NMR Prediction Accuracy Validation

ACD/CNMR Predictor
Version 10.05

Kirill Blinov, Mikhail Kvasha, Brent Lefebvre, Ryan Sasaki and Antony Williams
Advanced Chemistry Development, Inc.
Toronto, ON, Canada
www.acdlabs.com

Introduction

The validation of performance of NMR chemical shift prediction algorithms is a challenging problem for a number of reasons. These will be discussed only at a general level in this technical evaluation since they have been discussed elsewhere.[1] The central challenge associated with the validation of NMR shift prediction algorithms is obtaining a quality data set for validation of the prediction accuracy.

If the validation data set is mainly simple structures, or structures that are well represented in the database used as the basis of the prediction algorithms then the validation exercise will not truly represent the challenges of prediction. The most valid test would be conducted on a validation set containing chemical structures which are very different from these contained within the training dataset. Ideally, an independent party without knowledge of the structures in the training set should choose the validation set, so as to avoid any bias.

The quality of a validation database is important but difficult to prove in most cases. The ideal validation set does not contain any errors in assignment and covers the whole range of structural diversity available in present chemistry and in all future diversity possibilities. While this is clearly impossible to attain, large diverse datasets do exist and, while not ideal, can be used for the purpose of validation. Every large dataset contains errors but for comparisons of prediction between different algorithms this is actually irrelevant since any errors remain challenging for all algorithms.

A resource is available on the Internet that has met the above criteria of size and quality to serve as a fair and reliable validation set to evaluate the performance of ACD/CNMR Predictor in terms of accuracy of NMR prediction. This resource is a database called NMRShiftDB[2] and is created as a collaborative effort by chemists and spectroscopists submitting data to the database. This document is an analysis of the performance of the ACD/CNMR predictor using the NMRSHIFTDB database as the validation set. Due to the availability of a comparison test issued by Wolfgang Robien we also have an opportunity to compare performance with another commercial product, NMR Predict provided by Modgraph Consultants, Ltd.

NMRShiftDB

The NMRShiftDB is an open source collection of chemical structures and their associated NMR shift assignments. The database is generated as a result of contributions by the public and has been described in detail elsewhere[3,4]. Currently[5], the database contains 19,958 structures with 214,136 assigned carbon chemical shifts. Based on a cursory examination of the structural diversity within the database these data represent a statistically relevant set to use in an evaluation of predictive accuracy and is the first large dataset available from an independent source which we could use for this purpose.

Robien has already published an analysis of performance of his neural network predictions[6]. This review provides an evaluation of the NMR prediction algorithms he has developed over many years. These algorithms have been the basis of a number of software products including a commercially available product, NMRPredict[6], offered by Modgraph Consultants, Ltd. Robien focused his analysis on the presence of a number of outliers but gave no specific review of the quality of the dataset focusing only on the problem assignments.

NMR Prediction Validation

The NMRShiftDB website offers visitors the opportunity to download a file in SDF format containing all of the structures and chemical shifts that compose the NMRShiftDB database[5]. This file was downloaded and the structures and shifts were imported into an ACD/Labs' format.

As a first step, an analysis of the degree of overlap between the structures in the training set within the ACD/CNMR Predictor and the validation set of NMRShiftDB was undertaken. It was found that 57% of the carbon chemical shifts in the NMRShiftDB were already in the ACD/Labs database. Using this information the NMRShiftDB database was then stripped of these chemical shifts, since they have been used as the basis of the prediction algorithms in ACD/CNMR Predictor. The statistics comparing the full dataset and the validation subset are shown below.

Results Summary

As mentioned above, 2 sets of results were obtained. The average deviation of the predicted vs. experimental values based on the entire data set, and the same statistics for the subset of chemical shifts that were unique. ACD/CNMR Predictor significantly outperforms Robien's program by a significant margin. The average deviation obtained by CNMR Predictor was 40% lower than that obtained by Robien.

Validation on Entire Dataset

Entire Dataset Comparison	Shift Count	Average Deviation (ppm)[9]	Standard Deviation (ppm)[9]	Outliers (ppm Difference)		
				>10 ppm	>25 ppm	>50 ppm
ACD/CNMR v10.05	**214,136**	**1.59**	**2.76**	**1,040 (0.5%)**	**141 (0.07%)**	**31 (0.01%)**
CSEARCH (Modgraph)	**209,412**	**2.22**	**N/A**	**N/A**	**194 (0.09%)**	**56 (0.03%)**

Only 203,284 unique carbon centers were represented in the database but some had multiple assignments. All redundancy was included in case there was disagreement between the assignments. And therefore over 214,000 assignments were considered. This is more than the number used by Robien and we assume this to be due to the fact that we downloaded the data file later than Robien and new data had been added. Robien used 209,412 chemical shifts. Consultation with Steinbeck indicates that some of the errors identified by Robien in his analysis have been corrected. At present the original source file utilized by Robien in his analysis is being sourced in order to allow a direct comparison of performance using the exact dataset and we will report on that comparison in a separate publication. The web posting of Robien[6] did not provide a measure of Standard Deviation or the number of chemical shift predictions that were more than 10 ppm from their experimental value and this is the reason for the absence of this parameter in the table.

