

An Improved Formalism for Assigning Proteins Using Nuclear Vector Replacement Framework

A thesis submitted to the
Graduate School of Natural and Applied Sciences

by

Şeyma ÇETINKAYA

in partial fulfillment for the
degree of Master of Science

in

Electronics and Computer Engineering



This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science in Electrical and Computer Engineering.

APPROVED BY:

Assist. Prof. M. Serkan Apaydın
(Thesis Advisor)



Prof. Bülent Çatay



Assoc. Prof. Vural Aksakallı

This is to confirm that this thesis complies with all the standards set by the Graduate School of Natural and Applied Sciences of İstanbul Şehir University:

DATE OF APPROVAL:

25 January 2016.

SEAL/SIGNATURE:

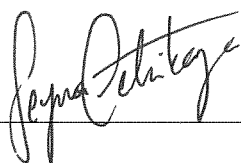


Declaration of Authorship

I, Şeyma ÇETINKAYA, declare that this thesis titled, 'Progress in Nuclear Vector Replacement for NMR Protein Structure-Based Assignments ' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:



Date:

04.01.2016

An Improved Formalism for Assigning Proteins Using Nuclear Vector Replacement Framework

Şeyma ÇETINKAYA

Abstract

Proteins are macromolecules in living systems used in crucial functions in all biological processes. In order to understand the function of a protein it is necessary to determine the structure of it. There are various techniques to obtain structural information and Nuclear Magnetic Resonance (NMR) Spectroscopy is one of the most important ones. In this technique, an essential step is the backbone resonance assignment and Structure Based Assignment (SBA) is a method solving this problem with the help of a template structure. NVR is an NMR protein SBA program, that takes as input ^{15}N and ^1H chemical shifts and unambiguous NOEs, as well as RDCs, HD-exchange and TOCSY data. To run NVR, there is a sequence of steps in obtaining the datafiles from NMR data and the template structure. In this study, the process of preparing these datafiles is simplified and automatized, which is an important practical step in running NVR on novel proteins. A method to distinguish NH_2 peaks from HSQC peaks is generated. Finally, rather than computing a single assignment, an ensemble of assignments is computed. Using this ensemble of assignment results, degree of reliability for individual peak-amino acid assignments is obtained and assignment accuracy is improved. The results show that these improvements bring NVR closer to a tool to be useful and practical tool, able to handle the input data automatically and analyze the reliability of assignments.

Keywords: Structural bioinformatics, NMR structure based protein assignment, NVR

NMR Protein Yapı Tabanlı Atamaları için NVR (Nükleer Vektör Yerdeğişimi) çerçevesini kullanan gelişmiş bir yaklaşım

Şeyma ÇETINKAYA

ÖZ

Proteinler, yaşayan sistemlerde bulunan ve temel biyolojik süreçlerde hayati fonksiyonları gerçekleştiren makromoleküllerdir. Proteinin işlevini anlamak için o proteinin yapısının belirlenmesi gereklidir. Yapısal bilgiyi elde etmek için çeşitli yöntemler vardır ve Nükleer Manyetik Rezonans (NMR) Spektroskopisi en önemli olanlardandır. Bu teknikte, gerekli olan bir adım omurga rezonans atamasıdır ve Yapı Tabanlı Atama kalıp protein yardımıyla bu sorunu çözmek için kullanılır. Nükleer Vektör Yerdeğişimi programı, ^{15}N ve H^N kimyasal kaymaları ve net NOE'lerin yanı sıra RDC'leri, HD değişimi ve TOCSY verilerini kullanan NMR protein Yapı Tabanlı Atama programıdır. NVR'ı çalıştırmak için, NMR verisinden ve kalıp yapıdan gelen data dosyalarını elde etmede bir dizi adım vardır. Bu çalışmada, NVR'ı yeni proteinlerde çalıştırmada önemli bir adım olan data dosyalarını hazırlama süreci basitleştirildi ve otomatikleştirildi. HSQC tepeciklerinden NH_2 tepeciklerini ayırt etmek için bir yöntem oluşturuldu. Son olarak, tek bir atama hesaplamak yerine bir takım atama hesaplandı. Bu takım atama sonuçları kullanılarak, tepecik-amino asit atamaları için bir güvenilirlik derecesi elde edildi ve atama doğruluğu yükseltildi. Sonuçlar gösteriyor ki, bu gelişmeler NVR'ı girdi dataalarını otomatik olarak halleden ve atama güvenilirliğini analiz edebilen pratik bir araç haline getirmektedir.

Anahtar Sözcükler: Yapısal biyoinformatik, NMR protein yapı tabanlı atama, Nükleer vektör Yerdeğişimi

Acknowledgments

I would like to express my greatest gratitude to Assist. Prof. Mehmet Serkan Apaydın for his guidance, support, and kindness. Without his counseling, I would have never completed this study.

I thank Assoc. Prof. Vural Aksakallı and Prof. Bülent Çatay for participating in my thesis committee and for their valuable comments. I want to thank Ewen Lescop for discussions and David Albachten for his helpful comments and feedback.

I would like to give special thanks to my old friends Betül Özateş, Esra Polat, Beyza Karakaya, Zeynep Dagnık, Esra Altnısık, Aysenur Çor and Zehra Bilgin who always believe in me.

I thank my friends Şeyma Nur Ekren, Sena Nur Günay, Nihal Vatandaş, Nafiye Polat, Seda Ediz, Fatma Zehra Kaçar and Şule Kütükde for their motivation and support.

Finally, I express my deepest gratitude to my dearest family. I am thankful to my father Hüseyin Çetinkaya and my mother Nermin Çetinkaya for their support and motivation and I would like to thank my sister Fatma Zehra and my brother Muhammed Yusuf for their friendship.

Contents

Declaration of Authorship	ii
Abstract	iii
Öz	iv
Acknowledgments	v
List of Figures	vii
List of Tables	viii
Abbreviations	ix
1 Introduction	1
2 Literature Review	6
2.1 Related Work	6
2.2 NVR Framework	7
2.2.1 Mathematical Formulation	8
2.2.2 Template Selection	9
3 Methodology	10
3.1 Automatization of the Data Preparation	10
3.2 Distinguishing NH_2 peaks	11
3.2.1 Experimental Analysis of Distinguishing NH_2 peaks	13
3.3 Providing a Measure of Reliability of Assignments	14
4 Results	17
4.1 Automatization of the Data Preparation Results	17
4.1.1 Test Results on Two Novel Proteins	17
4.2 Distinguishing NH_2 peaks Results	17
4.3 Reliability Results	19
5 Conclusion	21
A NH_2 Removal Scores of Randomly Selected Proteins	23
Bibliography	26

List of Figures

1.1	General Structure of Amino Acid	1
1.2	Protein Structure of Hemoglobin	2
1.3	Solving the Structure of a Molecule by X-ray Crystallography	3
1.4	HSQC Spectrum	4
3.1	Structure of Amino acids Asparagine and Glutamine	11
3.2	Computing CS Probabilities	12
3.3	ROC Curve for NH_2 Peak Classification.	14
3.4	ROC Curve for Various Thresholds to Determine Strong Assignments	15
3.5	Obtaining Final Assignment using Hungarian Algorithm	16
4.1	NH_2 Scores of 1UBI	18
4.2	NH_2 Scores of Prp	18
4.3	NH_2 Scores of S1	19
A.1	NH_2 Scores of 4183	23
A.2	NH_2 Scores of 6134	24
A.3	NH_2 Scores of 19047	24
A.4	NH_2 Scores of 19217	24
A.5	NH_2 Scores of GB1	25

List of Tables

3.1	Number of Strong Assignments and Accuracy for Different Thresholds for MBP	15
4.1	NH2 Prediction Results	19
4.2	NH2 Prediction Results	19



Abbreviations

ACO	Ant Colony Optimization
ASN	ASparagiNe
BIP	Binary Integer Programming
EM	Expectation Maximization
GLN	GLutamiNe
HD	Hydrogen Deuterium
HSQC	Heteronuclear Single Quantum Correlation
MBP	Maltose Binding Protein
NA	NOE Aware
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
NVR	Nuclear Vector Replacement
PDB	Protein Data Bank
RDC	Residual Dipolar Coupling
SBA	Structure Based Assignment
TOCSY	TOtal Correlated SpectroscopY
TS	Tabu Search

Chapter 1

Introduction

Proteins are macromolecules in living systems used in crucial functions. They are enzymes that catalyse chemical reactions, they are used in the processes of energy storage, defence of cells from antibodies, transportation of molecules inside or the outside of the cells, transmission of signals between cells, etc.. Proteins are made of 20 amino acids and amino acids composed of amino-group, carboxyl-group, and R-group that are attached to the central carbon (Figure 1.1). The 20 amino acids differ in the R-group and the properties of amino acids are different because their R-groups are different. Amino acids come together to construct a sequence with peptide bonds and form proteins (also called polypeptides).

