Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 25 (2006) 333–339
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Prediction of carbon-13 NMR chemical shift of alkanes with rooted path vector

L.P. Zhou [a,*], L.L. Sun [a], Y. Yu [a], W. Lu [a], Z.L. Li [b]

[a] *Chongqing Key Laboratory of Biochemistry and Molecular Pharmacology, Chongqing University of Medical Sciences, Chongqing 400016, People's Republic of China*

[b] *College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400044, People's Republic of China*

Received 10 October 2005; received in revised form 24 January 2006; accepted 24 January 2006
Available online 28 February 2006

Abstract

Systematic studies were further made on graph theory in quantitative structure-spectrum relationships (QSSR) for various areas of spectroscopies. Chemical shifts (CS) in alkanes for carbon-13 nuclear magnetic resonance (^{13}C NMR) were well correlated with a set of novel molecular graph indices, called the rooted path vector of various lengths, as several multivariate regression equations as following:

$$CS = 3.022 + 5.336P_1 + 7.356P_2 - 1.648P_3 + 0.83859P_4 + 0.210P_5 - 0.138P_6 - 0.506P_7 + 2.486P_8 - 1.669P_9;$$
$$n = 402,\ m = 9,\ R = 0.944,\ R_{CV} = 0.9413,\ \text{S.D.} = 3.333,\ F = 358.343,\ U = 35833.211,\ Q = 4355.422$$

for all types (primary, secondly, tertiary, quaternary as well as methane) of carbon atoms

$$CS = 0.983 + 6.811P_1 + 7.584P_2 - 2.029P_3 + 0.809P_4 + 0.106P_5 + 0.043P_6 - 0.124P_7 + 1.715P_8 - 1.101P_9;$$
$$n = 374,\ m = 9,\ R = 0.975,\ R_{CV} = 0.9737,\ \text{S.D.} = 2.303,\ F = 773.372,\ U = 36912.109,\ Q = 1930.363$$

for primary, secondly, tertiary (including methane) carbon atoms; and

$$CS = 27.819 + 2.351P_2 + 0.549P_3 - 0.440P_4 + 0.170P_5 - 0.050P_6;$$
$$n = 27,\ m = 5,\ R = 0.992,\ R_{CV} = 0.9674,\ \text{S.D.} = 0.324,\ F = 265.418,\ U = 138.891,\ Q = 2.198$$

for quaternary carbon atoms, respectively.

Quite good estimation and prediction results were obtained from the quantitative molecular modeling and the performance of multiple linear regression (MLR) equations were tested to work well through cross-validation (CV) with the leave-one-out (LOO) procedure.

Keywords: Quantitative structure–spectrum relationship (QSSR); Graph theory index; Rooted path vector; Carbon-13 nuclear magnetic resonance (^{13}C NMR) chemical shift; Alkanes

1. Introduction

The analysis of nuclear magnetic resonance of carbon-13 originated from the G–P method proposed by Grant and Paul [1] in 1964 and the L–A method by Lindemann and Adams [2] in 1971 for estimating carbon-13 nuclear magnetic resonance (^{13}C NMR) spectra of alkanes based on the addition of group contribution. In their equations of additional contribution, many empirical coefficients or modified parameters representing effects of substituent groups on the examined properties such as the chemical shift sum (CSS) are not easy to obtain. Afterwards, a modified simulation method by means of multiple linear regression (MLR) was developed and the scope

* Corresponding author. Tel.: +86 2368485808; fax: +86 2368485808.
E-mail address: lpzsh@hotmail.com (L.P. Zhou).

1093-3263/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2006.01.008

of application was expanded by Small and Jurs [3,4]. On the early research of ^{13}C NMR, a index of path number called Wiener index (W) which was defined as the total number of C–C bonds in the molecule of alkanes and a polarization index called P_3 was developed by Wiener H [5], P_3 was defined as the number of pathways in which the C–C–C–C fragment of an alkane can be superimposed on the alkane molecule and is equal to the path count of length three. Wiener approach was also applied to molecules containing heteroatoms. And recently, many mathematical, chemometric, statistic methods predicting ^{13}C NMR chemical shifts (CS) of organic compounds have been developed by means of artificial neural network [6–15] algorithm and/or multiple linear regression method [16]. Neural networks were also used to predict a group of alkanes [17–19]. Thereinto a lot of attention has been paid on employment of more general structural parameters, in particular, those derived from chemical graph theory [20–27].

