Pergamon

PII: S0097-8485(97)00019-3

*Computers Chem.* Vol. 21, No. 6, pp. 437–443, 1997

Printed in Great Britain
0097-8485/97 $17.00 + 0.00

# $^{13}C$ NMR chemical shift sum prediction for alkanes using neural networks

O. Ivanciuc[1], J.-P. Rabine[2] and D. Cabrol-Bass*[2]

[1]Department of Organic Chemistry, Faculty of Chemical Technology, University "Politehnica" of Bucharest, Splaiul Independentei 313, 77206, Bucharest, Romania and [2]LARTIC, University of Nice–Sophia Antipolis, Parc Valrose 06108, Nice, Cedex, France

(Received 14 October 1996; Accepted 7 March 1997)

**Abstract**—The $^{13}C$ NMR chemical shift sum (*CSS*) of alkanes was estimated with multi-linear regression (MLR) and multi-layer feed-forward artificial neural networks (ANN), using as structural descriptors the number of paths of length 1, 2, 3, and 4. The *CSS* prediction ability of both the MLR and ANN models was tested by the "leave-20%-out" (L20%O) cross-validation method. Four activation functions were tested in the neural model: the hyperbolic tangent, a bell-shaped function, a linear function and the symmetric logarithmoid function. The linear and symmetric logarithmoid functions were used only for the output layer. All combinations of activation functions give close results both in calibration and cross-validation, with somewhat lower performances for the networks with a bell-shaped output function. The best results were exhibited by the networks with the symmetric logarithmoid output function, followed by the networks with a linear output function. Because the results were very close, from a statistical point of view one could not definitively choose a particular combination of activation functions. The neural model provides better calibration and cross-validation results than the MLR model. © 1998 Elsevier Science Ltd

## 1. INTRODUCTION

Graph theoretical descriptors have been found useful in developing quantitative structure–property relationships (QSPR). In particular, the enumeration of paths of different lengths provided a basis for characterization of atomic environments in a molecule which was used by Grant and Paul (Grant and Paul, 1964) for the estimation of the $^{13}C$ chemical shifts in alkanes. By considering the sum of $^{13}C$ chemical shifts in alkanes Randic (Randic, 1980) defined a new molecular quantity which reflects some structural features, the "chemical shift sum" (*CSS*). Randic reported that the *CSS* for 18 octanes correlate with molecular path counts. He found that $P_2$ (the number of paths of length two) makes a positive contribution, while $P_3$ (the number of paths of length three) makes a negative one to the molecular isomeric variations of the *CSS*. In a subsequent study, Randic and Trinajstic (Randic and Trinajstic, 1988) have found that the difference $(P_2 - P_3)$ leads to a good correlation for the *CSS* in octanes and nonanes.

A very efficient QSPR correlation was found between the *CSS* for alkanes between $C_2$ and $C_{10}$ and a new topological descriptor $L$ (Miyashita *et al.*, 1989), which combines the paths $P_1$, $P_2$, and $P_3$: $L = 2P_1 + P_2 - P_3 - 2$. A simple, but conceptually significant, modification of the index $L$ generated the new graph theoretical descriptor $M$ (Miyashita *et al.*, 1991): $M = P_0 + P_1 + P_2 - P_3$. The index $M$ was used to obtain an excellent correlation for the *CSS* of alkanes.

Two other graph descriptors used to correlate with the *CSS* of alkanes were the molecular walks count of length 2 ($mwc_2$) and of length 3 ($mwc_3$) (Rücker and Rücker, 1993). The two indices, $mwc_2$ and $mwc_3$, gave good results for highly-branched alkanes, which were difficult to model with the previous descriptors, $L$ and $M$, respectively.

Considering the wide availability of chemical shifts, the *CSS* may be of considerable interest in many structural studies. The theoretical prediction of the *CSS* may help in structure elucidation and computer-assisted structural studies, in detecting systematic errors in experimental $^{13}C$ NMR chemical shifts, or as a structural parameter in QSAR studies. A related molecular descriptor, the mean $^{13}C$ NMR chemical shift for the aromatic carbons, was used to model the carcinogenicity of polycyclic and chlorinated monocyclic aromatic compounds (Sakamoto and Watanabe, 1986; Aoki *et al.*, 1996).

The scope of the present paper is to study the application of neural networks for the prediction of the *CSS* of alkanes. An artificial neural network (ANN) (Rumelhart *et al.*, 1986) is a model that can detect patterns and correlations in data. It can learn

* Author to whom correspondence should be addressed.

to recognize a pattern by increasing the emphasis placed on important information and ignoring irrelevant information or noise in the data. The success of this methodology in the recognition and classification of patterns as well as in QSPR studies has attracted much interest in recent years (Rojas, 1996).

Because the use of ANN in QSPR studies is relatively new, it is necessary to compare the capabilities of ANN to those of more established models. Such comparisons will bring out the advantages and/or disadvantages of neural networks.

