# Developing a Scoring Function for NMR Structure-based Assignments using Machine Learning

Mehmet Çağrı Çalpur[1], Hakan Erdoğan[1], Bülent Çatay[1], Bruce R. Donald[2] and Mehmet Serkan Apaydın[1]

[1] Sabanci University
Faculty of Engineering and Natural Sciences
Tuzla Istanbul, 34956, TURKEY
[2] Duke University, Department of Computer Science
Duke University Medical Center, Department of Biochemistry
Durham NC 27708, USA

**Abstract.** Determining the assignment of signals received from the experiments (peaks) to specific nuclei of the target molecule in Nuclear Magnetic Resonance (NMR[1]) spectroscopy is an important challenge. Nuclear Vector Replacement (NVR) ([2, 3]) is a framework for structure-based assignments which combines multiple types of NMR data such as chemical shifts, residual dipolar couplings, and NOEs. NVR-BIP [1] is a tool which utilizes a scoring function with a binary integer programming (BIP) model to perform the assignments. In this paper, support vector machines (SVM) and boosting are employed to combine the terms in NVR-BIP's scoring function by viewing the assignment as a classification problem. The assignment accuracies obtained using this approach show that boosting improves the assignment accuracy of NVR-BIP on our data set when RDCs are not available and outperforms SVMs. With RDCs, boosting and SVMs offer mixed results.

## 1 Introduction

The gold standard in determining the protein structure is wet-lab experiment, the primary ones being X-ray crystallography (XRC) and NMR spectroscopy. In order to understand a protein's function and do rational drug design, it is necessary to determine the protein structure.

Structure-based assignment (SBA) aims to determine the assignments using a template structure. This template is homologous to the target. Previous techniques for NMR SBA include NVR-EM [3], NVR-BIP [1], MARS [4], NOE-net [5], Hus et al. [6].

[1] Abbreviations used: NMR, Nuclear Magnetic Resonance; NVR, nuclear vector replacement; NOE, Nuclear Overhauser Effect; BIP, binary integer programming; SVM, Support Vector Machine; RDC, Residual Dipolar Coupling; XRC, X-ray Crystallography; PDB, Protein Data Bank; SBA, Structure-based Assignment; EM, Expectation Maximization.

NVR-BIP works comparably well on the proteins on which NVR-EM was tested. Furthermore it provides significantly better accuracies on four novel proteins. However the scoring function of NVR involves simple addition of the contribution of 7 different terms although these terms are not independent. The goal of this work is to explore machine learning techniques to learn optimal ways of combining these terms.

To the best of our knowledge, this is the first approach that uses classification techniques to develop a scoring function for NMR SBA. Our contributions in this paper are:

- the combination of the components of NVR-BIP's scoring function using SVMs and boosting
- incorporation of the novel scoring function into NVR-BIP and
- testing the novel scoring function on NVR-BIP's data set and comparison with the results reported in [1].

The rest of the paper is as follows: Section 2 describes the proposed algorithm, followed by the implementation in Section 3. Section 4 presents the experimental study and discusses the results. Finally, concluding remarks are given in the last section.

## 2 Methods

The data set is divided into two components: A training set and a test set.The training set consists of data corresponding to those proteins that are homologous to the target, except the template with which the SBA will be performed and which forms the test set.

The goal is to learn a classifier that distinguishes the correct peak-residue pair from incorrect ones. SVMs and boosting return scores corresponding to how confidently the corresponding classification is made. The output of the learning algorithm is used as the scoring function of the BIP model, which solves the SBA problem. After the initial assignments are made, an alignment tensor is computed and then the components of the scoring function corresponding to RDCs are included.

## 3 Implementation

The training data set belongs to two classes, +1 and -1. Positive label represents the correct peak-residue assignment pair and negative label represents incorrect assignments. Roughly there are 2000 positively labeled instances and 100,000 negatively labeled instances. SVMs require weighting adjustment to the data, in order not to classify all instances as -1. The +1 instances are multiplied by the weight factor (-1/+1 instance ratio). We solve the BIP problem using ILOG OPL Studio CPLEX solver.

The execution times for the BIP solution change according to the number of available peak-residue assignments. Without RDCs, using the boosting scores, the CPLEX solver runs for an average of 5 minutes to solve the system for ubiquitin. Adding RDC information to the process greatly reduces the number of available assignments, therefore reducing the problem size, and the average execution time for boosting scores becomes 45 seconds on an Intel Celeron 560 computer with 2.13 Ghz processor with 2GB memory.