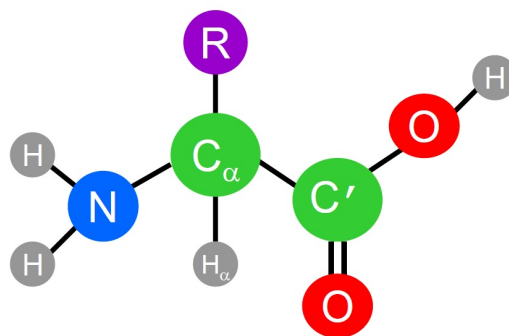


FIGURE 1.1: General Structure of Amino Acid

All proteins have four basic levels of structure: primary, secondary, tertiary, and quaternary. The primary structure of a protein is the sequence of amino acids (Figure 1.2(a)). With hydrogen bonds the polypeptide bends and forms α -helices and β -sheets that give secondary structure to the protein (Figure 1.2(b)). Tertiary structure is the

overall three-dimensional structure of the protein, that is the resulting shape after a protein folds (Figure 1.2(c)) and finally the combination of tertiary structures of multiple proteins results in quaternary structure (Figure 1.2(d)).

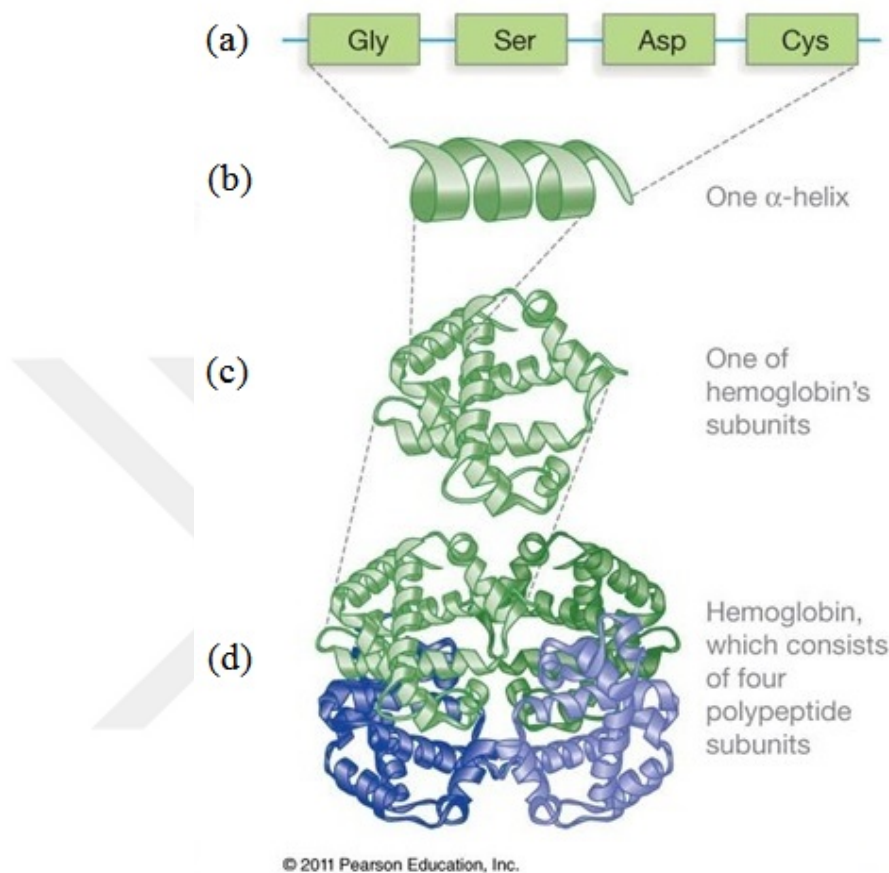


FIGURE 1.2: Protein Structure of Hemoglobin

To understand the functions of proteins, it is necessary to determine their 3D-structure. Additionally, the knowledge of a protein's structure is important to understand why some proteins misfold or partially fold causing some diseases such as Parkinson's disease and Huntington's disease, find structural similarities between proteins, design new drugs, predict how proteins bind with other proteins, and so on.

In order to determine a protein's structure there are two main techniques: X-Ray Crystallography and NMR (nuclear magnetic resonance) spectroscopy. In X-Ray Crystallography the protein is crystallized and exposed to an intense beam of X-rays. The crystallized protein diffracts the X-ray beam when the light interacts with electrons of

the atoms. The diffraction pattern is determined using location and intensity of diffractions. Then the map of the distribution of electrons in the molecule is obtained and used to determine the location of each atom. Using this data the preliminary model is fit. Then this model is refined several times until there is no longer improvement to fit the map more accurately¹ (Figure 1.3). Most of the protein structures in the Protein Data

