The Impact of Available Experimental Data on the Prediction of ^{1}H NMR Chemical Shifts by Neural Networks

Yuri Binev,[†] Marta Corvo, and João Aires-de-Sousa*

REQUIMTE, CQFB, Departamento de Química, Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa, Quinta da Torre, 2829-516 Caparica, Portugal

Received October 22, 2003

Two different ways were explored to incorporate new available experimental data into previously trained ensembles of feed-forward neural networks, for the structure-based prediction of ^{1}H NMR chemical shifts of organic compounds. One approach used the new data as the memory of an associative neural network (ASNN) system. For an independent prediction set of 952 cases, a mean average error of 0.19 ppm was achieved (0.13 ppm for 90% of the cases). This approach advantageously avoids retraining the networks, and the predictions compared favorably with those obtained by available commercial software packages. Excellent predictions could also be achieved by retraining the networks with the new data, but only if the training sets were selected so as to be balanced or if the retraining started with the weights of the previously trained networks.

INTRODUCTION

Most successes and failures of neural networks are closely related to the quality and quantity of experimental data used for training. An inaccurate prediction from a neural network can be often traced back to the inexistence or underrepresentation of certain types of objects in the training set. A well-designed training set should be at the same time diverse and balanced, although in most applications it is limited by the availability of data. And, the amount of data that is available to a system is constantly increasing, and sometimes not well-defined, as new data are always being produced and stored, access to some proprietary data is often only possible after the system is developed, and predictive systems are increasingly delivered with the possibility to be improved with the user's own data. Neural networks should hence be able to easily incorporate new data, made available after the training, expanding their applicability and improving their accuracy.

This situation is well-represented by our studies of neural networks for the prediction of ^{1}H NMR chemical shifts, which were originally trained with a data set of 744 protons belonging to organic compounds, and further optimized with the same data.[1] Remarkable results could be achieved with relatively small training sets. Although the data were selected from the literature to be balanced and to cover as many different situations as possible,[2] it is definitely a small data set for the development of a general model. This observation becomes more apparent by considering other general approaches to estimate ^{1}H NMR chemical shifts, centered on databases with more than 1 million entries.[3]

When a larger data set from Chemical Concepts became available to us, we explored two different ways of incorporating new knowledge into previously trained feed-forward neural networks (FFNNs). Here we describe those studies. One way was to use the newly available data as additional memory of the previously trained FFNNs, now organized as associative neural networks.[4] Associative neural networks adjust the prediction obtained from an ensemble of NNs on the basis of the errors for the *k* nearest neighbors (KNN) in the memory. The KNNs are searched in the output space; i.e., they are the cases in the memory with the most similar output profile to the query proton. The term output *profile* is used because a number of outputs are obtained, for the same proton, from the different NNs of the ensemble. Definition of KNNs *in the output space* avoids considering the relative importance of the different proton descriptors to the model. If the search was performed in the input (descriptors) space, each descriptor would make the same contribution to the similarity measure, while their relative importance in the model is not certainly the same. With this approach, the larger data set could be incorporated in the memory and retraining of the neural networks was spared. Associative neural networks have been successful for example in the prediction of lipophilicity.[5]

In the second approach, the FFNNs were retrained with the new data.

METHODOLOGY

Descriptors of Hydrogen Atoms. Protons were represented by physicochemical, geometric, and topological descriptors, which were calculated with fast empirical methods. These were previously defined and implemented.[1,2] Protons were classified in four classes: aromatic, nonaromatic π, rigid aliphatic, and nonrigid aliphatic. The four classes were treated separately; i.e., each used a different ensemble of NNs, and some descriptors were specific for one class.

Data Sets. The data set, obtained from Chemical Concepts GmbH, contained 5631 experimental chemical shifts and the

* Corresponding author phone: +351-21-2948575; fax: +351-21-2948550; e-mail: jas@fct.unl.pt.

[†] Permanent address: Institute of Organic Chemistry, Bulgarian Academy of Sciences, BG-1113 Sofia, Bulgaria.

