-affinity ligands for proteins: SAR by NMR. *Science*, 274(5292):1531–1534, Nov 1996.

[107] Aroop Sircar and Jeffrey J Gray. Snugdock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput Biol*, 6(1):e1000644, Jan 2010.

[108] Jaime L Stark and Robert Powers. Application of NMR and molecular docking in structure-based drug discovery. *Top Curr Chem*, Sep 2011.

[109] Tim Stevens. CcpNmr analysis tutorials. `http://www.ccpn.ac.uk/ccpn/software/ccpnmr-analysis/tutorials/three-day-course`, accessed January 2011.

[110] D. Stratmann, C. Heijenoort, and E. Guittet. NOEnet–use of NOE networks for NMR resonance assignment of proteins with known 3D structure. *Bioinformatics*, 25(4):474–481, 2009.

[111] Dirk Stratmann, Rolf Boelens, and Alexandre M J J Bonvin. Quantitative use of chemical shifts for the modeling of protein complexes. *Proteins*, 79(9):2662–2670, Sep 2011.

[112] Dirk Stratmann, Eric Guittet, and Carine van Heijenoort. Robust structure-based resonance assignment for functional protein studies by NMR. *J Biomol NMR*, 46(2):157–173, 2010.

[113] E.L. Ulrich, H. Akutsu, J.F. Doreleijers, Y. Harano, Y.E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C.F. Schulte, D.E. Tolmie, R.K. Wenger, H. Yao, and J.L. Markley. BioMagResBank. *Nucleic Acids Res.*, 36(Database issue):D402–D408, 2008.

[114] Utrecht NMR Research group. Analysis of NMR titration data and docking results in the study of biomolecular complexes. `http://www.nmr.chem.uu.nl/~abonvin/tutorials/Titration-Data/titration.html`, accessed January 2011.

[115] Julia Vaynberg, Tomohiko Fukuda, Ka Chen, Olga Vinogradova, Algirdas Velyvis, Yizeng Tu, Lily Ng, Chuanyue Wu, and Jun Qin. Structure of an ultraweak protein-protein complex and its crucial role in regulation of cell morphology and motility. *Mol Cell*, 17(4):513–523, Feb 2005.

[116] A.C. Wang and A. Bax. Determination of the backbone dihedral angles phi in human ubiquitin from reparametrized empirical Karplus equations. *J. Am. Chem. Soc.*, 118(10):2483–2494, 1996.

[117] Johan Weigelt, Mats Wikström, Johan Schultz, and Maria J P van Dongen. Site-selective labeling strategies for screening by NMR. *Comb Chem High Throughput Screen*, 5(8):623–630, Dec 2002.

[118] Mike P. Williamson. Applications of the NOE in molecular biology. volume 65, chapter 3, pages 77–109. Academic Press, 2009.

[119] D. S. Wishart and D. A. Case. Use of chemical shifts in macromolecular structure determination. *Methods Enzymol*, 338:3–34, 2001.

[120] J. Michael Word, Simon C. Lovell, Jane S. Richardson, and David C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of sidechain amide orientation. *Journal of Molecular Biology*, 285:1735–1747, 1999.

[121] Sitao Wu and Yang Zhang. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res*, 35(10):3375–3382, 2007.

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

[123] F. Xiong and C. Bailey-Kellogg. A hierarchical grow-and-match algorithm for backbone resonance assignments given 3D structure. In *BIBE 2007*, pages 403–410, 2007.

[124] F. Xiong, G. Pandurangan, and C. Bailey-Kellogg. Contact replacement for NMR resonance assignment. *Bioinformatics*, 24(13):205–213, 2008.

[125] Yingqi Xu and Stephen Matthews. TROSY NMR spectroscopy of large soluble proteins. *Top Curr Chem*, Sep 2011.

[126] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38, December 2006.

[127] Tallys Yunes, Ionut D. Aron, and J. N. Hooker. An integrated solver for optimization problems. *Oper. Res.*, 58(2):342–356, 2010.

[128] Huizhen Zhang, Cesar Beltran-Royo, and Miguel Constantino. Effective formulation reductions for the quadratic assignment problem. *Comput. Oper. Res.*, 37:2007–2016, November 2010.

[129] David Zuckerman. On unapproximable versions of NP-complete problems. *SIAM J. Comput.*, 25(6):1293–1304, 1996.

[130] Erik R P Zuiderweg. Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry*, 41(1):1–7, Jan 2002.