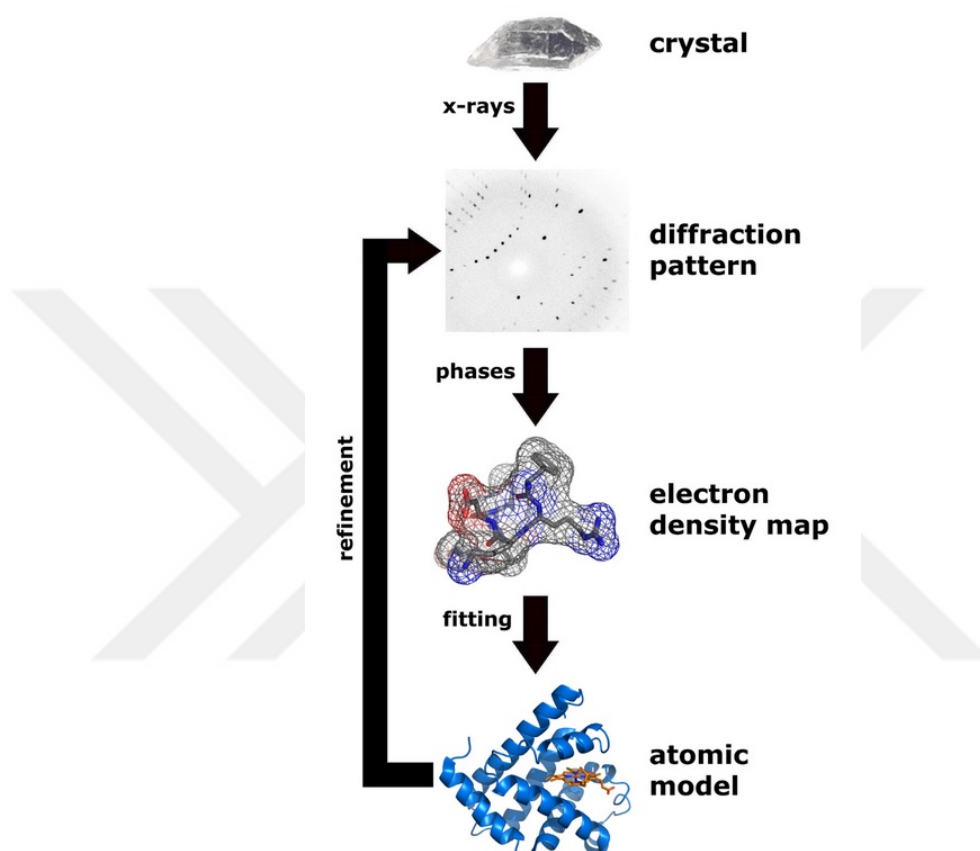


FIGURE 1.3: Solving the Structure of a Molecule by X-ray Crystallography

Bank are solved by X-Ray Crystallography. However, the crystalline form of the protein may be different than the form in solution and some proteins cannot be crystallized. For such proteins, NMR provides a good alternative. NMR studies the protein in an environment similar to the native environment. The protein is often examined in solution, therefore with NMR, information about the physical properties of the protein such as the geometry of atoms, bonds between them, and dynamics of the protein can be obtained. Moreover, for proteins that do not form crystals, NMR is the only alternative for atomic resolution structures. However, NMR is limited by protein size, larger proteins usually

¹Lawson, D., "A Brief Introduction to Protein Crystallography", <https://www.jic.ac.uk/staff/david-lawson/xtallog/summary.htm>, 09.09.2015

result in more missing and overlapping signals. NMR is different from X-ray crystallography in which instead of finding the atomic resolution structure directly, in NMR, various experiments are done and the structure is attempted to be found by combining the information coming from these experiments.

Atoms with an odd number of nucleons have a non-zero quantum mechanical property, called spin. NMR spectroscopy exploits the magnetic properties of the atoms whose spin is equal to $1/2$ (e.g., ^{13}C , ^{15}N and ^1H) [1]. Therefore, to study a protein using NMR, the protein is labeled with isotopes such as ^{13}C and ^{15}N . A magnetic field is applied to an NMR active nuclei and as a result, nuclei precess. Since the protein folded, every nucleus has a unique electronic environment. Thus, each nucleus has a unique precession frequency, so each nucleus can be identified by its frequency. Precession frequency gives a property called chemical shift (CS) and a tuple of chemical shifts form a peak which corresponds to an amino acid. In Figure 1.4 a sample of 2D HSQC spectrum is given. In HSQC spectrum each axis represents a type of atom. In this sample, x-axis gives ^1H CS and y-axis gives ^{15}N CS that corresponds to backbone atoms of an amino acid and all points on the HSQC spectrum correspond to peaks.

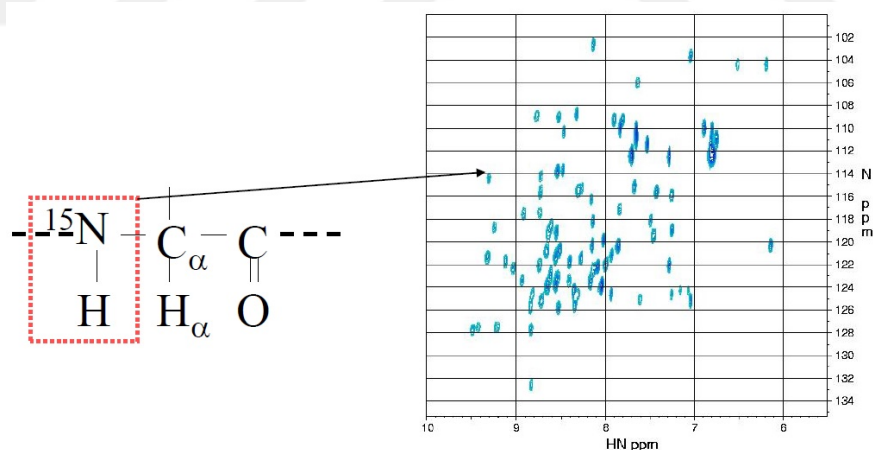


FIGURE 1.4: HSQC Spectrum

One of the important steps in determining the protein structure starting from NMR data is to assign these peaks to the corresponding amino acids. The assignment of proteins in NMR laboratories is a laborious process. Automatizing the assignment process with a high degree of accuracy is important in order to expedite the NMR protein structure determination. Structure Based Assignment (SBA) achieves this objective with the help

of a template structure that is homologous to the target protein. The knowledge of template provides prior information about the structure of the target protein and allows to obtain more reliable assignment results [2].

Nuclear Vector Replacement (NVR) is an approach that solves the protein SBA problem. NVR-EM [3] uses expectation maximization algorithm and finds a local optimum solution. NVR-BIP [4] uses binary integer programming to obtain the global solution for the problem. However, it is unable to obtain the assignments for larger proteins due to the considerable resources such an exact solution requires. For such proteins, metaheuristic approaches such as NVR-TS [5] and NVR-ACO [2] have been developed. NVR-TS uses tabu search and NVR-ACO uses ant colony optimization to arrive at a solution. Finally, the latest version of NVR is NA-NVR-ACO [6] that is NOE-aware version of NVR-ACO that distinguishes different types of NOEs.

In this thesis our contributions are:

- Automatization of input data preparation
- Removing NH_2 peaks from HSQC spectrum
- Providing a measure of reliability of assignments
- Improving assignment accuracy using an ensemble of assignment results

In the following chapter we give literature review and describe NVR and the problem formulation. In Chapter 3, we describe our contributions. The test results are given in Chapter 4. Finally, Chapter 5 presents the conclusions of the thesis and discusses the future work.

Chapter 2

Literature Review

2.1 Related Work

There are various software programs to help with protein structure based assignments semi-automatically (using e.g. Analysis [7]) or automatically (using e.g. Flya [8] or Mars [9]), and fully automated assignment of small proteins is possible [8]. Although there exist software to automate the assignment process, manual analysis of NMR spectra is the most reliable method. Manual verification of assignments are almost always done to handle possible errors, since automation is not trustworthy [1]. Moreover, for large proteins the assignment step can take weeks and even months [1], since the available data is incomplete and ambiguous [10]. Other challenges in obtaining the assignments is the spectra can be crowded, noisy, and there may be extra and missing peaks. In addition to these, there is another challenge of computational complexity, backbone resonance assignment problem is NP-Hard [5].

NVR is a framework for the NMR structure based assignment problem that tries to find the optimal matching between the set of amino acids and the set of peaks using only backbone amide proton and nitrogen chemical shifts, and backbone NOEs. It can also use RDCs, TOCSY, and Hydrogen-Deuterium exchange data, if available.

2.2 NVR Framework

There are different experiments of NMR. In NMR we observe CSs (Chemical Shifts) those are atomic property of specific atoms, NOEs (Nuclear Overhauser Effects) that are pairwise distance constraints between H atoms with up to about 5\AA (Angstrom), RDCs (Residual Dipolar Couplings) that give bond orientations dynamics, TOCSY (Total Correlation Spectroscopy) that is used to analyse scalar (J) coupling networks between protons, and Hydrogen-Deuterium exchange data that is used in the case of large proteins to understand the location of H atoms in the surroundings of the protein.

In NVR, assignment probabilities of the set of peaks to the set of amino acids are calculated using the difference between the observed and predicted CS values and if available RDC, TOCSY, and Hydrogen-Deuterium exchange data. It combines the informations obtained and give a score to all of the possible assignments. Then assignment of peaks to residues is solved using these scores and NOE constraints.

