

AUTOMATING THE USAGE OF UNAMBIGUOUS NOES IN NUCLEAR VECTOR
REPLACEMENT FOR NMR PROTEIN STRUCTURE-BASED ASSIGNMENTS

by
MURODZHON AKHMEDOV

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of the requirements for the degree of
Master of Science

Sabanci University

July 2013

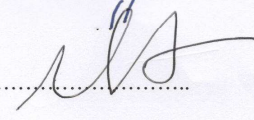
AUTOMATING THE USAGE OF UNAMBIGUOUS NOES IN NUCLEAR VECTOR
REPLACEMENT FOR NMR PROTEIN STRUCTURE-BASED ASSIGNMENTS

Approved by:

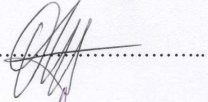
Assoc. Prof. Dr. Bülent Çatay
(Thesis Co-Supervisor)


.....

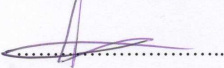
Assist. Prof. Dr. Mehmet Serkan Apaydın
(Thesis Co-Supervisor)


.....

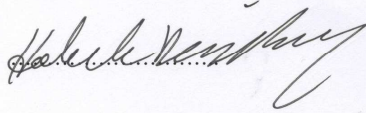
Prof. Dr. Uğur Sezerman


.....

Assist. Prof. Dr. Kemal Kılıç


.....

Assoc. Prof. Dr. Haluk Demirkan


.....

Date of Approval: 22.07.2013

© Murodzhon Akhmedov 2013

All Rights Reserved

AUTOMATING THE USAGE OF UNAMBIGUOUS NOES IN NUCLEAR VECTOR REPLACEMENT FOR NMR PROTEIN STRUCTURE-BASED ASSIGNMENTS

Murodzhon Akhmedov

Industrial Engineering, MS Thesis, 2013

Thesis Co-Supervisor: Assoc. Prof. Dr. Bülent Çatay

Thesis Co-Supervisor: Assist. Prof. Dr. Mehmet Serkan Apaydın

Keywords: Automated NMR Assignment, Tabu Search, NMR Structural Biology, Structural Bioinformatics, Nuclear Overhauser Effect (NOE), Ant Colony Optimization (ACO), Computational Biology, Metaheuristics.

Abstract

Proteins perform various functions and tasks in living organisms. The structure of a protein is essential in identifying the protein function. Therefore, determining the protein structure is of utmost importance. Nuclear Magnetic Resonance (NMR) is one of the experimental methods used to determine the protein structure. The key bottleneck in NMR protein structure determination is assigning NMR peaks to corresponding nuclei, which is known as the assignment problem. This assignment process is manually performed in many laboratories. In this thesis, we have developed methodologies and software to automate this process.

The Structure Based Assignment (SBA) is an approach to solve this computationally challenging problem by using prior information about the protein that is obtained from a template structure. NVR-BIP is an approach that uses the Nuclear Vector Replacement (NVR) framework to model SBA as a binary integer programming problem. NVR-TS is a tabu search algorithm equipped with a guided perturbation mechanism to handle the

proteins with larger residue numbers. NVR-ACO is an ant colony optimization approach that is inspired by the behavior of living ants to minimize peak-nuclei matching cost. One of the input data utilized in these approaches is the Nuclear Overhauser Effect (NOE) data. NOE is an interaction observed between two protons if the protons are located close in space. These protons could be amide protons (HN), protons attached to the alpha-carbon atom in the backbone of the protein (HA), or side chain protons. NVR only uses backbone protons. In the previous approaches using the NVR framework, the proton type was not distinguished in the NOEs and only the HN coordinates were used to incorporate the NOEs into the computation. In this thesis, we fix this problem and use both the HA and HN coordinates and the corresponding distances in our computations. In addition, in the previous studies within this context the distance threshold value for the NOEs was manually tuned for different proteins. However, this limits the application of the methodology for novel proteins. In this thesis we set the threshold value in a standard manner for all proteins by extracting the NOE upper bound distances from the data. Furthermore, for Maltose Binding Protein (MBP), we extract the NOE upper bound distances from the NMR peak intensity values directly and test this protein on real NMR data.

We tested our approach on NVR-ACO's data set and compared our new approaches with NVR-BIP, NVR-TS, and NVR-ACO. The experimental results show that the proposed approach improves the assignment accuracies significantly. In particular, we achieved 100% assignment accuracy on EIN and 80% assignment accuracy on MBP proteins as compared to 83% and 73% accuracies, respectively, obtained in the previous approaches.

NMR PROTEİN YAPI TABANLI ATAMA PROBLEMİ İÇİN BELİRLİ NOELERİ NÜKLEER VEKTÖR DEĞİŞTİRMEDE KULLANIMINI OTOMATİKLEŞTİRME

Murodzhon Akhmedov

Endüstri Mühendiliği, Yüksek Lisans Tezi, 2013

Tez Eş Danışmanı: Doç. Dr. Bülent Çatay

Tez Eş Danışmanı: Yard. Doç. Dr. Mehmet Serkan Apaydın

Anahtar Kelimeler: Otomatik NMR Atamaları, Tabu Arama (TA), NMR Yapısal Biyoloji, Yapısal Biyoinformatik, Nükleer Overhauser Etkisi, Karınca Kolonisi Optimizasyonu (KKO), Hesaplamalı Biyoloji, Metasezgiseller.

Özet

Proteinler canlı organizmalarda çeşitli işlevleri ve görevleri yerine getirirler. Protein yapısı proteinin fonksiyonunun belirlenmesinde gereklidir. Bu nedenle protein yapısının belirlenmesi çok önemlidir. Nükleer Manyetik Rezonans (NMR) protein yapısını belirlemek için geliştirilmiş yöntemlerden biridir. Atama problemleri olarak bilinen NMR tepelerine karşılık gelen amino asitlerin eşleştirilmesi NMR çalışmalarında önemli bir darboğaz oluşturmaktadır. Bu atama işlemi çoğu laboratuarda otomatikleşmemiş ve uzun süren bir süreç sonucunda elde edilir. Bu tezin amacı bu süreci hızlandırmak ve otomatikleştirmek için yeni yöntemler ve yazılım programları geliştirmektir.

Yapı Tabanlı Atama (YTA) bu zor problemi homolog protein yapısını kullanarak çözmek için geliştirilmiş bir yaklaşımdır. NVD-ITP, YTA'yi ikili tamsayı programlama

(ITP) problemi olarak modelleyen ve çözüm için Nükleer Vektör Değişirme (NVD) çerçevesi kullanan bir yaklaşımdır. NVD-TA ise NVD-ITP'in çözemediği daha büyük proteinlerin NMR rezonans verisini atamak için rehberli bir pertürbasyon mekanizması ile donatılmış tabu araması kullanan bir yaklaşımdır. NVD-KKO zirveleri çekirdeklere eşleştirme maliyetini en aza indirmek için doğal karıncalardan esinlenerek geliştirilmiş bir karınca kolonisi optimizasyonu yaklaşımıdır. Bu programlar tarafından kullanılan temel veri kaynaklarından birisi Nükleer Overhauser Etkisidir (NOE). NOE, belli bir yakınlıktaki proton çiftleri arasında ölçülen bir etkidir. Bu protonlar amid protonları (HN), protein omurgasındaki alfa-karbon atomuna bağlı protonlar (HA) veya yan zincir protonları olabilir. NVD sadece omurga protonlarını kullanır. Daha önce geliştirilen yaklaşımlarda NOE'lerde proton tipi ayırt edilmemişti ve sadece HN koordinatları NOE'leri hesaplamalara dahil etmek için kullanılmıştı. Bu tezde ise hesaplamalarda HA ve HN koordinatları ve ilgili uzaklıklar kullanılmıştır. Ayrıca, önceki çalışmalarda NOE etkisinin ölçülebileceği uzaklık eşik değeri her protein için ayrı ayrı belirlenmişti ve bu değerler belirlenirken pratikte mevcut olmayan veriler kullanılmıştı. Metodolojinin uygulama alanının sınırlayan bu yöntem bu tezde NOE mesafe üst sınırlarının hesaplamalara dahil edilerek eşik değerinin tüm proteinler için standart bir şekilde ayarlandığı bir yaklaşımla daha geliştirilmiştir. Ayrıca, Maltoz Bağlayıcı Proteini (MBP) için doğrudan NMR tepe yoğunluğu değerlerinden NOE üst sınır uzaklıkları elde edilerek gerçek NMR verisiyle sınanmıştır.

Geliştirilen yeni yaklaşımlar NVD-KKO verileri kullanarak sınanmış ve elde edilen sonuçlar NVD-ITP, NVD-TA ve NVD-KKO sonuçlarıyla karşılaştırılmıştır. Deneysel sonuçlar önerilen yaklaşımın atama doğruluklarını önemli ölçüde iyileştirdiğini göstermektedir. Önceki yaklaşımlarla EIN ve MBP için sırasıyla 83% ve 73% atama doğrulukları elde edilmişti. Yeni yaklaşımlarla EIN protein verisi için 100% atama doğruluğu ve MBP protein verisi için 80% atama doğruluğu elde edilmiştir.

Acknowledgements

I am glad to express my gratitude to my co-supervisors Assoc. Prof. Dr. Bülent Çatay and Assist. Prof. Dr. Mehmet Serkan Apaydın for their guidance to shape my masters studies and finalize my thesis. Furthermore, I am thankful them for sharing their expertise, skills and knowledge throughout this journey.

I would like to express my gratitude to my thesis committee, Prof. Dr. Uğur Sezerman, Assist. Prof. Dr. Kemal Kılıç and Clinical Prof. Dr. Haluk Demirkan for their review and valuable comments.

I am deeply thankful to my friends, Fardin Dashty Saridarq, Zhenishbek Zhakypov, Tarık Edip Kurt for their valuable discussions in certain scientific topics and their contributions to my personal maturity. Besides, I want to thank them for their friendship and being a source of motivation.

I am indebted to all my friends in the IE office. My special thanks to Burcu, Özge, Gürkan, Fikri, Mahir, Çetin, Canan, Merve, Utku, Behrooz, Daniel, Semih, Apıtunç and Berk. Thanks them for sharing the moments.

My great gratitude to my family that they believed in me and morally supported up to this moment, and presented their unconditional love.