An ANN consists of a number of simple, connected computational units that operate in parallel and it can be trained to map a set of input patterns onto a set of output patterns. Each processing unit or artificial neuron has the basic functionality of a biological neuron: to receive signals, to sum the signals, to transform the signal with an activation function, to produce a signal which is passed onto other units. The commonest architecture of ANN is the multi-layer feed-forward (MLF) network, in which the units are organized in three types of layers: a layer of input units, one or more layers of hidden units, a layer of output units. The activity of the input units represents the raw information that is fed into the network. Each unit in one layer is connected to all the units of the next one. The activity of each hidden unit is determined by the activities of the input units and the weights associated with the connections between the input and hidden units. Similarly, the behaviour of the output units depends on the activity of the hidden units and the weights between the hidden and output units.

The network learns by modifying the values of the weights in a well-defined manner, described by a learning rule. The general type of learning used in MLF ANN is supervised learning, which requires a knowledge of the desired responses to input signals (i.e. observations about the system) and incorporates an external teacher (algorithm to apply the learning rule). The aim is to minimize the errors between the desired and computed output unit values.

A great number of problems from diverse branches of chemistry have already been investigated by applying neural networks, and among them QSPR studies form the largest part of the chemical applications of neural networks (Gasteiger and Zupan, 1993; Burns and Whitesides, 1993). It should be pointed out that there are many methods that can be used as alternatives to ANN models, and which have been used successfully for many years in QSPR studies. Many chemical applications in which neural networks are used could probably be solved equally well by using statistical and pattern-recognition methods such as multi-linear regression (MLR), clustering methods, principal component analysis (PCA) and partial least squares (PLS). In many cases these methods should be able to deliver results that are as good as those from a neural network. However, the use of ANN has some important advantages, due to the fact that the mathematical form of the relationship between the input and output data does not need to be provided, and ANNs are able to represent in a very simple way non-linear relationships between input and output patterns. Moreover ANNs are not very sensitive to noise in data.

The neural model was applied for the $^{13}C$ NMR shift prediction for a wide range of classes of chemical compounds: alkanes and cycloalkanes (Panaye *et al.*, 1994; Svozil *et al.*, 1995; Ivanciuc, 1995), $sp^2$ carbon in alkenes (Ivanciuc *et al.*, 1996), $sp^3$ carbon in alkenes (Ivanciuc *et al.*, 1997) monosubstituted benzenes (Kvasnicka *et al.*, 1992), keto-steroids (Anker and Jurs, 1992), halomethanes (Miyashita *et al.*, 1994), and monosaccharides (Mitchell and Jurs, 1996).

## 2. EXPERIMENTAL METHOD

In the present study, MLF networks provided with a single hidden layer and a bias neuron connected to all neurons in the hidden and output layers were used. Biases are useful for scaling the neurons responses to the desired values. The size of the input layer of the network is determined by the path length code which describes the structure of the alkanes. After examining the effect of the maximum path included in the code, it was found that use of the four path counts, $P_1$, $P_2$, $P_3$ and $P_4$, offers good results, which could not be greatly improved by using higher paths. Therefore, the input layer contains four neurons and the output layer has one neuron which provides the calculated value of the $^{13}C$ *CSS*. The number of neurons in the hidden layer was selected on the basis of systematic empirical trials in which ANNs with an increasing number of hidden neurons were trained to estimate the experimental $^{13}C$ *CSS* of alkanes.

### 2.1. Data Set

The structure and experimental $^{13}C$ NMR *CSS* (in ppm) of the 66 alkanes used in the present investigation were taken from the literature (Lindeman and Adams, 1971; Lachance *et al.*, 1979; Kalinowski *et al.*, 1984) and are reported in Table 1.

### 2.2. Learning Method

The training of the ANNs was performed with the standard back-propagation method, until the convergence was obtained, i.e. the correlation coefficient between experimental and calculated *CSS* values improved by less than $10^{-5}$ in 100 epochs. One epoch corresponds to the presentation of one complete set of examples. The patterns were presented randomly to the network and the weights updated after the presentation of each pattern. Random values between $-0.1$ and 0.1 were used as initial weights. In order to evaluate the effect of the initial random weights ten different sets were generated for each network investigated. Two parameters intervene in the correction of weights by the back-propagation method: the "learning rate" $r$ and the "momentum" $\mu$. Each weight is corrected by a fraction $r$ of the derivative of the total error function with respect to this weight. If this fraction is large the changes in weights are fast, but the risk that the optimal weight values are overrun increases. The learning rate was maintained at a constant level during training but depends on the activation function used (see

Section 2.3). The momentum is used to introduce an averaging term from one iteration to the next which is useful in dealing with fluctuations in the gradient of the errors. The momentum was set to 0.8 for all trials. In all cases the networks converged in a few thousands epochs and the final results were slightly influenced by the initial random set of weights.

