ELSEVIER

Carbohydrate Research 271 (1995) 65–77

CARBOHYDRATE RESEARCH

Simulation of the ^{13}C nuclear magnetic resonance spectra of trisaccharides using multiple linear regression analysis and neural networks

Deborah L. Clouser, Peter C. Jurs *

Department of Chemistry, 152 Davey Laboratory, The Pennsylvania State University, University Park, PA 16802, USA

Received 26 October 1994; accepted 19 December 1994

Abstract

Predictive models are developed for the ^{13}C NMR chemical shifts of the carbon atoms comprising the central rings of 46 trisaccharide compounds. Thirty-nine trisaccharides are used as a training set for development of models using regression analysis and computational neural networks, and seven compounds are used as an external prediction set. The descriptors used in the models are developed directly from the molecular structures of the trisaccharides. Three different methods of descriptor selection are compared. The dependence of the models on the geometries of the trisaccharides is explored. The models developed with geometric descriptors are better than those developed without geometric descriptors, although the latter models are still of a comparable quality. Overall, the best model found is a neural network based on descriptors selected by multiple linear regression.

Keywords: NMR spectroscopy; Trisaccharides; Multiple linear regression analysis; Neural networks

1. Introduction

Carbohydrates comprise an important class of chemical compounds, playing many important roles, such as the source of energy for mammals and the primary structure of plants [1]. ^{13}C NMR spectroscopy has proven itself as one of the most useful techniques for elucidation of the often complex conformation of these compounds [2]. Used in conjunction with structure generation programs, the simulation of these shifts can be

* Corresponding author.

SSDI 0008-6215(95)00051-8

used to judge the quality of the generated structures. In this paper, the shifts of some trisaccharides are simulated and compared to the observed shifts.

Empirical modeling is a method that can be used to simulate the ^{13}C NMR chemical shifts of compounds. This method uses linear models developed from a set of compounds with known chemical shifts. These models relate the chemical shift to a set of atom-based descriptors using the following equation

$$[S = b_0 + b_1 D_1 + b_2 D_2 + \ldots + b_n D_n]$$

where S is the chemical shift of the carbon, b_n is the coefficient as determined by multiple linear regression, and D_n is the value of an atom-based descriptor. After these models have been developed, they can be used to predict the shifts of carbons in external compounds, or compounds not used in model development.

Another method that can be used to simulate the ^{13}C NMR spectrum of a compound is the computational neural network. A quasi-Newton training algorithm based on the work of Broyden, Fletcher, Goldfarb, and Shanno [3–8] is used in this work, as it has been incorporated into our methodology in preference over the more common back-propagation training algorithm [9]. The quasi-Newton algorithm has been shown to give slightly better results and trains more quickly than back-propagation [10].

In most modeling studies that use atom-based descriptors, three types of descriptors are used: topological, electronic and geometrical. As will be discussed in the following sections, the models developed and their results are somewhat dependent on the geometry of the trisaccharides used in this data set. The results of models built without geometric descriptors will also be presented.

In this study, only the shifts for carbon atoms comprising the central ring of each trisaccharide are being investigated. This limitation was imposed for several reasons. First, the ^{13}C NMR chemical shift of a carbon is sensitive to the chemical environment as much as four or five bonds distant, so using trisaccharides provides sufficient surroundings only for the carbons of the central ring. Second, this study is our first attempt to model the ^{13}C NMR shifts of saccharides that could be useful for larger structures, and trisaccharides could be snipped from longer oligosaccharide chains for prediction.

2. Experimental

The software used in this work is part of the ADAPT software package [11,12]. This software is installed on a Sun 4/110 workstation in operation at the Pennsylvania State University. The neural network software, the quasi-Newton algorithm [13], is installed on a DEC 3000 AXP 500 workstation running field-test software.

Data set. — The trisaccharide compounds used for this study are compiled from a number of references [14–19] and are listed in Table 1. All rings are pyranoses. The ^{13}C NMR spectra for these compounds were recorded at 30–40°C. For refs [14–19], the shifts at 30°C were calculated from the information given in the paper. Compounds **1–39** were used as a training set, the set of compounds with which the regression models were built. Compounds **40–46** were used as an external prediction set, which