However, the conditions that a new graph-theoretical index of the compound should at least have both good discrimination for diverse isomers and high correlation with physico-chemical property/biological activity are relatively difficult to fulfill simultaneously. In this paper, we have developed a rooted path vector base on two most fundamental structural variables, one for the distance between atoms in the molecular graph and another for the edge of the adjacency in the molecular graph. This rooted path vector will be used to illustrate the chemical environment of the examined atom in the alkane molecule, which is composed of various atoms. We expect to obtain a practical structure–property equation for the representative physicochemical properties such as the chemical shift of ^{13}C NMR spectra. Then, those quantitative structure–property correlation equations can be used to predict the values of ^{13}C NMR chemical shifts for various alkane compounds.

2. Principle and methodology

A ^{13}C NMR spectrum reflects a local property of a compound through the chemical shift signals of various chemically or magnetically equal carbon-13 atoms. There the property can be attributed to the corresponding atom in the molecule structure. So to relate NMR chemical shift with the molecule structure, we should get the information of the atom environment and characterize the structural feature of all examined atoms.

2.1. Variable description using root path vector

To properly characterize chemical structure of molecules or to select variable is one of the key fundamental problems in QSPR/QSAR studies. Many various topological indices are developed and used in quantitative structure–property/activity relationships (QSPR/QSAR). Various molecular structures can be conveniently expressed as different chemical structural graphs or carbon skeleton graph in chemistry. First, the path with length 1–9 in alkanes are accounted as C–C, C–C–C. . . and C–C–C–C–C–C–C–C–C segment, respectively. In other words, when the examined atom is taken as the rooted atom, P_1 is the number of path length 1 (C–C). In the similar way, the rest ones, P_i (i = 2, 3, . . ., 9) may be deduced. Accordingly, P_1–P_9 composed the rooted path vector describing the examined atom. $\boldsymbol{P} = (P_1, P_2, \ldots, P_9)'$ or $\boldsymbol{x} = (X_1, X_2, \ldots, X_9)'$. Taking 2,2,3-trimethypentane (Fig. 1) as an example, three carbons C^1, C^7 and C^8 in our case are chemically equivalent while the other five carbons, C^2, C^3, C^4, C^5, C^6 are all chemically non-equivalent, corresponding six values of chemically shift. The procedure of creating a rooted path vector for the total chemically non-equivalent carbon atoms in the examined molecular graph is illustrated as follows: for each atom, no.1, no.7 or no.8: the number of the path length 1 is 1, $P_1 = 1$; the number of the path length 2 is 3, $P_2 = 3$; the number of the path length 3 is 2, $P_3 = 2$; the number of the path length 4 is 1, $P_4 = 1$; the number of the path length more than 4 are all 0, $P_i = 0$ (i = 5, 6, 7, 8, 9). Therefore, the rooted path vector is written as $\boldsymbol{p}_1 = \boldsymbol{p}_7 = \boldsymbol{p}_8 = (1, 3, 2, 1, 0, 0, 0, 0, 0)'$. By following the same way, one can obtain the various rooted path vectors for the other five chemically non-equivalent atoms

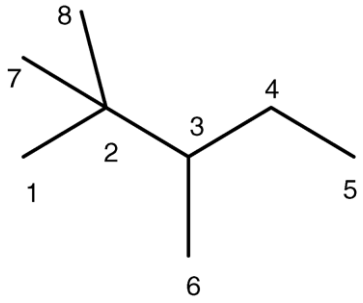

Fig. 1. The carbon skeleton graph of 2,2,3-trimethypentane.