TABLE OF CONTENTS

Table of Contents

Acknowledgements	VIII
TABLE OF CONTENTS	IX
LIST OF FIGURES	XI
LIST OF TABLES	XII
LIST OF ABBREVIATIONS	XIII
1 INTRODUCTION.....	14
2 LITERATURE REVIEW	20
2.1 Related Work.....	20
2.2 Background	22
3 PROBLEM DESCRIPTION AND FORMULATION	24
3.1 Problem Definition.....	24
3.2 NVR Framework	26
3.3 NOE Usage in the NVR Framework.....	26
3.4 Mathematical Formulations.....	28
3.4.1 Distinguishing the type of NOE	29
3.4.2 Using the NOE upper bounds extracted from the data.....	30
4 SOLUTION METHODOLOGY.....	32
4.1 NA-NVR-TS	33
4.1.1 Distinguishing the type of NOE	33
4.1.2 Using the NOE upper bounds extracted from the data.....	34
4.2 NA- NVR-ACO.....	34

5	EXPERIMENTAL STUDY	36
5.1	Data Sets	36
5.2	Computational Results	37
5.2.1	Distinguishing the type of NOE	37
5.2.2	Using the NOE upper bounds extracted from the data.....	41
6	CONCLUSION AND FUTURE WORK.....	46
	Bibliography	48
	Appendix A	51
	Appendix B.....	53

LIST OF FIGURES

Figure 1.1: Peptide bond formation.....	14
Figure 1.2: Secondary structure and colored surface of ubiquitin	15
Figure 1.3: Protein structure determination by NMR.....	16
Figure 1.4: Structure-Based Assignment of NMR peaks to amino acids.....	17
Figure 3.1: NOE relations and Assignment of NMR peaks to amino acids.....	25
Figure 3.2A: Structural Formula	27
Figure 3.2B: Ball and Stick	27
Figure 3.3: Portion of protein	27

LIST OF TABLES

Table 5.1: Assignment accuracies for ubiquitin when distinguishing NOE type.....	38
Table 5.2: Assignment accuracies for SPG when distinguishing NOE type.....	38
Table 5.3: Assignment accuracies for lysozyme when distinguishing NOE type.....	39
Table 5.4: Assignment accuracies for other proteins when distinguishing NOE type ...	39
Table 5.5: Assignment accuracies for large proteins when distinguishing NOE type ...	40
Table 5.6: Assignment accuracies for ubiquitin when using the NOE upper bounds	42
Table 5.7: Assignment accuracies for SPG when using the NOE upper bounds	43
Table 5.8: Assignment accuracies for lysozyme when using the NOE upper bounds ...	43
Table 5.9: Assignment accuracies for other proteins when using the NOE upper bounds.....	44
Table 5.10: Assignment accuracies for large proteins when using the NOE upper bounds.....	44

LIST OF ABBREVIATIONS

ACO: Ant Colony Optimization

BIP: Binary Integer Programming

EIN: Amino Terminal Domain of Enzyme I from Escherichia Coli

EM: Expectation Maximization

ff2: The FF Domain 2 of human transcription elongation factor CA150
(RNA polymerase II C-terminal domain interacting protein)

hSRI: Human Set2-Rpb1 Interacting Domain

ILP: Integer Linear Programming

IP: Integer Programming

LP: Linear Programming

LS: Local Search

MBP: Maltose Binding Protein

NA: NOE aware

NMR: Nuclear Magnetic Resonance

NOE: Nuclear Overhauser Effect

NVR: Nuclear Vector Replacement

PDB: Protein Data Bank

RA: Resonance Assignment

RDC: Residual Dipolar Coupling

RNA: Ribonucleic Acid

SBA: Structure-Based Assignments

SPG: Streptococcal Protein G

TS: Tabu Search

Chapter 1

INTRODUCTION

Proteins are large biological molecules that consist of one or more than one amino acid combinations in the chain form. Proteins vary from one another primarily in their sequence of amino acids that is dictated by the nucleotide sequence of their genes. There are 20 types of amino acids. The amino acid structure includes alpha-carbon, carboxylic group, amino group, and a side chain (see Figure 1.1). The side chain is specific to each amino acid and determines the physical and chemical properties of amino acid. Amino acids come together to form the peptide bonds in a protein chain.

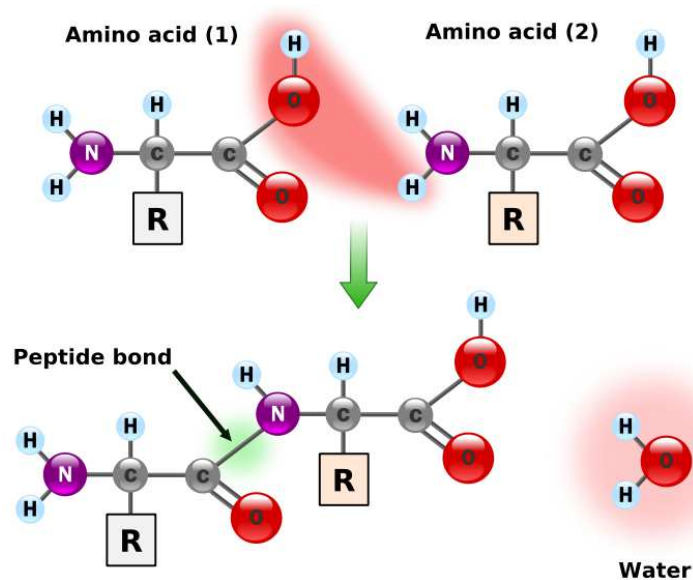


Figure 1.1: Peptide bond formation

The peptide bond is formed between carboxyl group of one molecule (Amino acid 1) and amino group of other molecule (Amino acid 2), causing a release of water molecule. This process is demonstrated in Figure 1.1.

The human body consists of 45% proteins and proteins have large range of important functions in living organisms. Some of these functions include building and repairing the body, water balancing processes, transporting the information, replicating DNA, catalyzing metabolic reactions, responding to stimuli, and helping the immune system. There is a strong relationship between the three dimensional structure and the function of the protein. Furthermore, 3D structure and surface of protein plays a vital role in protein-protein interactions and protein–ligand binding affinity analysis. Therefore, identifying the protein structure is essential to understand and analyze the functional behavior of proteins as well as their dynamics, for protein redesign, diagnosis and treatment of medical diseases. In Figure 1.2A, the backbone fold of ubiquitin is demonstrated with the secondary structure elements [17]. The surface of ubiquitin is displayed in Figure 1.2B and it is colored by residue type. The color scheme is gray for non polar, green for polar (uncharged), red for acidic, and blue for basic amino acids. Clearly, the chemical and physical properties are tightly related to 3D structure and surface of protein.

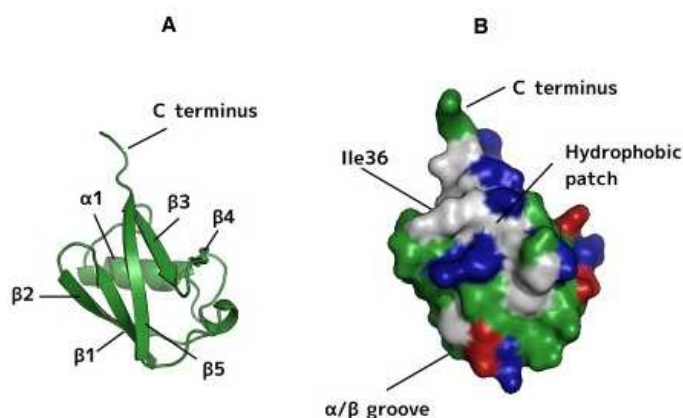


Figure 1.2: Secondary structure and colored surface of ubiquitin

There exist two major methods for protein structure determination in the literature. The first method is X-Ray Crystallography. It is a process by which x-rays are passed through the molecular lattice of a crystal and reveal the crystal's underlying atomic structure. It was introduced by Von Laue in 1912, and since that time, x-ray diffraction has grown to encompass crystallography of DNA structure, proteins, various molecules, and complex structures [10]. This method requires crystallized protein form to obtain the structure. However, it can take a long time to crystallize some proteins. For this and other several reasons, the Nuclear Magnetic Resonance (NMR) method has been recently developed in the literature. NMR is ideally suited for detailed studies of protein-protein and protein-ligand interactions as well as dynamics of protein. Furthermore, it is well suited for probing and analyzing changes to the local electronic environment of the protein [5]. NMR does not yield a 3D structure of a protein directly. Instead, it gives high throughput data related to the structure and the 3D structure can be calculated through intensive data analysis. The protein structure determination steps using NMR spectroscopy in solution can be divided as follows (Figure 1.3): preparation of the protein solution, the NMR experiments and measurements (identification of conformation constraints, e.g. distances between hydrogen atoms), the assignment of NMR signals to individual atoms in the protein, the calculation of 3D structure of the protein [6].

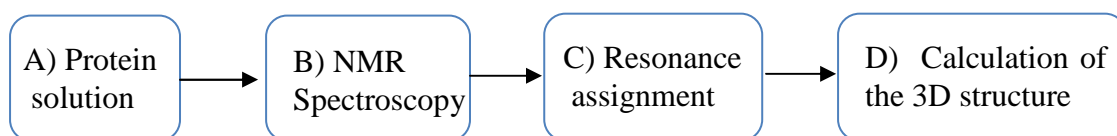


Figure 1.3: Protein structure determination by NMR

In the NMR experiment, the protein solution is prepared (Figure 1.3A), the protein atoms in solutions are irradiated via magnetic wave frequency, and irradiation is recorded, and converted to a spectrum (Figure 1.3B). In the spectrum, each peak corresponds to one amino acid in the protein sequence and it should be assigned to continue the structure determination process further (Figure 1.3C). This process is a bottleneck in the NMR approach and is still manually done in many laboratories. Our efforts are dedicated to resolve this bottleneck and to automatically assign NMR signals

to individual atoms by using prior structural information from the template structure. This process is shown in Figure 1.4 and explained below.

