Using Amino Acid Typing to Improve the Accuracy of NMR Structure Based Assignments

Halit Erdoğan and Mehmet Serkan Apaydın
Sabanci University
Faculty of Engineering and Natural Sciences
Orhanli Tuzla, 34956,
Istanbul, TURKEY
Email: {halit,apaydin}@sabanciuniv.edu

***Abstract*—Nuclear Magnetic Resonance (NMR[1]) spectroscopy is an important experimental technique that allows one to study protein structure in solution. An important challenge in NMR protein structure determination is the assignment of NMR peaks to corresponding nuclei. In structure-based assignment (SBA), the aim is to perform the assignments with the help of a homologous protein. NVR-BIP [1] is a tool that uses Nuclear Vector Replacement's (NVR) ([9], [10]) scoring function and binary integer programming to solve SBA problem. In this work, we introduce a method to improve NVR-BIP's assignment accuracy with amino acid typing. We use CRAACK that takes the chemical shifts of C, N and H atoms and returns the possible amino acids along with their confidence scores. We obtain improved assignment accuracies and our results show the effectiveness of integrating amino acid typing with NVR.**

I. 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.

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 signals are recorded. After processing these signals, these 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 the Structure-Based Assignment (SBA). SBA is analogous to molecular replacement in X-ray Crystallography [16]. 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 [5], MATCH [19], [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 ([9], [10]), 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.

Amino acid typing refers to the determination of the amino acid type based on the chemical shifts. It can be used as a filter to help in NMR assignments. CRAACK [4] is an amino acid typing tool that combines multiple programs to help determine the amino acid type. The main contributions of this work are as follows:

1) We use amino acid typing software CRAACK to predict the amino acid groups that each NMR peak belongs to;
2) We integrate CRAACK's output with NVR-BIP; and
3) We test our approach on NVR-BIP's data set and compare our results with NVR-BIP.

The rest of the paper is organized as follows: In Section II, we review some of the scoring functions for NMR-SBA defined in the literature. In Section III, we review amino acid typing and CRAACK [4] software that will be useful for our scoring function, which is explained in Section IV. Data preparation is in Section V and the results are in Section VI. We conclude and discuss future work in Section VII.

II. NMR SBA SCORING FUNCTIONS

In an NMR assignment, the problem is to find the correspondence between a set $\mathcal{P}$ of peaks and a set $\mathcal{A}$ of residues. A scoring function determines the score associated

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

978-1-4244-5969-8/10/$26.00 ©2009 IEEE

HIBIT April 20-22, 2010
Antalya, Turkey

with assigning each NMR peak p to each amino acid a. The scoring functions in SBA make use of the template structure to compute this function. Improving the accuracy of scoring function is of paramount importance for NMR-SBA.

There exists several scoring functions in the literature. In [1], the authors mainly use NVR ([9], [10]) scoring function with additional sources of data and obtain accurate results in SBA. In NVR's scoring function each peak-residue pair assignment has a corresponding score contributed by 7 sources of information:

1) Chemical shift (CS) probabilities as computed from Biological Magnetic Resonance Bank [17] statistics,
2) Probabilities obtained from the difference between observed and predicted chemical shifts (predictions made with SHIFTS [20]),
3) Probabilities obtained from the difference between observed and predicted chemical shifts (predictions made with SHIFTX [14]),
4) Probabilities obtained from sidechain chemical shifts measured by TOCSY (a type of NMR experiment),
5) Probabilities obtained from Hydrogen-Deuterium exchange data,
6) Probabilities obtained by residual dipolar couplings (RDCs) in one medium, and
7) Probabilities obtained by another set of RDCs measured in the same or a different medium.

NVR-BIP combines these probabilities by taking the negative logarithm and then summation to obtain a scoring function. NVR-BIP computes an assignment whose total score is minimum subject to NOE constraints. For details please refer to [1].

Alternatives to NVR-score include [13] which uses CSs, NOEs and RDCs and combines them with empirically determined weights. In MARS [7], a scoring function based on the differences between experimental and measured CSs is introduced. In [6], RDCs are also incorporated into this scoring function, again with an empirically determined weighting constant.

III. Amino Acid Typing

Amino acid typing involves identifying the type of an amino acid based on the chemical shifts. Example programs include TATAPRO II [2], which takes in CA and CB chemical shifts and outputs one out of 8 categories to which the amino acid may belong to. Alternative to typing is the HADAMAC [11] experiment which enables to successfully distinguish the type of the amino acid in a couple of hours.