> **Note** Robien quoted an Average Deviation of 2.19 PPM after correction of some errors, but for comparison purposes, we have used the 2.22 PPM value with no corrections since no corrections were made to dataset that was run through the CNMR Predictor.

Validation on Completely Novel Chemical Shifts

Obtaining a good result with the full dataset was a useful exercise yet a more rigorous comparison was conducted. The data used to train ACD/Labs NMR prediction algorithms include those collected from recent literature articles and an overlap with a significant number of structures in the NMRShiftDB was expected. In order to compare the predictive accuracy of the algorithm and provide an estimate of the performance of the predictors on novel structures, the NMRShiftDB was cross-referenced with the internal database of ACD/CNMR Predictor to remove duplicate structures. This exercise revealed that 57% of the compounds in the NMRShiftDB were also found in the ACD/Labs database. This left 43% of the compounds, a total of

92,927 chemical shifts in NMRShiftDB to use as the dataset for the second validation study. The results are shown below.

Dataset Comparison	Shift Count	Average Deviation (ppm)[9]	Standard Deviation (ppm)[9]	Outliers (ppm Difference)		
				>10 ppm	>25 ppm	>50 ppm
Data Subset	**92,927**	**1.74**	**3.22**	**695 (0.7%)**	**89 (0.1%)**	**24 (0.03%)**
Entire Dataset	**214,136**	**1.59**	**2.76**	**1,040 (0.5%)**	**141 (0.07%)**	**31 (0.01%)**

The average deviation associated with the data subset has increased only slightly. The question as to whether the comparison of different datasets leads to significant differences in performance has been examined. The expectation would be that the correction of a few data points in a dataset of over 200,000 shifts would have a very small impact on the statistics presented in the table above. In fact, ignoring all data points with an error of >25ppm reduces the average deviation to a value of 1.56ppm, a difference of 0.03ppm. Clearly the removal of a few points in error only makes a small difference to the overall statistics.

Conclusion

The NMRShiftDB is an excellent resource for the purpose of evaluating chemical shift prediction accuracy as evidenced by this work and the previous work of Robien. As identified by Robien initially, and later in this work, there are certainly outliers in the dataset requiring review and correction. Our previous work has shown that the literature itself contains about 8% errors in the form of mis-assignments, transcription errors and incorrect structures. The obvious errors in NMRShiftDB are certainly below this level and this is a testament to the value of this resource. The NMRShiftDB dataset is large and structurally diverse and continues to grow as scientists contribute.

Despite a large overlap between the NMRShiftDB and the ACD/Labs carbon NMR database, a statistically relevant validation set of over 92,000 chemicals shifts was extracted from the NMRShiftDB and used to test the algorithms. The data presented here shows that the ACD/Labs prediction algorithms have an average deviation of less than 1.8 ppm on the validation set and significantly outperforms the algorithms of Robien presented in his review. This work will be discussed in further detail in a future publication and validation is presently being performed on the proton NMR shift data.

Acknowledgements

We thank Christoph Steinbeck and the members of his team for the provision of the NMRShiftDB service and dataset. We would also like to thank all of the contributors to the NMRShiftDB. This has provided an invaluable resource for the testing of NMR prediction algorithms.

References

1. Jens Meiler et al, *J. Magn. Res.*, **157,** 242–252 (2002).
2. NMRShiftDB, http://nmrshiftdb.ice.mpg.de/
3. C. Steinbeck, S. Krause, and S. Kuhn, *J. Chem. Inf. Comput. Sci.* **43**, 1733-1739 (2003)
4. C. Steinbeck and S. Kuhn, *Phytochemistry* 65, 2711–2717 (2004)
5. Information available on NMRShiftDB as of May 7th, 2007.
6. http://nmrpredict.orc.univie.ac.at/csearchlite/enjoy_its_free.html
7. http://www.modgraph.co.uk/product_nmr.htm
8. http://nmrshiftdb.pharmazie.uni-marburg.de/nmrshiftdbhtml/NmrshiftdbWithSignals.sdf.zip

9.

$$\text{Standard Deviation} = \sqrt{\frac{\sum (\delta_{expi} - \delta_{calci})^2}{n-1}}$$

$$\text{Average Deviation} = \frac{\sum |\delta_{expi} - \delta_{calci}|}{n}$$