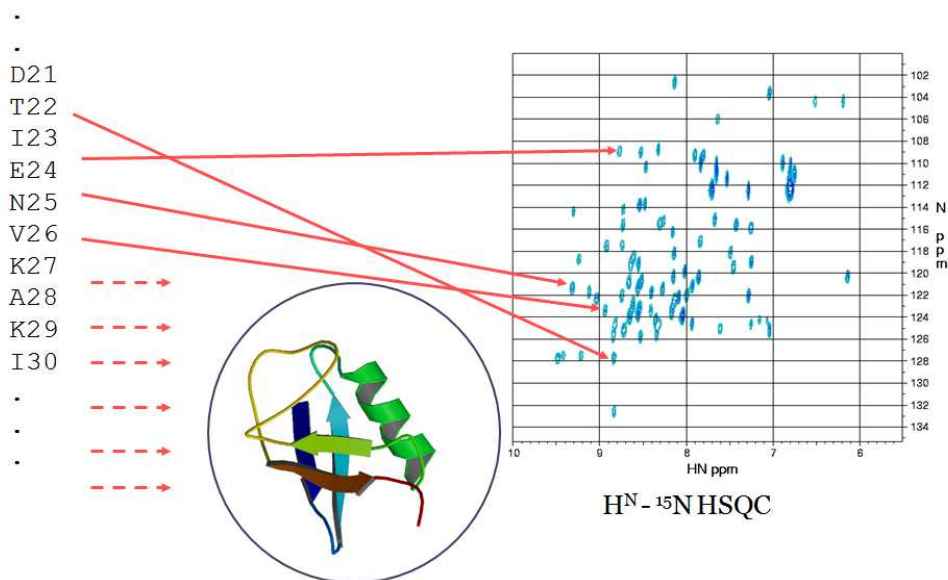


Figure 1.4: Structure-based assignment of NMR peaks to amino acids

On the right side of Figure 1.4, the $H^N-^{15}N$ HSQC 2D data of protein is presented. The horizontal and vertical values demonstrate the chemical shifts of hydrogen atoms and nitrogen atoms of the protein residues, respectively. In the spectrum, each peak corresponds to one residue in the protein sequence and it should be assigned. There exist the sequential and the structure-based resonance assignment methods for assigning NMR peaks to corresponding nuclei. The Structure-Based Assignment (SBA) is an approach that uses the homologous structure while making peak-nuclei assignment. SBA resembles the molecular replacement technique in x-ray crystallography which determines the structure rapidly and accurately with the help of template structure.

The NMR methodology intensively uses NOESY- ^{15}N - HSQC experiment. NOE is the effect measured between protons when a pair of protons close in space is irradiated. That effect is independent from the direct connection of the protons by

chemical bonds. The NOE is observed from the nuclei in a vicinity roughly less than 6 Å; therefore it can be used to determine inter and intra molecular distances.

The molecular size of the protein is very important, its largeness, thus constitutes a limitation in the NMR technique. It is more difficult to analyze NMR experiment data for proteins larger than 30 kDa due to the huge amount of signals that overlap each other. However, a novel technique called Transverse Relaxation Optimized Spectroscopy (TROSY) overcomes this challenge. This technique is enriched with different types of NMR experiments and reduces the signal loss. Therefore, it allows analyzing the molecular data corresponding to proteins larger than 100 kDa [6].

In [18, 20, and 22], the threshold value is manually tuned for NOE relations for each protein and this threshold is the same for all NOEs. However, this limits the applicability of the approach. The threshold value for NOE relation should be determined automatically by an approach. We incorporated the NOE distance upper bounds into the computations as threshold values on NOE relations for all proteins. Furthermore, in the previous approaches only HN-HN NOEs were incorporated into computations. The NOEs between HN-HA and HA-HN protons were treated as HN-HN NOEs. However, this causes a problem in determining correct distances between protons since only HN coordinates were used for pair-wise proton distance computations. Thus, the proton type distinction in NOEs becomes necessary. In this thesis, we overcame these challenges. We distinguish the proton types in NOEs and use both HN and HA protons coordinates, and incorporate corresponding distances into computations. We call our modified approaches as NOE aware NVR-BIP (NA-NVR-BIP), NOE aware NVR-TS (NA-NVR-TS) and NOE aware NVR-ACO (NA-NVR-ACO).

Our contributions are:

- Formulation of NVR-BIP model to incorporate HN and HA coordinates and utilizing the upper bound of NOE relations as a threshold value
- Formulation of NVR-TS algorithm to incorporate HN and HA coordinates and utilizing the upper bound of NOE relations as a threshold value

- Formulation of NVR-ACO algorithm to incorporate HN and HA coordinates and utilizing the upper bound of NOE relations as a threshold value
- Extraction of NOE upper bound values automatically from the NOE distances
- Testing the NA-NVR-BIP, NA-NVR-TS, and NA-NVR-ACO on NVR-ACO's data set.
- Test on a large protein with real NMR data

The remainder of the thesis is organized as follows: in the next chapter we present the literature review. Then, we give problem definition, NVR framework, NOE usage in NVR framework and mathematical formulations of the problem. We describe the NA-NVR-TS and NA-NVR-ACO algorithms under the solution methodology in chapter 4. The next chapter consists of information about the data sets and computational results. Finally, we present the conclusion and future work in chapter 6.

Chapter 2

LITERATURE REVIEW

2.1 Related Work

There are several software programs that perform resonance assignments in the literature. MARS [2] facilitates an automatic backbone assignment of proteins by using ^{13}C and ^{15}N labeled protons. MARS simultaneously optimizes the local and global quality of assignment and combines the secondary structure information from PSIPRED [27]. However, it uses triple-resonance experiments and makes an exhaustive search while processing the assignments. The program is tested on maltose binding protein with 370 residues and 96% error-free assignment is obtained. In [7], authors target an enhanced backbone resonance assignment by matching experimental Residual Dipolar Coupling (RDC) to back computed values from a known 3D structure. Furthermore, RDC is helpful in reducing chemical shift degeneracy in sequential connectivity experiments. Besides, the combination of sequential connectivity information and RDC matching can improve the performance of MARS against missing data.

There are several SBA algorithms in the literature. Some algorithms require Residual Dipolar Coupling (RDCs) and triple resonance experiments as an input. Nuclear Vector Replacement (NVR) [5] is a molecular replacement-like approach for SBA. NVR performs backbone resonance assignment as a combinatorial optimization problem by employing geometric and topological constraints of prior 3D homologous structure, such that all NMR data should satisfy the existing constraints. In [5] the NVR algorithm is proposed to perform the resonance assignments in polynomial time for proteins with known structures or homologous structures. NVR processes an unassigned NOESY- ^{15}N - HSQC spectra, HN - ^{15}N RDCs, and sparse HN-HN NOEs and uses uniform ^{15}N -labeling of the protein. The algorithm is tested on ubiquitin (76 residues)

and lysozyme (129 residues) proteins and 90% and 98% assignment accuracies are achieved, respectively. Previous algorithms that utilize homologous structure require ^{13}C -labeling to perform resonance assignment. On the other hand, NVR uses only ^{15}N -labeling which is much less expensive to obtain and does not require triple resonance experiments.

NVR-EM [4] has a polynomial time complexity and uses a greedy expectation maximization (EM) algorithm to perform the assignments. RDC data gives a global orientation about inter molecular bound vectors in space. NVR-EM is an RDC-based approach to determine alignment tensors and to perform the resonance assignments by correlating chemical shifts of $\text{HN}-^{15}\text{N}$ – HSQC peak spectra with homologous structure. Furthermore, the method can handle the missing data in RDCs and resonances.

In [9], the authors propose a fully automated RDC-based NMR resonance assignment strategy for rapidly determining the tertiary structure of RNA.

In [23], the authors proposed HANA that uses RDCs and Hausdorff- based pattern matching technique to analyze the similarity between experimental and back-computed NOE spectra and to assign peaks to pairs of protons. The algorithm is tested on human ubiquitin, domain of human DNA Y-polymerase Eta (pol η) and human Set2-Rpbl interacting domain (hSRI) and over 90% assignment accuracies are obtained.

In general, it is known that two similar protein sequences are most likely to have a similar 3D structure and sequence-based structural homology prediction methods could be used for structure determination. On the other hand, it is hard to predict the structural similarity of two dissimilar protein sequences for sequence-based homology predictors. [8] addresses the challenge of structural homology detection of dissimilar protein sequences. The authors propose HD algorithm in NVR framework for detecting the structural homology likelihood from sparse and unassigned NMR data. The advantage of their method is its independence from sequence homology and requirement of less time to acquire the experimental protein NMR data. HD is tested on 3 proteins and successful homology detection is reported, and no false positives or false negatives are reported for sequences with less than 30% similarities.

2.2 Background

Today in many laboratories, the assignment problem is performed manually which is a time consuming process. Our aim is to develop methods to automatically solve the assignment problem. The SBA problem was formulated as a binary integer programming in NVR-BIP [18], under the scope of NVR. Since this problem is the NP-hard, NVR-TS and NVR-ACO metaheuristic approaches are developed to obtain a solution for large proteins.

Tabu search (TS) is a metaheuristic algorithm that was created by Fred W. Glover and it is widely used in combinatorial optimization problems. TS uses the neighborhood search procedure to iteratively move from one solution to another solution in order to improve the objective function.

The NVR-TS is a tabu search based approach with well equipped perturbation mechanism. Starting from an initial solution, TS investigates the neighbors of the existing solution at each iteration in an attempt to improve the incumbent best solution. It avoids the repetition of the same solutions by maintaining a mechanism called tabu list. The tabu list keeps the information of the latest moves or solutions and prevents the search from returning to those solutions for a specified number of iterations since they guide either to local optimal solutions or to solutions that have already been explored. TS accepts a tabu move only if it satisfies a predefined aspiration criterion. NVR-TS allows the NOE violations by penalizing each of them with predetermined penalty score in the objective function.

Ant colony optimization (ACO) is a probabilistic technique to solve computational problems. It is inspired by natural behavior of ants while they search for food. Ants find shortest path between their nest and the food source in a reasonable time by using pheromone level. Greater level of pheromone on the path increases the probability of following that path by ants. The level of pheromone on the path is negatively proportional to the length of the paths. Intuitively, all ants will follow the shortest path in time. The behavior of real ants is simulated by artificial ants in ACO to solve combinatorial problems. Artificial ants obtain a solution on a graph using constructive mechanism guided by pheromone update and greedy heuristic known as visibility. Pheromone trail τ_{ij} intensity values between node i and j are proportional to quality of generated solution and show the collective memory of ants. The visibility n_{ij}

is heuristic information that represents the attractiveness of moving from node i to j . Furthermore, the artificial ant can use local search heuristics in order to improve solution quality.

NVR-ACO is an ant colony optimization approach to solve the SBA problem. It is inspired by the efficiency of food gathering in ant behavior. Ants explore the shortest path from their nest to food source by using information known as pheromone. In a similar fashion, NVR-ACO assigns peaks to amino acids by minimizing matching cost and penalizing assignments with NOE violation.