Table 1
List of compounds

1	*O*-α-D-mannopyranosyl-(1 → 2)-*O*-α-D-mannopyranosyl-(1 → 2)-α-D-mannopyranose
2	*O*-α-D-mannopyranosyl-(1 → 3)-*O*-α-D-mannopyranosyl-(1 → 2)-α-D-mannopyranose
3	*O*-β-L-fucopyranosyl-(1 → 4)-*O*-[β-L-fucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
4	*O*-α-L-fucopyranosyl-(1 → 4)-*O*-[α-L-fucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
5	*O*-α-L-fucopyranosyl-(1 → 4)-*O*-[α-L-fucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
6	*O*-α-D-glucopyranosyl-(1 → 4)-*O*-[α-D-glucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
7	*O*-β-D-glucopyranosyl-(1 → 4)-*O*-[α-D-glucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
8	*O*-β-D-glucopyranosyl-(1 → 4)-*O*-[β-D-glucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
9	*O*-α-D-glucopyranosyl-(1 → 4)-*O*-[β-D-glucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
10	*O*-α-D-glucopyranosyl-(1 → 4)-*O*-[α-L-fucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
11	*O*-β-D-glucopyranosyl-(1 → 4)-*O*-[α-L-fucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
12	*O*-α-D-glucopyranosyl-(1 → 4)-*O*-[β-L-fucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
13	*O*-β-D-glucopyranosyl-(1 → 4)-*O*-[β-L-fucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
14	*O*-α-L-fucopyranosyl-(1 → 4)-*O*-[α-D-glucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
15	*O*-α-L-fucopyranosyl-(1 → 4)-*O*-[β-D-glucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
16	*O*-β-L-fucopyranosyl-(1 → 4)-*O*-[β-D-glucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
17	*O*-α-D-glucopyranosyl-(1 → 2)-*O*-β-D-glucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
18	*O*-β-D-glucopyranosyl-(1 → 2)-*O*-α-D-glucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
19	*O*-β-D-glucopyranosyl-(1 → 2)-*O*-β-D-glucopyranosyl-(1 → 3)-methyl α-D-galactopyranoside
20	*O*-α-L-rhamnopyranosyl-(1 → 2)-*O*-[α-D-glucopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
21	*O*-β-D-glucopyranosyl-(1 → 2)-*O*-[β-D-galactopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
22	*O*-α-L-rhamnopyranosyl-(1 → 2)-*O*-[α-D-glucopyranosyl-(1 → 3)]-methyl α-D-mannopyranoside
23	*O*-α-D-mannopyranosyl-(1 → 2)-*O*-[α-D-glucopyranosyl-(1 → 3)]-methyl α-D-mannopyranoside
24	*O*-β-L-fucopyranosyl-(1 → 2)-*O*-[β-L-fucopyranosyl-(1 → 3])-methyl α-D-mannopyranoside
25	*O*-α-L-fucopyranosyl-(1 → 2)-*O*-[β-L-fucopyranosyl-(1 → 3])-methyl α-D-mannopyranoside
26	*O*-α-D-mannopyranosyl-(1 → 2)-*O*-[α-D-mannopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
27	*O*-β-D-glucopyranosyl-(1 → 2)-*O*-[α-D-mannopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
28	*O*-β-D-glucopyranosyl-(1 → 2)-*O*-[β-D-glucopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
29	*O*-α-D-mannopyranosyl-(1 → 2)-*O*-[α-L-rhamnopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
30	*O*-α-D-mannopyranosyl-(1 → 2)-*O*-[β-L-fucopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
31	*O*-β-D-glucopyranosyl-(1 → 2)-*O*-[α-L-rhamnopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
32	*O*-β-D-glucopyranosyl-(1 → 2)-[β-L-fucopyranosyl-(1 → 3])-methyl α-L-rhamnopyranoside
33	*O*-α-L-rhamnopyranosyl-(1 → 2)-*O*-[α-L-rhamnopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
34	*O*-α-L-fucopyranosyl-(1 → 2)-*O*-[β-L-fucopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
35	*O*-β-L-fucopyranosyl-(1 → 2)-*O*-[β-L-fucopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
36	*O*-α-L-fucopyranosyl-(1 → 2)-*O*-[α-D-mannopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
37	*O*-α-L-fucopyranosyl-(1 → 2)-*O*-[β-D-glucopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
38	*O*-β-L-fucopyranosyl-(1 → 2)-*O*-[α-D-mannopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
39	*O*-β-L-fucopyranosyl-(1 → 2)-*O*-[β-D-glucopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
40	*O*-α-D-mannopyranosyl-(1 → 6)-*O*-α-D-mannopyranosyl-(1 → 6)-α-D-mannopyranose
41	*O*-β-L-fucopyranosyl-(1 → 4)-*O*-[α-L-fucopyranosyl-(1 → 3)]-methyl α-galactopyranoside
42	*O*-β-L-fucopyranosyl-(1 → 4)-*O*-[α-D-glucopyranosyl-(1 → 3)]-methyl α-D-galactopyranoside
43	*O*-α-D-glucopyranosyl-(1 → 2)-*O*-α-D-glucopyranosyl-(1 → 3)-methyl α-D-galactopyranoside
44	*O*-α-D-glucopyranosyl-(1 → 2)-*O*-[α-L-rhamnopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
45	*O*-α-D-mannopyranosyl-(1 → 2)-*O*-[β-D-glucopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside
46	*O*-β-L-fucopyranosyl-(1 → 2)-*O*-[α-L-rhamnopyranosyl-(1 → 3)]-methyl α-L-rhamnopyranoside

was used to test the predictive ability of the regression models. The seven compounds for the external prediction set were chosen randomly from the references so that at least one compound from each reference was placed in this set so that any slight peculiarities

from reference to reference would be represented. Compound **2** was originally a tetrasaccharide. Since only the central ring of the saccharides is being predicted, this compound had to be broken down into two trisaccharides. One of the trisaccharides formed was already present in that reference, and it is compound **1**. Breaking down the tetrasaccharide into two compounds should not present a problem when predicting the chemical shifts, as chemical effects are not usually seen in atoms more than five bonds from the query carbon.

