Incorporating HADAMAC experiment into NMR Structure-Based Assignments

Halit Erdoğan
Sabanci University, Istanbul-Turkey
halit@sabanciuniv.edu

Mehmet Serkan Apaydın
Sabanci University, Istanbul-Turkey
Istanbul Sehir University, Istanbul-Turkey
apaydin@sehir.edu.tr

Abstract

Protein structure determination is crucial to understand a protein's function and to develop drugs against diseases. Nuclear Magnetic Resonance (NMR[1]) spectroscopy is an experimental technique that allows one to study protein structure in solution. In NMR Structure-based assignment problem, the aim is to assign experimentally observed peaks to the specific nuclei of the target molecule by using a template protein and it is an important computational challenge. NVR-BIP is a tool that utilizes a scoring function based on NVR's [5, 6] framework and computes assignments for given NMR data. In this paper, we incorporate HADAMAC experiment—which helps predict an amino acid class for each peak— with NVR-BIP's scoring function. Experiments show that the new scoring function results in higher assignment accuracies compared to the previous approaches.

1 Introduction

Proteins are one of the major macromolecules that are present in all biological organisms. They serve as enzymes, are used as storage molecules, are needed for the immune system and have many other functions in the cell. Therefore, determining the functions of proteins is crucial to understand important biological processes and to develop drugs against diseases.

[1]Abbreviations used: NMR, Nuclear Magnetic Resonance; CS, Chemical Shift; RDC, Residual Dipolar Coupling; NOE, Nuclear Overhauser Effect; SBA, Structure-Based Assignment; NVR, Nuclear Vector Replacement; BIP, Binary Integer Programming.

The function of a protein depends on its 3-D structure. There are two main experimental methods to determine the protein structure. These are X-ray crystallography and Nuclear Magnetic Resonance (NMR) Spectroscopy. About 85% of the protein structures in the Protein Data Bank were determined using X-ray Crystallography, on the other hand approximately 15% were solved using NMR. NMR allows one to study protein structure in solution. In addition, not all proteins can be crystallized. Therefore, NMR spectroscopy is an important experimental technique for protein structure determination.

In NMR, several experiments are performed on the protein and the signals are recorded. After processing these signals, the experiments result in various NMR spectra. The initial stage is to pick the peaks in the NMR spectrum and this stage is largely automated. The second stage is to find the mapping between the peaks and the atoms. This is called the assignment problem and is an important computational challenge. An existing structure (the "template") can be used to help assign a target protein. This is called Structure-Based Assignment (SBA). SBA is analogous to molecular replacement in X-ray Crystallography [8]. In NMR SBA, the data coming from NMR spectroscopy and the template protein are analyzed. The available programs use a scoring function that maps each (peak, amino acid) pair to a real number that corresponds to the likelihood of the corresponding assignment. Then various methods (such as Monte Carlo Simulation, memetic algorithm or integer programming) are employed to find the assignments corresponding to the optimum or near-optimum of this scoring function (see, e.g., MONTE [4], MATCH [9], NVR-BIP [1]).

In [1], the authors developed a tool called NVR-BIP which can be used to solve the SBA problem. NVR-BIP uses the Nuclear Vector Replacement (NVR) framework [5, 6], with additional sources of data, to determine the scoring function, and binary integer programming (BIP) to find the assignment. In NVR-BIP, the assignment problem is formulated as an integer linear model with additional Nuclear Overhauser Effect (NOE) constraints. In [1], the authors present their results on several proteins.

The accuracy of NVR-BIP is highly related to the quality of the scoring function. Therefore, improving the scoring function will improve the assignment accuracies. This can be achieved by incorporating additional experimental data into NVR. For instance, additional chemical shifts obtained from triple resonance experiments can be added to NVR's data types. These chemical shifts could then be used with amino acid typing to help determine the type of the amino acids or reduce the possibilities, therefore act as a filter. With this motivation, in [3], we integrated {N,HN,HA,HB,CA,CB,CO} chemical shifts and used CRAACK [2], an amino acid typing software to modify NVR's scoring function and obtained improved assignment accuracies.