Table 1. Alkanes used in the QSPR study, experimental $^{13}$C NMR chemical shift sums (*CSS*), calibration and L20%O cross-validation residuals for the MLR and ANN models*

| Alkanes | *CSS* | Experimental MLR calibration residuals | MLR cross-validation residuals | ANN calibration residuals | ANN cross-validation residuals |
|---|---|---|---|---|---|
| Propane | 48.5 | 3 | 5 | −0.5 | −5.7 |
| n-Butane | 76 | 2.3 | 2.8 | −0.2 | −3.4 |
| 2-Methylpropane | 97.1 | −0.8 | −1.6 | 2 | −5.5 |
| n-Pentane | 106.1 | 2.1 | 3.5 | −2 | −1.8 |
| 2-Methylbutane | 116.8 | 2.6 | 3.2 | 2.9 | 0 |
| 2,2-Dimethylpropane | 154 | −8.6 | −7.1 | 1.7 | −5.3 |
| n-Hexane | 136.2 | 1.9 | 2.2 | −1.7 | −2.3 |
| 2-Methylpentane | 150.3 | 3.7 | 4.1 | 3.3 | 4.3 |
| 3-Methylpentane | 137.1 | 4.5 | 5.5 | 4 | 3.9 |
| 2,2-Dimethylbutane | 161 | −6 | −7.5 | −1.7 | −4.5 |
| 2,3-Dimethylbutane | 144.8 | 2 | 2.8 | 3.8 | 3.7 |
| n-Heptane | 165.6 | 1 | 0.4 | −0.6 | −4.4 |
| 2-Methylhexane | 178.1 | 1.2 | 1.5 | 1.2 | −0.4 |
| 3-Methylhexane | 166.6 | 1.6 | 1.8 | 3 | 3.6 |
| 2,2-Dimethylpentane | 199.6 | −1.9 | −1.5 | 0.7 | 0.4 |
| 2,3-Dimethylpentane | 163.1 | −0.2 | 0.7 | 3 | 2.8 |
| 2,4-Dimethylpentane | 191.2 | −0.1 | 0 | 0.9 | 3.4 |
| 3,3-Dimethylpentane | 165.7 | −7.8 | −9.8 | −5.9 | −9.2 |
| 2,3,3-Dimethylbutane | 187 | 3.3 | 4.8 | 2.4 | −0.1 |
| 3-Ethylpentane | 149.5 | −3.6 | −3.8 | −0.2 | 3 |
| n-Octane | 195.6 | 0.7 | 0.2 | −0.5 | −7.6 |
| 2-Methylheptane | 208.4 | 1.2 | 1.6 | 1.5 | 1.5 |
| 3-Methylheptane | 198.6 | 3.3 | 3.9 | 3.4 | 5.1 |
| 4-Methylheptane | 199.2 | 1.8 | 2.5 | 2.1 | 1.5 |
| 2,2-Dimethylhexane | 226.4 | −5.4 | −6.2 | −1.3 | −3.1 |
| 2,3-Dimethylhexane | 195.6 | −0.1 | 0.5 | 1.5 | 3.2 |
| 2,4-Dimethylhexane | 209.4 | −0.3 | −0.1 | 0.3 | −1.9 |
| 2,5-Dimethylhexane | 220.2 | 0.7 | 1 | 2.8 | 1.5 |
| 3,3-Dimethylhexane | 204.6 | −3.4 | −4.6 | −2.3 | −2.8 |
| 3,4-Dimethylhexane | 183.4 | −0.4 | 0.4 | 5.3 | 5.1 |
| 3-Ethylhexane | 182.5 | −3 | −3.6 | 0.5 | −1 |
| 2,2,3-Trimethylpentane | 210.4 | 4.1 | 4 | 4.5 | 4.8 |
| 2,2,4-Trimethylpentane | 248.6 | 0.3 | 0.3 | 3 | 1.7 |
| 2,3,3-Trimethylpentane | 191.3 | −1 | −0.6 | −0.2 | 2.4 |
| 2,3,4-Trimethylpentane | 194.3 | −1.8 | −1.2 | 1.4 | 2.3 |
| 2-Methyl-3-ethylpentane | 183.5 | −2.3 | −4.6 | 4.3 | 0.7 |
| 3-Methyl-3-ethylpentane | 172.3 | −9.8 | −13.1 | −3.1 | −9.2 |
| 2,2,3,3-Tetramethylbutane | 223.6 | 10.9 | 14.7 | 3 | 8.6 |
| n-Nonane | 225.4 | 0.2 | 0.4 | 0.9 | −3.1 |
| 2-Methyloctane | 237.2 | −0.3 | −0.9 | 2 | −0.7 |
| 3-Methyloctane | 226.9 | 1.3 | 1.2 | 1.4 | −0.7 |
| 4-Methyloctane | 228.6 | 0.9 | 0.4 | 0.1 | 0.4 |
| 2,2-Dimethylheptane | 256.2 | −5.9 | −5.1 | −0.3 | −4.5 |
| 2,3-Dimethylheptane | 225.1 | −0.9 | −0.6 | −1.4 | 1 |
| 2,4-Dimethylheptane | 240.9 | −1.2 | −0.8 | −3.8 | −6.2 |
| 2,5-Dimethylheptane | 238.4 | 0.5 | 0.7 | 1.4 | 2.9 |
| 2,6-Dimethylheptane | 250 | 0.2 | 0.3 | 4.1 | 4.7 |
| 3,3-Dimethylheptane | 232.6 | −5.7 | −6.7 | −6.2 | −7.5 |
| 3,4-Dimethylheptane | 216.3 | 0.1 | 1.2 | 0.2 | 0.1 |
| 3,5-Dimethylheptane | 228.8 | 0.7 | 1 | −0.6 | −1.9 |
| 4,4-Dimethylheptane | 240.8 | −1.6 | −2.3 | −5.6 | −5.4 |
| 3-Ethylheptane | 211.8 | −4 | −3.7 | −4.5 | −7.7 |
| 2,2,4-Trimethylhexane | 267.7 | 1 | 0.5 | −0.9 | −0.4 |
| 2,2,5-Trimethylhexane | 267.8 | −6.6 | −6 | 0.8 | 2.8 |
| 2,3,3-Trimethylhexane | 227.2 | 0.5 | 1.4 | −1.1 | −2.2 |
| 2,3,4-Trimethylhexane | 222.5 | 6 | 6.5 | 6.6 | 5.3 |
| 2,3,5-Trimethylhexane | 240.5 | 0.1 | −0.6 | −1.6 | −1.8 |
| 2,2,3,3-Tetramethylpentane | 229.1 | 5.8 | 7.4 | 1.6 | −1.2 |
| 2,2,3,4-Tetramethylpentane | 247.3 | 6.2 | 6.4 | 1.9 | 1.7 |
| 2,3,3,4-Tetramethylpentane | 214.4 | 1.3 | 3.3 | 2.3 | 0.2 |
| 2,2,4,4-Tetramethylpentane | 312.1 | 4.8 | 7.7 | 1.2 | 16.7 |
| 2,4-Dimethyl-3-ethylpentane | 228.4 | 7.7 | 10.1 | 4.5 | 7.8 |
| 3,3-Diethylpentane | 173.9 | −18.8 | −23.2 | −2.9 | −19 |
| n-Decane | 260 | 4.5 | 4.7 | 6.3 | 6.7 |
| 3-Methylnonane | 259.4 | 3.5 | 3.5 | 3.2 | 4 |
| 4-Methylnonane | 261.6 | 3.6 | 3.1 | 1.1 | 6.2 |