The 315 shifts of the carbon atoms in this group of 46 compounds cluster into five groups. One group, with 19 shifts from 17.60 to 18.20 ppm, consists of the shifts for C-6 on the ring when no hydroxy group is present. Another group, with 39 shifts from 55.90 to 56.40 ppm, is made up of the shifts for the carbon in methyl side chains of C-1 in the ring. A third group, with 26 shifts from 60.93 to 62.47 ppm, contains the shifts of C-6 on the ring when a hydroxy group is present. A fourth group, with 185 shifts from 66.95 to 82.42 ppm, contains the bulk of the shifts of the data set, and these are the shifts for carbons 2, 3, 4, and 5 in the ring. The final group, with 46 shifts from 97.18 to 103.68 ppm, are the shifts for C-1 in the ring. This clustering of shifts into groups presented some problems which will be discussed later.

Structure entry. — The structures for the trisaccharides used in this study were entered into the computer by sketching. This was done by first drawing the individual monosaccharides present in the trisaccharides and then modeling them using MM2 [20,21]. Once the monosaccharides were modeled into energy minimized three-dimensional conformations, they were connected to the appropriate monosaccharides in the correct linkages to form the trisaccharides. These trisaccharides were then modeled using MM2 as well.

Unique carbon atom perception. — Every atom in the central ring of the trisaccharides was unique. In this data set there were 315 carbon atoms present in the 46 central rings, so there were 315 unique carbon atoms with shifts to be predicted. Of these, there were 268 carbon atoms in the training set of compounds **1–39**, and 47 carbon atoms present in the external prediction set of compounds **40–46**. The number of atoms used in the external prediction set is somewhat greater than 10% of the number of atoms in the training set, which is the number often used as a guideline as to how large the external prediction set should be. More atoms were preferred for this external prediction set due to the large number of compounds being used in the data set.

Descriptor calculation. — Once the data set was defined, descriptors, which are numerical representations of the environment surrounding each carbon atom, were calculated using a number of descriptor development routines [22]. The types of descriptors calculated were topological, electronic and geometrical. Once calculated, the descriptors were then screened to insure that only meaningful descriptors were used to develop regression equations. Descriptors with high pairwise correlations, $r > 0.95$, and descriptors that contained more than 80% identical values were eliminated. A pool of 160 topological, electronic and geometrical descriptors for each carbon atom remained after the screening process. The individual descriptors that were found to be useful in this study will be in the table detailing the specific model.

Because of the clustering present in the shifts, it was noted that two groups of shifts, 17.60 to 18.29 ppm and 55.90 to 56.40 ppm, covered such narrow ranges that the two

groups of points could adversely affect the model building process. Indicator variables were developed for these two groups of points. These variables are descriptors that simply contain ones for atoms in these two groups, and zeros for atoms not in these groups. These variables were added to the above descriptor pool, and if they represent information not already encoded by some other descriptors, then they would appear in the models along with the regularly calculated descriptors. These indicator variables were also forced into any models that were developed in which they did not appear, to insure that the information contained within them was encoded by the descriptors already present.

3. Results and discussion

Multiple linear regression. — After the descriptor screening process, regression equations could be developed. The regression method found to be most effective is the leaps-and-bounds method which both selects descriptors and builds regression models [23]. The leaps-and-bounds algorithm finds the best subset of descriptors from the pool of 160 using the *R*-squared value as the criterion for selection. During the model building process a patterning of the calculated shifts was noted. Models that had the lowest overall errors, in the range of 1.00 to 1.50 ppm, had shifts that were clustered in horizontal groups along the regression line in a calculated versus observed plot. In other words, atoms having different observed shifts were calculated to have nearly identical shifts. An example of this patterning is shown in Fig. 1. The predictive model shown

