Poster Presented at ENC 2006, Pacific Grove, CA, April 27, 2006

The Effect of Structure Description Schemes on Chemical Shift Prediction by Incremental and Neural Network Approaches

Yegor D. Smurnyy, Kirill A. Blinov,
Brent A. Lefebvre, and Antony J. Williams
Advanced Chemistry Development, Inc. (ACD/Labs),
Toronto, ON, Canada

INTRODUCTION

Accurate and fast chemical shift prediction is a significant task that has been challenging chemists for over twenty years. However, the investment in chemical shift prediction algorithms can benefit many areas, including chemical structure verification, structure elucidation, and education.

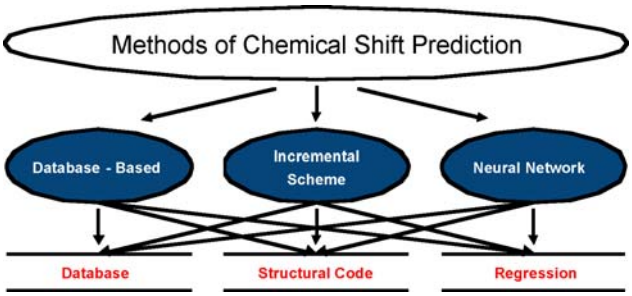

Figure 1: *The three major approaches to chemical shift prediction (Blue) and the three major components of these algorithms (Red).*

The three main methods that have emerged to produce chemical shift predictions are the following:

1. **Database-Based** methods rely on a wide collection of previously acquired assigned spectra. A user's structure is compared with references in the database and the result is the weighted average of chemical shifts from chemically similar structures.
2. **Incremental Schemes** split the structure around a particular nucleus into the set of pre-defined "increments". These are defined as chemically meaningful groups of atoms. Each of these groups is assigned an increment value and the chemical shift prediction result is the sum of these values.
3. **Neural networks** consist of artificial neurons. Each neuron is a function that non-linearly transforms the sum of input variables. The neural network is trained with a database of assigned chemical shifts. The result is called the "neuron output" and is treated either as the final prediction result or as an input to a peer neuron.

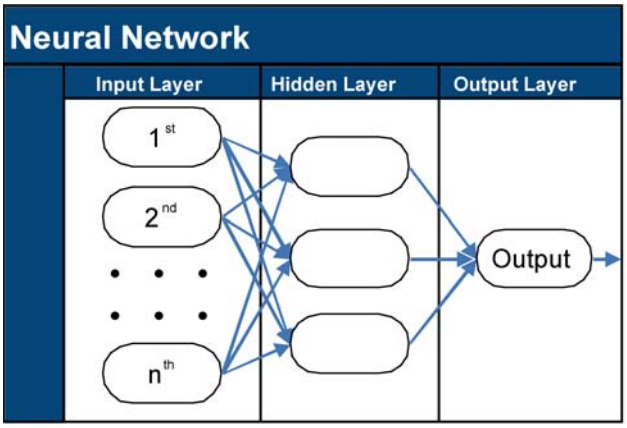

Figure 2: *The structure of a simple neural net. The input layer is fed with N inputs, then the values are transformed by the hidden layer and the output neuron produces the final output value.*

All of these methods use three main components:

1. A Database is the set of known shifts needed to either train a neural net or estimate incremental values. In the database-based approach, this is the most important part of the algorithm.
2. Structural Code is the way to convert a chemical structure into a set of numbers. In the database approach, HOSE codes are traditionally used. In the rules-based and

neural net algorithms, increment groups are counted and the result is considered to be the numerical description of a chemical topology.

3. Regression is the mathematical technique used to calculate the output from the structural code. In rules-based schemes, the result depends linearly on the quantity of a particular substituent. In neural nets, the result is obtained from weight coefficients assigned to particular neurons.

EXPERIMENTAL DETAILS

We have focused here on neural networks and incremental approaches. The third method, database-based, was not included in the comparison and evaluation of algorithms for this study.

Database

In the current work, two databases were used—the first was used for neural net training and incremental scheme regression and consists of 190,000 unique structures and nearly 2 million ^{13}C chemical shifts; the second was used for validation of the algorithms, and is smaller with only 8,500 structures and 118,000 chemical shifts.