Chapter 3

PROBLEM DESCRIPTION AND FORMULATION

In this chapter, the definition of the assignment problem in the SBA scope is given. Furthermore, the NVR framework and the NOE usage in the NVR framework are explained in detail. In addition, we present two mathematical formulations to the assignment problem that are adapted from NVR-BIP [18]. These formulations take into account NOE type distinction and extraction of the NOE upper bound distance information from the data.

3.1 Problem Definition

In NMR experiment, the protein atoms are irradiated via magnetic wave frequency then irradiation is recorded, and converted to the spectrum. In the spectrum, every peak corresponds to one amino acid in a protein sequence and it should be assigned to further proceed with the structure determination process. This problem is known as the assignment problem and it is a bottleneck in the NMR approach.

One of the experiment types that are extensively used in NMR methodology is the NOESY- ^{15}N - HSQC experiment. This experiment yields the NOE which is observed between the nearby pairs of backbone protons. NOE is an effect that is measured between protons when a pair of protons close in space is irradiated. The NOE effect is independent from the direct connection of the protons by chemical bonds. The NOE is in general observed from nuclei in vicinity less than 6 Å; therefore it can be used to determine inter and intra-molecular distances.

The NOE relation between the protons and the assignment problem are demonstrated in Figure 3.1.

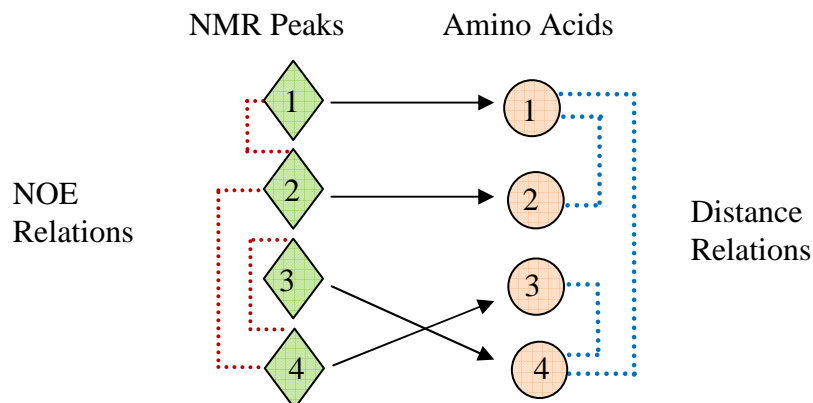


Figure 3.1: NOE relations and assignment of NMR peaks to amino acids

There is a set of NMR peaks that should be assigned to a set of amino acids. An arc between a pair of NMR peaks demonstrates the NOE relation between the corresponding peaks. An arc between a pair of amino acids shows that the distance between corresponding protons is smaller than NOE distance threshold (NTH) value and that the amino acids are located in the vicinity of each other. The peaks associated with NOE relations should be mapped to amino acids that have a distance relation. For instance, there is an NOE between peak 1 and peak 2. If peak 1 is mapped to amino acid 1 and peak 2 is mapped to amino acid 2, as shown in Figure 3.1, then this assignment is feasible, because the distance between amino acid 1 and amino acid 2 is less than NTH. However, if peak 2 is mapped to amino acid 2 and peak 4 is mapped to amino acid 3, then this assignment would be infeasible due to distance between the corresponding amino acids. Here, the assignment problem is to find a maximum bipartite graph mapping of peaks and atoms with the minimum matching cost by utilizing the NOE and distance constraints, and penalizing the infeasible assignments.

3.2 NVR Framework

NVR is a SBA framework where the goal is to find a matching between the peaks and amino acids. At the same time it minimizes the mapping cost while satisfying the NOE constraints and distance constraints between the amino acids. Since NOE constraints are between a pair of peaks, they limit the available amino acid assignments to the corresponding peak pairs.

NVR uses the following data types: HN-¹⁵N – HSQC, NOESY-¹⁵N – HSQC (observed between nearby pairs of backbone protons), HN-¹⁵N RDCs in two media (which provide global orientational restraints on bond vectors), ¹⁵N TOCSY (for the side-chain chemical shifts), and amide exchange HSQC (to identify, probabilistically, solvent exposed amide protons). NVR associates an assignment probability with each peak to amino acid match. Interested readers may refer to [18] for detailed information.

3.3 NOE Usage in the NVR Framework

NOE is one of the input data types that are used in NVR framework. It is an effect between a pair protons close in 3D space. This effect is highly related to the distance between the protons. However, it is independent of any chemical bonding between protons and it can be observed with or without any interaction between protons. Thus, NOE is useful to determine inter and intra- molecular distances.

In [18, 20 and 22], the NOE type was not distinguished and only HN-HN NOE type was utilized and incorporated into computations. HN-HA and HA-HN NOE types were considered as HN-HN, and only HN proton coordinate was used to incorporate these NOEs. However, this could create errors due to mismatch of the NOE type and proton coordinate. Thus, to obtain more realistic solutions the distinction of NOE type and employment of correct proton coordinates is unavoidable. It also improves the robustness of the models and approaches. The proton coordinates and NOE relations are explained in details in the following figures.

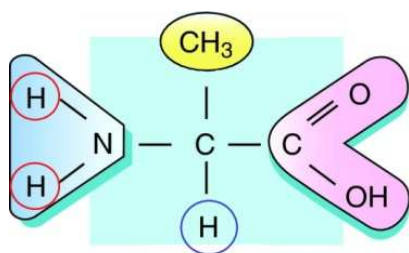


Figure 3.2A: Structural Formula

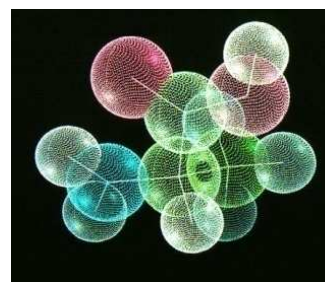


Figure 3.2B: Ball and Stick

In Figure 3.2A and Figure 3.2B the structural formula and the ball-and-stick model of alanine is presented. In Figure 3.2A, the hydrogens circled in red are HN protons and the hydrogen circled in blue is HA proton. In Figure 3.2B, green, red, aqua, and lime represent the carbon, oxygen, nitrogen, and hydrogen, respectively.

In Figure 3.3 the small portion of the protein is demonstrated containing a glycine in the middle. The green line represents the HN-HA NOE and the purple line shows the HN-HN NOE between the residue i and residue $i-1$.

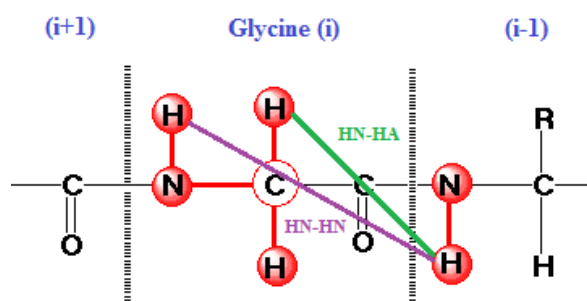


Figure 3.3: Portion of protein

By examining these figures, it is easy to recognize that there may be a significant difference between using HN proton coordinates instead of HA proton coordinates to incorporate NOEs into computations. Furthermore, this can lower the assignment accuracies and make them less reliable. To remedy this drawback, the proton type distinction in NOEs becomes unavoidable for the robustness of the algorithm. In this thesis, we distinguish HN-HA and HA-HN NOEs from HN-HN NOEs and three

different intra-proton distance matrices are calculated for HN-HN, HA-HN, and HN-HA NOE relations from the template structure.

3.4 Mathematical Formulations

In NVR-BIP [18], the SBA problem was formulated as a binary integer programming. The formulated problem was implemented in ILOG OPL environment which employs a CPLEX solver engine. In NVR-BIP, the type of NOE input data was not distinguished and only HN proton coordinate was used in calculations. Also, the threshold value for NOE was manually set for each protein. In order to automate the assignment process of NMR peaks to amino acids, the distinction of NOE type and correct proton coordinate usage is necessary. The automatic threshold value selection on distances among protons is also necessary for automating the assignment process. These modifications will expand the application of the approach to novel proteins. Besides, the distinguished NOE data type may improve the assignment accuracies. We obtained a new formulation of the NVR-BIP problem in this thesis in order to achieve these goals. We incorporated all these changes into the mathematical model in two steps. As a first step, we reformulated NVR-BIP to distinguish the proton types in NOEs. Here, the correct proton coordinates are employed along with NOE input data type distinction. This approach is considered under the “Distinguishing the type of NOE” model. In the next step, we reformulated the NVR-BIP to distinguish the type of NOE and automatically set the NOE threshold value. In this context, we utilize the NOE upper bounds as the threshold values on intra-proton distances and use the correct proton type in our calculations. The NOE upper bound distance information is extracted directly from the input data. This approach is considered under “Using the NOE upper bounds extracted from the data” model.

These reformulated models are named as NA-NVR-BIP and the details regarding the models are provided in the following sections.

3.4.1 Distinguishing the type of NOE

Distinguishing the type of NOE is the first step to automate the NMR peak assignment process within our approach. In this mathematical model the proton type in NOEs is differentiated and corresponding proton coordinate is employed. However, the threshold value on distances between protons is obtained manually and is the same for all NOE constraints. For each tested protein, the threshold value is manually adjusted in the sense that the solution without NOE violation could be achieved. The threshold is determined as a value that is greater than exact distances between protons which are correctly assigned to pair of peaks that have NOE relation. The notation and the formulation of the model are described below.

Notation:

P : set of peaks

A : set of amino acids

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 t coordinate type, $t \in T = \{HN - HN, HN - HA, HA - HN\}$

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

NTH : The threshold value for intra – amide proton distances

$$b_{jlt} = \begin{cases} 1 & \text{if } d_{jlt} \geq NTH \\ 2 & \text{otherwise} \end{cases} \quad \forall j, l \in A, \forall t \in T$$

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:

$$\text{Minimize} \quad \sum_{i \in P} \sum_{j \in A} s_{ij} x_{ij} \quad (1)$$

$$\text{s. t.} \quad \sum_{i \in P} x_{ij} \leq 1 \quad \forall j \in A \quad (2)$$

$$\sum_{j \in A} x_{ij} \leq 1 \quad \forall i \in P \quad (3)$$

$$\sum_{i \in P} \sum_{j \in A} x_{ij} = N \quad (4)$$

$$x_{ij} + x_{kl} \leq b_{jlt} \quad \forall j, l \in A, \forall i, k \in P, \forall t \in T, \forall k \in \text{NOE}(i) \quad (5)$$

$$x_{ij} \in \{0,1\} \quad \forall i \in P, \forall j \in A \quad (6)$$

In this model, the objective function (1) minimizes the total cost of mapping peaks to amino acids. Constraint set (2) satisfies that each amino acid is assigned to at most one peak and constraint set (3) ensures that each peak is mapped to at most one amino acid. Constraint (4) equalizes the total number of assignments to the number of peaks to be assigned. This constraint will be redundant if the number of peaks is equal to the number of amino acids and “ \leq ” sign is replaced by “ $=$ ” sign in constraints (2) and constraints (3). The parameter b_{jlt} is determined according to d_{jlt} and threshold value. Constraint set (5) satisfies the NOE relations between peaks and the constraint set (6) forces the decision variables to be binary.