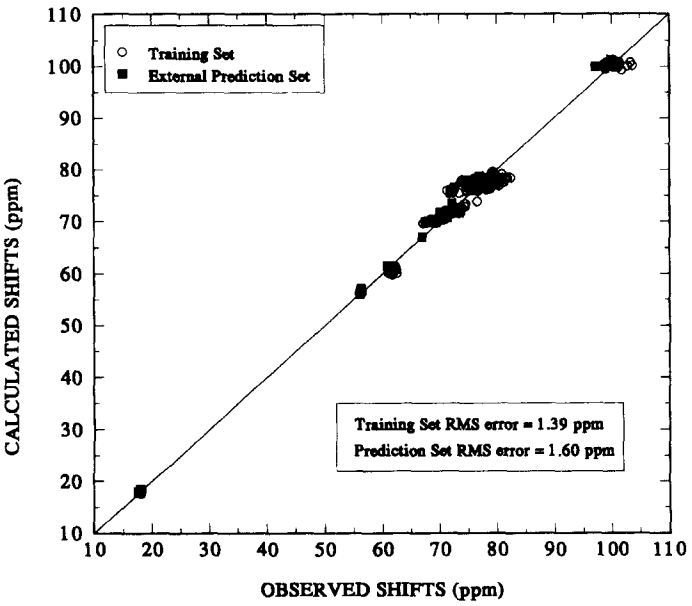

Fig. 1. An example of the data patterning encountered in this study.

Table 2
Regression equation relating 10 calculated descriptors to ^{13}C NMR chemical shifts for the trisaccharide data set

Descriptor [a]	Mean	SD [b]	Coefficient	Mean effect [c] (ppm)
ICNC 3	1.35	0.509	21.2 ± 0.83	28.6 ±1.1
ACNC 3	0.288	0.0209	−321 ± 13.2	−92.4 ±3.8
ACON 2	0.420	0.0382	89.0 ± 6.0	37.3 ±2.5
NTHC 4	−0.407	0.776	2.10± 0.17	−0.85±0.069
CHCG 1	0.491	0.224	78.5 ± 0.99	38.5 ±0.49
CO4D 2	0.0815	0.115	−8.23± 1.3	−0.67±0.10
HRD3 4	0.0513	0.0305	−28.9 ± 5.7	−1.48±0.29
HXI3 2	0.105	0.0469	−24.4 ± 5.8	−2.56±0.61
SMTA 1	0.0175	0.00201	−1260 ±150	−22.0 ±2.6
ATAS 1	0.822	0.304	−2.00± 0.39	−1.64±0.32
Intercept			88.5	
	$n = 268$		$s = 1.67$ ppm	$R = .996$
	$n = 47$		$s = 2.33$ ppm	$R = .992$

[a] Descriptor definition ("heavy atom" denotes all non-hydrogen atoms). ICNC 3, the corrected connectivity index over bonds three bonds away from the carbon center; ACNC 3, the average corrected connectivity index over bonds three bonds away from the carbon center; ACON 2, the average connectivity index over bonds two bonds away from the carbon center; NTHC 4, the sum of the extended Hückel charges for all heavy atoms four bonds away from the carbon center; CHCG 1, the extended Hückel charge on the carbon center; CO4D 2, the inverse throughspace distance from the carbon center to carbons two bonds away with the bonding configuration of two heavy-atom connections with two single bonds; HRD3 4, the sum of the inverse cubed throughspace distance for the carbon center to hydrogens attached to carbons two bonds away; HXI3 2, the sum of the inverse throughspace distance from hydrogens attached to the carbon center to all heavy atoms four bonds away; SMTA 1, the smallest inverse torsional angle involving the carbon center (this is actually the largest torsional angle); ATAS 1, the average strain associated with torsional bonds involving the carbon center.

[b] SD = standard deviation.

[c] Mean effect = the average shielding and deshielding contribution of each descriptor on the predicted chemical shift.

here is an early model and was not used further. Since many models with quite low errors produced patterned plots, our search for the best models focused on those models with a balance between low error and no patterning. To achieve this, the data were divided into two subsets. One subset contained all the atoms with shifts in the range 17.60–56.40 ppm and 97.18–103.68 ppm. The second subset contained all the atoms with shifts in the range 60.93–82.42 ppm. The second subset contained the atoms whose predicted shifts produced the patterned plots. Using leaps-and-bounds regression, a model was found for each subset that did not produce patterned plots, and these descriptors were then combined to form one model. The result was a 10-descriptor model, which is described in detail in Table 2. The error for the training set is 1.67 ppm with an external prediction set error of 2.33 ppm. A plot of the calculated shifts versus the observed shifts is presented in Fig. 2. There still appears to be a small amount of patterning with the shifts between 60 and 90 ppm, but the regression equation is satisfactory. Of the 10 descriptors in the model, three are topological, two are electronic, and five are geometrical. The presence of five geometrical descriptors and two electronic

