

Simulation of ^{13}C nuclear magnetic resonance spectra of lignin compounds using principal component analysis and artificial neural networks

M. Jalali-Heravi*, S. Masoum, P. Shahbazikhah

Department of Chemistry, Sharif University of Technology, P.O. Box 11365-9516, Tehran, Iran

Received 16 May 2004; revised 8 July 2004

Available online 22 September 2004

Abstract

Theoretical models relating atom-based structural descriptors to ^{13}C NMR chemical shifts were used to accurately simulate ^{13}C NMR spectra of lignin model compounds (poly-substituted phenols). The structure–activity relationship (SAR) studies for 15 lignins using pattern recognition methods of principal component analysis (PCA) and artificial neural networks (ANNs) were performed in this work. The most important parameters affecting the ^{13}C chemical shifts of different carbons were descriptors consisting of the charge density of the atoms at different distances from the center carbon. Among the large number of parameters, these descriptors were selected using PCA and were used as ANN input. The least square regression analyses of the results indicate correlation coefficient (R) values in excess of 0.983 for the total data set.

© 2004 Elsevier Inc. All rights reserved.

Keywords: Neural networks; Principle component analysis; ^{13}C chemical shift; Lignin compounds; Poly-substituted phenols

1. Introduction

Lignins are the most abundant natural aromatic polymers on earth. The chemistry [1] and analysis of these compounds [2] have been the subject of vast researches, related to pulping [3], bleaching [4], and biodegradation [5]. Among the myriad instrumental methods that have been used in attempting to decipher the structure and reactivity of lignin, a large body of information has been developed from nuclear magnetic resonance spectroscopy, with particular emphasis on the use of ^{13}C techniques [6].

Spectral simulation techniques for ^{13}C NMR spectroscopy can provide aid in the solution of complex structure elucidation problems and in the verification of chemical shift assignment, particularly when suitable

reference spectra are not available. There are different basic approaches of ab initio, empirical, linear regression, and neural networks to calculation and prediction of the ^{13}C NMR chemical shift.

The ab initio methods are able to calculate the magnetic properties of any molecular structure, such as shielding tensors, shielding anisotropy, and isotropic chemical shifts with respect to an applied magnetic field and the nuclear magnetic moment [7–9]. It is noteworthy that the resulting chemical shift values obtained using these methods are not biased by previous experimental results. However, obtaining accurate chemical shifts using ab initio methods requires the correct configuration and conformation of the molecules. The real three-dimensional structure is often unknown and multiple conformations have to be taken into account for small and flexible molecules, in particular. These problems make such calculations very time consuming and expensive.

* Corresponding author. Fax: +98 21 6012983.

E-mail address: jalali@sharif.edu (M. Jalali-Heravi).

The empirical approaches rely on the knowledge of chemical shifts from large sets of known molecular structures [10,11]. In this approach empirical parameters are derived for a wide variety of structural substituents using large databases of carbon atoms and their chemical shifts. The hierarchically ordered spherical description of environment (HOSE) code, developed by Bremser is suitable for this approach [12]. However, the accuracy of the ^{13}C NMR chemical shift prediction strongly depends on the similarity between the new and known HOSE codes and on the quality of the database.

Linear regression modeling relates a carbon atom's chemical shift to atom-based features (descriptors). These relationships are derived through multiple linear regression analysis techniques. Since the chemical shifts of different atoms in a molecule may show a nonlinear characteristic, therefore, a trend to obtain better results with non-linear methods such as neural networks as compared to regression methods have been observed [13,14].

One active area of research involves substituting linear regression by neural network. Artificial neural networks (ANNs) have been reported for use in a few analytical chemical studies including ^1H NMR [15] and ^{13}C NMR spectroscopy [16–21]. To ensure the robustness and generality of an ANN model, various types of descriptors should be used as its input. In order to develop more compact models, it is common to use selected subsets of descriptors, instead of all possible descriptors. There are two basic approaches to solve this problem such as genetic algorithm [15] and principal component analysis [22].

The aim of the present work was to reveal the capabilities of PCA-ANN method for the calculation of ^{13}C chemical shifts of a number of lignin model compounds. Success of such techniques will allow the prediction of chemical shifts of unknown compounds, and serve as an aid for verification of the structure of the molecules.

2. Theoretical background

2.1. Principal components analysis

Principal components analysis (PCA) is a multivariate procedure, which rotates the data such that maximum variabilities are projected onto the axes [23,24]. The main use of PCA is to reduce the dimensionality of a data set while retaining as much information as is possible. It computes a compact and optimal description of the data set.

The first principal component is the combination of variables that explains the greatest amount of variation. The second principal component defines the next largest amount of variation and is independent to the first principal component. There can be as many possible principal components as there are variables.

It can be viewed as a rotation of the existing axes to new positions in the space defined by the original variables. In this new rotation, there will be no correlation between the new variables defined by the rotation. The first new variable contains the maximum amount of variation, the second new variable contains the maximum amount of variation unexplained by the first one and is orthogonal to the first.

A set of multivariate data describes objects and their features. Objects are characterized by a predefined set of features. A feature is a numerical variable that describes an aspect of the objects. Such data are best described by an $n \times p$ matrix \mathbf{X} , containing a row for each of the n objects, and a column for each of the p features.