3.4.2 Using the NOE upper bounds extracted from the data

In addition to distinguishing the proton type, in this section we also obtain the NOE upper bound information from the data. This reduces the number of manually tuned parameters the system relies on and makes the approach more general. As a result, it yields more realistic solutions.

In the context of using the NOE upper bound distances, there is a different threshold for every pair of peaks that have NOE relation between them. In other words, the number of threshold values is equal to the number of NOE constraints used for the assignments in the tested proteins. Each NOE relation has its own predetermined threshold value from the data. In this way we generalized the method by automatically determining the threshold values. The notation and the formulation of the model are expressed below.

$$b_{ijklt} = \begin{cases} 1 & \text{if } d_{jlt} \geq UB_{ik} \\ 2 & \text{otherwise} \end{cases} \quad \forall j, l \in A, \forall i, k \in P, \forall t \in T$$

Where

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

$$x_{ij} + x_{kl} \leq b_{ijklt} \quad \forall j, l \in A, \forall i, k \in P, \forall t \in T, \forall k \in NOE(i) \quad (7)$$

Here, the objective function and some constraints the same as in distinguishing the type of NOE model. Minor changes are; the parameter b_{jlt} is replaced by b_{ijklt} and the constraint set (5) is replaced by constraint set (7). In using NOE upper bounds extracted from the data model, the threshold values over interproton distances are gathered from the input data. There exists a unique threshold value on each pair of peaks that has an NOE relation between them. Similarly, NOE relations between pair of peaks are also updated and expressed in constraint set (7).

Chapter 4

SOLUTION METHODOLOGY

In this chapter, two formerly developed metaheuristic approaches are adapted to the models of NA-NVR-BIP by relaxation of NOE relation constraints. Since the backbone resonance assignment problem is an NP-hard problem, NVR-BIP found results for only small proteins. To fix this drawback and obtain assignment solutions for novel proteins NVR-TS [20] and NVR-ACO [22] metaheuristic algorithms were developed. In this thesis, we adapted the metaheuristic algorithms to incorporate the proton type distinctions in NOEs and NOE upper bound information utilization and we refer to these approaches as NA-NVR-TS and NA-NVR-ACO.

In these proposed approaches the correct proton coordinate is used to incorporate NOEs into computations and proton type in NOE is distinguished. On top of it, the NOE upper bounds are utilized as a threshold over the interproton distances. We had the NOE upper bound relations as a distance magnitude for all proteins except MBP; those are directly taken from the input data. For the MBP, we had the intensity values for NOE relations between the peak pairs. We converted the intensity values to the upper bound distance limits by using the simple protocol in Clore and Gronenborn work [24, 25, and 26]. The peak intensities are ranked and binned into the 4 categories. The peak intensities in the range of 0-20% are considered as a very weak, 20-50% considered as a weak, 50-80% considered as a medium, 80-100% considered as a strong and they have an upper bound distance limit of 6.0 Å, 5.0 Å, 3.3 Å, and 2.7 Å, respectively. 0.5 Å is added to all upper bounds in order to correct for the experimental error and intensity of methyl crosspeaks that are larger than expected. These upper bounds are used throughout calculations.

4.1 NA-NVR-TS

We adapted NVR-TS [20] for distinguishing the type of NOE and utilizing NOE upper bound distances that are extracted from the input data in this approach. The implementation of the algorithm is based on relaxation of NOE constraints in NA-NVR-BIP models. In NA-NVR-BIP, the NOE constraints are considered as of hard type and do not allow NOE violations in solutions. On the other hand, NOE violations are allowed in relaxed models by penalizing them in objective function. Constraint set (5) in distinguishing the type of NOE model and constraint set (7) in using NOE upper bounds extracted from the data model are removed and added to the objective functions with corresponding NOE violation penalties. Minimization models avoid NOE violations since they have positive multipliers in objective function. The corresponding models are adapted in the following sections.

4.1.1 Distinguishing the type of NOE

The NA-NVR-BIP's distinguishing the type of NOE model adaptation is presented as a quadratic relaxation formulation below.

$$\text{Minimize} \quad \sum_{i \in P} \sum_{j \in A} s_{ij} x_{ij} + \sum_{i \in P} \sum_{k \in \text{NOE}(i)} \sum_{j \in A} \sum_{l \in A} \sum_{t \in T} p_{jlt} x_{ij} x_{kl} \quad (8)$$

$$p_{jlt} = \begin{cases} s' & \text{if } d_{jlt} > \text{NTH} \\ 0 & \text{otherwise} \end{cases} \quad \forall j, l \in A, \forall t \in T \quad (9)$$

Where

$$s' = \max \{s_{ij} : i \in P, j \in A\} \quad (10)$$

The objective function (8) minimizes the total mapping cost of peaks to amino acids and simultaneously minimizes the number of NOE violations. The NOE relation constraint set (5) is added to the objective function. Each NOE violation is penalized with p_{jlt} constant and plays a vital role in the procedure. If penalty is a very small number then the model ignores NOE violations and concentrates on mapping cost. In

contrast, if a big number is chosen then the model neglects the matching cost and NOE violations get higher priority. After serious preliminary tests, the penalty is determined as in (9) and (10). Any NMR peak to amino acid assignment that satisfies the constraint set (2)-(4) is an initial solution. The algorithm starts from an initial solution and iteratively improves it. The interested reader is referred to [20] for detailed information of the algorithm and its working mechanism.

4.1.2 Using the NOE upper bounds extracted from the data

Using the NOE upper bounds extracted from the data model adaptation is similar to distinguishing the type of NOE. The differences are, the objective function (8) is replaced by (11) and violation penalty coefficient (9) is replaced by (12). The quadratic relaxation formulation of the model is as follows:

$$\text{Minimize} \quad \sum_{i \in P} \sum_{j \in A} s_{ij} x_{ij} + \sum_{i \in P} \sum_{k \in \text{NOE}(i)} \sum_{j \in A} \sum_{l \in A} \sum_{t \in T} p_{ijklt} x_{ij} x_{kl} \quad (11)$$

$$p_{ijklt} = \begin{cases} s'' & \text{if } d_{jlt} > UB_{ik} \\ 0 & \text{otherwise} \end{cases} \quad \forall j, l \in A, \forall i, k \in P, \forall t \in T \quad (12)$$

Where

$$s'' = \max \{s_{ij} : i \in P, j \in A\} \quad (13)$$

The objective function (11) minimizes the total matching cost of peak to residue assignments and NOE violation cost. The violation penalty coefficient is updated as in (12) after using NOE upper bound distance limits as threshold values.

4.2 NA-NVR-ACO

In this approach we modified NVR-ACO by distinguishing the proton types in NOEs and provided the algorithm with the corresponding input data. The implementation of the algorithm relies on the NA-NVR-BIP model. The algorithm became sensitive to NOE types with this modification. In the NA-NVR-ACO algorithm,

the correct coordinates of protons are used to incorporate NOEs into calculations. Furthermore, NA-NVR-ACO utilizes NOE upper bound distance limits that are obtained from the data as a threshold value. The interested reader may refer to [22] for detailed information of the algorithm and its mechanism.

Chapter 5

EXPERIMENTAL STUDY

5.1 Data Sets

We tested the performance of NA-NVR-BIP, NA-NVR-TS and NA-NVR-ACO on the data set used in NVR-BIP since the scores obtained by solving NVR-BIP are optimal. NVR-BIP data set includes lysozyme, human ubiquitin, hSRI, GB1, ff2, SPG and pol η . Furthermore, we tested the algorithms on two novel proteins which were not included in NVR-BIP's data set: Amino Terminal Domain of Enzyme I from Escherichia Coli (EIN) with 243 residues and Maltose-binding protein (MBP) with 348 residues. The proteins we tested all have NOE data where the source of the NOE is distinguished. The remaining proteins in NVR-BIP are tested by means of simulated NOE data.

For most cases, the templates used correspond to the x-ray structures of the proteins. The NMR backbone resonance assignments are performed for 13 structural homologous models in lysozyme protein family and a total of 534 NOE constraints are used, including HN-HN, HN-HA and HA-HN NOE types. For ubiquitin protein family, the NMR data assignments for five homologous models are computed and 270 NOE constraints are employed in total. The backbone resonance assignments for three structural homologous models are computed using 204 NOE constraints in SPG protein family. For large proteins, 1021 NOE constraints for EIN and 474 NOE constraints for MBP are utilized. For the rest of the proteins, 266, 260, 234, 156 NOE constraints are employed for hSRI, GB1, ff2 and pol η , respectively.

5.2 Computational Results

As stated before, in previous approaches [18, 20, 22], the proton type was not distinguished in handling NOE data. In addition, the threshold values were manually tuned on the distances among amide protons which are obtained from homologous structures. In this thesis, we distinguished the proton types in NOEs and also utilized the NOE relation upper bound data as the threshold values on the distances among protons.

In this section we compare the results obtained by previous approaches with NA-NVR-BIP, NA-NVR-TS, and NA-NVR-ACO. First, the results from [18, 20, 22] are compared with those obtained by proton type distinction in NOEs. Next, we compare the results from the previous approaches with the results achieved by the combination of proton type distinction in NOEs and the automatic usage of threshold values obtained from the data.

The implementation of NA-NVR-BIP is realized in ILOG OPL whereas NA-NVR-TS and NA-NVR-ACO are implemented in Java programming language environment. We tested all three algorithms on an Intel(R) Core (TM)2 Quad CPU Q8200 machine with 8 2.33GHz processors each with total of 8GB RAM memory. We performed 10 runs of NA-NVR-TS and NA-NVR-ACO for each protein and the best assignment accuracy obtained with the lowest score is presented in the section 5.2.1 and 5.2.2. The average accuracy results are provided in the Appendix A.