NVR-BIP is a binary integer programming based approach that computes the assignments using CPLEX. NVR-BIP minimizes the score of the assignment subject to NOE constraints and finds the optimal solution for small proteins (less than approximately 150 amino acids) and has high assignment accuracies. However, NVR-BIP is unable to compute a solution for large proteins due to the large number of constraints. For such proteins, metaheuristic based approaches within the NVR framework, such as NVR-TS and NVR-ACO have been developed. NVR-TS is a tabu search-based approach to the problem. Instead of applying hard constraints and disallowing NOE violations, NVR-TS uses a penalty term for NOE violations and can find a solution for large proteins. NVR-ACO is the first application of ant colony optimization to the problem and is based on the observation of the behavior of real ant colonies searching for food sources [2]. NVR-ACO finds the optimal solution for small proteins and can find solutions for large proteins with high accuracies. NVR-ACO uses backbone NOEs, however does not differentiate between HN-HA and HN-HN NOEs, and sets NOE distance thresholds (UB value in the formulation below) manually. NOE-aware version NA-NVR-ACO differentiates the type of backbone NOE and uses the appropriate coordinates from the template structure, and also obtains the NOE upperbound information directly from the NOE intensities in the data.

2.2.1 Mathematical Formulation

The latest version of NVR is NA-NVR-ACO and its mathematical model is as follows [6]:

Notation:

P : set of peaks

A : set of amino acids

T : set of distance types, $T = \{HN - HN, HN - HA, HA - HN\}$

s_{ij} : score associated with assigning peak i to amino acid j

N : number of peaks to be assigned ($N \leq |P|$)

d_{jlt} : distance between amide protons of amino acids j and l by using distance type t

$NOE(i)$: set of peaks that have an NOE with peak i

UB_{ik} : NOE upper bound distance limit between peaks i and k

$$b_{ijklt} = \begin{cases} 1, & \text{if } d_{jlt} \leq UB_{ik} \\ 0, & \text{otherwise} \end{cases}$$

Decision variables:

$$x_{ij} = \begin{cases} 1, & \text{if peak } i \text{ is assigned to amino acid } j \\ 0, & \text{otherwise} \end{cases}$$

Mathematical model:

$$\text{Minimize } \sum_{i \in P} \sum_{j \in A} s_{ij} x_{ij} \quad (2.1)$$

$$\text{s.t. } \sum_{i \in P} x_{ij} \leq 1, \forall j \in A \quad (2.2)$$

$$\sum_{i \in A} x_{ij} \leq 1, \forall j \in P \quad (2.3)$$

$$\sum_{i \in P} \sum_{j \in A} x_{ij} = N \quad (2.4)$$

$$x_{ij} + x_{kl} - 1 \leq b_{ijklt} \forall j, l \in A, \forall i \in P, \forall t \in T, \forall k \in NOE(i) \quad (2.5)$$

$$x_{ij} \in \{0, 1\}, \forall i \in P, \forall j \in A \quad (2.6)$$

In this model, the objective function (1) minimizes the total score of assigning peaks to amino acids. Constraints (2) ensure each amino acid is assigned to at most one peak and constraints (3) guarantee each peak is mapped to at most one amino acid. Constraint (4) determines the number of peaks that are going to be assigned. This allows us to obtain a partial assignment. Constraint set (5) requires peaks i and k which have an NOE between them of type t to be assigned to amino acids j and l if the distance between these amino acids (d_{jlt}) is less than UB_{ik} . Constraint set (6) forces the decision variables to be binary.

2.2.2 Template Selection

NA-NVR-ACO requires a template structure in order to compute the scoring matrix and obtain the distance constraints to be used with NOE data. This template structure can be an X-ray structure corresponding to the same protein, or could be a structural homolog. In this study, X-ray structure is used as the template. Previous work [11] involved using more distant templates and improving the assignment accuracy of NVR-EM.

Chapter 3

Methodology

3.1 Automatization of the Data Preparation

In order to run NA-NVR-ACO, a sequence of steps should be followed to prepare input files from the NMR data and the template structure. This procedure includes computing distances between protons in the PDB structure, obtaining the scoring matrix by using chemical shift prediction programs such as SHIFTS and SHIFTX, and combining the NMR data coming from different sources corresponding to the same peak. We simplified this process by automating these steps and enabled running NA-NVR-ACO on novel proteins faster. The pseudo-code is as follows:

```
/* Parsing steps */  
parsedResonancesFile ← parseResonanceFile(experimentalShiftFileName)  
parsedPDBfile ← parsePDB_File(PDBbaseName)  
secondaryStructureFile ← parseSSE_Info(parsedPDBfile)  
NHvectorsFile ← parseVectors_N - H(parsedPDBfile)  
SHIFTXFile ← shiftx(PDBbaseName)  
parsedSHIFTXFile ← parseSHIFTX_File(SHIFTXFile)  
SHIFTSFile ← shifts(PDBbaseName)  
parsedSHIFTSFile ← parseSHIFTS_File(SHIFTSFile)  
  
/* Assembly step */
```

InputFileOfNVR ← *assembleInput*(*parsedResonancesFile*, *NHvectorsFile*,
secondaryStructureFile, *parsedSHIFTXFile*, *parsedSHIFTSFile*)

parseResonanceFile parses the resonance file and extracts H^N , N chemical shifts.
parsePDB_File parses the template PDB file to extract H^N , N , H_α and C_α coords.
parseSSE_Info parses the secondary structure information of the template protein.
parseVectors_N - H calculates N-H bond vectors from PDB file.

SHIFTX [12] and SHIFTS [13] are chemical shift prediction tools, *parseSHIFTX_File*
and *parseSHIFTS_File* read the output of these and extract N , H^N and C_α chemical
shifts.

Finally, *assembleInput* combines all of the files that are extracted.

3.2 Distinguishing NH_2 peaks

Amino acids Asparagine (ASN) and Glutamine (GLN) differ from others in which they
have an extra Nitrogen atom binding two Hydrogen atoms in their side chains (Figure
3.1).

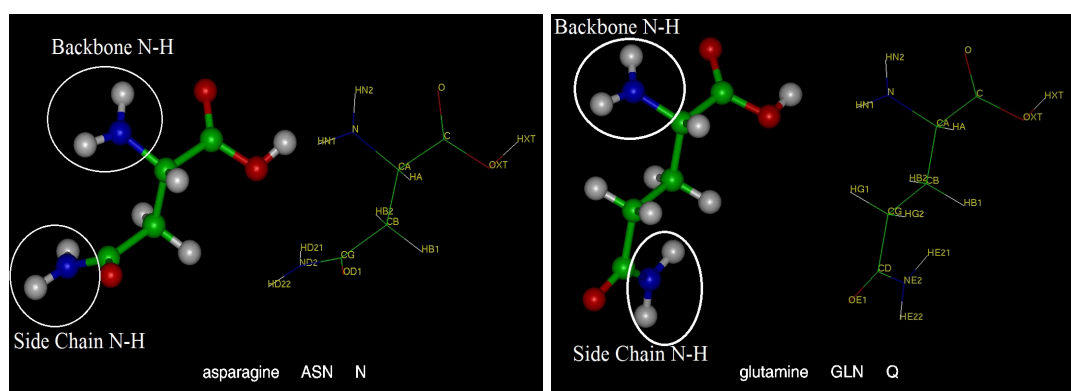


FIGURE 3.1: Structure of Amino acids Asparagine and Glutamine

Therefore, for each *ASN* or *GLN* in a protein, there are two extra peaks in the HSQC
spectrum with no corresponding amino acid to these peaks. Before performing the as-
signments, there is a need to remove these peaks from HSQC spectrum and this process
is usually done by spectroscopists manually.

