Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 26 (2007) 86–92
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Quantitative structure spectroscopy relationships of carbon-13 nuclear magnetic resonance chemical shifts of steroids

Jianbo Tong [a], Shuling Liu [b], Peng Zhou [c], Shengwan Zhang [a,*], S. Zhiliang Li [c,**]

[a] College of Chemistry and Chemical Engineering, Shanxi University, Taiyuan 030006, China
[b] College of Chemistry and Chemical Engineering, Shandong University, Jinan 250100, China
[c] College of Chemistry and Chemical Engineering, Key Laboratory of Biomedical Engineering of Educational Ministry and Chongqing Municipality, Chongqing University, Chongqing 400044, China

Received 22 January 2006; received in revised form 27 September 2006; accepted 27 September 2006
Available online 30 September 2006

Abstract

Quantitative structure spectroscopy relationships (QSSRs) are systematically studied for carbon-13 nuclear magnetic resonance (^{13}C NMR) spectroscopic simulation of steroid compounds. Both the atomic electronegativity interaction vector (AEIV) and the atomic hybridization state index (AHSI) are used for the expression of local chemical microenvironment and atomic hybridization state of 4434 resonance carbon atoms in 203 steroid molecules. A multiple linear regression (MLR) model is built after screening some insignificant parameters with the stepwise multiple regression (SMR) technique. Correlation coefficients of the developed model are $R^2_{cum} = 0.9341$ and $Q^2_{LOO} = 0.9336$ for classical estimation of molecular modeling and the cross-validation with leave-one-out (LOO) procedures, respectively, primarily indicating that the MLR model has good modeling stability and prediction ability. Furthermore, the superior performance of the MLR model is tested by the leave-33%-out (L33%O) cross-validation method, where the mean correlation coefficients of three test sets are $Q^2 = 0.9310$ and $Q^2_{ext} = 0.9196$ for both internal and external sets. In conclusion, AEIV and AHSI descriptors can be used for estimating and predicting ^{13}C NMR chemical shifts of steroids.

Keywords: Quantitative structure spectroscopy relationship (QSSR); Steroid; Carbon-13 nuclear magnetic resonance (^{13}C NMR) chemical shift; Atomic electronegativity interaction vector (AEIV); Atomic hybridization state index (AHSI)

1. Introduction

Nuclear magnetic resonance (NMR) spectroscopy is undoubtedly one of the most important methods for elucidating complicated structures and processes, including structural configuration [1,2], reaction mechanisms, dynamic processes, chemical equilibrium and even three-dimensional structures of protein molecules in aqueous solution [3]. However, it is commonly not enough to obtain all structural parameters from experiments, due to the diverse natures of the structures and reactivities. Therefore, the NMR simulation technique has attracted increasing interest. ^{13}C NMR simulation parameterizes ^{13}C atoms in different chemical microenvironments in organic molecules by employing knowledge of chemistry, mathematics and computer technology to build quantitative structure spectroscopy relationships (QSSRs). Such work was first studied by Grant and Paul [4] and Lindeman and Adams [5] for alkanes. Later, it was used to study compounds containing heteroatoms and/or rings. Over the last four decades, much effort has been focused on the development of ^{13}C NMR simulation, specifically for the following aspects: first, the development of computer science and computational chemistry have made ^{13}C NMR simulation to be an active field [6–11]; secondly, *ab initio* and density functional calculations for chemical shifts of organic molecules have recently emerged as one of the most promising new approaches for structure elucidation [12–20]; moreover, many computer systems have been developed for the prediction of NMR chemical shifts and for the elucidation of molecular structures. The combination of NMR chemical shift prediction systems and structural elucidation systems is becoming a powerful tool for automatic structural determination and identification [21,22]; in addition,

* Corresponding author. Tel.: +86 351 7011787; fax: +86 351 7011688.
** Corresponding author. Tel.: +86 23 65106677; fax: +86 23 65106677.
E-mail addresses: jianbotong@yahoo.com.cn (J. Tong), zswan@sxu.edu.cn (S. Zhang), liszyx@126.com, zlli2662@163.com (S.Z. Li).

1093-3263/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2006.09.011

both graph theory and topological theory are very powerful tools in NMR spectroscopic studies [23,24].

Steroids are very important compounds in extracts from natural plants [25,26]. These compounds show a wide spectrum of biological activities [27] and are widely used for synthesizing derivatives in the pharmaceutical industry. It is very important to determine the molecular structures of these compounds, by simulating their ^{13}C NMR spectroscopy. In this work, based on the two-dimensional structures of all organic molecules, we employed both the atomic electronegativity interaction vector (AEIV) and the atomic hybridization state index (AHSI) to characterize the chemical microenvironment and atomic hybridization state of 4434 equivalent resonance carbon atoms in 203 molecules of steroids. A four-parameter multiple linear regression (MLR) model was established to estimate and predict chemical shifts of 4434 equivalent resonance carbon atoms.