5.2.1 Distinguishing the type of NOE

The assignment accuracies obtained with the former approaches as well as with the proposed new approaches are provided in tables below. These tables contain the best results obtained from the 10 runs with the proposed approaches having the lowest total assignment scores for each protein. The assignment accuracy is defined as the ratio of the number of correctly assigned peaks to the total number of assigned peaks. In previous work [18] the results without and with RDCs have been provided. RDC is a type of NMR experiment which NVR can use if it is available. We provide the results which are obtained by proton type distinction in NOEs and compare it with the results of NVR-BIP [18] in Table 5.1 through Table 5.4. In these tables, the column named

“NVR-BIP” labels the assignment accuracies obtained by NVR-BIP. In NVR-BIP, the proton type was not distinguished. HN-HA and HA-HN NOEs were considered as HN-HN NOEs and HN-HN proton coordinates were used in calculations. The columns with the names of “NA-NVR-BIP”, “NA-NVR-TS”, and “NA-NVR-ACO” refer to the assignment accuracies obtained by distinguishing the type of NOE. The threshold value over the distances among amide protons is manually tuned for each protein. This means, for each tested protein, a value is selected as threshold that is greater than the distances between amide protons assigned to pair of peaks that have NOE relations between them. For example, it is chosen as 7 Å for 1AAR protein by analyzing the distances between protons assigned to peak pairs with NOE relations.

Table 5.1: Assignment accuracies for ubiquitin when distinguishing NOE type

PDB ID	No of Residues	Accuracy							
		NVR-BIP		NA-NVR-BIP		NA-NVR-TS		NA-NVR-ACO	
		Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
1AAR	72	79%	97%	91%	100%	91%	100%	91%	100%
1G6J		87%	97%	100%	100%	100%	100%	100%	100%
1UBI		87%	97%	100%	100%	100%	100%	100%	100%
1UBQ		87%	97%	100%	100%	100%	100%	100%	100%
1UD7		81%	97%	97%	97%	97%	97%	97%	97%

According to the results in Table 5.1, the assignment accuracies are improved in NA-NVR-BIP, NA-NVR-TS and NA-NVR-ACO for all proteins except 1UD7 with RDC case. The NA-NVR-TS and NA-NVR-ACO achieved optimal solutions for all tested proteins, and performed equally in ubiquitin protein test.

Table 5.2: Assignment accuracies for SPG when distinguishing NOE type

PDB ID	No of Residues	Accuracy							
		NVR-BIP		NA-NVR-BIP		NA-NVR-TS		NA-NVR-ACO	
		Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
1GB1	55	100%	100%	100%	100%	100%	100%	100%	100%
1PGB		100%	100%	100%	100%	100%	100%	100%	100%
2GB1		96%	100%	100%	100%	100%	100%	100%	100%

For SPG protein, NA-NVR-BIP, NA-NVR-TS and NA-NVR-ACO provided better accuracies than NVR-BIP. Furthermore, the new approaches attained 100%

accuracy for 2GB1 without RDC. Both NA-NVR-TS and NA-NVR-ACO obtained optimal solutions and demonstrated same accuracies in SPG protein test.

Table 5.3: Assignment accuracies for lysozyme when distinguishing NOE type

PDB ID	No of Residues	Accuracy							
		NVR-BIP		NA-NVR-BIP		NA-NVR-TS		NA-NVR-ACO	
		Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
193L	126	78%	100%	92%	100%	92%	100%	92%	100%
1AKI		78%	98%	83%	100%	83%	100%	83%	100%
1AZF		74%	94%	90%	100%	90%	100%	90%	100%
1BGI		75%	97%	90%	100%	90%	100%	90%	100%
1H87		77%	100%	92%	100%	92%	100%	92%	100%
1LSC		74%	100%	90%	100%	90%	100%	90%	100%
1LSE		75%	98%	94%	100%	94%	100%	94%	100%
1LYZ		79%	82%	94%	100%	94%	100%	94%	100%
2LYZ		75%	91%	92%	100%	92%	100%	92%	100%
3LYZ		79%	90%	94%	100%	94%	100%	94%	100%
4LYZ		75%	91%	94%	98%	94%	98%	94%	98%
5LYZ		75%	91%	94%	98%	94%	98%	94%	98%
6LYZ	75%	96%	92%	100%	92%	100%	92%	100%	

The results in Table 5.3 indicate that the assignment accuracies are higher in NA-NVR-BIP, NA-NVR-TS, and NA-NVR-ACO compared to NVR-BIP. The optimal solutions are obtained by NA-NVR-TS and NA-NVR-ACO for all proteins. The performance of the NA-NVR-TS and NA-NVR-ACO is the same for lysozyme protein test.

Table 5.4: Assignment accuracies for other proteins when distinguishing NOE type

PDB ID	No of Residues	Accuracy							
		NVR-BIP		NA-NVR-BIP		NA-NVR-TS		NA-NVR-ACO	
		Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
pol η	31	100%	100%	100%	94%	100%	94%	100%	94%
GB1	55	96%	100%	100%	100%	100%	100%	100%	100%
ff2	80	85%	93%	87%	93%	87%	93%	87%	93%
hSRI	96	73%	89%	79%	94%	79%	94%	79%	94%

In Table 5.4, we also observe that the assignment accuracies are higher in the new approaches. Both the NA-NVR-TS and NA-NVR-ACO achieved identical assignment accuracies and the optimal solutions.

The results in the tables clearly show that distinguishing the proton types in NOE relations improves the backbone resonance assignment accuracies in all proteins. Note that both the NVR-BIP and NA-NVR-BIP return the optimal solutions. In all tested proteins the NA-NVR-TS and the NA-NVR-ACO achieved the optimal solutions since the assignment accuracy and total score of their solutions are same as NA-NVR-BIP. Thus, this will guarantee the robustness of the NA-NVR-TS and NA-NVR-ACO in testing new proteins.

NVR-BIP could not find a solution in a reasonable time for large proteins EIN and MBP because of the exponential time complexity of the problem. For this reason, we compare NVR-TS [20] with NA-NVR-TS and NVR-ACO [22] with NA-NVR-ACO in Table 5.5. The columns named “NVR-TS” and “NVR-ACO” label the assignment accuracies obtained by NVR-TS [20] and NVR-ACO [22], respectively. In NVR-TS and NVR-ACO, the proton type is not distinguished. HN-HA and HA-HN NOEs were considered as HN-HN NOEs and HN-HN proton coordinates were used in calculations. The columns with the names “NA-NVR-TS” and “NA-NVR-ACO” demonstrate the assignment accuracies which are acquired by proton type differentiation in NOEs.

Table 5.5: Assignment accuracies for large proteins when distinguishing NOE type

		Accuracy							
		NVR-TS		NVR-ACO		NA-NVR-TS		NA-NVR-ACO	
PDB ID	No of Residues	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
EIN	243	24%	83%	67%	100%	93%	100%	93%	100%
MBP	348	49%	63%	49%	73%	65%	74%	64%	66%

In Table 5.5, the assignment accuracies are improved in NA-NVR-TS and NA-NVR-ACO compared to NVR-TS and NVR-ACO for EIN. In particular, the assignment accuracy for the case without RDC is increased from 24% to 93% in NA-NVR-TS and from 67% to 93% in NA-NVR-ACO. NA-NVR-TS and NA-NVR-ACO demonstrated equal performance in assignment accuracies for EIN protein.

All real NOE relations for MBP are HN-HA type. In NVR-TS and NVR-ACO, these NOEs are considered as HN-HN NOEs and HN-HN proton coordinates are used. The distinction in NOE type and correct proton coordinate usage are strong requirements to automate the assignment process. It is the first phase within our approach and realized in NA-NVR-TS and NA-NVR-ACO. The falls in assignment accuracies for some proteins are tolerated to automate the process in this phase. This is the case in NA-NVR-ACO for MBP with RDC compared to NVR-ACO.

NA-NVR-TS enhanced the assignment accuracies compared to NVR-TS for both with and without RDC in MBP. However, both NVR-TS and NA-NVR-TS failed to obtain a solution without NOE violations. On the other hand, NA-NVR-ACO obtained a solution without NOE violation for MBP.

5.2.2 Using the NOE upper bounds extracted from the data

An automatic threshold value determination is performed in this approach. The NOE upper bounds gathered from the input data are utilized as the threshold value. Meanwhile, the type of protons in NOEs are distinguished and correct proton types are used. The NOE upper bound information is directly taken from the input data for all proteins except MBP. For the MBP, the intensity values for NOE relations between peak pairs are converted to the upper bound distance information.

The NA-NVR-BIP model that uses NOE upper bound distance information is solved in ILOG OPL environment employing CPLEX solver engine. The assignment problem was infeasible for some protein data sets. This infeasibility was originated from the NOE constraints. When the NOE upper bounds are used as threshold over the distances between protons that are assigned to peak pairs, distance violations may arise even for the correct assignment. In other words, the NOE upper bound extracted from the data could be smaller than the exact distance between the protons assigned to the corresponding peak pairs that have NOE between them. In this case, an NOE violation occurs which prevents us to find a feasible assignment scheme.

While NA-NVR-BIP cannot find any solutions due to NOE violations, NA-NVR-TS and NA-NVR-ACO allow the NOE violations during the search and can provide assignments. In these approaches, the NOE violations are penalized during the search process in an attempt to construct a solution without NOE violation in the end. The

higher number of distance violation may cause even lower assignment accuracies. For lysozyme family, there are between 29 and 73 distance violations and total of 534 NOE constraints. There are between 1 and 3 distance violations and 270 NOE constraints present in ubiquitin family. The SPG family has between 9 and 11 distance violations and 204 NOE constraints. For the large proteins, 126 distance violation and 1021 NOE constraints exist for EIN and 1 distance violation and 474 NOE constraints exist for MBP. For the rest of the proteins, 6, 2, 5 distance violations and 260, 234, 156 NOE constraints are present for GB1, ff2 and pol η , respectively.

We compare the results obtained by NVR-BIP [18] with NA-NVR-TS and NA-NVR-ACO in Tables 5.6 - 5.9. The column “NVR-BIP” reports the assignment accuracies obtained by [18]. The columns “NA-NVR-TS” and “NA-NVR-ACO” show the assignment accuracies obtained by using the NOE upper bounds extracted from the data. In this section, the tables contain the best results obtained with the lowest total score out of the 10 runs for each protein for the proposed approaches.