Structural Code

In general, two techniques can be employed to split a structure into smaller pieces ("incremental groups"). One of them divides the molecule into "big" parts, each of which is a substituent in the traditional chemical meaning, i.e., nitro group, phenyl ring, double bond, etc. Typically, several thousand individual increments are required to describe 99% of the molecules found in a substantial structure database.

An alternative way is to shrink the size of substituents to only a single atom. In the current study, we use only 32 main types of atoms. The neighborhood of the atom is divided into several spheres (3-6), each containing all of the atoms separated from the center by that number of covalent bonds. We then compared our results to those previously obtained with sophisticated encoding schemes and saw no principal difference. We found it necessary to add "cross-increments" to improve the quality of prediction by the rules-based approach. These cross-increments are included in the encoding if two increments are present simultaneously in the vicinity of the atom.

To account for the cis/trans geometry of a double bond, the third sphere was extended by three to treat double bonds of each geometry differently.

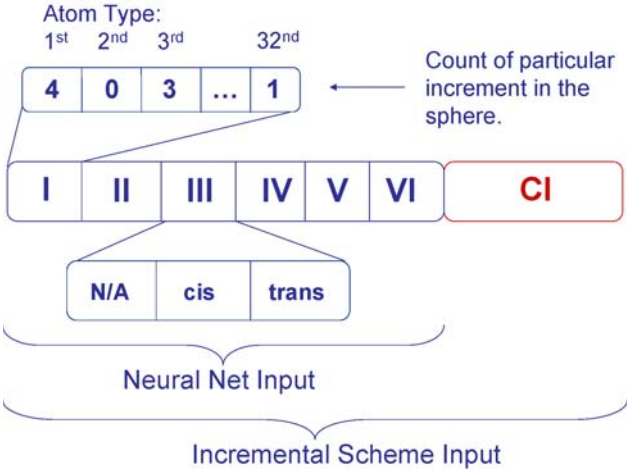

Figure 3: *The hierarchical structure of the input vectors used in the current study. Spheres are numbered with Roman numerals, each consisting of 32 cells filled with counts of the substituents. The third sphere is expanded into three to take into account the double bond geometry. CI stands for "Cross-Increments". These are additional inputs used for the rules-based calculations.*

Computation Details

All calculations were performed on a PC workstation with a 3.2 GHz Intel processor, 1 GB of RAM, running Windows 2000. Programming was done in Delphi 5.0, with the LAPACK library. Training of neural net on the whole database takes approximately 40-50 hours.

RESULTS

Splitting the Database

It is difficult to design a single neural network or incremental scheme that can cover the entire chemical shift range of 0-200 PPM typically found in ^{13}C NMR spectra. For this reason, the

whole database was split into several smaller ones, according to the nature of the central atom. Here we report the results obtained when splitting the single database into 6 and 15 sub-databases.

Table 1: *The results obtained by calculations on a neural network (3 hidden layers with 100, 25, and 5 neurons, respectively) using 6 and 15 sub-databases. Errors given are the Mean Errors in PPM on the test data. The "Het" designation denotes a heteroatom attached to the central atom.*

6 Centers		15 Centers		
Center	Error	Center	Error	Mean Error
Alkene	2.53	Alkene C	3.51	2.43
		Alkene CH	2.88	
		Alkene CH2	1.74	
		Alkene CO	1.45	
Alkyne	2.43	Alkyne	3.03	3.03
Aliphatic	2.01	Aliphatic C	2.25	1.81
		Aliphatic CH	2.58	
		Aliphatic CH2	1.77	
		Aliphatic CH3	1.20	
Aliphatic (Het)	2.00	Aliphatic C (Het)	2.36	1.94
		Aliphatic CH (Het)	2.17	
		Aliphatic CH2 (Het)	1.59	
Aromatic	1.69	Aromatic C	2.11	1.77
		Aromatic CH	1.54	
Aromatic (Het)	1.99	Aromatic C (Het)	1.98	1.98

The Optimal Number of Sub-Databases

From Table 1 it could be concluded that splitting the data into either 6 or 15 sub-databases doesn't make a significant difference since the results are very similar. Not shown are the results from splitting the database into an even greater number of sub-databases (up to 118). However, this turned out to be a losing strategy. In these cases, the algorithm had too little data to learn from and was unable to predict accurately outside of the test set, which greatly increased the prediction errors.