## 4 Results

The experimental results are reported in Table 1. Both the results without and with RDCs are provided. In addition, the results obtained by the addition of scoring function components, which are used in NVR-BIP [1] are given as a reference and is labeled the addition method. It can be seen that, without RDCs, for most of the proteins SVM accuracies are about the same as the accuracies obtained using the addition method. Boosting accuracies are 7-16% higher than addition method. On the other hand, with RDCs, boosting and addition method's accuracies are similar. The SVM results given in the following table are obtained with RBF Kernel. The boosting results given in Table 1 are the results achieved by Gentle AdaBoost algorithm. Results on more proteins are available in our technical report.

**Table 1.** Results on Ubiquitin without and with RDCs.

| | without RDCs | | | with RDCs | | |
|---|---|---|---|---|---|---|
| PDB ID | addition | SVM | boosting | addition | SVM | boosting |
| 1UBI | 87% | 84% | 97% | 97%$^a$<br>100%$^b$ | 97%$^a$<br>97%$^b$ | 97%$^a$<br>100%$^b$ |
| 1UBQ | 87% | 87% | 100% | 97%$^a$<br>100%$^b$ | 97%$^a$<br>100%$^b$ | 100%$^a$<br>100%$^b$ |
| 1G6J | 87% | 87% | 100% | 97%$^a$<br>93%$^b$ | 94%$^a$<br>93%$^b$ | 100%$^a$<br>90%$^b$ |
| 1UD7 | 81% | 81% | 97% | 97%$^a$<br>97%$^b$ | 89%$^a$<br>89%$^b$ | 97%$^a$<br>100%$^b$ |
| 1AAR | 79% | 83% | 86% | 97%$^a$<br>100%$^b$ | 93%$^a$<br>93%$^b$ | 93%$^a$<br>93%$^b$ |

$^a$ with NH RDCs in two media
$^b$ with NH and CH RDCs.

## 5 Conclusion

In this study, we combine the SVM and boosting techniques with BIP within NVR's framework to perform SBA. The tests without RDCs show that boosting has better assignment accuracy than the addition method. With RDCs, our accuracies are comparable to the addition method for both SVM and boosting. This may be explained by the fact that with RDCs, the RDCs dominate the feature vectors and they don't allow separating the positive examples from negative ones. Boosting method is therefore especially suitable for use when RDCs are not available. When RDCs are available, our method could be used to accelerate converging to the best assignment by providing a better assignment from which a better alignment tensor estimate could be obtained.

Our results also indicate that, the training set for a protein from the homologous protein data provides good assignment accuracies. However this limits our approach to those proteins for which homologous proteins and their corresponding assignments are known. As future work we are interested in developing a Bayesian scoring function for SBA that does not have this requirement.

## Acknowledgments

We thank Dr. Nanjiang Shu for discussions. We thank Dr. Pei Zhou for providing us with the NMR data. We acknowledge funding from the Scientific and Technical Research Council of Turkey (TUBITAK)[program code 1001; 109E027] to M.S.A. and from the NIH [GM-65982] to B.R.D.

## References

1. Apaydın, M. S., Çatay, B., Patrick N. and Donald, B. R.: 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: doi:10.1093/comjnl/bxp120.
2. Langmead, C., Yan, A., Lilien, R., Wang, L., and Donald, B.: A Polynomial-Time Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments (2003) In Proc. The Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB) Berlin, Germany, April 1013: ACM Press. appears in: J. Comp. Bio. (2004), 11 (2-3), pp. 277-98 pp. 176-187.
3. Langmead, C. and Donald, B.: An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments (2004), Journal of Biomolecular NMR. 29(2), 111-138.
4. Jung, Y. and Zweckstetter, M.: Mars – robust automatic backbone assignment of proteins (2004), Journal of Biomolecular NMR 30(1), 11-23.
5. Stratmann, D., van Heijenoort, C. and Guittet, E.: NOEnet-Use of NOE networks for NMR resonance assignment of proteins with known 3D structure, Bioinformatics (2009), 25(4):474–481
6. Hus, J., Prompers, J., and Bruschweiler, R.: Automated NMR assignment and protein structure determination using sparse dipolar coupling constraints (2002), J. Mag. Res. 157(1), 119-125.

This article was processed using the LaTeX macro package with LLNCS style