CRAACK [4] is a tool that takes chemical shifts $\{N, HN, HA, HB, CA, CB, CO\}$ as input and outputs a list of amino acid types. Each predicted amino acid type has a confidence score. CRAACK uses different amino acid type prediction tools such as RESCUE [15], RESCUEN [3], RESCUE2 [12], PLATON [8], and SVMTYPING [4]. CRAACK gets the prediction values of these tools and uses two approaches to compute a single consensus score value for the amino acid type corresponding to the chemical shift values. In the first approach, the amino acid types are categorized into eight groups and support vector machines (SVM) is used to determine the confidence score of the amino acid group. In the second approach, the consensus score is computed by voting in which each source (e.g., the aforementioned prediction tools and consensus score of SVM) has experimentally predetermined weights. We used the consensus scores which range between 0 and 6.8 in our experiments.

IV. NVR+CRAACK Scoring Function

The main motivation of this work is to investigate whether amino acid typing can be used to improve the accuracy of NVR-BIP. We provide chemical shifts to CRAACK and obtain amino acid predictions along with confidence scores. This results in a matrix (CRAACK score) that has for each (peak, residue) pair the consensus score associated by CRAACK. We integrate this matrix with NVR's score matrix using two approaches.

Our notations for the score matrices is as follows: Let S_n be the scoring matrix of NVR and S_c be the scoring matrix of CRAACK. Then, $S_n[i][j] = s_n$ corresponds to the NVR score of assigning peak i to amino acid j. The lower this value, the higher is the probability of assignment according to NVR. Similarly, $S_c[i][j] = s_c$ corresponds to CRAACK score of assigning peak i to amino acid j. Unlike S_n, this value is proportional to the assignment probability according to CRAACK. S_n is equal to ∞ if the assignment of peak i to residue j is impossible according to any of the scoring functions. S_n ranges between 4.5 and 760 (for ubiquitin) otherwise. S_c is 0 if the corresponding amino acid is not among the list of residues returned by CRAACK.

A. Approach 1

This approach uses CRAACK as a filter to eliminate the possibility of certain assignments. If the type of the considered residue is not amongst the set of amino acid possibilities returned by CRAACK, the corresponding score is assigned an infinite value and that assignment possibility is eliminated. More formally, for each peak i and for each amino acid j the combined score matrix that is derived from this approach (S^1_{nc}) is defined as follows:

$$S^1_{nc}[i][j] = \begin{cases} S_n[i][j] & \text{if } S_c[i][j] > 0 \\ \infty & \text{otherwise} \end{cases}$$

B. Approach 2

The idea of this approach is to reward the assignments whose CRAACK score is positive. Therefore, we subtract CRAACK score from NVR score. But if the CRAACK score is 0 then the corresponding assignment possibility is eliminated. More formally, for each peak i and for each amino acid j the combined score matrix that is derived from this approach (S^2_{nc}) is defined as follows:

$$S^2_{nc}[i][j] = \begin{cases} S_n[i][j] - S_c[i][j] & \text{if } S_c[i][j] > 0 \\ \infty & \text{otherwise} \end{cases}$$

V. Data Preparation

We test our approach on the data set of NVR-BIP using the chemical shifts collected from various sources. NVR-BIP only requires $^{15}\mathrm{N}$ and $\mathrm{H^N}$ chemical shifts. Although CRAACK can run with this minimal set of data, the predictions are not accurate. Therefore we provided CRAACK with the full list of chemical shifts. We collected this data using SHIFTS [20] and SHIFTX [14] which are chemical shift prediction tools. For some proteins we also used experimental chemical shifts collected from BMRB [17]. 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 proteins (template pdb ids: 193L, 1AKI, 1AZF, 1BGI, 1H87, 1LSC, 1LSE, 2LYZ, 3LYZ, 4LYZ, 5LYZ, 6LYZ), human Set2-Rpb1 interacting domain (hSRI), the FF Domain 2 of human transcription elongation factor CA150 (RNA polymerase II C-terminal domain interacting protein) (ff2), Y-polymerase Eta (pol η), B1 domain of streptococcal protein G (GB1). Note that in NMR community experimental results on multiple proteins is considered adequate [21]. More details on these proteins can be found in [1].

VI. Experimental Results

In this section, we present the results of the experiments that are performed on several proteins using NVR score matrix and the new score matrices defined in the previous section.

The results are shown in Table IV. The tested proteins are in the first column. Third column shows the best accuracy results obtained by NVR score matrix. Fourth and fifth columns show the accuracies obtained by the approaches explained in Subsection IV-A and IV-B respectively.

The assignment accuracies improve by up to 15% with the first approach. On the other hand, the assignment accuracies improve by up to 21% with the second approach. The only exceptions are 4LYZ and 5LYZ for which the accuracies of the assignments of NVR-BIP are 91% but with the first and second approach they decrease by 4%.