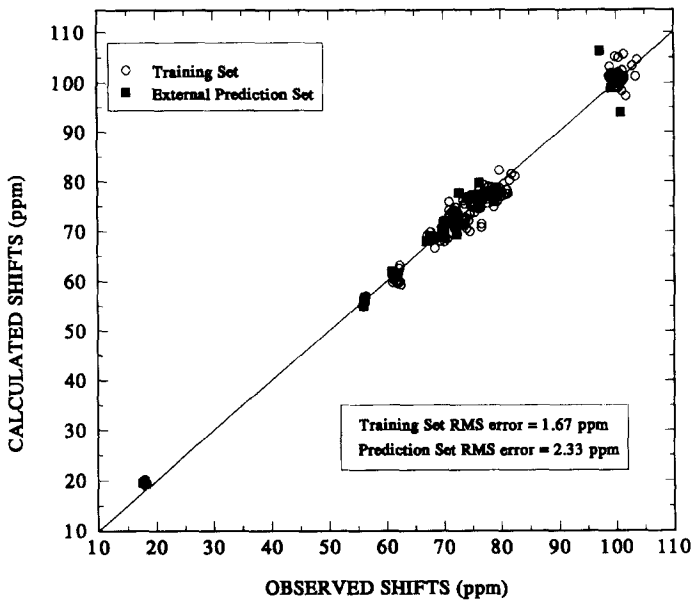

Fig. 2. Calculated versus observed plot for the leaps-and-bounds regression model.

descriptors, whose values are affected by geometry, raised some concerns about this model being overly dependent on the conformations of the trisaccharides.

In addition to using leaps-and-bounds regression analysis, two alternative methods of descriptor selection recently investigated in our group were utilized with this data set. The first method of descriptor selection used simulated annealing [24,25]. This method of descriptor selection was based on the physical process of annealing. The specific algorithm used by our group has been described in detail elsewhere [26]. The second method involved the use of the genetic algorithm [27,28]. The genetic algorithm generates subsets of descriptors and chooses the offspring with lowest RMS error based on the model of evolution. The genetic algorithm approach to descriptor selection is under further investigation in our laboratory.

Using simulated annealing, a 10-descriptor model was found. This model gave a training set error of 1.09 ppm and a external prediction set error of 1.85 ppm. This model contained three topological, four electronic, and three geometrical descriptors. Of the three topological descriptors present in this model, one was similar to a descriptor in the leaps-and-bounds model, ACNC. Two of the electronic descriptors are also similar to NTHC and one of the geometrical descriptors is identical to HRD3.

With the genetic algorithm descriptor selection program, a nine-descriptor model was found. The training set error for this model was 1.37 ppm and the external prediction set error was 1.82 ppm. This model had four topological, four electronic, and one geometrical descriptor. This model also contained descriptors similar to those found in the leaps-and-bounds model. One of the topological descriptors is similar to ACNC, one

Table 3
Summary of models developed during this study

Descriptor selection method	Regression analysis		Neural networks		
	Training set error (ppm)	External prediction set error (ppm)	Training set error (ppm)	Cross-validation set error (ppm)	Prediction set error (ppm)
A. All descriptors					
1. Leaps-and-bounds	1.67	2.33	0.75	0.83	1.37
2. Simulated annealing	1.09	1.85	0.93	0.94	1.30
3. Genetic algorithm	1.37	1.82	1.14	1.09	1.69
B. Without geometric descriptors					
1. Leaps-and-bounds	1.16	1.90	1.00	1.12	1.85
2. Simulated annealing	1.32	1.65	1.19	1.15	1.74
3. Genetic algorithm	1.36	1.67	1.19	1.18	1.61

of the electronic descriptors is identical, CHCG, and the one geometrical descriptor is similar to HRD3. The results for this model and all others are summarized in Table 3.

In an attempt to find models that were less dependent on geometry than the leaps-and-bounds model, a reduced pool of descriptors was used to develop regression equations that did not contain any geometrical descriptors. Electronic descriptors were included as their dependence on the geometry was weak enough that their use would not adversely affect the results. Using leaps-and-bounds regression analysis, a nine-descriptor model was found. This model has a lower error than the previous model with a training set error of 1.16 ppm and an external prediction set error of 1.90 ppm. The error may be lower since the criterion of having models without patterning was not as strictly followed. The calculated shifts versus the observed shifts were still checked to find a model with as little patterning as possible. As this model shows, it is possible to generate accurate models without geometric descriptors, but the model does not encode as much information about the compound as a model with geometric descriptors. However, a model without them is simpler and relatively free of the problem that the trisaccharides can assume multiple conformations.

Testing of conformational dependence. — Modeling carbohydrates such as trisaccharides is a formidable task. As stated by Brant and Christ, when modeling carbohydrates, averaging of the different conformational states is required to accurately account for observed properties [29]. Averaging of the conformations was relatively easy with small compounds such as tetrahydropyrans [30], but with the trisaccharides, this would have proven to be an extremely time-consuming task. Even without averaging the conformations the models generated are close to the 1.00 ppm error that we strive for to predict the chemical shifts. However, to check the conformational dependence of our models, a test was performed which involved changing the conformations of the trisaccharides and repeating the predictions. To perform this test, compound **3** from Table 1 was chosen. Using a molecular mechanics modeling program that supports structure manipulations, each one of the four bonds involved in the two glycosidic linkages was rotated a specific

number of degrees (30°, 60°, 90°, 120°, 150°, 180°), and the new conformation obtained from each rotation was stored. This resulted in 24 new conformations for this one compound. These conformations were not re-minimized but were left in their altered conformations. Since each one of these new conformations may give different values for the seven geometry-dependent descriptors (five geometrical and two electronic), each descriptor was recalculated for these new conformations, resulting in 24 new values for each one of the descriptors for each atom. Inspection of the values for these 24 descriptors for each atom shows that some of the seven geometry-dependent descriptors change very little and some have substantial changes. For example, for atom 3 and the descriptor SMTA 1, 13 of the values are identical and the remaining nine vary by as much as 40%. Overall, the values for these geometry-dependent descriptors are fairly strong functions of the geometry of each trisaccharide.