10.1021/ci034229k CCC: $27.50
Published on Web 02/13/2004

corresponding hydrogen atoms from 482 structures. Stereochemistry was not assigned. Only data from spectra obtained in $CDCl_3$ were considered. The collection was restricted to CH_n protons in compounds containing elements C, H, N, O, S, F, Cl, Br, or I. Chemical shifts of hydrogen atoms bonded to heteroatoms were not included since they are strongly influenced by experimental conditions (such as the concentration of the sample).

A prediction set of 952 experimental chemical shifts and the corresponding protons was extracted from the initial set. This set of protons is from 100 structures. The selection of the prediction set was done after mapping the initial 482 structures (encoded with RDF codes) on a 10 × 10 Kohonen self-organizing map. One structure was randomly taken from each of the occupied neurons of the map, at the end of training.[6]

The prediction set was not included in the memory of associative neural networks, nor was it used for retraining neural networks.

Associative Neural Networks.[4] An associative neural network (ASNN) consists of an ensemble of neural networks and a memory. The ensemble is a set of neural networks, which contribute to a single prediction.[7,8] The prediction is calculated as the average of the outputs from the nets of the ensemble. Fully connected feed-forward neural networks (FFNNs) with an input layer (including a bias equal to 1), one hidden layer (also including a bias equal to 1), and one output neuron had been trained and optimized, as described in the preceding paper. The number of FFNNs in an ensemble, the number of hidden neurons, and the number of descriptors were the following for the four classes of protons: 50 neural network ensemble, 9 hidden neurons, 50 descriptors for aromatic protons; 50 neural network ensemble, 6 hidden neurons, 61 descriptors for nonaromatic π protons; 25 neural network ensemble, 10 hidden neurons, 70 descriptors for rigid aliphatic protons; 50 neural network ensemble, 10 hidden neurons, and 53 descriptors for nonrigid aliphatic protons.

The ensemble of NNs is combined with a *memory* into a so-called *associative neural network*.[4] In our studies, the memory consists of a list of protons, represented by their descriptors, and the corresponding experimental chemical shifts. The ASNN scheme is employed for composing a prediction of the chemical shift from (a) the outputs from the ensemble of NNs and (b) the data in the memory. When a query proton is submitted to an ASNN, the following procedure takes place to obtain a final prediction:

(1) The descriptors of the query proton are presented to the ensemble, and a number of output values are obtained from the different NNs of the ensemble—the output profile of the query proton.

(2) The average of the values in the output profile is calculated. This is the uncorrected prediction of the chemical shift for the query proton.

(3) Every proton of the memory is presented to the ensemble to obtain an output profile.

(4) The memory is searched to find the *k* nearest neighbors of the query proton. The search is performed in the output space; i.e., the nearest neighbors are the protons with the most similar output profiles (calculated in step 3) to the query proton (calculated in step 1). Similarity is here defined as the Spearman correlation coefficient between output profiles.

Table 1. Mean Absolute Errors (ppm) of the Predictions of ^1H NMR Chemical Shifts for the Prediction Set, Obtained by Feed-Forward Neural Networks (FFNN), Associative Neural Networks (ASNN), ACD, and ChemDraw

class (no. of cases)	FFNN	ASNN[a]	ACD	ChemDraw
aromatic (247)	0.26	0.18	0.17	0.27
nonaromatic π (93)	0.35	0.19	0.21	0.22
nonrigid aliphatic (375)	0.19	0.13	0.13	0.19
rigid aliphatic (237)	0.43	0.30	0.33	0.44
total (952)	0.29	0.19	0.20	0.27

[a] The memory of the ASNNs consisted of 1287 aromatic, 645 nonaromatic π, 1704 nonrigid aliphatic, and 1043 rigid aliphatic protons chemical shifts.

(5) For each of the KNN protons, a prediction (uncorrected) is also obtained—the average of its output profile.

(6) The uncorrected predictions for the KNN protons (calculated in step 5) are compared with the experimental chemical shifts. The mean error is computed.