*The ANN used to obtain the results was provided with five hidden neurons and a tanh–symlog pair of activation functions.

### 2.3. Activation Functions

The most commonly used activation function in chemical applications of neural networks has a sigmoidal shape and takes values between 0 and 1; for large negative arguments its value is close to 0, and practice demonstrated that learning might be difficult in such conditions. To overcome this deficiency of the logistic function, in the present study it was used hyperbolic tangent (tanh) which takes values between $-1$ and 1. For the hidden layer neurons it was also investigated a bell-shaped activation function, defined as $\mathrm{Act}(z) = 1/(1 + z^2)$. The bell function offered better ANN models than a sigmoidal function in QSPR studies where a highly non-linear relationship exists between structural parameters and the investigated property (Ivanciuc *et al.*, 1996). The use of the bell-shaped activation function has a theoretical basis, because recently Kreinovich (Kreinovich, 1991) demonstrated that an arbitrary non-linear activation function in the hidden layer is sufficient to represent all functions by neural networks. His theorem opened the possibility of using new activation functions in the neural model. In previous studies (Ivanciuc *et al.*, 1996) we have obtained good results with a bell-shaped hidden activation function; for the output layer the tanh and linear functions provided better results than the bell-shaped function.

The tanh activation function (and also the sigmoid function) is very flat when the absolute value of its argument is greater than 10. Therefore, its derivative has an extremely small value and this type of activation function has a poor sensitivity to large positive or negative arguments. This is an important cause of the very slow rates of convergence during the training of neural networks with algorithms that use the derivative of the activation function (e.g. the back-propagation algorithm). In such situations, a linear-output function provides better quantitative estimations than a sigmoidal one; therefore, for the output layer, we have investigated both the linear and the tanh activation functions. Another new type of activation function which overcomes the limitations of the sigmoid and tanh functions is the symmetric logarithmoid (Bulsari and Saxén, 1991a, 1991b), defined by the formula: $\mathrm{Act}(z) = \mathrm{sign}(z)\ln(1 + |z|)$. The symmetric logarithmoid (symlog) is a monotonically increasing function with the maximum sensitivity near zero and with a monotonically decreasing sensitivity away from zero but, because its output is not restricted to the range between $-1$ and 1, this function is not insensitive to large positive or negative arguments. The hidden layer learning rate was 0.05 for the bell function and 0.01 for the tanh function. The output-layer learning rate was 0.01 for all activation functions investigated.