Table 5.6: Assignment accuracies for ubiquitin when using the NOE upper bounds

PDB ID	No of Residues	Accuracy					
		NVR-BIP		NA-NVR-TS		NA-NVR-ACO	
		Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
1AAR	72	79%	97%	100%	100%	100%	100%
1G6J		87%	97%	97%	97%	97%	97%
1UBI		87%	97%	100%	100%	100%	100%
1UBQ		87%	97%	100%	100%	100%	100%
1UD7		81%	97%	97%	97%	97%	100%

The results clearly show that the assignment accuracies are improved in NA-NVR-TS and NA-NVR-ACO for all proteins except 1G6J and 1UD7 with RDC case. The NA-NVR-TS and NA-NVR-ACO show a similar performance in ubiquitin protein test.

Table 5.7: Assignment accuracies for SPG when using the NOE upper bounds

PDB ID	No of Residues	Accuracy					
		NVR-BIP		NA-NVR-TS		NA-NVR-ACO	
		Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
1GB1	55	100%	100%	100%	100%	100%	100%
1PGB		100%	100%	100%	100%	100%	100%
2GB1		96%	100%	100%	100%	100%	100%

For SPG protein tests, the NA-NVR-TS and NA-NVR-ACO obtained the same accuracies.

Table 5.8: Assignment accuracies for lysozyme when using the NOE upper bounds

PDB ID	No of Residues	Accuracy					
		NVR-BIP		NA-NVR-TS		NA-NVR-ACO	
		Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
193L	126	78%	100%	100%	100%	100%	100%
1AKI		78%	98%	100%	100%	100%	98%
1AZF		74%	94%	100%	100%	100%	100%
1BGI		75%	97%	100%	100%	100%	100%
1H87		77%	100%	100%	100%	100%	100%
1LSC		74%	100%	100%	100%	100%	100%
1LSE		75%	98%	100%	100%	98%	96%
1LYZ		79%	82%	100%	100%	85%	85%
2LYZ		75%	91%	100%	100%	98%	98%
3LYZ		79%	90%	100%	100%	98%	98%
4LYZ		75%	91%	100%	100%	97%	95%
5LYZ		75%	91%	100%	100%	97%	95%
6LYZ		75%	96%	100%	100%	100%	100%

According to the results in the Table 5.8, the new approach improved the assignment accuracies. NA-NVR-TS demonstrated better a performance on lysozyme protein test since it obtained higher accuracies for numerous tests compared to NA-NVR-ACO.

Table 5.9: Assignment accuracies for other proteins when using the NOE upper bounds

PDB ID	No of Residues	Accuracy					
		NVR-BIP		NA-NVR-TS		NA-NVR-ACO	
		Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
pol η	31	100%	94%	100%	100%	94%	94%
GB1	55	96%	100%	100%	100%	100%	100%
ff2	80	85%	93%	65%	93%	75%	89%
hSRI	96	73%	89%	79%	94%	79%	94%

In Table 5.9 the assignment accuracies are higher for all proteins in new approach compared to NVR-BIP except ff2. This decrease is tolerable since the NOE upper bound distance parameters are automatically obtained in the new approach.

Since NVR-BIP could not find a solution for large proteins EIN and MBP, we compare the NVR-TS with NA-NVR-TS and NVR-ACO with NA-NVR-ACO in Table 5.10. The columns named “NVR-TS” and “NVR-ACO” show the assignment accuracies obtained by NVR-TS [20] and NVR-ACO [22], respectively. In [20] and [22], the proton type is not distinguished. HN-HA and HA-HN NOEs were considered as HN-HN NOEs and HN-HN proton coordinates were used in calculations. The columns with names “NA-NVR-TS” and “NA-NVR-ACO” refer the assignment accuracies which are acquired by using the NOE upper bounds extracted from the data.

Table 5.10: Assignment accuracies for large proteins when using the NOE upper bounds

PDB ID	No of Residues	Accuracy							
		NVR-TS		NVR-ACO		NA-NVR-TS		NA-NVR-ACO	
		Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
EIN	243	24%	83%	67%	100%	100%	100%	90%	88%
MBP	348	49%	63%	49%	73%	67%	76%	67%	80%

By observing the Table 5.10, it easy to monitor that result improves in NA- NVR-TS and NA-NVR-ACO for both proteins. Nevertheless, NA-NVR-ACO failed to exceed NVR-ACO for EIN with RDC. This due to the large number of NOE distance

violations (126). The NA-NVR-TS outperformed NA-NVR-ACO according to results for large protein test.

It can be seen that the new approach which distinguishes the proton type and incorporates both HN and HA coordinates and corresponding distances into computations, and determines the threshold values in a standard manner improved the backbone resonance assignment accuracies in all proteins. In most tested proteins, the assignment accuracies with and without RDCs are higher with the new approach. Note that the best assignment accuracy is equal to the average accuracy for almost all tested proteins. This emphasizes the robustness and stability of NA-NVR-TS and NA-NVR-ACO algorithms. Besides, the increases in assignment accuracies are not the only contribution. In addition, we automate the usage of NOE data by means of new approaches. The NVR suite of programs no longer need hand coded parameters for handling the NOE data. This makes the approach more reliable and gives way to more realistic solutions on novel proteins.

Chapter 6

CONCLUSION AND FUTURE WORK

In the previous studies [18, 20, 22], NOE type was not differentiated and the threshold values on NOE relations were manually set for all tested proteins. This approach brings some drawbacks such as the lower assignment accuracies and restricts the application range of methods on novel proteins. In this thesis, we reformulated NVR-BIP and we adapted NVR-TS and NVR-ACO in order to distinguish the type of backbone NOEs and set the threshold values in a standard manner. We made these modifications by reformulating NVR-BIP [18] in two new mathematical models. In Model 1, we distinguished the proton types in NOEs and incorporated the correct proton coordinates into the computations. On top of proton type distinction, we utilized the NOE upper bound distance limits as a threshold values in Model 2. We tested the new approaches on 7 small proteins and two large proteins, namely E1N and MBP. The NOE upper bound distance limits are gathered from the data for all proteins while it is automatically extracted from the NOE peak intensity values for MBP by using simple protocol.

Our results show that the incorporation of HN and HA proton coordinates and using NOE relation upper bounds as a threshold value in both models improved the assignment accuracy compared to the previous approach. In particular, we achieved 100% assignment accuracy with the NA-NVR-TS on the large protein E1N by distinguishing the type of NOE. However, NA-NVR-TS which takes the distinguished NOE input data did not find any feasible solution on MBP real data. The NA-NVR-ACO that was adapted for distinguished NOE input data gave a feasible solution for MBP with 67% and 80% assignment accuracies for without RDC and with RDC, respectively.

According to the outcomes of the two models, the new approaches significantly improved the solutions compared to the NVR-BIP. Both models had similar performance in the experiments. However, Model 2 is more reliable and realistic due to the NOE upper bound usage as a threshold value. In addition, the Model 2 has a wider application scope. Furthermore, NA-NVR-TS displayed more accurate performance compared to NA-NVR-ACO. This superiority is more visible in the results obtained by using the NOE upper bounds extracted from the data model.

A similar structure such as the one obtained by x-ray crystallography is used as template structure in these tests. It would be interesting to study the effect of using more distant templates. All NMR peaks are assigned in entire tests. The partial NMR peak assignment could be modeled and implemented as a future work in order to assign the NMR data of a protein to the more distant template structure. Since the NMR data will be assigned to more distant template structure there will be a considerable number of NOE violations and this would be a good test for partial NMR peak assignment formulation.

For MBP test, NA-NVR-TS could not succeed to obtain the solution without NOE violations. We directly used the parameter settings of previous studies in the experimental study and this could be the main reason. Further studies can focus on fine tuning the parameters of NA-NVR-TS and NA-NVR-ACO to further improve the quality of the results.

Further studies can focus on improving the TS algorithm neighborhood search where multiple neighborhood search structure could be used. Moreover, other metaheuristic algorithms may be employed to compare the performance of NA-NVR-TS and NA-NVR-ACO algorithms.

Finally, the output of our NA-NVR-TS and NA-NVR-ACO could be tested in HADDOCK [28, 29] NMR protein docking software to make further analysis of 3D structure of proteins and protein-ligand binding affinity.

BIBLIOGRAPHY

- [1] M.G. Rossmann and D.M. Blow, “The Detection of Sub-Units within the Crystallographic Asymmetric Unit,” *Acta Crystallographica*, vol. 15, no. 1, pp. 24–31, Jan. 1962.
- [2] Y.S. Jung & M. Zweckstetter, “Mars – robust automatic backbone assignment of proteins”, *Journal of Biomolecular NMR*,30: 11–23, 2004.
- [3] S. Neal, A.M. Nip, H. Zhang, D.S. Wishart, “Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts”, *Journal of Biomolecular NMR*, 26: 215–240, 2003.
- [4] X. P. Xu and D.A. Case “Automated prediction of 15N, 13C α , 13C β and 13C γ chemical shifts in proteins using a density functional database” , *Journal of Biomolecular NMR*, 21:321–333, 2001.
- [5] C.J. Langmead, A. Yan, R. Lilien, L. Wang ,B.R. Donald, “A Polynomial-Time Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments”, RECOMB’03, April 10–13, 2003
- [6] K.R. Pervushin, G.Wider Riek, and K. Wüthrich, “Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution”, *Proc. Nat. Acad. Sci*, USA 94:12366–12371, 1997
- [7] Y.S. Jung & M. Zweckstetter, “Mars – robust automatic backbone assignment of proteins Backbone assignment of proteins with known structure using residual dipolar couplings”, *Journal of Biomolecular NMR*,30: 25–35, 2004.
- [8] C.J. Langmead and B.R. Donald, “High-Throughput 3D Structural Homology Detection via NMR Resonance Assignment”, Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference
- [9] H.M. Al-Hashimi, A. Gorin, A. Majumdar, Y. Gosser, and D.J. Patel, “Towards Structural Genomics of RNA: Rapid NMR Resonance Assignment and