2. Principle and methodology

2.1. Calculation of atomic electronegativity interaction vector (AEIV)

It is well known that chemical shifts of all NMR spectroscopy are influenced by many factors, of which the most important ones are the local chemical microenvironment and hybridization state of atoms. Chemical microenvironment is related with distribution of electron cloud, while the distribution of electron cloud is related to the kind and number of atoms or groups bonded to the observed atom. For most organic molecules there are mainly several atoms including H, C, N, P, O, S, F, Cl, Br and I. They can be classified as five types of atoms (shown in Table 1) according to families of the Periodic Table of Elements. It was shown that the chemical microenvironment was related closely to atomic electronegativity and bond distance [28]. Considering atomic type, element electronegativity and bond length, a new vector, the atomic electronegativity interaction vector (AEIV) was developed to characterize the local chemical microenvironment. In AEIV, each atomic type is denoted by k (k = 1–5), which means there are k types of atoms linking to the examined atom. The five elements in the AEIV vector for the chosen carbon atom are defined as Eq. (1):

$$v_{i,k} = \sum_{j \in k, j \neq i}^{\mathrm{all}(j)} \frac{\chi_j}{d_{i,j}^6} \quad (1 \leq k \leq 5) \tag{1}$$

Table 1
Division of atomic type of atoms in organic compounds

	Type of atoms				
	1	2	3	4	5
Families of Periodic Table	IA	IVA	VA	VIA	VIIA
Atoms	H	C	N, P	O, S, Se	F, Cl, Br, I

Here, i represents the examined atom, j represents all atoms belonging to atomic type k, x is atomic relative electronegativity, being $x_A = X_A/X_C$, based on Pauling's scale (Table 2), defined as the ratio of atomic electronegativity to that of carbon atom, for example, relative electronegativity of oxygen atom is 3.44/2.55 = 1.349, d_{ij} represents the relative distance between the ith and jth atoms along the shortest bond connecting pathway, calculated as sum of relative bond lengths (Table 3) between atoms defined as the ratio of the bond length to the single C–C bond length, for example, relative bond length of C=O is 0.122/0.154 = 0.792. The interaction effect between atoms i and j is in proportion to sixth power of bond distance between five different atoms and target atoms [28], so the denominator in Eq. (1) is raised to the sixth power. The five elements in the AEIV vectors are noted as v_H, v_C, v_N, v_O and v_X, respectively.

AEIV is used to describe the effect of atoms and groups bonded to the central carbon. It distinguishes different effects of different atoms according to the relative bond length but not the number of chemical bonds [29]. AEIV includes effects of all atoms bonded to the chosen atom.

2.2. Calculation of atomic hybridization state index (AHSI)

The atomic hybridization state index (AHSI) descriptor is introduced to characterize the hybridization state of the examined atom itself. Based on the atomic intrinsic state

Table 2
Pauling's electronegativity and relative electronegativity of ordinary atoms in organic compounds

Atom	Pauling's electronegativity[a]	Relative electronegativity
C	2.55	1.0000
H	2.20	0.8627
Si	1.90	0.7451
N	3.04	1.1922
P	2.19	0.8588
As	2.00	0.7843
O	3.44	1.3490
S	2.58	1.0118
Se	2.45	0.9608
F	3.98	1.5608
Cl	3.16	1.2392
Br	2.96	1.1608
I	2.66	1.0431
Fe	1.90	0.7451
Na	0.90	0.3529
K	0.80	0.3137
Mg	1.20	0.4706
Mn	1.50	0.5882
Cu	2.00	0.7843
Zn	1.60	0.6275
Al	1.50	0.5882
Cr	1.60	0.6275
Co	1.80	0.7059
Ca	1.00	0.3922
Al	1.50	0.5882
B	2.00	0.7843

[a] Taken from Refs. [36,37].