Each feature can be considered as a coordinate of a point; each object then corresponds to a point in a p -dimensional feature space. The fundamental hypothesis for multivariate data interpretation is the existence of relationships between the locations or the distances of points (objects) and relevant properties. The essential concept of multivariate data analysis is the use of so-called latent variables as plot coordinates. The goal of many chemometric methods is to find a mathematical function or a more general algorithm to define appropriate latent variables. The guiding principle is a representation of the p -dimensional multivariate data by a minimum number of latent variables.

A particular direction that defines a linear latent variable in a p -dimensional feature space is described by a vector \mathbf{b} (b_1, b_2, \dots, b_p) which is usually scaled to length one. The value of the corresponding latent variable u for an object \mathbf{x} (x_1, x_2, \dots, x_p) is obtained by projecting the object point onto a straightline which is defined by the direction \mathbf{b} (Fig. 1). Mathematically this is a linear com-

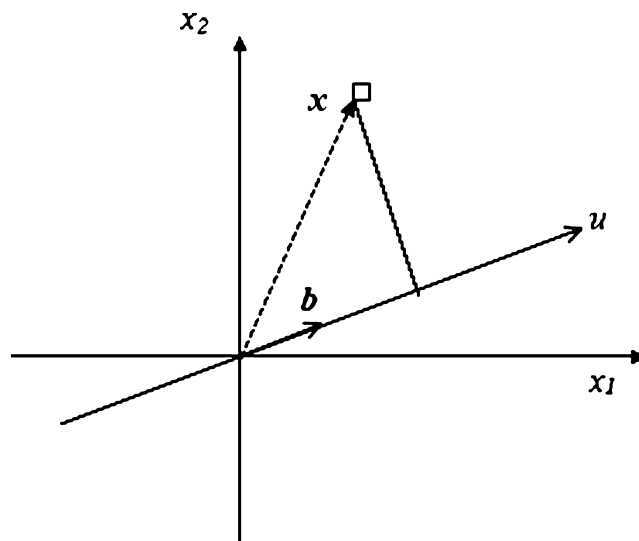


Fig. 1. Projection of an object vector \mathbf{x} onto a straightline which defines a latent variable by vector \mathbf{b} . The value (score) of the latent variable is u .

bination of the features x_j of the object and the vector component b_j ; an equivalent notation is the scalar product of the vectors \mathbf{b}^T and \mathbf{x} .

$$u = \mathbf{b}^T \mathbf{x} = b_1 x_1 + b_2 x_2 + \cdots + b_p x_p. \quad (1)$$

The value of a latent variable is called a score; scores are often used as plot coordinates. The vector components b_j are called loadings; they define the direction of the latent variable in the feature space and they describe the contributions of the individual features to the scores. Usually two orthogonal directions \mathbf{b}_1 and \mathbf{b}_2 are used as projection axes (the product $\mathbf{b}_1 \cdot \mathbf{b}_2$ becomes zero) to define a projection plane.

The vectors \mathbf{b}_1 and \mathbf{b}_2 can be arranged in a matrix \mathbf{B} . Scores \mathbf{U} for objects \mathbf{X} are calculated by a matrix multiplication.

$$\mathbf{U} = \mathbf{X} \cdot \mathbf{B}. \quad (2)$$

Two fundamental types of plots can be generated (Fig. 2). In a score plot each point corresponds to an object; the coordinates are given by the scores. The distances between objects in the score plot are approximations of the distances in the multivariate feature space; groups (clusters) of similar objects can be detected visually. In a loading plot each point corresponds to a feature; the coordinates are given by the loadings of the features for the same axes as used in the corresponding score plot. The loading plot indicates the similarities and correlations between features. Furthermore this plot makes evident which features are responsible for the relative positions of the objects in the score plot. Features with small loadings are located near the origin; they have on the average only little influence on the data structure. A feature with a high loading for a latent variable causes that objects are placed in the corresponding region of the score plot if this feature has a large value.