- Simultaneous RNA Tertiary Structure Determination Using Residual Dipolar Couplings,” *J. Molecular Biology*, vol. 318, no. 3, pp. 637-649, 2002.
- [10] T. J. Dubose and R. Ludwig, “Crystallography and the Gene Helix Discovery : An Interdisciplinary Educational Project”, *Journal of Diagnostic Medical Sonography*, vol. 19, no. 6, pp. 347-357, 2003
- [11] J. Hus, J. Prompers, and R. Bruschweiler, “Assignment Strategy for Proteins of Known Structure,” *J. Magnetic Resonance*, vol. 157, no. 1, pp. 119-125, 2002.
- [12] C.J. Langmead and B.R. Donald, “An Expectation/Maximization Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments,” *J. Biomolecular NMR*, vol. 29, no. 2, pp. 111-138, June 2004.
- [13] G. Kochenberger and F. Glover, “A Unified Framework for Modeling and Solving Combinatorial Optimization Problems: A Tutorial,” *Multiscale Optimization Methods and Applications*, pp. 101-124, Springer, 2006.
- [14] S. de Vries, A. van Dijk, M. Krzeminski, M. van Dijk, A. Thureau, V. Hsu, T. Wassenaar, and A. Bonvin, “HADDOCK versus HADDOCK: New Features and Performance of HADDOCK2.0 on the CAPRI Targets,” *Proteins*, vol. 69, no. 4, pp. 726-733, 2007.
- [15] M.S. Apaydin, V. Conitzer, and B.R. Donald, “Structure-Based Protein NMR Assignments Using Native Structural Ensembles,” *J. Biomolecular NMR*, vol. 40, no. 4, pp. 263-276, 2008.
- [16] D. Stratmann, E. Guittet, C. Heijenoort, “Robust structure-based resonance assignment for functional protein studies by NMR,” *J. Biomolecular NMR*, vol. 46, no. 2, pp. 157–173, 2010.
- [17] Jason M. Winget and Thibault Mayor, “The Diversity of Ubiquitin Recognition: Hot Spots and Varied Specificity”, *Molecular Cell*, vol. 38, no. 5, pp. 627 - 635, 2010
- [18] M.S. Apaydin, 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.*, vol. 54, no. 5, pp. 708-716, 2011.
- [19] R. Jang, X. Gao, and M. Li, “Towards Fully Automated Structure- Based NMR Resonance Assignment of ¹⁵N-Labeled Proteins from Automatically Picked Peaks,” *J. Computational Biology*, vol. 18, no. 3, pp. 347-363, 2011.

- [20] G. Cavuslar, B. Catay , and M.S. Apaydın, “A Tabu Search Approach for the NMR Protein Structure-Based Assignment Problem,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 6, pp. 1621-1628, 2012
- [21] M. Akhmedov, B. Çatay, M.S. Apaydın, “Distinguishing the type of NOE for NMR Protein Structure-Based Assignments” *SIU 2013*, April 24-26, 2013.
- [22] J. Aslanov, B. Çatay, M. S. Apaydın, “An Ant Colony Optimization Based Approach for Solving the Nuclear Magnetic Resonance Structure Based Assignment Problem”, *GECCO 2013*, July 06-10, 2013.
- [23] J. Zeng, C. Tripathy, P. Zhou, B.R. Donald, “A hausdorff-based NOE assignment algorithm using protein backbone determined from residual dipolar couplings and rotamer patterns”, *Comput Syst Bioinformatics Conf. 2008* ; 2008: 169–181.
- [24] G.M. Clore and A.M. Gronenborn, “Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magnetic resonance spectroscopy”. *Crit. Rev. Biochem. Mol. Biol*, 24:479–564, 1989.
- [25] G.M. Clore and A.M. Gronenborn, “Applications of three- and four-dimensional heteronuclear NMR spectroscopy to protein structure determination”, *Progr. Nucl. Magn. Reson. Spectroscopy*, 23:43–92, 1991.
- [26] G.M. Clore and A.M. Gronenborn, “Two, three and four dimensional NMR methods for obtaining larger and more precise three-dimensional structures of proteins in solution”, *Ann. Rev. Biophys. Biophys. Chem*, 20:29–63, 1991.
- [27] L.J. McGuffin, K. Bryson, and D.T. Jones, “The PSIPRED protein structure prediction server”, *Bioinformatics*, 16, 404–405, 2000.
- [28] S.J. de Vries, A.D.J. van Dijk, M. Krzeminski, M. van Dijk, A. Thureau, V. Hsu, T. Wassenaar and A.M.J.J. Bonvin, "HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets", *Proteins: Struct. Funct. & Bioinformatic* **69**, 726-733, 2007.
- [29] C. Dominguez, R. Boelens and A.M.J.J. Bonvin, “HADDOCK: a protein-protein docking approach based on biochemical and/or biophysical information”, *J. Am. Chem. Soc.*, 125, 1731-1737, 2003.

APPENDIX A

Table A.1: Average accuracies of 10 runs for ubiquitin

		Accuracy											
		NVR-BIP		NA-NVR-BIP		Distinguishing the type of NOE NVR-TS		Distinguishing the type of NOE NVR-ACO		Using the NOE upper bounds extracted from the data NVR-TS		Using the NOE upper bounds extracted from the data NVR-ACO	
PDB ID	No of Res.	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
1AAR	72	79%	97%	100%	100%	91%	100%	91%	100%	100%	100%	100%	100%
1G6J		87%	97%	100%	100%	100%	100%	100%	100%	97%	97%	97%	97%
1UBI		87%	97%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
1UBQ		87%	97%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
1UD7		81%	97%	97%	97%	97%	97%	97%	97%	97%	97%	97%	100%

Table A.2: Average accuracies of 10 runs for lysozyme

		Accuracy												
		NVR-BIP		NA-NVR-BIP		Distinguishing the type of NOE NVR-TS		Distinguishing the type of NOE NVR-ACO		Using the NOE upper bounds extracted from the data NVR-TS		Using the NOE upper bounds extracted from the data NVR-ACO		
PDB ID	No of Res.	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	
193L	126	78%	100%	92%	100%	92%	100%	92%	100%	100%	100%	100%	100%	
1AKI		78%	98%	83%	100%	83%	100%	83%	100%	100%	100%	100%	98%	
1AZF		74%	94%	90%	100%	90%	100%	90%	100%	100%	100%	100%	100%	
1BGI		75%	97%	90%	100%	90%	100%	90%	100%	100%	100%	100%	100%	
1H87		77%	100%	92%	100%	92%	100%	92%	100%	100%	100%	100%	100%	
1LSC		74%	100%	90%	100%	90%	100%	90%	100%	100%	100%	100%	100%	
1LSE		75%	98%	94%	100%	94%	100%	94%	100%	100%	100%	100%	98%	96%
1LYZ		79%	82%	94%	100%	94%	100%	94%	100%	100%	100%	100%	85%	85%
2LYZ		75%	91%	92%	100%	92%	100%	92%	100%	100%	100%	100%	98%	98%
3LYZ		79%	90%	94%	100%	94%	100%	94%	100%	100%	100%	100%	98%	98%
4LYZ		75%	91%	94%	100%	94%	98%	94%	98%	100%	100%	100%	97%	95%
5LYZ		75%	91%	94%	100%	94%	98%	94%	98%	100%	100%	100%	97%	95%
6LYZ		75%	96%	92%	100%	92%	100%	92%	100%	100%	100%	100%	100%	100%

Table A.3: Average accuracies of 10 runs for SPG

		Accuracy											
		NVR-BIP		NA-NVR-BIP		Distinguishing the type of NOE NVR-TS		Distinguishing the type of NOE NVR-ACO		Using the NOE upper bounds extracted from the data NVR-TS		Using the NOE upper bounds extracted from the data NVR-ACO	
PDB ID	No of Res.	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
1GB1	55	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
1PGB		100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
2GB1		96%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table A.4: Average accuracies of 10 runs for other proteins

		Accuracy											
		NVR-BIP		NA-NVR-BIP		Distinguishing the type of NOE NVR-TS		Distinguishing the type of NOE NVR-ACO		Using the NOE upper bounds extracted from the data NVR-TS		Using the NOE upper bounds extracted from the data NVR-ACO	
PDB ID	No of Res.	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
pol η	31	100%	100%	100%	94%	100%	94%	100%	94%	100%	100%	94%	94%
GB1	55	96%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
ff2	80	85%	93%	87%	93%	85%	93%	87%	93%	65%	93%	75%	89%
hSRI	96	73%	89%	79%	94%	76%	92%	77%	94%	79%	94%	79%	94%

Table A.5: Average accuracies of 10 runs for large proteins

		Accuracy											
		NVR-TS		NVR-ACO		Distinguishing the type of NOE NVR-TS		Distinguishing the type of NOE NVR-ACO		Using the NOE upper bounds extracted from the data NVR-TS		Using the NOE upper bounds extracted from the data NVR-ACO	
PDB ID	No of Res.	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC	Without RDC	With RDC
EIN	243	5%	83%	67%	100%	93%	100%	93%	100%	100%	100%	88%	88%
MBP	348	49%	63%	49%	73%	55%	68%	64%	66%	60%	72%	64%	77%

APPENDIX B

Algorithm 1 Tabu Search Algorithm

Initialization: Obtain an initial solution s , $s' \leftarrow s$, $s'' \leftarrow s$, $s^* \leftarrow s$, $ctr' \leftarrow 0$, $ctr'' \leftarrow 0$, $ctr^* \leftarrow 0$

```
while  $ctr^* < Iter_3$  do
  while  $ctr'' < Iter_2$  do
    while  $ctr' < Iter_1$  do
      Move from  $s$  to  $s_l$ 
      if  $score(s_l) < score(s')$  then
        Update tabu list;  $s' \leftarrow s_l$ 
         $ctr' \leftarrow 0$ 
      else
         $ctr' \leftarrow ctr' + 1$ 
      end if
    end while
    Perturb( $s'$ )
    if  $score(s') < score(s'')$  then
       $s'' \leftarrow s'$ 
       $ctr'' \leftarrow 0$ 
    else
       $ctr'' \leftarrow ctr'' + 1$ 
    end if
  end while
  if  $score(s'') < score(s^*)$  then
     $s^* \leftarrow s''$ 
     $ctr^* \leftarrow 0$ 
  else
     $ctr^* \leftarrow ctr^* + 1$ 
  end if
   $s \leftarrow$  Perturb( $s''$ )
end while
Return  $s^*$ 
```

Algorithm 2 Ant Colony Optimization

```
initialize pheromone trails
while (stopping condition not satisfied) do
  for all ant do
    for  $i = 1 \rightarrow |P|$  do
      select a peak (using constrained peak selection)
      assign amino acid (using random selection rule)
    end for
  end for
  elitist (2-opt) local search
  elitist pheromone update
end while
```