HADAMAC [7] experiment uses Hadamard encoded amino acid type editing scheme. In Hadamard encoded type editing, the twenty amino acids are grouped into seven classes. The main contributions of this work are as follows:

1. We simulate the HADAMAC experiment to predict the amino acid class that each NMR peak belongs to;
2. We incorporate the HADAMAC experiment into NVR-BIP's scoring function; and
3. We test our approach on NVR-BIP's dataset and compare our results with the previous works [1] and [3].

The rest of the paper is organized as follows: In Section 2, we review the previous approaches, NVR-BIP and integration of CRAACK with NVR-BIP. In Section 3, we review the HADAMAC experiment and integration of HADAMAC experiment with NVR-BIP. Data preparation is in Section 4 and the experimental comparison of the scoring functions: NVR, NVR+CRAACK and NVR+HADAMAC is in Section 5. We conclude and discuss future work in Section 6.

2 Previous Work

NVR-BIP uses the Nuclear Vector Replacement (NVR) framework [6, 5], and incorporates additional sources of data, to determine the assignments. NVR-BIP uses NVR's scoring function which provides a score for assigning each peak to each amino acid. These scores are derived from the difference between the predicted and experimental data values; if the difference is too high then the assigned score is $+\infty$. NVR-BIP formulates the problem as a binary integer program where the objective is to find the assignment whose total score is minimum subject to the NOE constraints. NVR-BIP uses a BIP solver to find the minimum scoring assignment. In [1], NVR-BIP is tested on 7 proteins with 25 templates and results in better accuracies than the previous approaches.

In [3], we incorporated additional chemical shifts into NVR and used amino acid typing software CRAACK to improve NVR's scoring function. CRAACK predicts a set containing amino acid names with corresponding probabilities for each peak. This set contains the predictions that CRAACK makes based on the chemical shifts of the peak. The chemical shifts contain {HA,HB,CA,CB,CO} atoms in addition to the amide proton and nitrogen chemical shifts that NVR uses, these additional chemical shifts are assumed to be obtained with triple resonance experiments. CRAACK's output is used to modify the score values for peak-amino acid pairs. The resulting scoring function of NVR+CRAACK computes more accurate assignments compared to NVR-BIP.

3 NVR+HADAMAC Scoring Function

HADAMAC [7] experiment uses Hadamard encoded amino acid type editing scheme. In Hadamard encoded type editting, first, the twenty amino acids are grouped into seven classes. The different classes correspond to Gly (1), Val, Ile (2), Ala (3), Thr (4), Asn, Asp (5), Phe, Tyr, Trp, His, Cys, Ser (6), and Arg, Glu, Lys, Pro, Gln, Met and Leu (7) side chains. Then each peak is assigned to one of these

seven classes which represents the type of the *previous* residue of the residue corresponding to the peak.

We simulate the HADAMAC experiment. We assign each peak i to one of the seven classes according to the type of the residue $j-1$, where j is the residue that is to be assigned to peak i. We use $H(i)$ to represent the set that contains the amino acid types corresponding to peak i according to the HADAMAC experiment, and we use $type_j$ to represent the type of the residue j.

Given the NVR scoring function $S_n(i,j)$ which is defined for each peak-residue pair, we compute the new scoring function, S_{nh}, using the HADAMAC experiment as follows:

$$S_{nh}(i,j) = \begin{cases} S_n(i,j) & \text{if } type_{j-1} \in H(i) \\ \infty & \text{otherwise} \end{cases}$$

This new scoring function is similar to NVR's scoring function where some of the peak-residue assignments are pruned.

4 Data Preparation

We tested our approach on the data set of NVR-BIP so as to compare the results of both approaches. The proteins we have tested our approach on are: ubiquitin (template pdb ids: 1UBI, 1UBQ, 1G6J, 1UD7, 1AAR), streptococcal protein G (template pdb ids: 1GB1, 2GB1, 1PGB), lysozyme (template pdb ids: 193L, 1AKI, 1AZF, 1BGI, 1H87, 1LSC, 1LSE, 2LYZ, 3LYZ, 4LYZ, 5LYZ, 6LYZ), human Set2-Rpb1 interacting domain (hSRI, template pdb id: 2A7O), the FF Domain 2 of human transcription elongation factor CA150 (RNA polymerase II C-terminal domain interacting protein) (ff2, template pdb id: 2E71), Y-polymerase Eta (pol η, template pdb id: 2I5O), B1 domain of streptococcal protein G (GB1, template pdb id: 3GB1). Note that in NMR community experimental results on multiple proteins is considered adequate [10]. More details on these proteins can be found in [1].