Table 3
The practical bond length (nm) and relative bond length of ordinary chemical bonds in organic compounds

Bond type	Bond length[a]	Relative bond length
C–C	0.154	1.0000
C=C	0.134	0.8701
C≡C	0.120	0.7792
C≈C (alkene)[b]	0.144	0.9351
C≈C (benzene)[b]	0.139	0.9026
C–O	0.143	0.9286
C=O	0.122	0.7922
C≈O[b,c]	0.137	0.8896
C–S	0.182	1.1818
C=S	0.161	1.0455
C≈S[b,d]	0.171	1.1104
C–N	0.147	0.9545
C=N	0.130	0.8442
C≡N	0.116	0.7532
C≈N[b,e]	0.134	0.8701
C–P	0.181	1.1753
C–F	0.142	0.9221
C–Cl	0.178	1.1558
C–Br	0.191	1.2403
C–I	0.213	1.3831
N≈O[b,f]	0.122	0.7922
N≈N[b,g]	0.130	0.8442
N–N[h]	0.137	0.8896
P–O	0.156	1.0130
P=O	0.149	0.9675
N=N	0.124	0.8052
C–H	0.110	0.7143
N–H	0.103	0.6688
O–H	0.097	0.6299
S–H	0.134	0.8702

[a] Taken from Refs. [36,37].
[b] '≈' representation of conjugation bond.
[c] Representation in furan.
[d] Representation in thiophene.
[e] Representation in pyridine
[f] Representation in nitryl.
[g] Representation in pyridazine.
[h] Representation in pyrazole.

index (I) proposed by Hall and Kier [30], AHSI is calculated as Eq. (2):

$$\mathrm{AHSI} = \frac{\sqrt{v/4}((2/n)^2\delta_{\sigma+\pi}+1)}{\delta_\sigma} \quad (2)$$

In Eq. (2), v is the number of electrons in valence shell of the target atom, n the principal quantum number, $\delta_{\sigma+\pi}$ the number of total electrons forming σ and π bonds, and δ_σ is the number of only electrons forming σ bonds. The electrons of lone pairs do not participate in the formation of the covalent bond, and atoms in the same period of the Periodic Table of Elements, which form the same σ bond and π bond with non-hydrogen atoms and/or groups, will have the same $\delta_{\sigma+\pi}$ and δ_σ. Then a coefficient $(v/4)^{1/2}$ where 4 is the number of valence electrons of carbon atom, is introduced to make the atoms located in the same period, but a different main group such as $–CH_3$, $–NH_2$, –OH and –F having the same local chemical environment have different AHSI values. Differing from the primary definition by Hall and Kier [30], AHSI includes the coefficient of $(v/4)^{1/2}$ and does not detract electrons forming bonds with hydrogen atoms from the calculation of $\delta_{\sigma+\pi}$ and δ_σ. Table 4 displays AHSIs for different hybridization states of carbon, oxygen and nitrogen atoms.

3. Results and discussion

3.1. Data sets and model validations

^{13}C NMR chemical shifts of 4434 carbon atoms in 203 steroids were obtained from Ref. [31] including 72 androstanes, 32 lactones, 26 trihydroxylcholans, 7 pregnanes, 14 estranes, 29 bile acids, 8 spirostanes, 9 ergostanes and 6 other steroids. Their molecular structural schemes are given in Appendix Fig. 1. Carbon atoms of the 203 molecules are coded in the following way: carbon atoms 1–19 of the first molecule are coded 1–19; carbon atoms 1–19 of the second molecule are coded 20–38; and so on, until carbon atoms 1–27 of the 203rd molecule are coded 4408–4434.

The multiple linear regression (MLR) is a classic modeling technique. MLR was used to build a QSSR model of all 203 steroids with its estimation ability and prediction power examined. In recent years, the statistical parameter correlation coefficient (Q_{LOO}) [32], for the leave-one-out (LOO) cross-validation, has been used as a means of indicating the predictive ability of a model. Generally, many authors consider high a Q_{LOO} value as an indicator or even as the ultimate proof of the high predictive power of a QSAR model. However, the recent study of Tropsha and co-workers showed that there are no evident relationships between the value of Q_{LOO} and actual predictive power of a QSAR model, and external validation is required [33–35]. The predictive power of a model on external dataset can be expressed by Q_{ext} in contrast to the original Q_{LOO}:

$$Q_{\mathrm{ext}} = \sqrt{1 - \frac{\sum_{i=1}^{\mathrm{test}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{\mathrm{test}}(y_i - \bar{y}_{\mathrm{tr}})^2}} \quad (3)$$

In Eq. (3), both y_i and $\hat{y}_i$ are the observed and calculated values of the test dataset, and $\bar{y}_{\mathrm{tr}}$ is the mean value of observed values of the training dataset.