(7) The mean error computed in step 6 is added to the uncorrected prediction of the query proton (computed in step 2) to yield the corrected prediction of the chemical shift for the query proton.

Because stereochemistry was not assigned in the data sets used as memory, and also in the prediction set, the experimental chemical shifts for the two diastereotopic protons of CH_2 groups were averaged. The predictions obtained for two diastereotopic protons were also averaged.

The parameter *k* was empirically determined for each class of protons. It is the value that gives rise to the best predictions for the protons *of the memory*, calculated with the leave-one-out procedure. The following values of *k* were obtained: 11 for aromatic protons, 2 for nonaromatic π protons, 9 for rigid aliphatic protons, and 7 for nonrigid aliphatic protons.

For our investigations of associative neural networks, the ASNN program[4] was kindly provided by Dr. Igor Tetko.

Retraining of Feed-Forward Neural Networks. The same network architectures, the same selected variables, and ensembles with the same sizes were used as those were found to be optimal in the studies with smaller training sets described in the preceding paper in this issue. The networks were now trained with the new and the old data combined. Previous to the training, the training set was randomly split on equal-sized cross-validation and training sets.[9] The ASNN program[4] was used to train the networks, with the Levenberg–Marquardt algorithm.[10] Again, four different ensembles were trained with protons belonging to the four classes of protons. After training, the networks were tested with the prediction set.

Additionally, the networks were trained not with the entire data sets, but with subsets of these, chosen in order that no region of the diversity space was over-represented. This was done for each class of protons, by mapping the initial data set of protons, represented by their descriptors, on a Kohonen self-organizing map (SOM), and then taking randomly one proton from each of the occupied neurons.

RESULTS AND DISCUSSION

Incorporation of the new data into associative neural networks, previously trained with small data sets, resulted in a dramatic improvement of the predictions (Table 1). A