$$\boldsymbol{p}_2 = (4,2,1,0,0,0,0,0,0)', \quad \boldsymbol{p}_3 = (3,4,0,0,0,0,0,0,0)',$$
$$\boldsymbol{p}_4 = (2,2,3,0,0,0,0,0,0)', \quad \boldsymbol{p}_5 = (1,1,2,3,0,0,0,0,0)',$$
$$\boldsymbol{p}_6 = (1,2,4,0,0,0,0,0,0)'$$

2.2. Regression analysis using MLR

For each of four types of carbon atoms, a simple relationship between the rooted path vector and chemical shift of ^{13}C NMR spectra can be represented by

$$y = \sum_{i=0}^{i=m} x_i \times b_i = b_0 + \sum_{i=0}^{i=m} x_i \times b_i = bx \tag{1}$$

where $\boldsymbol{b} = (b_0, b_1, \ldots, b_m)$ refers to the vector of regression coefficient, $\boldsymbol{x} = (1, P_1, \ldots, P_m)$ is the vector of descriptor coefficient, in which, b_0 is constant, i.e. the intercept term. b_i is slope which can be regarded as the contribution value of the element P_j in $\boldsymbol{P}$ vector for the examined atom. $\boldsymbol{b}'$ is the transpose of $\boldsymbol{b}$. They can be attained by multiple linear regression method (MLR). The stability and prediction capacity for the external samples are tested by cross-validation technique. For ^{13}C NMR spectra, y is the chemical shift (CS). i.e. CS = $\boldsymbol{b}'\boldsymbol{p}$.

3. Algorithm section

There are 964 chemically non-equivalent carbon atoms distributed in 152 alkanes under examination of those ^{13}C NMR

chemical shifts of 402 non-equivalent carbon atoms are known. The experimentally measured chemical shifts relative to tetramethyl silane (TMS) of the alkanes compounds including the alkanes with the number of carbon per alkane spanned from one through 10, undecane and 2,2,4,6,6-pentamethylheptane work out at 402 carbon atoms are taken from reference [1,4]. The range of the experimental chemical shift values was located from −2.3 ppm for methane to 31.39 ppm for 2,2,4,6,6-pentamethylheptane. First, we count the path number of the examined 964 carbon atoms in these 152 compounds to get the nine elements P_1–P_9 and ignore the rare element P_{10} since only undecane have it. A true-basic program of MLR in our laboratory was employed for molecular modeling and a cross-validation with leave-one-out (LOO) procedure was done to test stability of the modeling equation.

4. Results and discussion

First, all known primary, secondly, tertiary, quaternary as well as methane carbon atoms were composed of a working data set for molecular modeling to estimate and to predict the chemical shifts of all types of the carbon atoms. The multiple linear regression equation and the related statistic are as follows:

$$\begin{aligned} \mathrm{CS} &= 3.022 + 5.336P_1 + 7.356P_2 - 1.648P_3 \\ &\quad + 0.839P_4 + 0.210P_5 - 0.138P_6 - 0.506P_7 \\ &\quad + 2.486P_8 - 1.669P_9; \\ n &= 402, m = 9, R = 0.944, \mathrm{S.D.} = 3.333, F = 358.343, \\ U &= 35833.211, Q = 4355.422 \end{aligned} \quad (2)$$

where n, m, R, S.D., U and Q are the sample number, variable number, regression coefficient, standard deviation, regression square sum and total residual deviation, respectively. F is F-statistic value, which is defined as: $F = (U/m)/[Q/(n - m - 1)]$. To test the stability and performance of the model, a cross-validation testing was done through the leave-one-out procedure and the results of cross-validation were given by $R^2_{CV} = 0.886$, $\mathrm{S.D.}_{CV} = 3.415$, $F_{CV} = 339.311$, $U_{CV} = 35616.699$, $Q_{CV} = 4571.934$.