In order to prove the validity and stability of the model, the whole data set is systematically divided into three subsets and each subset is predicted by using the other two subsets as the

Table 4
AHSI values of three kinds of atoms of carbon, oxygen and nitrogen with different hybridization state

Atomic type	AHSI
C sp^3	1.2500
C sp^2	1.6667
C sp	2.5000
O sp^3	1.8371
O sp^2	3.6742
N sp^3	1.4907
N sp^2	2.2361
N sp	4.4721

Table 5
Analysis of variables by SMR for ^{13}C NMR chemical shifts of 4434 carbon atoms

m	a_0	a_1	a_2	a_3	a_4	a_5	R^2_{cum}	RMSE	Q^2_{LOO}	$RMSE_{LOO}$
1	−305.7968	–	–	–	–	274.2865	0.7850	18.1694	0.8858	18.1763
2	−260.7482	–	–	12.0582	–	236.1416	0.8819	13.4618	0.9389	13.4965
3	−214.0839	−1.6089	–	10.2723	–	212.7742	0.9293	10.4260	0.9638	10.4496
4	−168.0513	−3.9130	−14.6520	3.6475	–	224.1326	0.9341	10.0590	0.9336	10.0917
5	−170.0189	−3.8166	−14.0482	3.9266	15.0502	223.6962	0.9341	10.0431	0.9336	10.0861

training set. Samples 1–1500 are chosen as the first subset; samples 1501–3000 are chosen as the second subset; the remained samples 3001–4434 are the third subset (Appendix Table 1).

3.2. Modeling with multiple linear regression

MLR is used to obtain an optimal result in the least squares (LS) significance by linear fitting. In order to assure their statistical significance, the descriptors are screened before being submitted to MLR analysis. Only information-rich descriptors pass the screening step onto regression analysis. The forward stepwise multiple regression (SMR) method was employed for variables screening (see Table 5 for details).

Table 5 indicated that the values of the correlation coefficients, R^2_{cum} and Q^2_{LOO}, increased gradually with the increase in the number of the variables (m). However, the values of root-mean-square errors, RMSE and $RMSE_{LOO}$, decreased with the increase of m. In order to assure their statistical significance, only information-rich descriptors pass the screening step onto regression analysis. And descriptors with greater than 80% identical values were eliminated since those descriptors were not encoding the structural differences between steroids that account for their chemical shift differences. Besides, it was found that when m was equal to 4, $RMSE_{LOO}$ reached the relative least value (10.0917), R^2_{cum} reached the largest value (0.9341), and Q^2_{LOO} the high value (0.9336). In addition, a t-test of Eq. (4) was performed. The results showed that all variables were distinctly significant (the minimum t value 7.155 was bigger than the critical value $t_{crit} = 2$) and did not have obvious multicollinearity (the biggest variance inflation factor was 43.464, less than VIF = 50, the empirical value). Therefore, a four-parameter MLR model was established to study relationships between AEIV, AHSI and ^{13}C NMR chemical shifts of 4434 samples, which was presented as follows

$$CS = -168.0513 - 3.9130\nu_H - 14.6520\nu_C + 3.6475\nu_O + 224.1326\,AHSI \quad (4)$$

$$\text{model fitting}: \; n = 4434, \; m = 4, \; R^2_{cum} = 0.9341, \; RMSE = 10.0590, \; F = 15,681; \quad \text{cross-validation}: \; Q^2_{LOO} = 0.9336, \; RMSE_{LOO} = 10.0917, \; F_{LOO} = 15,572$$

where ν_H is the effect item of hydrogen atoms on the central carbon atom, ν_C the effect item of other carbon atoms on the central carbon atom, ν_O the effect item of oxygen and sulfur atoms on the central carbon atom, n the number of samples used for model building, m the number of variables, R the multiple correlation coefficient, RMSE the root-mean-square error, F the Fisher statistic, LOO means leave-one-out cross-validation.

In Eq. (4), the regression coefficient of AHSI is far larger than those of the three AEIV descriptors, which indicate that the change of atomic hybridization state has obviously a greater effect on the shielding electric cloud than those exerted by other atoms. Observed results help us to understand that chemical shifts of sp^3 hybridization C atoms are in the range of 10–90 ppm, while those of sp^2 hybridization C atoms are above 120 ppm. MLR, SMR and cross-validation are performed by SPSS 12.0 software and visual Basic program written by the authors, respectively.