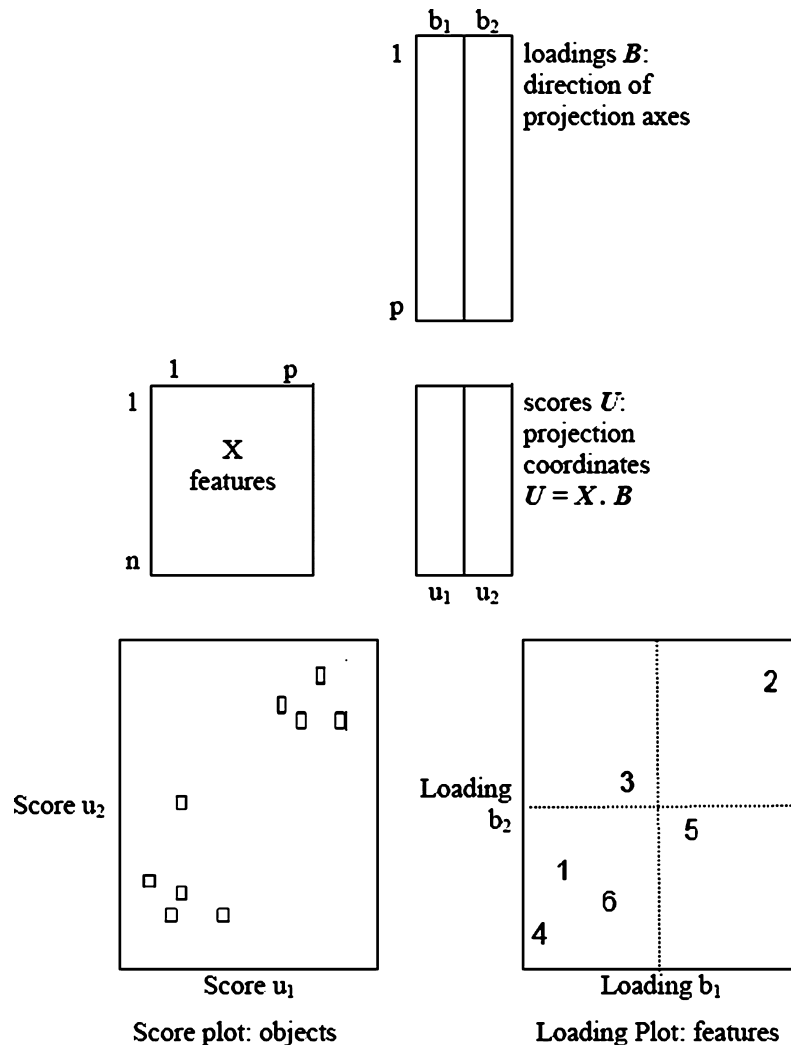


Fig. 2. Score plot and loading plot. The score plot in this demo example indicates two clusters of objects and one outlier. From the loading plot follows that feature with number 2 is characteristic for objects that are located in the right upper corner of the score plot; features 1, 4, and 6 are characteristic for objects in the left lower corner; features 3 and 5 are near the origin of the loading plot and therefore have only little influence.

2.2. Artificial neural networks

Artificial neural network (ANN) is a computer-based system derived from the simplified concept of the brain in which a number of nodes, called processing elements or neurons, are interconnected in a netlike structure. The ANN characteristics have been found to be nonlinear making them suitable for data processing in which the relationship between cause and results cannot be linearly defined. Three components constitute an ANN: the processing elements, the topology of connection between the nodes, and the learning rules. The PCA selected descriptors by loading plot were processed by ANN which was trained with the back-propagation of errors learning algorithm. Its basic theory and application to chemical problems can be found in the literature [25,26]. The structure of the network comprised of three node layers: an input, a hidden and an output layer, represented by i , h , and o , where they, respectively, indicate the number of nodes in the input layer, hidden layer, and output layer. The absorbance data versus the time were centered and normalized as the input for ANN. The input nodes transferred the weighted input signals to the nodes in the hidden layer, and the same as the hidden nodes for the output layer. A connection between the nodes of different layers was represented by a weight w_{ij} , and during the training process, the correction of weight Δw_{ij} was defined as the following:

$$\Delta w_{ij(n+1)} = \eta \delta_j o_j + \alpha \Delta w_{ij(n)}, \quad (3)$$

where δ_j is the error term, o_j is the output of node j , η is the learning rate, α is the momentum, and n is the iteration number. The iteration would be finished when the error of prediction reached a minimum.

A non-linear transformation which named the sigmoidal function was applied between the input and output of a node. The optimum of η and α was taken as those which made the error of prediction minimum.

3. Methodology

In an effort to focus on the neural network aspects of a ^{13}C NMR spectral prediction study, we decided to examine a data set of 15 model compounds [27,28].

For developing the ANNs usually three sample sets are necessary. A calibration or training set must be used to estimate the parameters of the model and a test set is needed for the optimization process of ANN. A third set, the so-called control set of data is needed for the evaluation of the prediction ability of the ANN model. Therefore the sample set was divided into three subsets of training, test, and control sets for the ANN calibration.

The chemical shifts for the carbon atoms in lignin model compounds fall in two distinct separated subsets; carbons of the side chains (Subset 1) and carbons in the benzene rings (Subset 2).

To prevent bias towards one type of carbon atom, only carbon atoms that are structurally unique are included in the development of PCA. For the data set, 100 unique carbon atoms were identified. Of these, 73 carbon atoms were considered as the training set, 20 carbon atoms were in the test set, and 7 carbon atoms were considered as control set. For the two distinct subsets, 30 of the unique carbons were in the first subset, and the remaining 70 unique carbon atoms were in the benzene rings.

Once the data set was defined, numerical representation of the environmental surrounding each carbon atom was calculated. This was done using some programs designed specially for the purpose of calculating descriptors. These programs calculate three different types of descriptors, i.e., geometric, electronic, and topological.

Geometric descriptors were calculated using optimized Cartesian coordinate and van der Waals radius of each atom in the molecules using Microsoft Excel 2002 [29]. The three-dimensional structure of each molecule was optimized using the semi-empirical molecular orbital method of AM1 implemented in the MOPAC package (version 6) [30]. Electronic descriptors were also calculated using AM1 Hamiltonian implemented in this program. Topological descriptors were calculated using two-dimensional representation of the molecules.