In the case of existence of *ASN*, NH_2 peaks are pairs of atoms ND2-HD21 and ND2-HD22, and for *GLN* these extra peaks are NE2-HE21 and NE2-HE22. Since for an NH_2 peak pair, N atom is the same atom (ND2 for *ASN* or NE2 for *GLN*), in HSQC spectrum their N CS's are almost the same. Therefore, in our approach we first find all of the peak pairs whose N CS's are 0.01 ppm close to each other. Then a score is given to all of the peak pairs by using chemical shift statistics. A couple of peaks is labeled as NH_2 peak pair if their score is under a threshold.

To avoid any confusion, remember that in this study, a peak means a couple of atoms (N and H atoms) and a peak pair means a couple of peaks.

The score of a peak pair i corresponding to an NH_2 is calculated as follows:

$$S_i = -\log(\max(p_{i,ASN}, p_{i,GLN})) \quad (3.1)$$

Here, $p_{i,ASN}$ is the probability of peak pair i to be NH_2 peak pair of *ASN* according to the CS values it has and it is computed by converting the difference between the experimental CS values of peak pair i (call as N1, H1 and N2, H2) and the expected CS values of ND2, HD21 and HD22 atoms (obtained from BMRB statistics of amino acid *ASN*) to a probability using a Gaussian distribution (Figure 3.2). Since the H and N CS

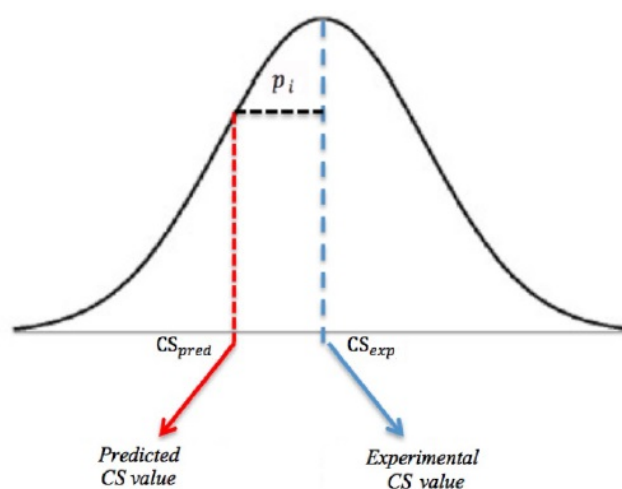


FIGURE 3.2: Computing CS Probabilities

values are independent from each other, we multiplied the probabilities of the examined atoms and got a resulting probability. However, since there are different matches of

Hydrogens (the first one is H1-HD21, H2-HD22, and the second one is H1-HD22, H2-HD21), two different resulting probabilities are obtained as follows:

$$p_{i,1} = p_{i,N1,ND2} * p_{i,N2,ND2} * p_{i,H1,HD21} * p_{i,H2,HD22} \quad (3.2)$$

$$p_{i,2} = p_{i,N1,ND2} * p_{i,N2,ND2} * p_{i,H1,HD22} * p_{i,H2,HD21} \quad (3.3)$$

After obtaining these probabilities, the maximum of them is chosen and it gives $p_{i,ASN}$.

$$p_{i,ASN} = \max(p_{i,1}, p_{i,2}) \quad (3.4)$$

Next, $p_{i,GLN}$ is calculated in the same way using expected CS values of NE2, HE21 and HE22 atoms obtained from BMRB statistics of *GLN*. Then, maximum of $p_{i,ASN}$ and $p_{i,GLN}$ is chosen and result is converted into a score for each peak pair by taking its negative logarithm. According to this score they are assigned as NH_2 peak pair or not.

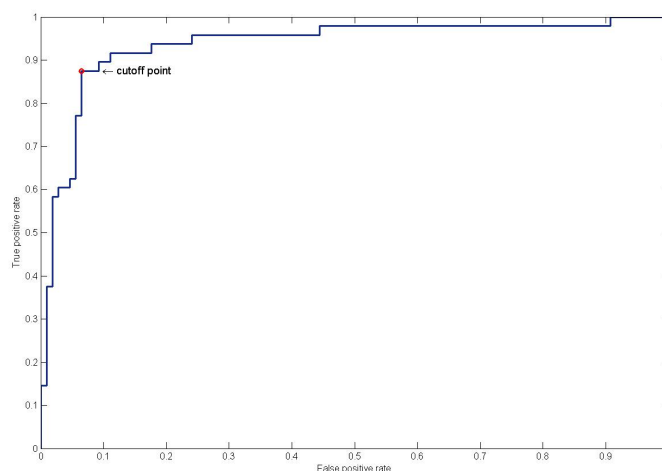
3.2.1 Experimental Analysis of Distinguishing NH_2 peaks

HSQC information of protein 1UBI and two novel proteins Prp and S1 is obtained from CNRS. In addition to these, five proteins are randomly selected from BMRB to test the approach.

To determine how well a binary classifier system performs and the threshold a system, a graphical plot named Receiver Operating Characteristic (ROC) curve is used. ROC curve is obtained as a plot of the true positive rate against the false positive rate for various thresholds.

First, we extracted all possible NH_2 peaks and calculated their score, and then, to determine the threshold of the classification and the quality of our approach, we draw ROC curve (Figure 3.3) using 80% of the results we obtained. (20% of the data is used to test the approach with the obtained threshold.)

The threshold is determined using ROC curve (Figure 3.3) and a peak pair is assigned as NH_2 if its score is under 7.71.

FIGURE 3.3: ROC Curve for NH_2 Peak Classification.

After assigning all of the peak pairs as NH_2 or not, we selected peaks to be deleted. A peak is removed from HSQC if it is always distinguished as NH_2 in the classification system.

3.3 Providing a Measure of Reliability of Assignments

NA-NVR-ACO can find the optimal solution for small proteins. However, for large proteins, the assignment results are distinct in different runs due to a lack of convergence to a global minimum in a very large search space. In that case, the individual result of a single assignment run is unreliable. In this thesis, it is hypothesized that in the lack of convergence, the assignments that are more likely to be correct will occur many times in multiple runs whereas the incorrect assignments will differ. Therefore, for such large proteins, rather than computing a single assignment, we computed an ensemble of assignments and we calculated how many times a peak is assigned to the same amino acid.

The assignment of a peak is determined as strong, if it is assigned to the same amino acid more than a percent of the time in all the runs. In order to determine this threshold as a percentage, we used the assignment results of MBP (Maltose Binding Protein), we computed the sensitivity and the specificity for different percentages and plot these points (Figure 3.4) as it is done in [11].

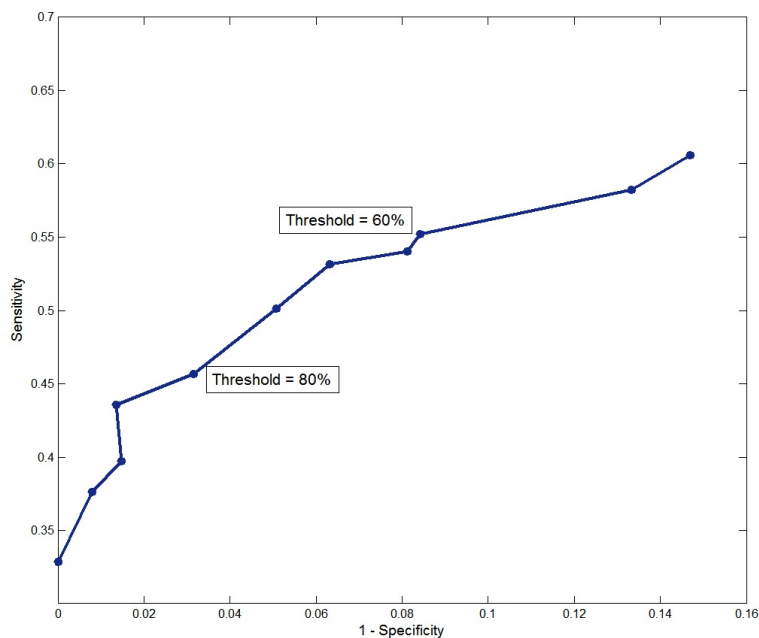


FIGURE 3.4: ROC Curve for Various Thresholds to Determine Strong Assignments

As the threshold increases, we expect the number of strong assignments to decrease and the accuracy of strong assignments to increase. For MBP, the effect of the threshold to the number of strong assignments and accuracy is calculated and it is given in Table 3.1.