Using the coefficients of the original model, but substituting in the new descriptor values, each one of the shifts was predicted, resulting in the prediction of the shifts for the carbon atoms in the entire central ring. The topological descriptors were left unchanged since they are not dependent on geometrical conformations. The overall error for each of the 24 new conformations was then compared to the original prediction. The 24 new conformations yielded errors that were larger than the original error of 0.92 ppm. Some of the errors for the new predictions were close to the original error, while others were very large. The errors that were large usually were the result of one or two calculated shifts among the seven that were huge, and when averaged in with the other calculated shifts resulted in a large overall error. The unusually large errors with some of the calculated shifts can be attributed to the changes in the torsional angles. After some of the rotations, unreasonable torsional angles were present, such as for one of the 90° rotations. In the original compound, atom-3 has an observed shift of 78.07 ppm and a calculated shift of 77.90 ppm. In the new conformation after a 90° rotation, atom 3 has a calculated shift of −2518.58 ppm. This huge error is the result of two torsional angles involving atom-3. In the original compound, these angles were 150.1 and 90.5 degrees, respectively. After the rotation, these angles were 119.9 and 0.5 degrees, respectively. With a geometrical difference such as that, the descriptor values change substantially, therefore changing the prediction. However, if the compounds are modeled reasonably, then torsional angles such as these would not be present, and the predictions should be relatively accurate.

Another factor that our models may be dependent upon was the force-field method used to model the conformations of the trisaccharides. Only MM2 was used for this work. Some systematic bias due to MM2 may exist in the conformations which may add to the geometrical dependence of the models. Such bias might overshadow any error introduced into the work by not averaging the conformations.

Neural networks. — After the linear regression analyses had been completed, the descriptors in the model found were then submitted as inputs to computational neural networks. A quasi-Newton training algorithm was used. The details of the algorithm have been described in detail previously [13]. When using neural networks, the 39 compound training set used for linear regression was divided into two groups, while the external prediction set remained the same. The first group, which contains 32 of the 39 compounds used for linear regression, is still called a training set. The network is trained

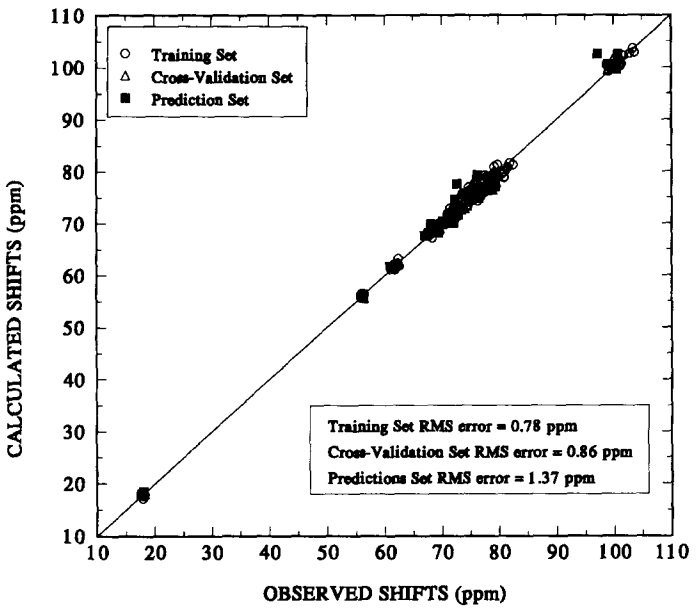

Fig. 3. Calculated versus observed plot for model formed with descriptors found using leaps-and-bounds regression and developed with a computational neural network.

on this set of compounds. The second group, which contains 7 compounds (**3, 8, 12, 13, 22, 27** and **39** from Table 1) is called the cross-validation set. Seven compounds were chosen for the cross-validation set so the number of atoms in the cross-validation set, 49, was similar to the number of atoms in the external prediction set, 47. The cross-validation set is used to avoid overfitting the data using the neural network. Periodically during training the neural network's adjustable parameters (weights and biases) are fixed and the carbon atoms of the cross-validation set are predicted. Then training resumes with the training set compounds. The improvement in RMS error for the training set and the cross-validation set are noted as a function of training effort. When the error of the cross-validation set fails to improve, the network is beginning to encode information specific to the individual training set compounds. This is called overfitting the data, which leads to a decrease in its external predictive ability. Therefore, training of the network is stopped when the cross-validation set error fails to improve.