5 Experimental Results

We performed experiments to compare the results of the scoring functions NVR, NVR+CRAACK and NVR+HADAMAC on the dataset of NVR-BIP mentioned in Section 4.

Tables 1, 2, 3, and 4 show the results of the experiments. NVR+HADAMAC consistently outperforms NVR-BIP. The assignment accuracies improve by up to 21% when we use NVR+HADAMAC instead of NVR-BIP. On the other hand, for all proteins except 1G6J, NVR+HADAMAC results in higher accuracies than NVR+CRAACK when we use RDCs. When we do not use RDCs, NVR+HADAMAC results in higher accuracies than NVR+CRAACK except for a small group of proteins. The reason why CRAACK results in higher accuracies for some proteins could be that CRAACK uses more information coming from the chemical shifts of many different atom types, whereas HADAMAC approach only crudely classifies each peak into seven amino acid types. The assignment accuracies improve by up to 17% when we use NVR+HADAMAC instead of NVR+CRAACK.

6 Conclusion

In this paper, we proposed an approach to integrate HADAMAC experiment with NVR's data types. The experimental results shown in the previous section indicate that the proposed approach leads to better accuracies than NVR-BIP and NVR+CRAACK. With the addition of the HADAMAC experiment, NVR becomes a more useful and practical tool that can be used in an NMR laboratory. Furthermore, HADAMAC experiment distinguishes the type of the amino acid in about 30 minutes; whereas conventional 3D experiments needed to acquire the data used by CRAACK take hours to complete.

On the other hand, HADAMAC experiment has some limits. These are as follows:

1. In order to measure HADAMAC data, we need to have reasonably well resolved HSQC crosspeaks. There can be partially overlapping peaks but there will be trouble for exactly overlapped 2D crosspeaks.
2. HADAMAC works well only for reasonably small proteins (up to about 15kDa)
3. The protein needs to be fully protonated, at least for the beta position.
4. The protein has to be ^{13}C and ^{15}N labeled.

Table 1: Results on Ubiquitin

Protein	RDCs	NVR's Score Function [1]	NVR+CRAACK 1^{st} Approach	NVR+CRAACK 2^{nd} Approach	NVR+HADAMAC
1UBI	without RDCs	87%	97%	97%	96%
	with RDCs	100%	100%	100%	100%
1UBQ	without RDCs	87%	97%	100%	96%
	with RDCs	100%	100%	100%	100%
1G6J	without RDCs	87%	93%	97%	91%
	with RDCs	93%	93%	100%	96%
1UD7	without RDCs	81%	87%	90%	90%
	with RDCs	97%	97%	97%	99%
1AAR	without RDCs	79%	94%	100%	96%
	with RDCs	100%	100%	100%	100%

Table 2: Results on streptococcal protein G

Protein	RDCs	NVR's Score Function [1]	NVR+CRAACK 1^{st} Approach	NVR+CRAACK 2^{nd} Approach	NVR+HADAMAC
1GB1	without RDCs	100%	100%	100%	100%
	with RDCs	100%	100%	100%	100%
2GB1	without RDCs	100%	100%	100%	100%
	with RDCs	100%	100%	100%	100%
1PGB	without RDCs	96%	96%	96%	100%
	with RDCs	100%	100%	100%	100%

5. HADAMAC experiment does not provide information for the last residue in protein sequence and for residues preceding proline residues since they are not followed by a residue with the H^N moiety.

Note that the experiments were performed on theoretical HADAMAC data except for ubiquitin. As future work, we plan to test our approach on real data for these proteins. We also plan to combine the CRAACK and HADAMAC results in NVR's scoring function.

Another interesting area of future work is to incorporate additional types of real data into NVR, such as NOEs, and use the intensity field of the NOEs to extract more useful information to perform the assignments.

Acknowledgments

We thank Dr. Ewen Lescop for discussions.

Funding

This work was supported by following grants to M.S.A.: The Scientific and Technical Research Council of Turkey research support program (program code 1001) [109E027] and EU Marie Curie Grant PIRG05-GA-2009-249267.