TABLE 3.1: Number of Strong Assignments and Accuracy for Different Thresholds for MBP

Threshold (%)	No of Strong Assignments	Percent of Strong Assignments	Accuracy
50	238	71.0	85.3%
55	225	67.2	86.7%
60	202	60.3	91.6%
65	197	58.8	91.9%
70	190	56.7	93.7%
75	177	52.8	94.9%
80	158	47.2	96.8%
85	148	44.2	98.6%
90	135	40.3	98.5%
95	127	37.9	99.2%
100	110	32.8	100%

From these thresholds, we chose 60%, and determined an assignment of a peak as strong, if it is assigned to the same amino acid more than 60% of the time in all the runs. With this method, we also derived information about the reliability of our assignments, as the

ratio of the number of times a peak has been assigned to a given residue over the total number of runs. This is similar to [11], but instead of using multiple templates to obtain different assignments, the assignment results of multiple runs are used.

Moreover, by using an ensemble of assignment results, we combined all assignments by obtaining a bipartite graph where a set of nodes corresponds to the peaks and the other set corresponds to the residues. The edges between peaks and residues are associated with a score corresponding to the number of times a peak is assigned to the corresponding amino acid in the assignment ensemble (Figure 3.5) [11].

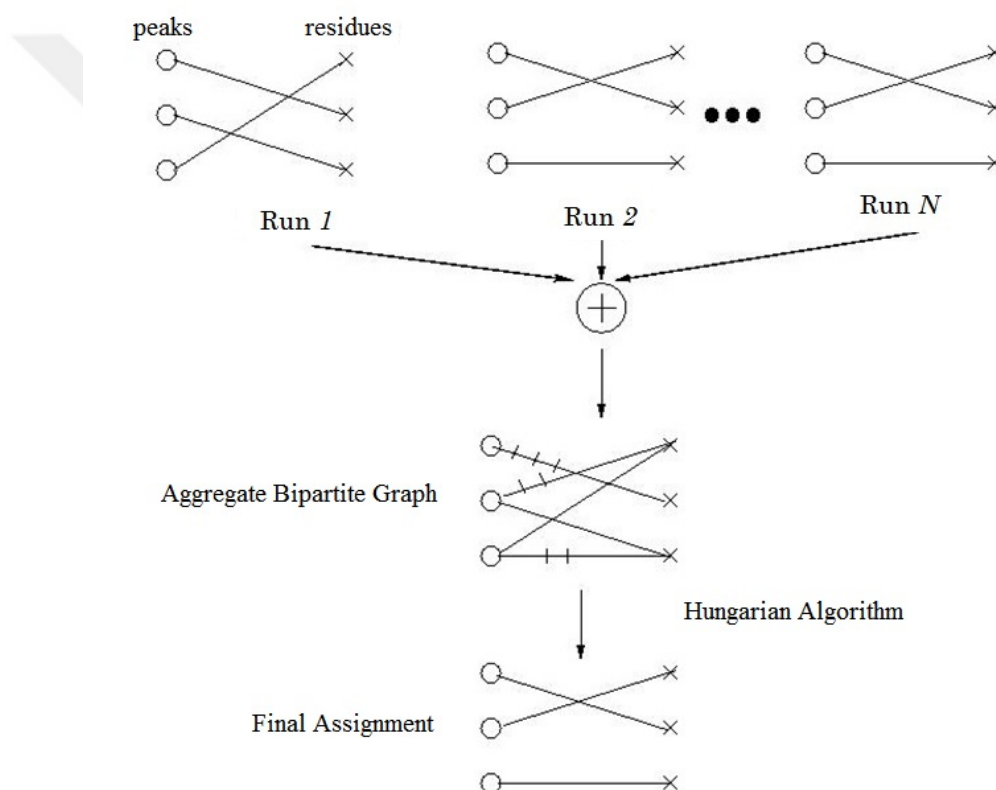


FIGURE 3.5: Obtaining Final Assignment using Hungarian Algorithm

After obtaining aggregate bipartite graph, using the scores of this matrix, the final assignment is calculated using Hungarian algorithm. With this method, we obtained our final assignment aggregate the results from all of the assignments we had.

Chapter 4

Results

4.1 Automatization of the Data Preparation Results

We studied a new protein molecular-weight-protein tyrosine phosphatase A (MptpA, 150 amino acids) that is not in the set of test proteins of NA-NVR-ACO. The process of extracting the input data of NVR required almost a week to complete. Then, we automatized this process using a combination of bash, perl and matlab scripts. With the automatization, we could obtain our datafiles in a few minutes.

4.1.1 Test Results on Two Novel Proteins

We simulated unambiguous NOEs of MptpA and computed its assignments using NA-NVR-ACO. We obtained an assignment accuracy of 100.0%.

We also computed test results of beta lactamase NDM1 (134 amino acids), whose data was obtained by CNRS. We simulated unambiguous NOEs of it and computed its assignments using NA-NVR-ACO. We obtained an assignment accuracy of 80.6%.

4.2 Distinguishing NH_2 peaks Results

At the bottom, there are plots of the test results of the proteins 1UBI, Prp and S1 (Figure 4.1, Figure 4.2, Figure 4.3). In x-axis, peaks those are possibly NH_2 peaks are

numbered. In these plots, NH_2 peak pairs are shown as green squares and others are red triangle.