Table 1
Path enumeration and chemical shift for some of the alkanes

NA	NB	Compound	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	CS (obs)	All (exp)	Err.	Err (%)	0 − 3 + 4(28)	Err.	Err (%)	1234 (27)	Err.	Err (%)	1 + 0	Err.	Err (%)
001	1	1	0	0	0	0	0	0	0	0	0	−2.3	3.022	5.322	−231.41	0.983	3.283	−142.73				−2.300	0.000	0.00
002	2	2	1	0	0	0	0	0	0	0	0	5.7	8.358	2.658	46.64	7.794	2.094	36.74	7.536	1.836	32.21	7.536	1.836	32.22
003	3	3	1	1	0	0	0	0	0	0	0	15.4	15.715	0.315	2.04	15.378	−0.022	−0.14	15.683	0.283	1.84	15.683	0.283	1.84
004	3	3	2	0	0	0	0	0	0	0	0	15.9	13.694	−2.206	−13.87	14.606	−1.294	−8.14	16.701	0.801	5.04			
005	4	4	1	1	1	0	0	0	0	0	0	13.1	14.066	0.966	7.38	13.349	0.249	1.90	12.846	−0.254	−1.94	12.846	−0.254	−1.94
006	4	4	2	1	0	0	0	0	0	0	0	24.9	21.051	−3.849	−15.46	22.190	−2.710	−10.88	24.205	−0.695	−2.79			
007	5	2M3	1	2	0	0	0	0	0	0	0	24.3	23.071	−1.229	−5.06	22.963	−1.337	−5.50	23.829	−0.471	−1.94	23.829	−0.471	−1.94
008	5	2M3	3	0	0	0	0	0	0	0	0	25.0	19.030	−5.970	−23.88	21.417	−3.583	−14.33	24.309	−0.691	−2.76			
009	6	5	1	1	1	1	0	0	0	0	0	13.5	14.906	1.406	10.41	14.158	0.658	4.88	13.573	0.073	0.54	13.573	0.073	0.54
010	6	5	2	1	1	0	0	0	0	0	0	22.2	19.402	−2.798	−12.60	20.161	−2.039	−9.19	22.067	−0.133	−0.60			
011	6	5	2	2	0	0	0	0	0	0	0	34.1	28.407	−5.693	−16.70	29.774	−4.326	−12.69	31.708	−2.392	−7.01			
012	7	2M4	1	2	1	0	0	0	0	0	0	21.9	21.422	−0.478	−2.18	20.934	−0.966	−4.41	20.992	−0.908	−4.15	20.992	−0.908	−4.14
013	7	2M4	3	1	0	0	0	0	0	0	0	29.9	26.387	−3.513	−11.75	29.001	−0.899	−3.01	29.682	−0.218	−0.73			
014	7	2M4	2	2	0	0	0	0	0	0	0	31.6	28.407	−3.193	−10.11	29.774	−1.826	−5.78	31.708	0.108	0.34			
015	7	2M4	1	1	2	0	0	0	0	0	0	11.5	12.418	0.918	7.98	11.320	−0.180	−1.56	10.009	−1.491	−12.97	10.009	−1.491	−12.97
016	8	2M2M3	1	3	0	0	0	0	0	0	0	31.6	30.427	−1.173	−3.71	30.547	−1.053	−3.33	31.976	0.376	1.19	31.976	0.376	1.19
017	8	2M2M3	4	0	0	0	0	0	0	0	0	28.0	24.366	−3.634	−12.98	28.123	0.123	0.44	27.819	−0.181	−0.65			
018	9	6	1	1	1	1	1	0	0	0	0	13.7	15.115	1.415	10.33	14.264	0.564	4.12	13.617	−0.083	−0.61	13.617	−0.083	−0.61
019	9	6	2	1	1	1	0	0	0	0	0	22.7	20.242	−2.458	−10.83	20.970	−1.730	−7.62	22.914	0.214	0.94			
020	9	6	2	2	1	0	0	0	0	0	0	31.7	26.758	−4.942	−15.59	27.745	−3.955	−12.48	29.570	−2.130	−6.72			