Once a suitable pool of descriptors had been found, some of them should be selected to generate the model. Testing every possible combination of descriptors is far too time-consuming process, therefore, a method for descriptor selection must be used. In this work, the principal component analysis (PCA) technique was chosen as feature selection method. This method is an extremely useful explorative tool which maps samples through scores and individual variables by loadings in a new vector space defined by the principle components (PC). From loading plots the more important variables can be easily identified as well as the correlation patterns among them (see Fig. 3). The first PC is generated in a way that it has maximum correlation with all of the variables and accounts for a large portion of the total data variance. From the remaining data variance (after the removal of the first PC) a second PC is extended which is completely uncorrelated (orthogonal) with the first one and accounts for the maximum possible remaining data set variance. This procedure is then repeated until all PCs are generated. The PCA study was carried out using the program package PLS_Toolbox_303a for MATLAB which contains PCA and related methods [31]. Then, in order to

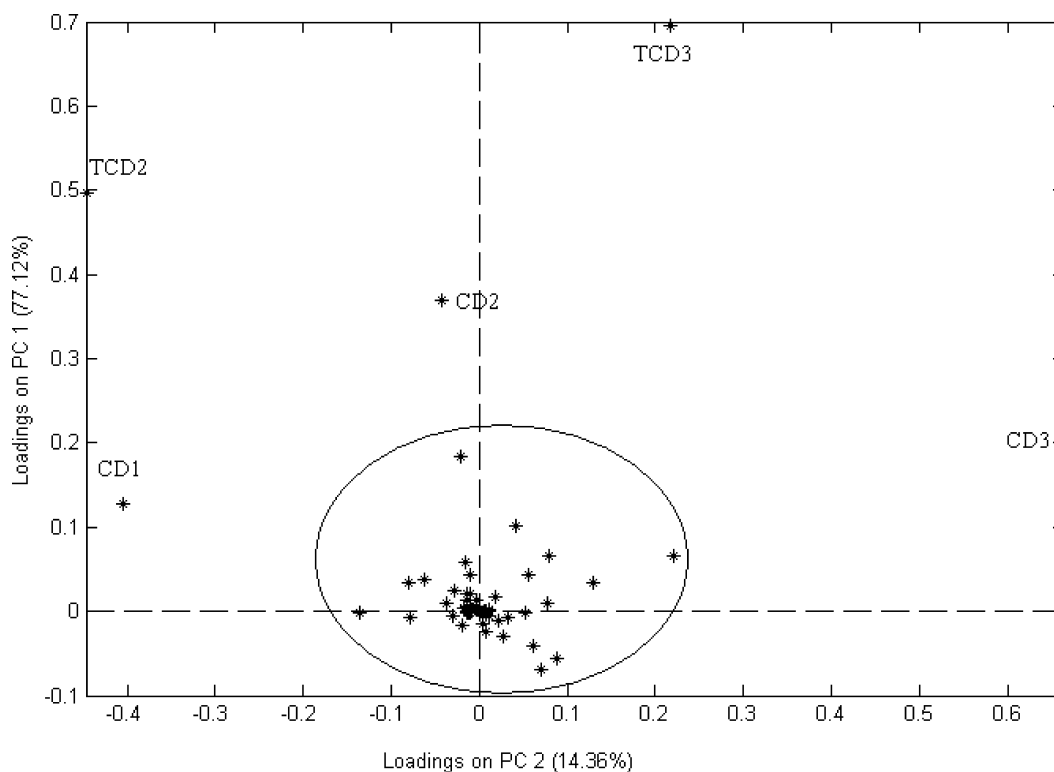


Fig. 3. The loading graph for the five descriptors selected by PCA.

develop a model for predicting ^{13}C chemical shift a neural network was generated using the PCA selected descriptors as inputs. A back-propagation neural network having three layers was created using Visual-Basic software package [32].

4. Results and discussion

A total of 62 descriptors were used for PCA investigation. These descriptors are listed in Table 1. Fig. 3 shows the loading plot for these descriptors. Descriptors with small loadings are located near the origin; they have on the average only a little influence on the data structure. Among these descriptors, CD1, CD2, CD3, TCD2, and TCD3 are electronic descriptors and show high loading values to cause largest effect on the ^{13}C chemical shifts of different carbons. It can be seen from Table 1 that these parameter are principally charge density surrounding a carbon center. It is noteworthy, that the calculation of these descriptors is very easy and nowadays, the tools for this type of calculation, i.e., molecular orbital packages such as MOPAC and HyperChem are available in most laboratories.

To investigate the prediction ability of the ANN, output layer should be considered as a single node,

corresponding to the chemical shift. The specifications of the ANN model are summarized in Table 2. Optimization process was carried out by trial and error procedure. Before training, the input and output values were normalized between -2 to 2 and 0 to 1 , respectively. In order to obtain the best network structure, several ANN systems with different number of hidden nodes were tested. The proper number of nodes in the hidden layer was determined by training the network with different number of nodes in the hidden layer. The root-mean-square error (RMSE) value measures how good outputs are in comparison with the target values. It should be noted that for evaluating the overfitting, the training of the network for the prediction of ^{13}C chemical shift must stop when the RMSE of the test set begins to increase while RMSE of calibration set continues to decrease. Therefore, during the training of the network, it is desirable that iteration be stopped when overtraining begins. The chemical shifts values of 100 unique carbon atoms of lignin model compounds were available both as experimental [27] and as results of ab initio calculation [28]. Table 3 presents the results of the chemical shifts values obtained using PCA-ANN of this work and the values obtained using ab initio methods for these unique carbon atoms. It can be seen that in most cases the values obtained by PCA-ANN are closer to the