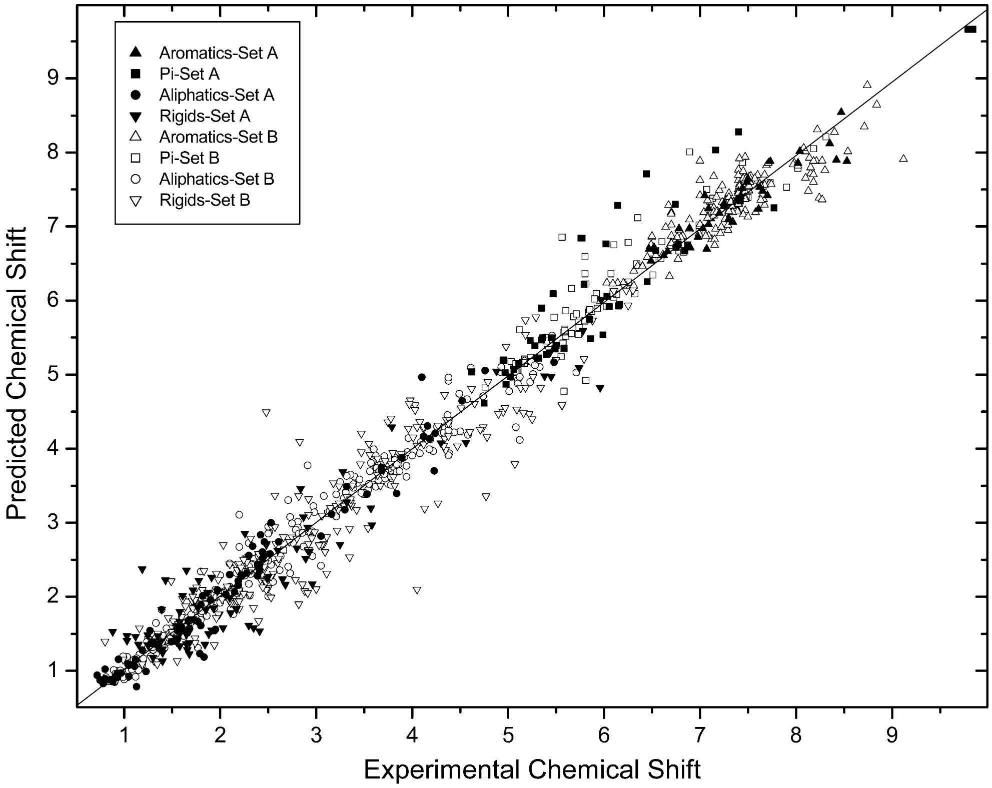

Figure 1. Experimental ^{1}H NMR chemical shifts vs predictions from ASNNs for the prediction set (set B) and for a smaller independent prediction set used in the preceding paper[1] (set A). ($R^2 = 0.9830$).

mean absolute error of 0.19 ppm was obtained for the entire prediction set (Figure 1) and 0.13 ppm for 90% of the cases. Comparing our results with those obtained with ACD I-Lab[3] in August 2003, and with ChemDraw 7,[11] for the same prediction set of 952 protons, the neural networks and the ACD software show a comparable level of accuracy, with neural networks showing a particular advantage in the class of rigid aliphatic protons. One reason for this result may reside in our geometrical descriptors. The predictions obtained by ChemDraw consistently came behind the other two methods. These results cannot be seen as a definitive comparison of the methods, but it can surely give some indications. It is hard to distinguish the quality of a method itself from the data on which it was based. Furthermore, an ideal test set should include only examples that were new to all the methods. And the population of each type of proton in such a set would require several subjective decisions. In any case, the results obtained by the neural networks compared with ACD are quite remarkable since they are based on a lot less data (the ACD database includes more than one million chemical shifts[3]), and we observed that 8% of the cases in the prediction set were present in the ACD database.

In terms of speed, the entire procedure from the connection table (.mol file) to the final predictions of chemical shifts using the ASNNs took ca. 1 s/structure. The structures had on average 17 hydrogen atoms attached to carbon atoms. The speed was measured with a nonoptimized system of different software pieces, on a PC equipped with a 1.6 GHz Celeron CPU and running a Linux kernel 2.4 operating system.

Another important feature of ASNN is the possibility to report which protons were used to compose the prediction. The result from submitting a proton to an ASNN is the prediction of its chemical shift, *and its* k *nearest-neighbors in the memory* (whose experimental chemical shifts were involved in the final prediction). This can be seen as a partial explanation of the prediction. For example, inspection of the top five outliers of Figure 1 showed that these predictions were always related to one of the following situations: (a) there was an absence of similar protons in the available data (quantitatively measured by the Spearman correlation coefficient between output profiles), (b) there were doubtful assignments of experimental chemical shifts, or (c) only one of the k nearest neighbors was in fact similar to the query proton and its contribution to the prediction was diluted with the other $k - 1$ neighbors.

Interesting results were obtained by *retraining* the neural networks with more data (Table 2). The predictions were clearly improved by comparison to the networks trained with the small data sets, but some precautions were required. The networks simply trained with all the available data consistently performed worse (Table 2, first column). This should be due to unbalanced distribution of different types of protons in the training set, which results in focusing the learning of the networks on certain types of objects—stacking in non-optimum solutions. The only case that did not show such behavior was the class of nonaromatic π protons, which significantly has fewer examples in the available data. A possible way to avoid this problem is to retrain the neural networks using (as starting point) weights calculated[1] and saved for the smaller data set. In fact such an experiment resulted in a significant improvement (Table 2, second column).[12] Another successful strategy was to carefully select a balanced training set from all the available data **(**Table 2, third column). These predictions could be even further

Table 2. Mean Absolute Errors (ppm) of the Predictions of ^{1}H NMR Chemical Shifts for the Prediction Set Obtained by FFNNs after Retraining with the New Available Data

	retrained with all available data		retrained with selected data[a] (starting from random weights)	
class (no. of cases)	starting from random weights	starting from old weights[b]	no ASNN	ASNN[c]
aromatic (247)	0.32	0.18	0.21	0.17
nonaromatic π (93)	0.19	0.19	0.24	0.19
nonrigid aliphatic (375)	0.54	0.15	0.16	0.14
rigid aliphatic (237)	0.57	0.25	0.28	0.26
total (952)	0.46	0.19	0.21	0.18