3.3. Validation of the QSSR model

An excellent model should not only have good estimation results of internal samples, but have fine prediction ability of external samples. Fig. 1 presents a plot of the observed chemical shifts versus predicted ones. Fig. 2 presents the calculated results related with residual distribution. Obviously both AEIV and AHSI were related with ^{13}C NMR chemical shifts of steroids, because in Fig. 1 one-to-one correlation line clearly demonstrated that the developed QSSR model was accurate, and in Fig. 2 most of residual distribution dots were within triple root-mean-square deviations (i.e. within the area enclosed with dotted line). A similar conclusion was obtained when comparing Cook's

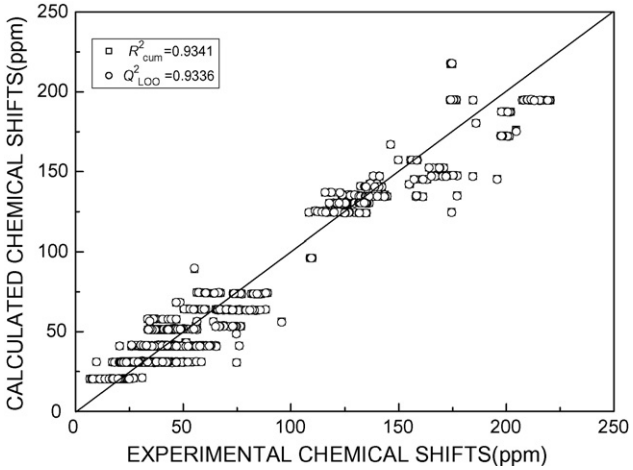

Fig. 1. Plot of estimated value (squares) as well as predicted value (circles) by LOO cross-validation for 4434 samples vs. observed value.

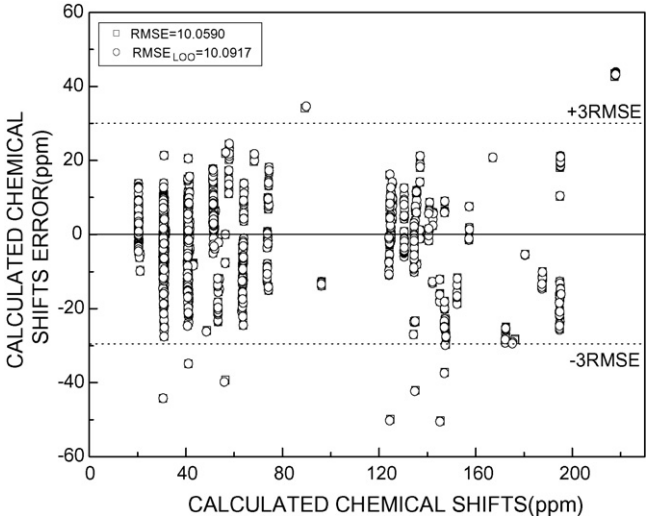

Fig. 2. Residual distribution for 4434 samples of estimated value (squares) and predicted value (circles) (The dashed is triple root mean square error.)

distance of training samples with the centered leverage values (see Fig. 3 for details). Samples 609, 647, 2377, 4390, 3068, 3069, 3110, 3135, 3160, 3185, 3210, 3235, 3260, 3285, 3310, 3335, 3360, 3385, 3410, 3435, 3460, 3485, 3510, 3535, 3560, 3585, 3610, 3635, 3660, 3685, 3710, 3735, 3760, 3785 and 3810 were a bit extraordinary, which might be attributed to different experimental conditions, or insufficiency of AEIV and AHSI descriptors representation the structural characterization, or insufficiency of the MLR model in establishing a linkage between structural parameters and property, or particularity of samples themselves as outliers. It was found in Fig. 2 that carbonyl carbon atoms, carbon atoms connected with sp^3 hybridization oxygen atoms and carbon atoms with complex dimensional structure had greater deviations. Since there exist mutual interactions between atoms in a molecule, all atoms around exert impact on the central carbon atom.

In steroid compounds, carbonyl carbon atoms have great chemical shifts of above 200 ppm in general. However, in AEIV, carbonyl carbon atoms considered as alkene carbon atoms are classified into the same type atoms, so the special effect of carbonyl on carbon atoms is ignored to some extent.

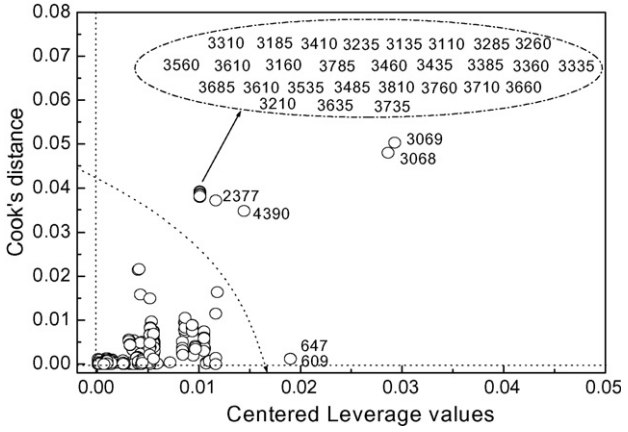

Fig. 3. Plot of Cook's distance vs. centered leverage values for 4434 samples.