References

[1] M. S. Apaydin, B. Catay, N. Patrick, and B. R. Donald. NVR-BIP: Nuclear Vector Replace-

Table 3: Results on lysozyme

Protein	RDCs	NVR's Score Function [1]	NVR+CRAACK 1^{st} Approach	NVR+CRAACK 2^{nd} Approach	NVR+HADAMAC
193L	without RDCs	78%	79%	79%	95%
	with RDCs	100%	100%	100%	100%
1AKI	without RDCs	78%	80%	80%	93%
	with RDCs	98%	98%	98%	98%
1AZF	without RDCs	74%	76%	78%	95%
	with RDCs	94%	95%	95%	95%
1BGI	without RDCs	75%	79%	83%	95%
	with RDCs	97%	97%	97%	100%
1H87	without RDCs	77%	79%	79%	95%
	with RDCs	100%	100%	100%	100%
1LSC	without RDCs	74%	78%	79%	95%
	with RDCs	100%	100%	100%	100%
1LSE	without RDCs	75%	78%	79%	95%
	with RDCs	98%	98%	98%	98%
1LYZ	without RDCs	79%	81%	79%	95%
	with RDCs	82%	87%	87%	95%
2LYZ	without RDCs	75%	79%	79%	95%
	with RDCs	91%	95%	95%	97%
3LYZ	without RDCs	79%	83%	83%	95%
	with RDCs	90%	90%	90%	97%
4LYZ	without RDCs	75%	79%	79%	95%
	with RDCs	91%	87%	87%	97%
5LYZ	without RDCs	75%	79%	79%	95%
	with RDCs	91%	87%	87%	97%
6LYZ	without RDCs	75%	79%	81%	95%
	with RDCs	96%	97%	97%	100%

ment using Binary Integer Programming for NMR Structure-Based Assignments. *The Computer Journal Advance Access published on January 6*, 2010. doi:10.1093/comjnl/bxp120.

[2] C. Benod, M. A. Delsuc, and J. L. Pons. CRAACK: Consensus program for NMR amino acid type assignment. *Journal of Chemical Information and Modelling*, 46:1517–1522, 2006.

[3] H. Erdogan and M. S. Apaydin. Using amino acid typing to improve the accuracy of NMR structure based assignments. In *Proc. of the 5th International Symposium on Health Informatics and Bioinformatics (HIBIT'10)*, 2010.

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

[5] C. Langmead, A. Yan, R. Lilien, L. Wang, and B.R. Donald. A Polynomial-Time Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments. In *Proc. of The Seventh Annual*

Table 4: Results on ff2, hSRI, pol η and GB1

.

Protein	RDCs	NVR's Score Function [1]	NVR+CRAACK 1^{st} Approach	NVR+CRAACK 2^{nd} Approach	NVR+HADAMAC
ff2	without RDCs	85%	93%	93%	92%
	with RDCs	93%	93%	93%	98%
hSRI	without RDCs	73%	73%	81%	88%
	with RDCs	89%	89%	94%	97%
pol η	without RDCs	100%	100%	100%	100%
	with RDCs	100%	100%	100%	100%
GB1	without RDCs	96%	100%	100%	100%
	with RDCs	100%	100%	100%	100%

International Conference on Research in Computational Molecular Biology (RECOMB), pages 176–187, Berlin, Germany, April 10–13, 2003. ACM Press. appears in: *J. Comp. Bio.* (2004), **11** (2-3), pp. 277-98.

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

[7] E. Lescop, R. Rasia, and B. Brutscher. Hadamard amino-acid-type edited NMR experiment for fast protein resonance assignment. *Journal of the American Chemical Society*, 130(15):5014–5015, 2008.

[8] M. G. Rossman and D. M. Blow. The detection of sub-units within the crystallographic assymetric unit. *Acta Crystal (D)*, 15:24–31, 1962.

[9] J. Volk, T. Herrmann, and K. Wüthrich. Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. *Journal of biomolecular NMR*, 41(3):127–138, 2008.

[10] J. Zeng, P. Zhou, and B. R. Donald. A markov random field framework for protein side-chain resonance assignment. In *Proc. of The Fourteenth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, Lisbon, Portugal, April 2010.