### 2.4. Pre-processing of the Data

Each component of the input and output patterns was scaled between $-0.9$ and 0.9, with the exception of the networks provided with a bell-output function, in which case the output data were scaled between 0.01 and 0.99.

### 2.5. Performance Indicators

The performances of the neural networks were evaluated both for the model calibration and prediction. The quality of model calibration is estimated by comparing the calculated CSS during the training phase ($CSS_{calc}$) with the target values ($CSS_{exp}$), while the predictive quality was estimated by a cross-validation method by comparing the predicted ($CSS_{pr}$) and experimental values. In order to compare the performance of the ANN models with the statistical results of the MLR equation, we have used the correlation coefficient $r$ and the standard deviation $s$ of the linear correlation between experimental and calibration or prediction CSS: $CSS_{exp} = A + B \cdot CSS_{calc/pr}$.

### 2.6. Cross-validation

In QSPR studies it is very important to take into consideration that feed-forward neural networks are universal approximators: they are capable of arbitrarily accurate approximation to arbitrary mappings, when the network has sufficiently large number of hidden units. Recently, Cybenko (Cybenko, 1989) and Funahashi (Funahashi, 1989) demonstrated that any continuous function can be approximated on a compact set with the uniform topology by a layered network with one hidden layer. Hornik, Stinchcombe, and White (Hornik *et al.*, 1989, 1990) have shown that any measurable function can be approximated with a multi-layer feed-forward neural network. The properties of the intermediate layer activation function are not crucial, and the sigmoid activation function is not necessary for universal approximation, as demonstrated by Kreinovich (Kreinovich, 1991).

This great modelling power of MLF ANN represents a property which must be considered with caution when using neural networks for quantitative property estimation. The most important problem with the use of ANN in QSPR is that any physical, chemical or biological property of a set of chemical compounds can be approximated using random numbers as input data if the network contains a sufficiently large number of hidden neurons. Due to the fact that MLF networks are universal approximators, a network with too many connections (adjustable parameters) can offer excellent calibration results for the patterns in the training set, but will have poor performances for predicting the properties for new patterns, which were not present in the training set.

Because the scope of a QSPR study is to develop a model that gives reliable predictions for new patterns that were not used in the calibration of the mathematical model, it is necessary to estimate the prediction capabilities of the models with a cross-validation method. In the present study the "leave-20%-out" (L20%O) cross-validation method was used for this purpose. From the complete data set, one selects at random 20% of the patterns and forms with them the prediction set; then the ANN model is calibrated with a learning set consisting of the remaining 80% of the data. The neural model obtained in the calibration phase is used to predict the *CSS* values for the patterns in the prediction set.

This procedure is repeated five times, until all patterns are selected in a prediction set once and only once. A linear regression between experimental and predicted *CSS* allows one to compare the prediction capabilities of different ANN architectures and sets of activation functions. The L20%O cross-validation procedure was repeated ten times for each size of the hidden layer and for each combination of hidden-output activation functions.

### 2.7. Software Used

A general program for ANN learning and evaluation by cross-validation was written in Borland C language and run on a PC486 DX2 at 66 MHz. Typical time of execution for a complete training is about 5–10 min.

## 3. RESULTS AND DISCUSSION

The primary advantage of using neural networks in QSPR is their capability of providing non-linear mapping of the structural to the corresponding physico-chemical property. In order to compare the results of the ANN model with those obtained by the MLR model the experimental data from Table 1 and the four structural parameters ($P_1$, $P_2$, $P_3$, and $P_4$) were used to obtain the following MLR calibration equation:

$$CSS = -22.55 + 27.85P_1 + 12.29P_2 - 11.92P_3 + 2.08P_4$$
$$n = 66 \quad r = 0.996 \quad s = 4.62$$

The predictive power of the MLR model is estimated by the L20%O method with the same random partitioning of the patterns in five sets used in the cross-validation of the neural model. The following correlation between the experimental and MLR-predicted *CSS* values is obtained:

$$CSS_{exp} = 0.710 + 0.997CSS_{cv}$$
$$n = 66 \quad r_{cv} = 0.994 \quad s_{cv} = 5.47$$

Columns 6 and 7 in Table 1 contain the residuals of the *CSS* computed by the MLR model for calibration and L20%O cross-validation, respectively.

In order to determine the optimum number of hidden neurons their number was modified between 1 and 8, and excellent results were obtained in all cases investigated. Even for networks with one hidden neuron the calibration correlation coefficient $r$ is higher than 0.99 and the L20%O cross-validation correlation coefficient $r_{cv}$ is higher than 0.98. As a general conclusion, the networks provided with a linear or symlog output functions out-performed the networks with bell- and tanh-output functions. By comparing the calibration and cross-validation results the network with five hidden neurons was selected as giving the best results. Using more hidden neurons increases the complexity of the model without significantly improving the calibration and cross-validation results. Table 2 contains the mean statistical results for ten simulations (each one with different initial random weights) obtained with networks with five hidden neurons, for all eight combinations of activation functions.