The first model submitted to neural networks was the model found using leaps-and-bounds regression analysis. The network architecture used was 10-5-1, 10 input neurons, 5 hidden layer neurons, and 1 output neuron. This network gave a training set error of 0.75 ppm and a cross-validation set error of 0.83 ppm. The prediction set, which contains the same atoms as the external prediction set of regression, had an error of 1.37 ppm. The results for this network are shown in Fig. 3. The shifts of the anomeric carbons as predicted by this network gave an average error of 0.48 ppm for the training set, 0.23 ppm for the cross-validation set and 1.19 ppm for the prediction set. The error for the prediction set anomeric carbons is high due to the presence of one outlier. An

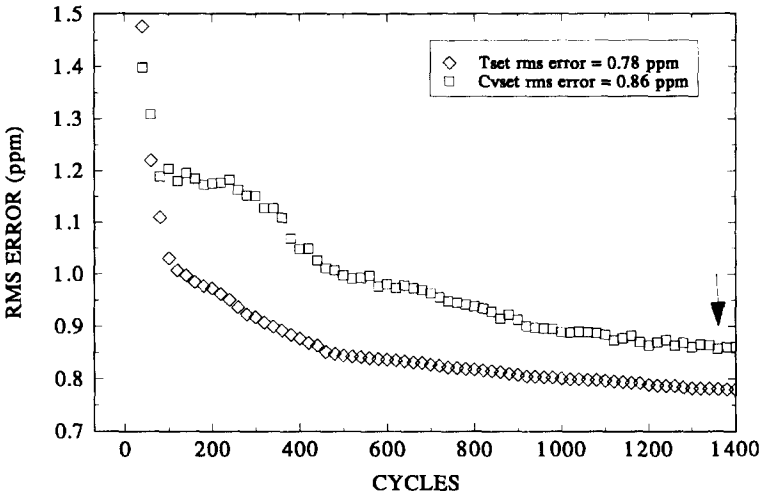

Fig. 4. Computational neural network training curve.

illustration of how the cross-validation set is used is shown in Fig. 4. During training, the RMS error for the training set is seen to fall continuously, as it must due to the mathematics of training. The RMS error of the cross-validation set is not so smooth, but it decreases overall. Eventually, the cross-validation set error reaches a minimum and begins to rise. This is the point at which training is terminated. When compared to the regression results, one can see that the neural networks have given drastically improved results, as has often been the case in previous studies. Because of the improvement shown by the neural networks, it has been demonstrated that having the goal in linear regression be a model with more consistent calculated shifts and non-patterned data instead of a model with low error and patterned data was justified and the model chosen was reasonable.

The models from the other two methods of descriptor selection were also submitted to neural networks. The atoms contained in the cross-validation set and the prediction set are the same for all three methods, leaps-and-bounds, simulated annealing, and the genetic algorithm. The 10-descriptor simulated annealing model used a 10-5-1 network and gave a training set error of 0.93 ppm, a cross-validation set error of 0.94 ppm, and a prediction set error of 1.30 ppm. The descriptors selected with the genetic algorithm gave a training set error of 1.14 ppm, a cross validation set error of 1.09 ppm and a prediction set error of 1.69 ppm. A summary of the neural network results is shown in Table 3. As compared to the leaps-and-bounds model, the error does not seem to improve very much for the two alternative descriptor sets, though it does improve more for the simulated annealing method than the genetic algorithm. Why one of these methods improves more than the other is not known, but it is probably simply the individual descriptors present in the model.

It was noticed in the beginning of this work that models with patterned data did not improve as much as models with less patterned data when they were submitted to neural

networks. This is shown above, where the leaps-and-bounds model improves greatly with neural networks, while the more patterned simulated annealing and genetic algorithm models do not improve as much. In fact, in regression these two different methods gave slightly better errors than leap-and-bounds, but after neural networks, the leaps-and-bounds model has the better overall RMS error. It is not clear why this is so. It may be that the descriptors in the patterned models have very little non-linear information contained within them, limiting their improvement in neural networks.

Also submitted to neural networks were the models developed without geometric descriptors. The leaps-and-bounds model improved to errors of 1.00, 1.12, and 1.85 ppm for the training set, cross-validation set, and prediction set, respectively. The simulated annealing model improved to 1.19, 1.15, and 1.74 ppm, while the genetic algorithm model improved to 1.19, 1.18, and 1.61 ppm. These results are also summarized in Table 3. The improvement of these models seems to be limited by the above factors as well.