VII. Conclusion & Future Work

The results from the previous section indicate that the approaches proposed in Section IV are potentially useful for SBA since in general they lead to better assignment accuracies. Especially with the second approach, we observe improvements over the assignments computed by the NVR score. Although the proposed approaches are simple, they lead to improvements which shows that more sophisticated integrations might lead to better accuracies. We also tested multiplying CRAACK score with a coefficient in our second approach, this resulted in small changes in assignment accuracy. A way of combining NVR score with CRAACK score is to use machine learning approaches such as support vector machines to find a hyperplane that divides the correct assignments and the corresponding score combinations from incorrect ones. We are currently working to achieve this goal.

It may be possible to tolerate the incorrect predictions of CRAACK by penalizing the corresponding (peak, residue) assignment, rather than eliminating that possibility as we have implemented in both of our approaches. This may make our tool more robust with respect to errors in chemical shifts. Note that our results suggest that amino acid typing is especially useful when RDCs are not available. There are many such proteins for which our approach could be useful.

As future work we plan to use a larger database of chemical shifts to extract statistics, use more advanced tools for chemical shift predictions such as SPARTA [18] and incorporate HADAMAC [11] experiment into NVR-BIP for amino acid typing. Our preliminary tests with HADAMAC on ubiquitin suggest that the assignment accuracy increases from 87% to 96% without RDCs and remains at 100% with RDCs. We plan to test HADAMAC data on more proteins. In addition, we plan to test TATAPRO II [2] for a more robust typing. Also, developing a system that computes a Bayesian scoring function might be beneficial.

Another scoring function we plan implement is Meiler&Baker's scoring function [13]. However this objective function has a quadratic term and our system needs to be extended to solve this quadratic assignment problem.

The chemical shifts for amino acid typing require NMR experiments in addition to those used by NVR-BIP, increasing the cost of data collection. However HADAMAC experiment is especially easy to acquire and this paper is a proof-of-principle that amino acid typing is potentially useful to improve the accuracy of NVR-BIP.

Acknowledgments

This work was supported by the Scientific and Technical Research Council of Turkey research support program (program code 1001) [109E027 to M.S.A]. We thank Dr. B. Çatay for the CPLEX models of the proteins in this study and Dr. Pei Zhou for providing us with data for some of the proteins.

References

[1] M. S. Apaydin, B. Catay, N. Patrick, and B. R. Donald. NVR-BIP: Nuclear Vector Replacement using Binary Integer Programming for NMR Structure-Based Assignments. *The Computer Journal Advance Access published on January 6*, 2010. doi:10.1093/comjnl/bxp120.

[2] H. S. Atreya, K. V. R. Chary, and G. Govil. Automated NMR assignments of proteins for high throughput structure determination: TATAPRO II. *Current Science Journal*, 83(11):1372–1376, 2002.

[3] D. Auguin, V. Catherinot, T. E. Malliavin, J. L. Pons, and M. A. Delsuc. Superposition of Chemical Shifts in NMR Spectra Can Be Overcome to Determine Automatically the Structure of a Protein. *Spectroscopy*, 17:559–568, 2003.

[4] 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.

[5] 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.

[6] Y. Jung and M. Zweckstetter. Backbone assignment of proteins with known structure using residual dipolar couplings. *Journal of Biomolecular NMR*, 30(1):25–35, 2004.

[7] Y. Jung and M. Zweckstetter. Mars - robust automatic backbone assignment of proteins. *Journal of Biomolecular NMR*, 30(1):11–23, 2004.

TABLE I
RESULTS ON UBIQUITIN

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

TABLE II
RESULTS ON STREPTOCOCCAL PROTEIN G

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

[8] D. Labudde, D. Leitner, M. Kruger, and H. Oschkinat. Prediction algorithm for amino acid types with their secondary structure in proteins (PLATON) using chemical shifts. *Journal of Biomolecular NMR*, 25:41–53, 2003.

[9] 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 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.

[10] 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.

[11] 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.

[12] A. Marin, T. E. Malliavin, P. Nicolas, and M. A. Delsuc. From NMR chemical shifts to amino acid types: investigation of the predictive power carried by nuclei. *Journal of Biomolecular NMR*, 30:47–60, 2003.

[13] J. Meiler and D. Baker. Rapid protein fold determination using unassigned NMR data. *PNAS*, 100(26):15404–15409, 2003. December.

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

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

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

[17] B. R. Seavey, E. A. Farr, W. M. Westler, and J. L. Markley. A relational database for sequence-specific protein NMR data. *Journal of Biomolecular NMR*, 1(3):217–236, September 1991.

[18] Y. Shen and A. Bax. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *Journal of Biomolecular NMR*, 38:289–302, 2007.

[19] 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.

[20] X. P. Xu and D. A. Case. Automated prediction of 15N, 13Cα, 13Cβ and 13C' chemical shifts in proteins using a density functional database. *Journal of Biomolecular NMR*, 21(4):321–333, 2001.

[21] 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. In press.

TABLE III
RESULTS ON LYSOZYME

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

TABLE IV
RESULTS ON FF2, HSRI, POL η AND GB1

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