As is apparent from Table 2, there is a small difference between the maximum and minimum correlation coefficient and standard deviation, which means that in all cases the networks converged to similarly good results, both in calibration and cross-validation. Because the results were very close for all combinations of activation functions, one could not definitively choose a particular combination of activation functions from a statistical point of view. However some partial conclusions can be drawn. The lower statistical indices are offered by the networks provided with bell-shaped output functions, which adds new evidence to our previous findings that the bell function gives good results when used in the hidden layer, but it is outperformed by the other activation functions when used in the output layer. The networks with a bell-hidden function offer results comparable with that offered by the networks with a tanh-hidden function. It can be conclude that this new type of activation function is a good alternative to the previous known functions for use in the hidden layers. The results presented in Table 2 show that the networks with linear- or symlog-output activation functions give better results both in calibration and cross-validation. This finding can lead to the

Table 2. Calibration (A) and L20%O cross-validation (B) statistical indices for the estimation of $^{13}$C NMR chemical shift sums (*CSS*) with neural networks*

| Pair of activation functions | $r_{min}$ | $r_{max}$ | $r_{med}$ | $s_{min}$ | $s_{max}$ | $s_{med}$ |
|---|---|---|---|---|---|---|
| A. Calibration | | | | | | |
| Bell–bell | 0.9966 | 0.9978 | 0.9972 | 3.22 | 4.06 | 3.63 |
| Bell–linear | 0.9978 | 0.9986 | 0.9982 | 2.62 | 3.22 | 2.96 |
| Bell–symlog | 0.9981 | 0.9987 | 0.9984 | 2.49 | 3.00 | 2.77 |
| Bell–tanh | 0.9976 | 0.9987 | 0.9982 | 2.51 | 3.40 | 2.93 |
| Tanh–bell | 0.9963 | 0.9968 | 0.9966 | 3.89 | 4.23 | 4.02 |
| Tanh–linear | 0.9957 | 0.9983 | 0.9975 | 2.82 | 4.53 | 3.38 |
| Tanh–symlog | 0.9977 | 0.9984 | 0.9981 | 2.78 | 3.31 | 3.01 |
| Tanh–tanh | 0.9973 | 0.9981 | 0.9976 | 3.05 | 3.60 | 3.36 |
| B. L20%O Cross-validation | | | | | | |
| Bell–bell | 0.9784 | 0.9847 | 0.9830 | 8.52 | 10.12 | 8.97 |
| Bell–linear | 0.9860 | 0.9937 | 0.9914 | 5.50 | 8.15 | 6.35 |
| Bell–symlog | 0.9899 | 0.9943 | 0.9920 | 5.21 | 6.95 | 6.15 |
| Bell–tanh | 0.9858 | 0.9912 | 0.9888 | 6.47 | 8.23 | 7.28 |
| Tanh–bell | 0.9822 | 0.9836 | 0.9829 | 8.84 | 9.21 | 9.02 |
| Tanh–linear | 0.9926 | 0.9943 | 0.9936 | 5.24 | 5.94 | 5.51 |
| Tanh–symlog | 0.9927 | 0.9945 | 0.9935 | 5.11 | 5.89 | 5.55 |
| Tanh–tanh | 0.9880 | 0.9899 | 0.9890 | 6.91 | 7.56 | 7.25 |

*All networks are provided with five hidden neurons and hidden-ouput activation functions as indicated in the first column. The mean of the correlation coefficient $r$ and standard deviation $s$ is computed for ten simulations with different random initial weights.

Table 3. Comparison of chemical shift sums of some linear alkanes

| Alkane | Chemical shift sums (*CSS*) | | |
|---|---|---|---|
| | Grant and Paul | Lindeman and Adams | ANN calibration |
| Propane | 45.8 | 48.5 | 49.0 |
| Butane | 74.4 | 76.0 | 76.2 |
| Pentane | 101.6 | 105.5 | 108.1 |
| Hexane | 134.6 | 136.2 | 137.9 |
| Heptane | 164.8 | 165.6 | 166.2 |
| Octane | 194.4 | 195.6 | 196.1 |
| Nonane | 224.1 | 225.4 | 224.5 |

conclusion that un-bounded output functions have better performances in QSPR studies, because these type of functions are not insensitive to large positive or negative arguments.

Because the tanh–symlog network offered best cross-validation predictions, this type of network was selected to illustrate the quality of the ANN estimation of *CSS* for alkanes. Columns 9 and 10 in Table 1 contain the residuals of the *CSS* computed by the tanh–symlog networks for calibration and L20%O cross-validation, respectively. Comparison of the results of the ANN and MLR models makes it clear that the neural model offers only slightly better results than the linear model, both for calibration and cross-validation. The same conclusion was recently obtained in a study concerning the prediction of $^{13}C$ NMR chemical shifts of alkanes (Svozil *et al.*, 1995). This result indicates that the relationship between the path counts and *CSS* has a small non-linear character and the ANN model could not provide much better results than the MLR model.