[a] The training set was obtained by mapping the protons of the new available data on the surface of a Kohonen map and taking one proton from each neuron (see Methodology). [b] The weights of the networks trained with smaller data sets[1] were used as the starting point (see Methodology). [c] The ensemble of FFNNs was incorporated in an ASNN with the memory containing all the available data (see Methodology).

improved if the corresponding networks were integrated in an ASNN using all the available data as memory (Table 2, last column). These are important results for the design of a system that is able to incorporate data from the user. The best overall results were obtained with networks retrained with selected data and integrated into ASNNs, or with networks retrained with all available data on top of previously calculated weights. Both results are practically the same as those obtained by the ASNNs simply with the networks trained on the basis of the smaller data sets and a memory with all the available data.

CONCLUSIONS

The amount of available experimental data played a decisive role in the quality of the predictions obtained by neural networks. By incorporating new experimental data in a memory of associative neural networks, the predictions were significantly improved, even without retraining the networks. Improved results could also be achieved after retraining the networks with the new data, but that approach required either careful selection of the training set or starting the training with the weights of the networks trained on the basis of the old smaller data sets. The excellent mean average errors are close to the experimental error—it is not uncommon to find data on chemical shifts for the same compound in the same solvent, but from different literature sources, with differences as large as 0.15 ppm.

The accuracy of the predictions for a test set of 952 cases was better for the neural networks than for any of the commercial software packages tried. This does not allow concluding that one method is generally better than the others. They should rather be viewed as complementary to each other. In this paper we emphasize the estimation of chemical shifts. If, for example, speed is a criterion, the ChemDraw method (a rule-based system) has an advantage compared to database-centered methods.

The descriptors of the hydrogen atoms used in our studies were revealed to be particularly suitable for the problem. These descriptors created the framework not only for training neural networks but also even for implementing associative neural networks, a task that was not planned at the time of their design.

ACKNOWLEDGMENT

The authors thank Dr. Igor Tetko who made available the ASNN program. The authors are also indebted to Prof. Johann Gasteiger for his continuous support, and access to software developed by his research group. Y.B. acknowledges Fundação para a Ciência e Tecnologia (Lisbon, Portugal) for a postdoctoral grant under the POCTI program (SFRH/BPD/7162/2001). Chemical Concepts GmbH is acknowledged for making available experimental data.

REFERENCES AND NOTES

(1) Binev, Y.; Aires-de-Sousa, J. Structure-Based Predictions of ^{1}H NMR Chemical Shifts Using Feed-Forward Neural Networks. *J. Chem. Inf. Comput. Sci.*, preceding paper in this issue.
(2) Aires-de-Sousa, J.; Hemmer, M.; Gasteiger J. *Anal. Chem.* **2002**, *74*, 80−90.
(3) Advanced Chemistry Development, Inc., http://www.acdlabs.com/.
(4) Tetko, I. V. Neural Network Studies. 4. Introduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717−728.
(5) Tetko, I. V.; Tanchuk, V. Y. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136−1145.
(6) For an example of this procedure in another application, see: Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429−434.
(7) Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S. On the Use of Neural Network Ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903−911.
(8) Dietterich, T. G. Ensemble Learning. In *The Handbook of Brain Theory and Neural Networks*; Arbib, M. A., Ed.; MIT Press: Cambridge, MA, 2002; pp 405−408.
(9) Tetko, I. V.; Livingstone, D. J.; Luik, A. I. Neural-Network Studies. 1. Comparison of Overfitting and Overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826−833.
(10) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: New York, 1994; p 998.
(11) CambridgeSoft Corp., http://www.cambridgesoft.com.
(12) The authors thank one reviewer for this suggestion.

CI034229K