Moreover, two-dimensional descriptors of AEIV and AHSI ignore three-dimensional isomers and cannot discriminate cis/trans isomers and chirality.

It was found from Fig. 3 that samples 609 and 647 had carbonyl carbon atoms (no. 3 carbon atoms of no. 33 and 35 molecules, respectively) in conjugated systems formed with a carbonyl group and two double carbon bonds. Though they are somewhat special in structure, their calculated chemical shifts are close to their experimental values, which is in accordance with shorter Cook's distance of these two samples. Samples 3068 and 3069, which are nos. 16 and 17 carbon atoms in two conjugated carbonyls of no. 150 molecule, are unique in 203 molecules. The descriptors in the proposed model do not contain the ν_X (X = F, Cl, Br, I) parameter, so they cannot express properly effects of fluorine atoms. Sample 2377 (no. 6 carbon atom of no. 115 molecule) is unique, because it connects with fluorine atom. Similarly, sample 4390 is no. 11 carbon atom in no. 202 molecule which is the number of three-membered cycle, has also unique structural feature. While samples 3110, 3135, 3160, 3185, 3210, 3235, 3260, 3285, 3310, 3335, 3360, 3385, 3410, 3435, 3460, 3485, 3510, 3535, 3560, 3585, 3610, 3635, 3660, 3685, 3710, 3735, 3760, 3785 and 3810 are carbon atoms of ester groups, indicating that the descriptors here do not express properly effects duo to their special structures. Considering the fact that they are not too exceptional and the 4434-sample set had only 29 such samples, they are retained in the developed model.

The squared cross-validation correlation coefficient Q^2_{LOO} is 0.9336, in comparison with the squared model-estimation correlation coefficient R^2_{cum} (0.9341), indicating good stability of the QSSR model. To further demonstrate the absence of chance correlation, the whole data set was divided into three subsets and each subset was predicted by using the other two subsets as training set. In this procedure, the same descriptors are retained in the MLR equation, but the coefficients are allowed to be varied. The results are shown in Table 6, with average training quality of $R^2_{cum} = 0.9354$, average validating quality of $R^2_{LOO} = 0.9310$, and average predicting $Q^2_{ext} = 0.9196$ for three test datasets. Table 7 presents the observed and predicted chemical shifts for all the carbons of two compounds. The data illustrated that the four-parameter model established with AEIV and AHSI descriptors has high precision for the estimation of ^{13}C NMR chemical shifts of steroids, so they are good descriptors to characterize carbon atoms in different chemical microenvironments and hybridization states.

In Fig. 1 we see several bandings of the predicted values, due to several reasons including insufficient encoding information contained in the descriptors and the actual non-linear relationship between the structural descriptors and the investigated ^{13}C NMR chemical shifts. MLR is not able to take advantage of non-linear information that may be contained within AEIV and AHSI descriptors.

4. Conclusion

Based on two-dimensional structures, AEIV and AHSI descriptors were proposed to characterize the external chemical

Table 6
Verification of statistical validity of the model

Training sets	R^2_{cum} [a]	RMSE[b]	Predicted sets	Q^{2c}	Q^2_{ext} [d]	$RMSE_{ext}$ [e]
1 and 2	0.9425	9.7807	3	0.9256	0.8976	10.8934
1 and 3	0.9268	9.7594	2	0.9436	0.9394	9.5189
2 and 3	0.9368	10.1305	1	0.9237	0.9218	9.1350
Average	0.9354	9.8902	Average	0.9310	0.9196	9.8491

[a] Cumulative multiple correlation coefficient of training set.
[b] Root-mean-square error of training set.
[c] Cumulative multiple correlation coefficient of test set.
[d] External Q^2 of test set.
[e] Root-mean-square error of test set.

Table 7
Experimental and calculated ^{13}C NMR chemical shifts of the first and second molecule

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	CS_{OBS}	38.3	21.9	26.5	28.8	47.0	28.8	31.7	35.0	56.5	36.9	37.5	215.3	54.9	54.6	24.8	19.5	31.9	11.9	17.7
	CS_{CAL}	30.6	30.9	30.9	30.8	40.9	30.7	30.7	40.9	40.8	51.3	31.1	194.8	51.8	40.8	30.8	30.9	30.7	20.4	20.4
2	CS_{OBS}	215.8	38.8	28.0	28.0	49.8	28.0	31.5	36.2	47.2	52.0	22.7	38.3	41.0	54.4	25.5	20.4	40.4	12.3	17.8
	CS_{CAL}	194.8	31.3	30.9	30.8	40.9	30.7	30.7	40.9	40.8	51.7	30.7	30.6	51.4	40.8	30.8	30.9	30.7	20.4	20.4