An analysis of the predictions of the ANN L20%O cross-validation residuals reveals the fact that highly-branched alkanes, with quaternary carbon atoms, usually exhibit large residuals: 2,2-dimethylpropane gave 5.3 ppm; 3,3-dimethylpentane, 9.2 ppm; 3-methyl-3-ethylpentane, 9.2 ppm; 2,2,3,3-tetramethylbutane, 8.6 ppm; 3,3-dimethylheptane, 7.5 ppm; 4,4-dimethylheptane, 5.4 ppm; 2,2,4,4-tetramethylpentane, 16.7 ppm; 3,3-diethylpentane, 19 ppm. This fact can be explained by the low number of quaternary carbons in the investigated population of alkanes, which can lead to a low quantitative prediction of *CSS* for alkanes with quaternary carbons. In order to obtain better neural models for the prediction of *CSS* we will consider the use of a larger population of alkanes, with a larger number of highly-branched molecules and new structural descriptors which can describe properly the effect of branching on the $^{13}C$ NMR chemical shift.

The approach applied here for the prediction of the $^{13}C$ NMR *CSS* of alkanes can be extended to molecules containing hetero-atoms and/or multiple bonds by considering all paths containing up to four bonds. The paths are classified according to the chemical nature of the atoms and bonds situated on the path. For example, in the case of aliphatic alcohols, the set of structural descriptors is made of the four paths containing carbon atoms, namely C–C; C–C–C; C–C–C–C; C–C–C–C–C; and the four paths containing the hydroxyl group, namely C–OH: C–C–OH; C–C–C–OH and C–C–C–C–OH. These eight parameters can be used to predict the *CSS* of aliphatic alcohols. In a similar way this method can be extended to other classes of chemical compounds.

The *CSS* can be used to detect systematic deviations in $^{13}C$ NMR spectra determined in different laboratories. In the case of systematic errors, the deviation for a single chemical shift might be too small to be noticeable; the chemical shift sum accumulate such systematic deviations, making more efficient and easy the detection of significant systematic errors in the $^{13}C$ NMR spectra. For example, in Table 3 we compare the $^{13}C$ *CSS* for seven linear alkanes from propane to nonane calculated from the values determined by Grant and Paul (Grant and Paul, 1964) (column 2), by Lindeman and Adams (Lindeman and Adams, 1971) (column 3), and computed by the neural network (column 4). Comparison of the data in Table 3 makes it is clear that there is a systematic difference between the chemical shifts reported by Grant and Paul and those reported by Lindeman and Adams, which constantly have a greater values. This difference is hardly detected when comparing the individuals $^{13}C$ chemical shifts values. The same conclusion can be obtained by comparison with the *CSS* computed by the ANN. In this way the *CSS* can be useful in detecting systematic deviations.

The structure–property relationship established between the path counts and the $^{13}C$ *CSS* can be extended to establish structure–activity relationships. As indicated in the introduction, the mean $^{13}C$ NMR chemical shift of aromatic carbons was used as a structural parameter for the prediction of the carcinogenicity of aromatic compounds (Sakamoto and Watanabe, 1986; Aoki *et al.*, 1996). This type of property–activity relationship requires the determination of the $^{13}C$ NMR spectra of the investigated compounds. On the other hand, our approach which uses MLF neural networks to reveal the relationship between path counts and *CSS* can be extended easily to develop structure-activity relationship model. This model suggests a more fundamental relationship between the molecular structure (expressed as path counts) and carcinogenicity with the main advantage that it would be possible to predict the carcinogenicity of a compound without the need to synthesize it and to determine its $^{13}C$ NMR spectrum.

## 4. CONCLUSIONS

A neural model for the estimation of the $^{13}C$ NMR *CCS* of alkanes, using as structural descriptors the counts of paths of length 1, 2, 3, and 4 has been developed and its performance evaluated. Use of

higher paths counts did not improve the model, suggesting that for the population of alkanes investigated the influence of carbon atoms separated by more than four bonds can be neglected.

The CSS prediction ability of the MLR and neural models was tested by the L20%O method, indicating that both models are stable and give fairly good predictions even with such a drastic perturbation of the model. Four activation functions were tested in the neural model: the hyperbolic tangent, the bell function $\text{Act}(z) = 1/(1 + z^2)$, the linear function and the symmetric logarithmoid function $\text{Act}(z) = \text{sign}(z)\ln(1 + |z|)$. To ensure consistent and reproducible results are obtained, all networks were simulated ten times, with different sets of random initial weights; in all cases the networks converged to closely final states, indicating that the results are not dependent on the initial weights. All combinations of activation functions give results in close agreement both in calibration and cross-validation, with lower results for the networks with a bell-output function and better results for the networks with a symmetric logarithmoid- and linear-output functions. Our results lead to the conclusion that in QSPR studies the bell function is a good alternative to the tanh function in the hidden layer but not in the output layer, and that, for quantitative computations, the unbounded symmetric logarithmoid and linear functions provide better estimations than the bounded functions (sigmoid or tanh). The neural model gives slightly better calibration and L20%O cross-validation results than the MLR results, indicating that in the case of alkanes population investigated the relationship between the paths counts and *CSS* is mainly linear.