Table 1
The definition of descriptors for PCA investigation

<i>PCH</i> : Partial charge of carbon center	
<i>CD</i> : Charge density of carbon center	
<i>PCHn</i> : Partial charge of atoms (carbon and oxygen) n bonds away from the carbon center ($n = 1-3$)	
<i>TPCHn</i> : Total partial charge of atoms (carbon and oxygen) one bond to n bonds away from the carbon center ($n = 2-3$)	
<i>CDn</i> : Charge density of atoms (carbon and oxygen) n bonds away from the carbon center ($n = 1-3$)	
<i>TCDn</i> : Total charge density of atoms (carbon and oxygen) one bond to n bonds away from the carbon center ($n = 2-3$)	
<i>PCHDn</i> : Partial charge of atoms (carbon and oxygen) n bonds away from the carbon center divided to n ($n = 2-3$)	
<i>CDDn</i> : Charge density of atoms (carbon and oxygen) n bonds away from the carbon center divided to n ($n = 2-3$)	
<i>PCHDnn</i> : <i>PCHDn</i> to power n .	
<i>CDDnn</i> : <i>CDDn</i> to power n .	
<i>TPCHDn</i> : Total partial charge of atoms (carbon and oxygen) one bond to n bonds away from the carbon center divided to n ($n = 2-3$)	
<i>MaxPCHn</i> : Maximum partial charge on the atoms (carbon and oxygen) that n bonds away from the carbon center ($n = 1-3$)	
<i>MaxCDn</i> : Maximum charge density on the atoms (carbon and oxygen) that n bonds away from the carbon center ($n = 1-3$)	
<i>MinPCHn</i> : Minimum partial charge on the atoms (carbon and oxygen) that n bonds away from the carbon center ($n = 1-3$)	
<i>MinCDn</i> : Minimum charge density on the atoms (carbon and oxygen) that n bonds away from the carbon center ($n = 1-3$)	
<i>PCHnn</i> : Partial charge of atoms (carbon and oxygen) n bonds away from the carbon center to power n ($n = 2-3$)	
<i>TPCHnn</i> : Total partial charge of atoms (carbon and oxygen) one bond to n bonds away from the carbon center to power n ($n = 2-3$)	
<i>NHEVn</i> : Number of heavy atoms in the n th shell	
n	Radial distance (Å)
1	2.3–3.1
2	3.1–3.9
3	3.9–4.6
4	4.6–5.4
<i>NHYDn</i> : Number of hydrogen atoms in the n th shell	
n	Radial distance (Å)
1	0.0–1.5
2	1.5–2.4
3	2.4–3.0
4	3.0–3.5
5	3.5–4.3
6	4.3–5.5
<i>DOX</i> : Distance to nearest oxygen	
<i>INVDON</i> : Inverse of <i>DOX</i> to power n ($n = 1-3$)	
<i>D4</i> : Distance to atom number 4 (This atom is same in all compounds)	
<i>INVD4n</i> : Inverse of <i>D4</i> to power n ($n = 1-3$)	
<i>CCIn</i> : Corrected connectivity index over bonds n bonds away from the carbon center ($n = 1-3$)	
<i>TCCIn</i> : Total corrected connectivity index over bonds n bonds away from the carbon center ($n = 2-3$)	
<i>AVTCCIn</i> : Average total corrected connectivity index over bonds n bonds away from the carbon center ($n = 2-3$)	

Table 2
Architecture of the ANN models and their specifications

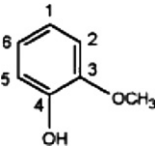
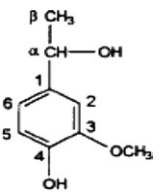
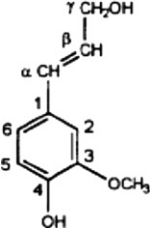
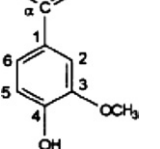
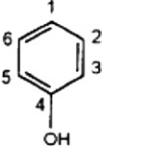
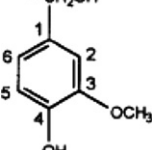
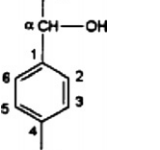
	Subset 1	Subset 2
No. of nodes in the input layer	5	5
No. of nodes in the hidden layer	3	4
No. of nodes in the output layer	1	1
Momentum	0.4	0.6
Learning rate	0.6	0.4
Transfer function	Sigmoid	Sigmoid