microenvironments together with their hybridization states of atoms in the examined molecules. Good results were obtained by simulating the ^{13}C NMR chemical shifts of steroids. The established model not only explained the relationships between ^{13}C NMR chemical shifts of steroids and molecular structural parameters, but also provided a new method to calculate ^{13}C NMR chemical shifts for some unknown steroids. The importance of this method is its simple calculation, fine precision and good effectiveness to reflect the shielding effect of atoms in different chemical environments. A limitation of the technique is that descriptors did not give explanation of three-dimensional effect, especially for chiral-molecular structures. Further studies are in progress.

Acknowledgments

The authors gratefully acknowledge financial support from both Shanxi province industry innovation foundation (2006031204) and State Key Laboratory of Chemobiosensing and Chemometrics Foundation (2005012).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2006.09.011.

References

[1] S. Witkowski, D. Maciejewska, I. Wawer, ^{13}C NMR studies of conformational dynamics in 2,2,5,7,8-entamethylchroman-6-ol derivatives in solution and the solid state, J. Chem. Soc., Perkin Trans. 2 (2000) 1471–1476.
[2] H. Neuvonen, K. Neuvonen, Correlation analysis of carbonyl carbon ^{13}C NMR chemical shifts, IR absorption frequencies and rate coefficients of nucleophilic acyl substitutions. A novel explanation for the substituent dependence or reactivity, J. Chem. Soc., Perkin Trans. 2 (1999) 1497–1502.
[3] K. Wüthrich, The way to NMR structures of proteins, Nat. Struct. Biol. 8 (2001) 923–925.
[4] 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.
[5] L.P. Lindeman, J.Q. Adams, Carbon-13 nuclear magnetic resonance spectrometry—chemical shifts for the paraffins through C9, Anal. Chem. 43 (1971) 1245–1252.
[6] 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.
[7] D.L. Clouser, P.C. Jurs, The simulation of ^{13}C nuclear magnetic resonance spectra of dibenzofurans using multiple linear regression analysis and neural networks, Anal. Chim. Acta 321 (1996) 127–135.
[8] O. Ivanciuc, J.P. Rabine, D. Cabrol-Bass, A. Panaye, J.P. Doucet, ^{13}C NMR chemical shift prediction of sp^2 carbon atoms in acyclic alkenes using neural networks, J. Chem. Inf. Comput. Sci. 36 (1996) 644–653.
[9] D.L. Clouser, P.C. Jurs, Simulation of the ^{13}C nuclear magnetic resonance spectra of ribonucleosides using multiple linear regression analysis and neural networks, J. Chem. Inf. Comput. Sci. 36 (1996) 168–172.
[10] O. Ivanciuc, J.P. Rabine, D. Cabrol-Bass, ^{13}C NMR chemical shift sum prediction for alkanes using neural networks, Comput. Chem. 21 (1997) 437–443.
[11] M. Nohair, D. Zakarya, Autocorrelation method adapted to generate new atomic environments: application for the prediction of 13-C chemical shifts of alkanes, J. Chem. Inf. Comput. Sci. 42 (2002) 586–591.
[12] R. Koch, B. Wiedel, C. Wentrup, ^{13}C NMR calculations on azepines and diazepines, J. Chem. Soc., Perkin Trans. 2 (1997) 1851–1859.
[13] J.W. Wiench, L. Stefaniak, E. Grech, E. Bednarek, Two amidine derivatives studied by ^{1}H, ^{13}C, ^{14}N, ^{15}N NMR and GIAO-CHF calculations, J. Chem. Soc., Perkin Trans. 2 (1999) 885–889.
[14] M. Barańska, K. Czarniecki, L.M. Proniewicz, Experimental and calculated ^{1}H, ^{13}C, ^{15}N NMR spectra of famotidine, J. Mol. Struct. 563–564 (2001) 347–351.
[15] S. Vázquez, GIAO-DFT study of ^{13}C NMR chemical shifts of highly pyramidalized alkenes, J. Chem. Soc., Perkin Trans. 2 (2002) 2100–2103.
[16] C. Bassarello, P. Cimino, L. Gomez-Paloma, R. Riccio, G. Bifulco, Simulation of 2D ^{1}H homo- and ^{1}H-^{13}C heteronuclear NMR spectra of organic molecules by DFT calculations of spin–spin coupling constants and ^{1}H and ^{13}C chemical shifts, Tetrahedron 59 (2003) 9555–9562.
[17] T. Zolek, K. Paradowska, I. Wawer, ^{13}C CP MAS NMR and GIAO-CHF calculations of coumarins, Solid State Nucl. Mag. 23 (2003) 77–87.

