

Large-scale testing of chemical shift prediction algorithms and improved machine learning-based approaches to shift prediction

K. Arun[†]
arunk@andrew.cmu.edu

Christopher James Langmead^{*,†,‡}
cjl@cs.cmu.edu

Abstract

The resonant frequencies, or chemical shifts, of nuclear magnetic resonance (NMR) active nuclei in proteins are determined by covalent and through-space interactions and, more generally, the electronic environment surrounding each nucleus. However, the precise nature of the correlation between protein three-dimensional (3D) structure and chemical shift remains largely unsolved. Thus, chemical shift prediction is a non-trivial task. This study tests the accuracy of three existing structure-based chemical shift prediction algorithms (SHIFTS, SHIFTX, PROSHIFT) against REFDB, a large database of experimentally determined, and manually re-referenced ¹H, ¹³C, and ¹⁵N chemical shifts. We report that the accuracy of backbone chemical shift predictions for each program is lower than that originally reported. This suggests these programs over-fit the data used in their construction. We then compare two novel methods for chemical shift prediction based on support vector machines (SVM) and bagging respectively. Each method was trained on REFDB using predictions made by SHIFTS, SHIFTX, and PROSHIFT as features. In cross-validated experiments, bagging is shown to be superior to SVMs, while both methods are substantially better than SHIFTS, SHIFTX, and PROSHIFT. Our results suggest that meta-methods for chemical shift prediction yield increased accuracy for chemical shift prediction.

1. Introduction and Overview

The chemical shifts of NMR-active nuclei are determined by the local electronic environment surrounding each nucleus. Thus, there is a strong relationship between a protein's three dimensional structure and the chemical shifts of its constituent nuclei. However, predicting and interpreting shifts, in the context of a structural model, remains a difficult problem. Existing algorithms for predicting chemical shifts from atomic-resolution structural models permit the quantitative exploration of the relationship between structure

and shift. In this study, we report on the prediction accuracies of three such algorithms, employing a large data set of carefully re-referenced experimentally determined chemical shifts - the REFDB database [4]. The prediction programs tested include SHIFTS [3], SHIFTX [2] and PROSHIFT [1].

Most approaches to chemical shift prediction from atomic co-ordinates employ either quantum mechanical, classical, or semi-classical mechanical calculations; or depend on empirical approaches utilizing experimentally determined data. The SHIFTS program takes a quantum mechanical approach and employs density functional calculations to predict ¹H, ¹³C, and ¹⁵N shifts, while SHIFTX uses empirically determined chemical shift hypersurfaces in combination with classical mechanical calculations. PROSHIFT applies a neural network trained on carefully chosen high-resolution structures.

The data used in this study are taken from REFDB, which contains almost 200,000 experimentally determined shifts from 601 different proteins. In contrast, SHIFTS and SHIFTX used much smaller training sets (about 20 and 60 proteins respectively), while the neural network used in PROSHIFT's implementation was trained on ~69,000 shifts from 292 proteins. When tested on the REFDB data, these programs have significantly lower accuracies for backbone chemical shift predictions than those originally reported. This suggests that these programs over-fit their respective training sets.

Our aim is to produce a better backbone chemical shift prediction program using techniques from machine learning. We compared the performance of support vector machines (SVM) and bagging for predicting backbone chemical shifts. Using cross-validation to compute the accuracy of the SVM and bagging regressors, both SVMs and bagging outperform each of the other programs. Moreover, the root mean square error (RMSE) values obtained using bagging approach the (over-fit) accuracies reported by SHIFTS, SHIFTX, and PROSHIFT.

2. Methods

The REFDB experimental chemical shift data along with provided secondary structure information were loaded into tables created using the PostgreSQL relational database management system (RDBMS). The REFDB also provides a mapping between its 601 entries and Protein Data Bank (PDB) structures which was used to obtain the PDB files that serve as input to the shift prediction programs. All three programs were obtained from their respective authors and run locally. The predicted chemical shifts were then also

* Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

† Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213

‡ Corresponding author: Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Phone: (412) 268-7571. Fax: (412) 268-5576. Email: cjl@cs.cmu.edu