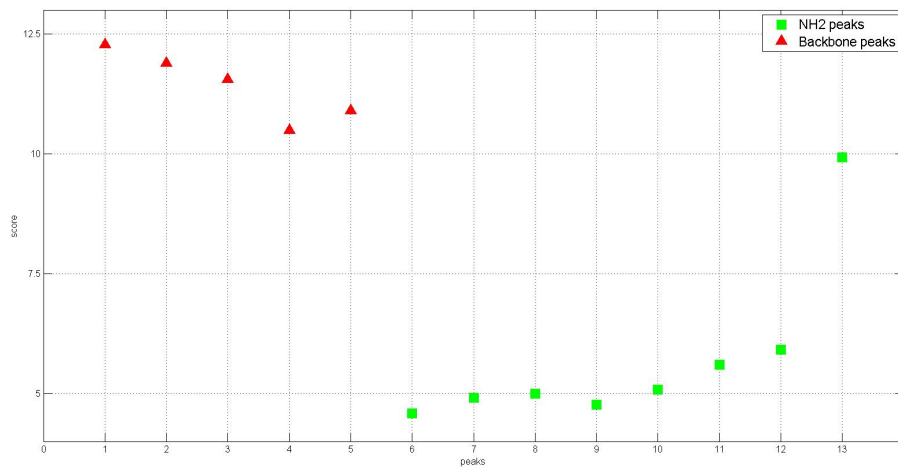


FIGURE 4.1: NH_2 Scores of 1UBI

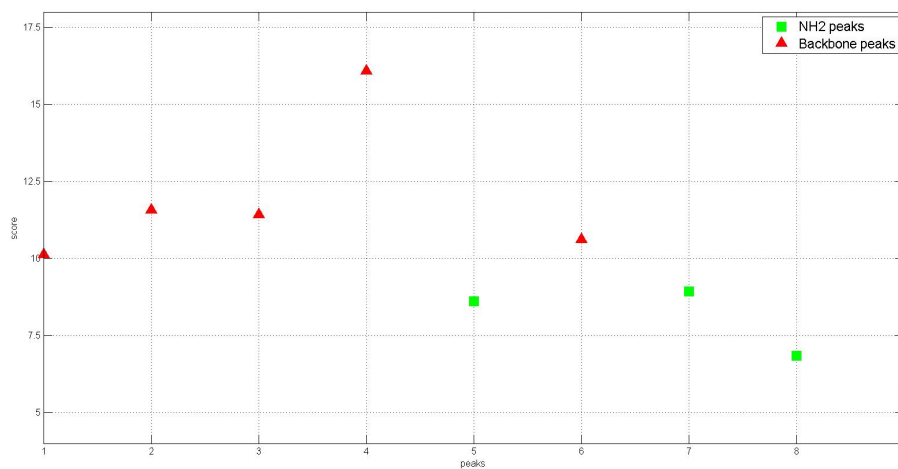
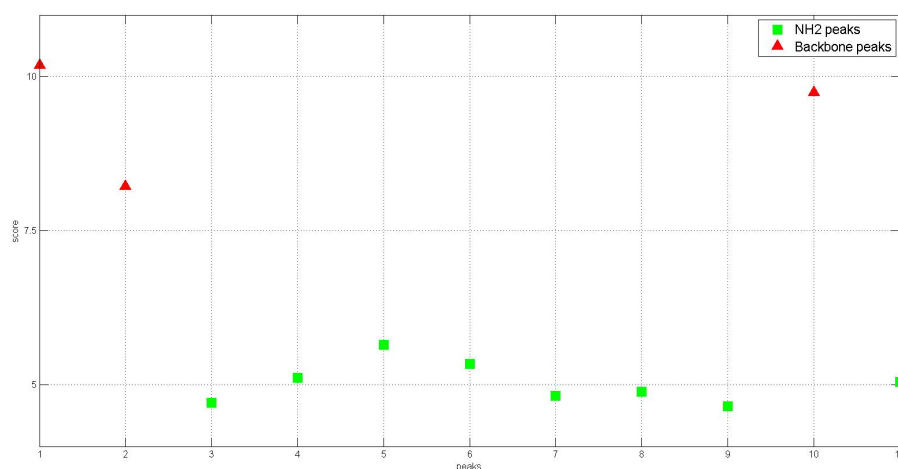


FIGURE 4.2: NH_2 Scores of Prp

In Table 4.1 and 4.2, the training and test set results that are obtained using our threshold are given.

FIGURE 4.3: NH_2 Scores of S1TABLE 4.1: NH_2 Prediction Results

	TP	TN	FP	FN	Accuracy
Training set	43	102	7	4	92.9%
Test set	16	19	1	3	89.7%

TABLE 4.2: NH_2 Prediction Results

	Precision	Recall	F Score
Training set	0.86	0.91	0.88
Test set	0.94	0.84	0.89

4.3 Reliability Results

We took 25 different assignment results of MBP using NA-NVR-ACO. Among these ensemble of assignment results, the assignment with minimum score has a 58.8% accuracy. The individual assignment accuracies range between 53.4% and 71.3% and the average assignment accuracy is 64.1%.

MBP has 335 peaks that are all assigned. By using our reliability measure, we found that 202 peaks (60% of the peaks) were assigned to the same amino acid in 25 runs in at least 60% of the runs, and these peaks had 91.6% accuracy. This information could

be used to partially assign the peaks with high accuracy. Additional experiments could be done for the remaining peaks to assign them correctly. Furthermore, by using the Hungarian algorithm we combined the assignment results of 25 runs and obtained an assignment accuracy of 72.8% for all the peaks.



Chapter 5

Conclusion

In this study, the following steps are performed to improve and automate NA-NVR-ACO.

- To facilitate the study on new proteins, input data preparation process is simplified. The time that is spent to obtain input data is reduced to a couple of minutes. Moreover, for two novel proteins MptpA and NDM1, test results of NVR are obtained.
- A method is generated to distinguish NH_2 peaks from HSQC peaks.
- The reliability of the assignments is determined using an ensemble of assignment results. A reliability degree of assignments is provided for the protein MBP.
- An ensemble based method is developed to enhance the assignment accuracy. This method is tested on MBP and the assignment accuracy is improved.

With these improvements, NVR becomes closer to being a practical tool useful in an NMR laboratory. The time it takes to obtain the assignments for a novel protein using NVR is significantly reduced. NVR can work with more noise in the data. It must be mentioned that the reliability information for peaks is available for large proteins for which the global optimal solution is not found. For such proteins, the assignment results differ from run to run.

One step that remains to increase the usability of NVR is to enable it to handle ambiguous NOEs. Obtaining enough unambiguous NOEs from raw data is a challenge and may require performing 4D NOESY experiments which are not always available. While

handling ambiguous NOEs, we will distinguish aromatic and aliphatic protons which have similar chemical shifts using the template structure information. Finally, we plan to assign larger proteins based on methyl group NOEs [14].