4. Conclusions

The ^{13}C NMR chemical shifts of the central-ring carbon atoms of 46 trisaccharides are predicted using linear regression and computational neural networks. The results of two different descriptor selection processes are also reported. These two methods give results that are overall not as accurate as the conventional method. Regression equations were also developed without the use of geometrical descriptors. The models generated are free of the conformational difficulties that we have shown are present. Even though we have shown that the shift predictions are dependent on the conformations of the trisaccharides, models with geometric descriptors are more robust. By not using the geometric descriptors the models are almost too simple, and not taking a large part of the information contained within the structure of the trisaccharide into consideration. If carbohydrates are modeled systematically, then the presence of the geometric descriptors should not pose a problem. The models found during this work are meant to be an example of the types of models that can be used to predict the carbon shifts of carbohydrates, but they may not be the only acceptable models.

Acknowledgments

The authors wish to thank A.J. Benesi and S.L. Dixon for their help with this work.

References

[1] R.W. Binkley, *Modern Carbohydrate Chemistry*, Marcel Dekker, New York, 1988, pp 1–4.
[2] E. Breitmaier and W. Voelter, *Carbon-13 NMR Spectroscopy*, VCH, Weinheim, 1987, p 379.
[3] C.G. Broyden, *J. Inst. Maths Appl.*, 6 (1970) 76–90.
[4] C.G. Broyden, *J. Inst. Maths Appl.*, 6 (1970) 222–231.

[5] R. Fletcher, *Comp. J.*, 13 (1970) 317–322.
[6] R. Fletcher, *Practical Methods of Optimization*, Vol. 1. Wiley, New York, 1979, pp 83–90.
[7] D. Goldfarb, *Math. Comp.*, 24 (1970) 23–26.
[8] D.F. Shanno, *Math. Comp.*, 24 (1970) 647–656.
[9] J.L. McClelland and D.E. Rumelhart, in *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs and Exercises*, MIT, Cambridge, MA, 1986.
[10] J.W. Ball and P.C. Jurs, *Anal. Chem.*, 65 (1993) 3615–3621.
[11] A.J. Stupper, W.E. Brugger, and P.C. Jurs, in *Computer-Assisted Studies of Chemical Structure and Biological Function*, Wiley-Interscience, New York, 1979, pp 83–90.
[12] P.C. Jurs, J.T. You, and M. Yuan, in E.C. Olsen and R.E. Christoffersen (Eds), *Computer-Assisted Drug Design*, American Chemical Society, Washington, D.C. 1979, pp 103–129.
[13] L. Xu, J.W. Ball, S.L. Dixon, and P.C. Jurs, *Environ. Toxicol. Chem.* 13 (1994) 841–851.
[14] A. Allerhand and E. Berman, *J. Am. Chem. Soc.*, 106 (1984) 2400–2412.
[15] H. Baumann, B. Ebring, P.-E. Jansson, and L. Kenne, *J. Chem. Soc., Perkin Trans. 1*, (1989) 2153–2165.
[16] H. Baumann, B. Ebring, P.-E. Jansson, and L. Kenne, *J. Chem. Soc., Perkin Trans. 1*, (1989) 2167–2178
[17] A. Adeyeye, P.-E. Jansson, L. Kenne, and G. Widmalm, *J. Chem. Soc., Perkin Trans. 2*, (1991) 963–973.
[18] N.E. Nifant'ev, A.S. Shashkov, G.M. Lipkind, and N.K. Kochetkov, *Carbohydr. Res.*, 237 (1992) 95–113.
[19] N.K. Kochetkov, G.M. Lipkind, A.S. Shashkov, and N.E. Nifant'ev, *Carbohydr. Res.*, 221 (1991) 145–168.
[20] U. Burkert and N.L. Allinger, *Molecular Mechanics*, ACS Monograph 177, American Chemical Society, Washington, D.C., 1982.
[21] T.A. Clark, *A Handbook of Computational Chemistry; A Practical Guide to Chemical Structure and Energy Calculations*, Wiley-Interscience, New York, 1985.
[22] P.C. Jurs, G.P. Sutton, and M.L. Ranc, *Anal. Chem.*, 61 (1989) 1115A–1122A.
[23] G.M. Furnival and R.W. Wilson, Jr., *Technometrics*, 16 (1971) 499–511.
[24] I.O. Bohachevsky, M.E. Johnson, and M.L. Stein, *Technometrics*, 28 (1986) 209–217.
[25] J.H. Kalivas, *J. Chemometrics*, 5 (1991) 37–48.
[26] J.M. Sutter and P.C. Jurs, *J. Chem. Inf. Comput. Sci.*, in press.
[27] D.B. Hibbert, *Chemo. Intell. Lab. Syst.* 19 (1993) 277–293.
[28] C.B. Lucasius and G. Kateman, *Chemo. Intell. Lab. Syst.* 19 (1993) 1–33.
[29] D.A. Brant and M.D. Christ, A.D. French and J.W. Brady (Eds), *Computer Modeling of Carbohydrate Molecules, ACS Symp. Ser.* 430, American Chemical Society Washington, D.C. 1990, pp 43–68.
[30] D.L. Clouser and P.C. Jurs, *Anal. Chim. Acta*, 295 (1994) 221–231.