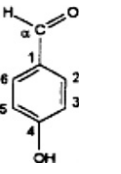
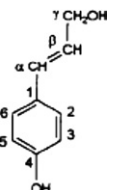
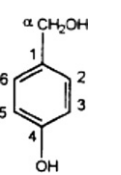
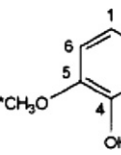
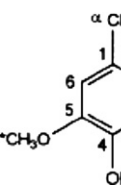
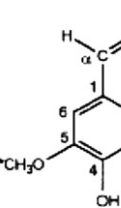
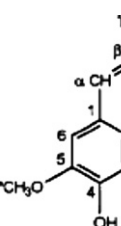
experimental values compared with those obtained using the ab initio methods. The statistical parameters obtained by these models for the training and the test sets are shown in Table 4. For the sake of comparison, the RMSE values and correlation coefficients for the ab initio results were calculated exactly in the same way as PCA-ANNs for the subsets 1 and 2 carbon centers. Inspection of these parameters

shows that although the correlation coefficients do not reveal a considerable improvement, but the RMSE values are much better for the PCA-ANN models presented in this work. The plot of the observed shifts versus the calculated values of the chemical shifts shows a good correlation (Fig. 4). The residuals of the ANN calculated values of chemical shift are plotted against the experimental values in Fig. 5. The propagation of the residuals in both sides of the zero line indicates that no systematic error exists in the development of the PCA-ANN. This plot also shows that there are no outliers in the calculated chemical shifts. An example of the observed spectrum compared to the simulated one for the compound 1-(4-hydroxy-3,5-dimethoxyphenyl) ethanol as control set is shown in Fig. 6. The RMS error for the prediction of this molecule was 2.759 ppm. The visual similarity between the two spectra is striking.

Table 3

Experimental and calculated values of chemical shifts (in ppm) for lignin compounds

	Carbon	Experimental	PCA-ANN prediction	Ab initio prediction
	1	120.15	121.54	124.39
	2	110.18	110.78	114.78
	3	145.7	145.67	149.78
	4	146.63	145.72	153.09
	5	114.6	114.66	120.83
	6	121.47	121.39	128.97
	OMe	55.86	56.09	50.74
	1	137.91	139.73	142.62
	2	108.05	106.67	112.62
	3	146.62	146.38	149.54
	4	144.96	145.68	152.18
	5	114.20 ^a	111.63	122.02
	6	118.31	118.94	126.43
	α	70.3	70.19	66.22
	β	25.07 ^a	26.27	30.48
	OMe	55.8	56.10	50.75
	1	129.3	129.97	134.5
	2	108.52	106.98	110.09
	3	146.75	146.42	150.32
	4	145.63	145.67	153.53
	5	114.57	111.57	120.68
	6	120.25	118.81	131.15
	α	131.24 ^a	131.3	134.9
	β	126.22	127.33	126.04
	γ	63.71	63.33	57.87
	OMe	55.88	56.10	50.78
	1	129.70 ^a	131.06	132.35
	2	109.02 ^a	109.32	117.27
	3	147.34 ^a	146.32	149.15
	4	151.99 ^a	145.71	161.55
	5	114.59 ^a	111.39	117.48
	6	127.57 ^a	121.29	138.55
	α	191.96 ^a	186.53	196.8
	OMe	56.07	56.10	51.65
	1	120.13	119.43	124.15
	2	129.71	124.36	138.26
	3	115.41	116.26	116.44
	4	155.25 ^a	158.32	164.89
	5	115.41	116.26	120.59
	6	129.71	124.36	139.59
	1	132.93	133.38	137.42
	2	109.94	109.75	112.71
	3	146.65	146.39	150.77
	4	145.26	145.68	151.39
	5	114.27	111.22	121.47
	6	120.22	121.12	123.72
	α	65.44	63.79	58.62
	OMe	55.90	56.10	51.06
	1	138.9	137.34	141.78
	2	127.31	127.6	136.37
	3	115.55 ^a	115.59	119.41
	4	156.98	158.39	163.31
	5	115.55 ^a	115.59	116.77
	6	127.31	127.6	135.77
	α	69.61	69.72	65.76
	β	26.10	26.45	30.54

	Carbon	Experimental	PCA-ANN prediction	Ab initio prediction
	1	129.81	130.42	132.45
	2	132.54	131.2	146.17
	3	116.01	115.56	113.11
	4	161.65	158.43	171.23
	5	116.01	115.56	119.48
	6	132.54	131.2	143.4
	α	191.28	189.83	195.66
	1	129.73	128.81	134.02
	2	128.33 ^a	126.82	141.12
	3	116.19	115.67	119.78
	4	157.78	158.4	163.8
	5	116.19	115.67	116.65
	6	128.33 ^a	126.82	133.2
	α	130.29	129.19	133.91
	β	127.67	126.72	126.4
	γ	63.47	63.13	57.82
	1	133.96	132.97	136.05
	2	129.05	130.37	134.65
	3	115.69	115.55	116.52
	4	157.23	158.38	162.7
	5	115.69 ^a	115.55	120.28
	6	129.05	130.37	134.83
	α	64.54	65.55	58.6
	1	119.06	124.17	126.76
	2	105.03	110.79	106.08
	3	147.31 ^a	147.04	153.21
	4	134.97	136.47	141.76
	5	147.31 ^a	147.04	155.12
	6	105.03	110.79	109.05
	OMe	56.27	56.10	51
	OMe*	56.27	56.10	50.68
	1	132.06	133.25	138.9
	2	103.88 ^a	109.08	102.84
	3	147.1	147.24	153.08
	4	134.19	137.03	139.49
	5	147.1	147.24	153.71
	6	103.88 ^a	109.08	102.2
	α	65.68	66.3	58.94
	OMe	56.26	56.10	51.15
	OMe*	56.26	56.10	47.67
	1	128.34	126.98	132.18
	2	106.81 ^a	108.83	117.75
	3	147.44	146.61	151.61
	4	141.03	137.36	149.39
	5	147.44	146.61	150.8
	6	106.81 ^a	108.83	109.64
	α	190.79	189.41	197.17
	OMe	56.48 ^a	56.10	50.82
	OMe*	56.48 ^a	56.10	51.66
	1	128.22	128.51	135.84
	2	103.35	106.71	100.97
	3	147.13	147.24	152.22
	4	134.80 ^a	134.8	140.88
	5	147.13	147.24	153.88
	6	103.35	106.71	111.81
	α	131.5	132.56	134.74
	β	126.58 ^a	127.01	127.65
	γ	63.76	63.96	57.92
	OMe	56.27	56.10	50.89
	OMe*	56.27	56.10	50.63