The Hardest to Predict Chemical Shifts

Our practice shows that multiple factors can produce a poor prediction of a particular chemical shift. Specifically, atoms that are part of conjugated systems will exhibit unusual chemical shifts which are hard to predict. Charged groups with many attached heteroatoms are also difficult to predict. Typically, these difficulties are handled by using large neural networks with a large number of considered spheres (5-6). A structure that represents one of the most challenging examples from our database is shown in Figure 4.

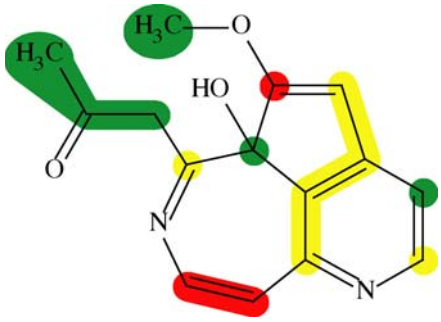

Figure 4: *An example of a structure with poorly predicted chemical shifts. Atoms with chemical shift prediction errors of more than 15 PPM are highlighted in red, more than 3 PPM in yellow, and less than 3 PPM in green. Atoms without ^{13}C chemical shifts are not colored.*

Neural Net Topology's Influence on Quality

During this study, we examined a number of neural network topologies. In most cases, we implemented nets with between one to four hidden layers. In all of the nets, all the neurons in a particular layer receive input from every neuron in the previous layer, i.e., the layers are fully linked with each other.

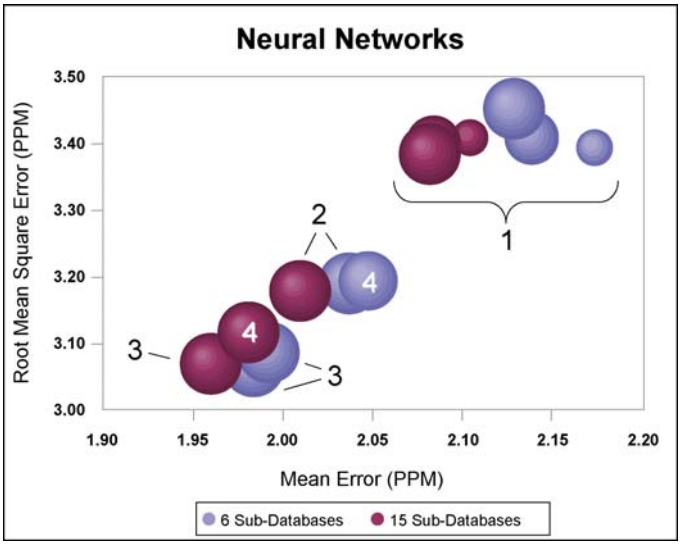

Figure 5: *Mean Errors vs. the Root Mean Square Errors, in PPM, of the test data for different neural net topologies. The numbers on the graph denote the number of hidden layers in the network. The size of the bubbles indicate the number of neurons in total (50 being the smallest and 130 the largest). From the data here, we can conclude that an optimal network consists of 120-150 hidden neurons, arranged in 3 layers. The most favorable results are also produced when 70-80% of neurons are in the first hidden layer. The number of sub-databases (6 or 15) did not seem to influence the prediction accuracy significantly.*

Neural Networks vs. Incremental Schemes and Database-Based Approaches

Table 2: *Comparison of the best results obtained with the four approaches—database-based, incremental (with multi-atomic increment), incremental with the same input as used for neural net, and neural net.*

Algorithm Approach	Mean Error (PPM)	Mean Square Error (PPM)
Database-based	2.08	3.75
Increments (multiatom)	2.46	3.51
Increments (the same input as for neural net)	1.81	2.69
Neural net (200-25-5 hidden neurons)	1.94	3.03

CONCLUSIONS

The results of the study show that both neural networks and incremental schemes can serve as fast and robust algorithms for chemical shift prediction. In the current work, we implemented a simple structure encoding routine as input for a neural net or rules-based algorithm. This encoding scheme is fast, yet powerful enough to handle minor details of the chemical environment. Despite the simple design of the neural net and incremental scheme, the results obtained are of the same or better quality than those obtained with the database-based approach. Such a finding suggests that future efforts should be focused more on developing the structural code to readily and precisely convert chemical topology into a set of numbers, as opposed to focusing on the details of the neural network.