Appendix A

NH_2 Removal Scores of Randomly Selected Proteins

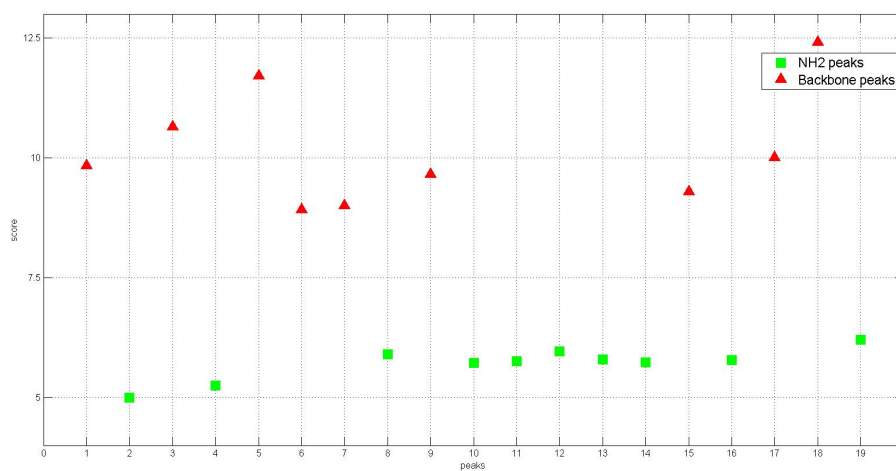
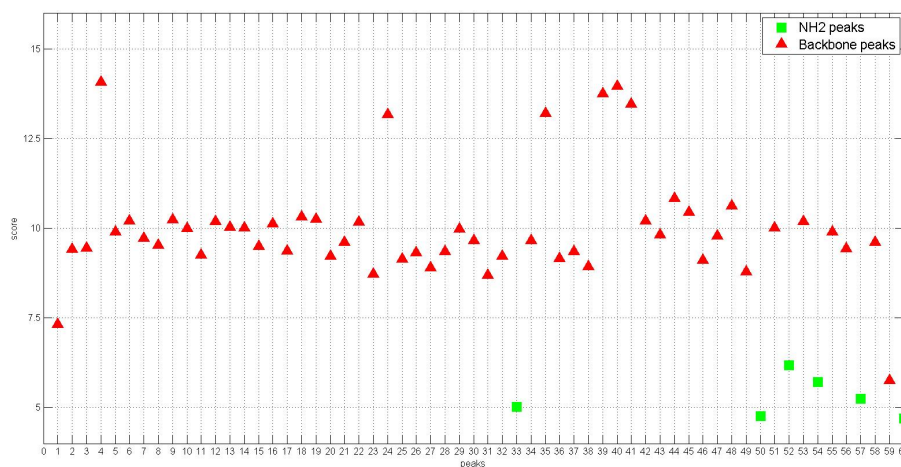
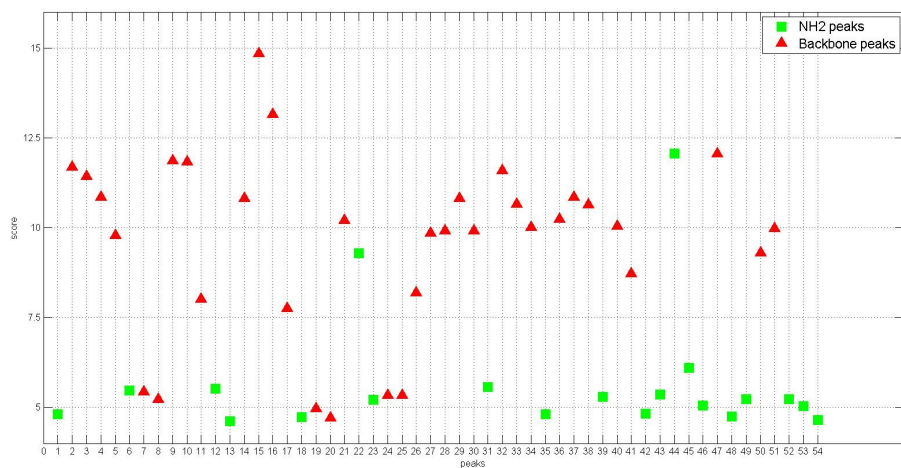
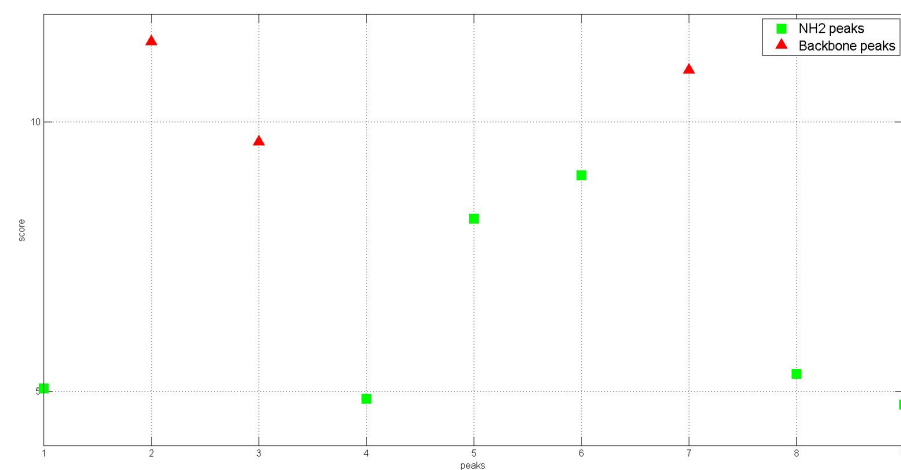
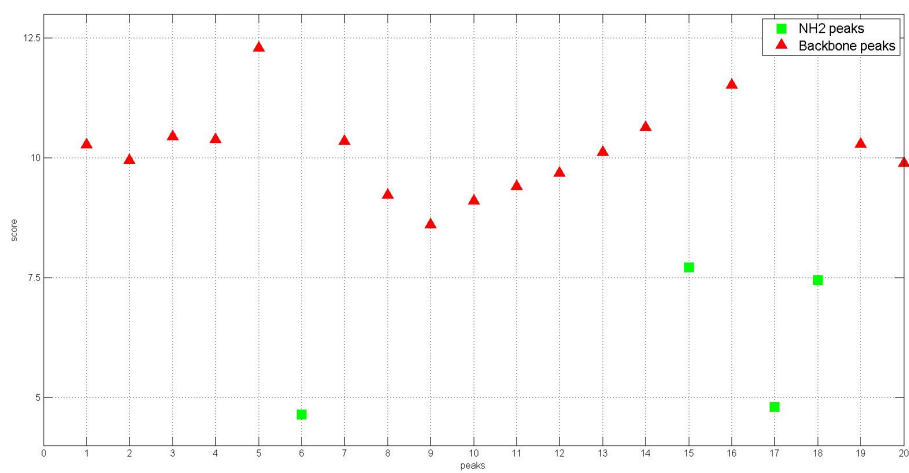


FIGURE A.1: NH_2 Scores of 4183

FIGURE A.2: NH_2 Scores of 6134FIGURE A.3: NH_2 Scores of 19047FIGURE A.4: NH_2 Scores of 19217

FIGURE A.5: NH_2 Scores of GB1

Bibliography

- [1] R. Jang. *Fast and Robust Mathematical Modeling of NMR Assignment Problems*. PhD thesis, University of Waterloo, Canada, 2012.
- [2] J. Aslanov, B. Çatay, and M.S. Apaydın. An Ant Colony Optimization Approach for Solving the Nuclear Magnetic Resonance Structure Based Assignment Problem. *GECCO*, 2013.
- [3] C.J. Langmead and B.R. Donald. An Expectation/Maximization Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments. *J. Biomolecular NMR*, 29(2):111–138, 2004.
- [4] M.S. Apaydın, B. Çatay, N. Patrick, and B.R. Donald. NVR-BIP: Nuclear Vector Replacement Using Binary Integer Programming for NMR Structure-Based Assignments. *The Computer J.*, 54(5):708–716, 2011.
- [5] G. Çavuslar, B. Çatay, and M.S. Apaydın. A Tabu Search Approach for the NMR Protein Structure-Based Assignment Problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(6):1621–1628, 2012.
- [6] M. Akhmedov, B. Çatay, and M.S. Apaydın. Automating Unambiguous NOE Data Usage in NVR for NMR Protein Structure-Based Assignments. *Journal of Bioinformatics and Computational Biology*, 2015.
- [7] W.F. Vranken, W. Boucher, T.J. Stevens, R.H. Fogh, A. Pajon, M. Llinas, E.L. Ulrich, J.L. Markley, J. Ionides, and E.D. Laue. The CCPN Data Model for NMR Spectroscopy: Development of a Software Pipeline. *Proteins: Structure, Function, and Bioinformatics*, 59(4):687–696, 2005.
- [8] E. Schmidt and P. Guntert. A New Algorithm for Reliable and General NMR Resonance Assignment. *J. Am. Chem. Soc.*, 134:12817–12829, 2012.

- [9] Y.S. Jung and M. Zweckstetter. Backbone Assignment of Proteins with Known Structure Using Residual Dipolar Couplings. *J. Biomolecular NMR*, 30(1):25–35, 2004.
- [10] C.A. MacRaid and R.S. Norton. RASP: Rapid and Robust Backbone Chemical Shift Assignments from Protein Structure. *J. Biomolecular NMR*, 58(3):155–163, 2014.
- [11] M.S. Apaydin, V. Conitzer, and B.R. Donald. Structure-Based Protein NMR Assignments Using Native Structural Ensembles. *J. Biomolecular NMR*, 40(4):263–276, 2008.
- [12] S. Neal, A.M. Nip, H. Zhang, and D.S. Wishart. Rapid and Accurate Calculation of Protein ^1H , ^{13}C and ^{15}N Chemical Shifts. *J. Biomolecular NMR*, 26(3):215–240, 2003.
- [13] X.P. Xu and D.A. Case. Automated Prediction of ^{15}N , $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$ and ^{13}C Chemical Shifts in Protein Using a Density Functional Database. *J. Biomolecular NMR*, 21(4):321–333, 2001.
- [14] F.A. Chao, J. Kim, Y. Xia, M. Milligan, N. Rowe, and G. Veglia. FLAMEnGO 2.0: An Enhanced Fuzzy Logic Algorithm for Structure-Based Assignment of Methyl Group Resonances. *Journal of Magnetic Resonance*, 245:17–23, Jan 2014.