^a The molecules included in the test set. The remaining molecules are considered in the calibration set.

Table 4

Statistical parameters obtained using ANN models^a

	Subset 1		Subset 2	
	PCA-ANN	Ab initio ^b	PCA-ANN	Ab initio ^b
RMSE _c	0.779	4.953	1.912	5.973
RMSE _t	2.502	4.46	3.118	7.504
R _c	0.999	0.998	0.992	0.985
R _t	0.999	0.998	0.983	0.977

^a c, calibration set; t, test set; RMSE, root mean square error; R, correlation coefficient.

^b The statistics are calculated using the results taken from [27].

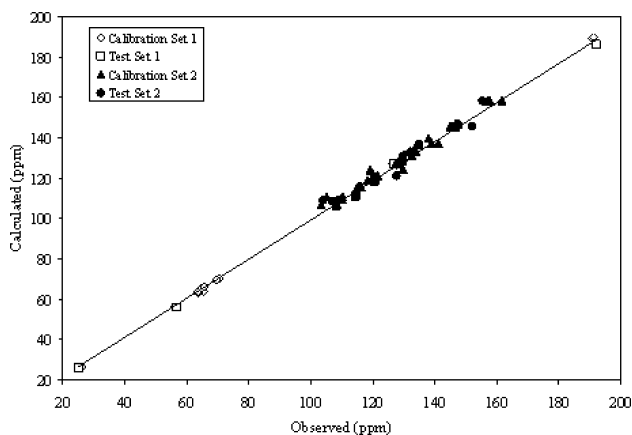


Fig. 4. Calculated vs. observed plot for the computational neural network models developed for the subsets one and two.

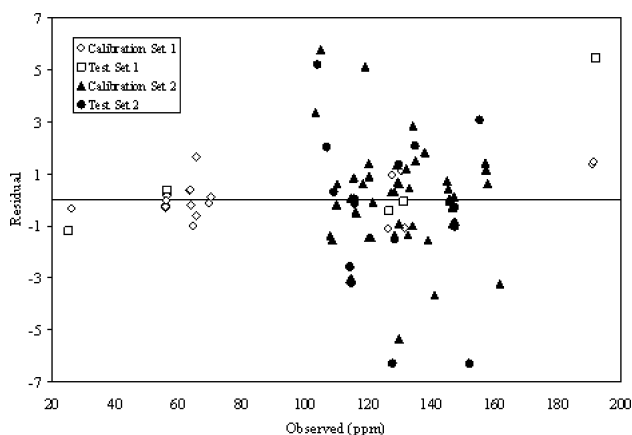


Fig. 5. Plot of the residual vs. observed values for the subsets one and two.

5. Conclusions

Over the dataset of 15 monomeric lignin model compounds, an accurate relationship is found between the experimental and calculated ¹³C NMR chemical shifts. Due to the data existing in two distinct groups of shifts, the data set was divided into two subsets and good fits of the data are found for each of them when treated individually. By accurately simulating

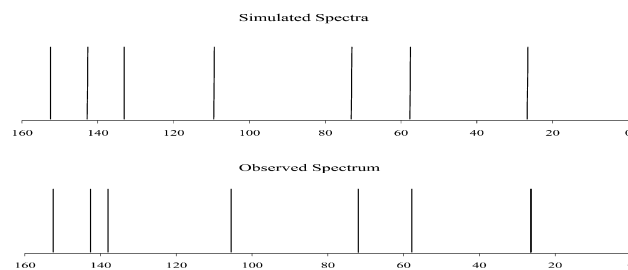
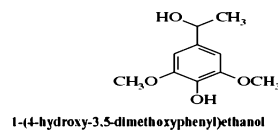


Fig. 6. Observed and simulated spectra of 1-(4-hydroxy-3,5-dimethoxyphenyl)ethanol.

the shifts of compounds such as lignins, researcher may be able to elucidate the structure of different lignin compounds.