Our study demonstrated that the four path counts $P_1$, $P_2$, $P_3$ and $P_4$ describe adequately the structural effect on the *CSS* of alkanes. We intend to extend our investigation by using new structural descriptors which can adequately describe the *CSS* of highly-branched alkanes and to determine the structural dependence of the $^{13}$C NMR *CSS* of other classes of organic compounds.

The chemical shift sum may be of considerable interest in detecting systematic errors in experimental $^{13}$C NMR chemical shifts and in computer-assisted structure elucidation studies.

*Acknowledgements*—O. Ivanciuc thanks the Ministry of Research and Technology for partial financial support of this research under Grant 381 TA10.

## REFERENCES

Aoki, T., Ohshima, S. and Sakamoto, Y. (1996) *Polycyclic Aromatic Compounds* **11**, 245.

Anker, L. S. and Jurs, P. C. (1992) *Analytical Chemistry* **64**, 1157.

Bulsari, A. B. and Saxén, H. (1991a) *Neurocomputing* **3**, 125.

Bulsari, A. B. and Saxén, H. (1991b) *Neural Network World* **4**, 221.

Burns, J. A. and Whitesides, G. M. (1993) *Chemical Reviews* **93**, 2583.

Cybenko, G. (1989) *Math. Control Signals Syst.* **2**, 303.

Funahashi, K. (1989) *Neural Networks* **2**, 183.

Gasteiger, J. and Zupan, J. (1993) *Angew. Chem. Int. Ed. Engl.* **32**, 503.

Grant, D. M. and Paul, E. G. (1964) *Journal of the American Chemical Society* **86**, 2984.

Hornik, K., Stinchcombe, M. and White, H. (1989) *Neural Networks* **2**, 359.

Hornik, K., Stinchcombe, M. and White, H. (1990) *Neural Networks* **3**, 551.

Ivanciuc, O. (1995) *Rev. Roum. Chim.* **40**, 1093.

Ivanciuc, O., Rabine, J.-P., Cabrol-Bass, D., Panaye, A. and Doucet, J. P. (1996) *Journal of Chemical Information and Computer Science* **36**, 644.

Ivanciuc, O., Rabine, J.-P., Cabrol-Bass, D., Panaye, A. and Doucet, J. P. (1997) *Journal of Chemical Information and Computer Science* **37**, 587.

Kalinowski, H.-O., Berger, S. and Braun, S. (1984) *Carbon-13 NMR Spectroscopy*, pp. 113. John Wiley & Sons, Chichester.

Kreinovich, V. (1991) *Neural Networks* **4**, 381.

Kvasnicka, V., Sklenák, Š. and Pospíchal, J. (1992) *Journal of Molecular Structure (Theochem)* **277**, 87.

Lachance, P., Brownstein, S. and Eastham, A. M. (1979) *Canadian Journal of Chemistry* **57**, 367.

Lindeman, L. P. and Adams, J. Q. (1971) *Analytical Chemistry* **43**, 1245.

Mitchell, B. E. and Jurs, P. C. (1996) *Journal of Chemical Information and Computer Science* **36**, 58.

Miyashita, Y., Okuyama, T., Ohsako, H. and Sasaki, S. (1989) *Journal of the American Chemical Society* **111**, 3469.

Miyashita, Y., Ohsako, H., Okuyama, T., Sasaki, S. and Randic, M. (1991) *Magnetic Resonance Chemistry* **29**, 362.

Miyashita, Y., Yoshida, H., Yaegashi, O., Kimura, T., Nishiyama, H. and Sasaki, S. (1994) *Journal of Molecular Structure (Theochem)* **311**, 241.

Panaye, A., Doucet, J. P., Fan, B. T., Feuilleaubois, E. and El Azzouzi, S. R. (1994) *Chemometrics and Intelligent Laboratory Systems* **24**, 129.

Randic, M. (1980) *Journal of Magnetic Resonance* **39**, 431.

Randic, M. and Trinajstic, N. (1988) *Theoretica Chimica Acta* **73**, 233.

Rojas, R. (1996) *Neural Networks, a systematic introduction*. Springer, Berlin.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) *Nature* **323**, 533.

Rücker, G. and Rücker, C. (1993) *Journal of Chemical Information and Computer Science* **33**, 683.

Sakamoto, Y. and Watanabe, S. (1986) *Bulletin of the Chemical Society of Japan* **59**, 3033.

Svozil, D., Pospíchal, J. and Kvasnicka, V. (1995) *Journal of Chemical Information and Computer Science* **35**, 924.