Atom type	Data points	SHIFTS	SHIFTX	PROSHIFT	SVM model	Bagging	Bagging (± 1)	Bagging (± 5)
H ^N	72606	0.684	0.643	0.616	0.56 (10%)	0.45 (27%)	0.47 (24%)	0.46 (25%)
N	49447	6.4	4.59	4.8	4.08 (11%)	3.13 (32%)	3.07 (33%)	3.12 (32%)
C ^{α}	44257	4.042	3.7	4.32	3.48 (6%)	1.27 (66%)	1.20 (67%)	1.17 (68%)
H ^{α}	59875	1.078	0.633	0.636	0.611 (3%)	0.56 (12%)	0.48 (24%)	0.38 (40%)

Table 1. RMSE values (in ppm) for individual chemical shift predictors and the SVM and bagging algorithms (percentage improvements over best individual predictor in parentheses)

loaded into the RDBMS in tabular form. Structured query language (SQL) code was used to merge experimental and predicted data tables, and the resultant table contained matched predictions from the three algorithms, the corresponding experimental shift, and secondary structure information for the amino acid residue. Given predicted and experimental shifts, RMSE values were calculated for each prediction algorithm, for each predicted atom type (¹H, ¹³C, ¹⁵N).

We next compared the performance of two different algorithms: support vector machines (SVM) and bagging, for predicting backbone chemical shifts. For each method, a different regressor was constructed for H^N, N, C ^{α} , and H ^{α} nuclei. The training data was taken from REFDB and contained about 72000, 49000, 44000 and 28000 shifts, respectively. The input features included the amino acid type; the secondary structure classification, taken from the structural model; and the SHIFTS, SHIFTX, and PROSHIFT predictions for the same nucleus. RMSE values were calculated for both the SVM and bagging regressors, for each predicted atom type (¹H, ¹³C, ¹⁵N).

3. Results and discussion

The results obtained for prediction accuracies of the individual predictors, and of the SVM and bagging algorithms are presented in table 1 in the form of RMSE values (units ppm) for each predicted atom type, and the percentage improvement for the latter two algorithms over the best individual predictor. The SVM models were trained and tested over the whole data set, and yield slightly improved prediction accuracies over the individual predictors. However, dramatic increases in the accuracy of the shifts predicted are observed when the bagging algorithm is employed (without windowing, and with windows of ± 1 and ± 5 residues in successive runs over the data).

It should be noted that a minimal number of input features were employed with both the SVM and bagging approaches - primarily, output from the three shift prediction algorithms and amino acid and secondary structural information from the PDB structure. Both algorithms' implementations support the use of many more features, the use of which is likely to improve these methods' prediction accuracies further. Such features may include (but are not limited to) larger window size

derived average chemical shift values for various partitions of the amino acid set, and additional chemical shift prediction component information available from algorithms such as SHIFTS.

4. Conclusions

This study was motivated by the desire to improve the accuracy of existing approaches to chemical shift prediction. It is clear from the results obtained that the most commonly used approaches do not yield prediction accuracies near those reported originally for proteins outside of the small training and testing sets used for the purpose of initial benchmarking. Applying ensemble learning techniques such as bagging in a layer over existing chemical shift predictors yields significantly improved accuracies for predictions. It is expected that more sophisticated input feature selection procedures will further improve these accuracy numbers, and will permit a more detailed exploration of both the relationship between 3D structure and chemical shifts, and the additional structural information inherent in shift data.

5. Acknowledgments

This work was supported in part by a young pioneer award to C.J.L. from the Pittsburgh Life Sciences Greenhouse. K.A. is supported by the Merck Computational Biology and Chemistry Program graduate fellowship. We would like to thank the authors and maintainers of the tested programs for making their code freely available - David Case for SHIFTS, David Wishart and Haiyan Zhang for SHIFTX and REFDB, and Jens Meiler for PROSHIFT.

References

- [1] J. Meiler. PROSHIFT: Protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR*, 26:25–37, 2003.
- [2] S. Neal, A. M. Nip, H. Zhang, and D. S. Wishart. Rapid and accurate calculation of protein ¹H, ¹³C and ¹⁵N chemical shifts. *J. Biomol. NMR*, 26:215–240, 2003.
- [3] X. Xu and D. Case. Automated prediction of ¹⁵N, ¹³C ^{α} , ¹³C ^{β} and ¹³C' chemical shifts in proteins using a density functional database. *J. Biomol. NMR*, 21:321–333, Dec 2001.
- [4] H. Zhang, S. Neal, and D. S. Wishart. RefDB: A database of uniformly referenced protein chemical shifts. *J. Biomol. NMR*, 25:173–195, 2003.