References

- [1] K.V. Sarkanen, C.H. Ludwig, Lignins-Occurrence, Formation, Structure and Reaction, Wiley, New York, 1971.
- [2] S.Y. Lin, C.W. Dence, Methods in Lignin Chemistry, Springer, Berlin, 1992.
- [3] S. Rydholm, Pulping Processes, Interscience, New York, 1965.
- [4] C.W. Dence, D. Reeve, Pulp Bleaching-Principles and Practice, TAPPI Press, Atlanta, GA, 1996.
- [5] T.K. Kirk, T. Higuchi, H.M. Chang, Lignin Biodegradation: Microbiology, Chemistry and Potential Applications, CRC Press, Boca Raton, CA, 1980.
- [6] H.H. Nimz, in: Proceedings of the Eighth International Symposium of Wood and Pulping Chemistry, Helsinki, Finland, 6–9 June, 1995, p. 1.
- [7] M. Schindler, W. Kutzelnigg, Theory of magnetic susceptibilities and NMR chemical shifts in terms of localized quantities. II. Application to some simple molecules, J. Chem. Phys. 76 (1982) 1919–1933.
- [8] J. Gauss, Accurate calculation of NMR chemical shifts, Ber. Bunsenges. Phys. Chem. 99 (1995) 1001–1008.
- [9] J.R. Cheeseman, G.W. Trucks, T.A. Keith, M.J. Frisch, A comparison of models for calculating nuclear magnetic resonance shielding tensors, J. Chem. Phys. 104 (1996) 5497.
- [10] D.M. Grant, E.G. Paul, Carbon-13 magnetic resonance. II. Chemical shift data for the alkanes, J. Am. Chem. Soc. 86 (1964) 2984–2990.
- [11] L.P. Lindeman, J.Q. Adams, Carbon-13 nuclear magnetic resonance spectroscopy, Anal. Chem. 43 (1971) 1245–1252.
- [12] W. Bremser, HOSE—a novel substructure code, Anal. Chim. Acta 103 (1978) 355–365.
- [13] L.S. Anker, P.C. Jurs, Prediction of carbon-13 nuclear magnetic resonance chemical shifts by artificial neural networks, Anal. Chem. 64 (1992) 1157–1164.
- [14] D.L. Clouser, P.C. Jurs, The simulation of ¹³C nuclear magnetic resonance spectra of dibenzofurans using multiple linear regression analysis and neural networks, Anal. Chim. Acta 321 (1996) 127–135.
- [15] J. Aires-de-Sousa, M.C. Hemmer, J. Gasteiger, Prediction of ¹H NMR chemical shifts using neural networks, Anal. Chem. 74 (2002) 80–90.

- [16] J.W. Ball, L.S. Anker, P.C. Jurs, Automated model selection for the simulation of carbon-13 nuclear magnetic resonance spectra of cyclopentanones and cycloheptanones, *Anal. Chem.* 63 (1991) 2435–2442.
- [17] M. Jalali-Heravi, M. Moosavi, Simulation of ^{13}C NMR spectra of nitrogen-containing aromatic compounds, *Aust. J. Chem.* 48 (1995) 1267–1275.
- [18] J. Meiler, M. Will, Automated structure elucidation of organic molecules from ^{13}C NMR spectra using genetic algorithms and neural networks, *J. Chem. Inf. Comput. Sci.* 41 (2001) 1535–1546.
- [19] J. Meiler, W. Maier, M. Will, R. Meusinger, Using neural networks for ^{13}C NMR chemical shift prediction-comparison with traditional methods, *Magn. Reson. Chem.* 157 (2002) 242–252.
- [20] J. Meiler, E. Sanli, J. Junker, R. Meusinger, T. Lindel, M. Will, W. Maier, M. Köck, Validation of structural proposals by substructure analysis and ^{13}C NMR chemical shift prediction, *J. Chem. Inf. Comput. Sci.* 42 (2002) 241–248.
- [21] J. Meiler, M. Will, Genius: a genetic algorithm for automated structure elucidation from ^{13}C NMR spectra, *J. Am. Chem. Soc.* 124 (2002) 1868–1870.
- [22] V.R. Coluci, R. Vendram, R.S. Braga, D.S. Galvão, Identifying relevant molecular descriptor related to carcinogenic activity of polycyclic aromatic hydrocarbons (PAHs) using pattern recognition methods, *J. Chem. Inf. Comput. Sci.* 42 (2002) 1479–1489.
- [23] K. Varmuza, Applied Chemometrics. Available from: <<http://www.lcm.tuwien.ac.at>>.
- [24] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier, Amsterdam, 1997.
- [25] B.J. Wythoff, Backpropagation neural networks A tutorial, *Chemom. Intell. Lab. Syst.* 18 (1993) 115–155.
- [26] J. Zupan, J. Gasteiger, Neural networks: anew method for solving chemical problems or just a passing phase?, *Anal. Chim. Acta* 248 (1991) 1–30.
- [27] T. Elder, Correlation of experimental and ab initio ^{13}C NMR chemical shifts for monomeric lignin model compounds, *J. Mol. Struct. (Theochem.)* 505 (2000) 257–267.
- [28] S.A. Ralph, J. Ralph, L.L. Landucci, *NMR database of lignin and cell wall model compounds*, 1996. Available from: <<http://www.dfrc.wisc.edu/software.html>>.
- [29] Microsoft® Excel 2002, Copyright© 1985–2003 Microsoft Corporation.
- [30] J.J.P. Stewart, MOPAC, Semiempirical Molecular Orbital Program, QCPE, No. 455 (1983), Version 6, 1990.
- [31] B.M. Wise; R. Bro. Available from: <www.eigenvector.com>, PLS_Toolbox_303aforMATLAB.
- [32] Y. Danon. Available from: <<http://www.geocities.com/sciware/winnn.htm>>.