[18] K. Tuppurainen, J. Ruuskanen, NMR and molecular modeling in environmental chemistry: prediction of ^{13}C chemical shifts in selected C_{10}-chloroterpenes employing DFT/GIAO theory, Chemosphere 50 (2003) 603–609.
[19] A. Balandina, V. Mamedov, X. Franck, B. Figadère, S. Latypov, Application of quantum chemical calculations of ^{13}C NMR chemical shifts to quinoxaline structure determination, Tetrahedron Lett. 45 (2004) 4003–4007.
[20] W.N. Moss, N.S. Goroff, Theoretical analysis of the ^{13}C NMR of iodoalkynes upon complexation with lewis bases, J. Org. Chem. 70 (2005) 802–808.
[21] J.S.L.T. Militão, V.P. Emerenciano, M.J.P. Ferreira, D. Cabrol-Bass, M. Rouillard, Structure validation in computer-supported structure elucidation: ^{13}C NMR shift predictions for steroids, Chemom. Intell. Lab. Syst. 67 (2003) 5–20.
[22] H. Satoh, H. Koshino, J. Uzawa, T. Nakata, CAST/CNMR: highly accurate ^{13}C NMR chemical shift prediction system considering stereochemistry, Tetrahedron 59 (2003) 4539–4547.
[23] S.S. Liu, Z.N. Xia, Y. Liu, S.X. Cai, Z. Li, An atomic electronegative distance vector and carbon-13 nuclear magnetic resonance chemical shifts of alcohols and alkanes, Chin. J. Chem. 18 (2000) 165–174.
[24] L.P. Zhou, L.L. Sun, Y. Yu, W. Lu, Z.L. Li, Prediction of carbon-13 NMR chemical shift of alkanes with rooted path vector, J. Mol. Graph. Model. 25 (2006) 333–339.
[25] N. Sarvilinna, H. Eronen, S. Miettinen, A. Vienonen, T. Ylikomi, Steroid hormone receptors and coregulators in endocrine-resistant and estrogen-independent breast cancer cells, Int. J. Cancer 118 (2006) 832–840.
[26] G.B. Dijksterhuis, B. Engel, P. Walstra, M. FontiFurnols, H. Agerhem, K. Fischer, M.A. Oliver, C. Claudi-Magnussen, F. Siret, M.P. Béague, D.B. Homer, M. Bonneau, An international study on the importance of androstenone and skatole for boar taint. II. Sensory evaluation by trained panels in seven European countries, Meat Sci. 54 (2000) 261–269.
[27] S. Shibata, New Natural Products and Plants Drugs with Pharmacological, Biological or Therapeutical Activity, Springer, Berlin, 1977.
[28] S.S. Liu, H. Liu, B.M. Yu, C.Z. Cao, S.Z. Li, Investigation on quantitative relationship between chemical shift of carbon-13 nuclear magnetic resonance spectra and molecular topological structure based on a novel atomic distance-edge vector (ADEV), J. Chemom. 15 (2001) 427–438.
[29] W. Bremser, HOSE—a novel substructure code, Anal. Chim. Acta 103 (1978) 355–365.
[30] L.H. Hall, L.B. Kier, Electrotopological state index for atom types: a novel combination of electronic, topological, and valence state information, J. Chem. Inf. Comput. Sci. 35 (1995) 1039–1045.
[31] D.Q. Yu, J.S. Yang, J.X. Xie, Fifth Handbook of Analytical Chemistry, Chemical Industry Press, Beijing, 1989, pp. 806–822 (in Chinese).
[32] S. Wold, Cross-validation estimation of the number of components in factor and principal components models, Technometrics 20 (1978) 897–903.
[33] A. Golbraikh, A. Tropsha, Beware of q^2! J. Mol. Graph. Model. 20 (2002) 269–276.
[34] P. Gramatica, P. Pilutti, E. Papa, Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling, J. Chem. Inf. Comput. Sci. 44 (2004) 1794–1802.
[35] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69–77.
[36] G.H. Aylward, T.J. Findlay, SI Chemical Data, (H.N. Zhou, Trans.), 2nd ed., High Education Press, Beijing, 1985, p. 94 (in Chinese).
[37] H.Q. Zhang, Z. Chen, Y.J. Lin, X.L. Ma, Y.H. Zhang, L.Z. Song, H. Yang, D.J. Wang, Handbook of Chemurgy, Science Press, Beijing, 2001, pp. 682–723 (in Chinese).