021	10	2M5	1	2	1	1	0	0	0	0	0	22.7	22.262	−0.438	−1.93	21.742	−0.958	−4.22	21.719	−0.981	−4.32	21.719	−0.981	−4.32
022	10	2M5	3	1	1	0	0	0	0	0	0	27.9	24.738	−3.162	−11.33	26.972	−0.928	−3.33	27.908	0.008	0.03			
023	10	2M5	2	3	0	0	0	0	0	0	0	41.9	35.763	−6.137	−14.65	37.358	−4.542	−10.84	39.212	−2.688	−6.42			
024	10	2M5	2	1	2	0	0	0	0	0	0	20.8	17.754	−3.046	−14.64	18.132	−2.668	−12.83	19.929	−0.871	−4.19			
025	10	2M5	1	1	1	2	0	0	0	0	0	14.3	15.745	1.445	10.11	14.967	0.667	4.66	14.300	0.000	0.00	14.3	0	0.00
026	11	3M5	1	1	2	1	0	0	0	0	0	11.4	13.257	1.857	16.29	12.129	0.729	6.40	10.736	−0.664	−5.82	10.736	−0.664	−5.82
027	11	3M5	2	2	1	0	0	0	0	0	0	29.4	26.758	−2.642	−8.99	27.745	−1.655	−5.63	29.570	0.170	0.58			
028	11	3M5	3	2	0	0	0	0	0	0	0	36.8	33.743	−3.057	−8.31	36.586	−0.214	−0.58	35.056	−1.744	−4.74			
029	11	3M5	1	2	2	0	0	0	0	0	0	18.7	19.774	1.074	5.74	18.904	0.204	1.09	18.156	−0.544	−2.91	18.156	−0.544	−2.91
030	12	2M2M4	1	3	1	0	0	0	0	0	0	28.7	28.778	0.078	0.27	28.518	−0.182	−0.63	29.139	0.439	1.53	29.139	0.439	1.53
031	12	2M2M4	4	1	0	0	0	0	0	0	0	30.3	31.723	1.423	4.69	30.241	−0.059	−0.20	30.171	−0.129	−0.43			
032	12	2M2M4	2	3	0	0	0	0	0	0	0	36.5	35.763	−0.737	−2.02	37.358	0.858	2.35	39.212	2.712	7.43			
033	12	2M2M4	1	1	3	0	0	0	0	0	0	8.5	10.770	2.270	26.70	9.291	0.791	9.31	7.172	−1.328	−15.62	7.172	−1.328	−15.62
034	13	2M3M4	1	2	2	0	0	0	0	0	0	19.2	19.774	0.574	2.99	18.904	−0.296	−1.54	18.156	−1.044	−5.44	18.156	−1.044	−5.44
035	13	2M3M4	3	2	0	0	0	0	0	0	0	34.0	33.743	−0.257	−0.76	36.586	2.586	7.60	35.056	1.056	3.11			
960	152	2M2M4M6M6M7	1	3	1	2	1	3	0	0	0	30.37	30.254	−0.116	−0.38	30.371	0.001	0.00	30.839	0.469	1.54	30.839	0.469	1.54
961	152	2M2M4M6M6M7	4	1	2	1	3	0	0	0	0	31.39	29.894	−1.496	−4.77	31.381	−0.009	−0.03	31.340	−0.050	−0.16			
962	152	2M2M4M6M6M7	2	5	1	3	0	0	0	0	0	54.32	51.345	−2.975	−5.48	52.924	−1.396	−2.57	54.622	0.302	0.56			
963	152	2M2M4M6M6M7	3	2	6	0	0	0	0	0	0	26.21	23.853	−2.357	−8.99	24.411	−1.799	−6.86	24.409	−1.801	−6.87			
964	152	2M2M4M6M6M7	1	2	2	6	0	0	0	0	0	25.29	24.811	−0.479	−1.90	23.758	−1.532	−6.06	22.518	−2.772	−10.96	22.518	−2.772	−10.96

NA is the sequence number of all the non-equivalent carbon atoms. NB is the sequence number of all the molecules.

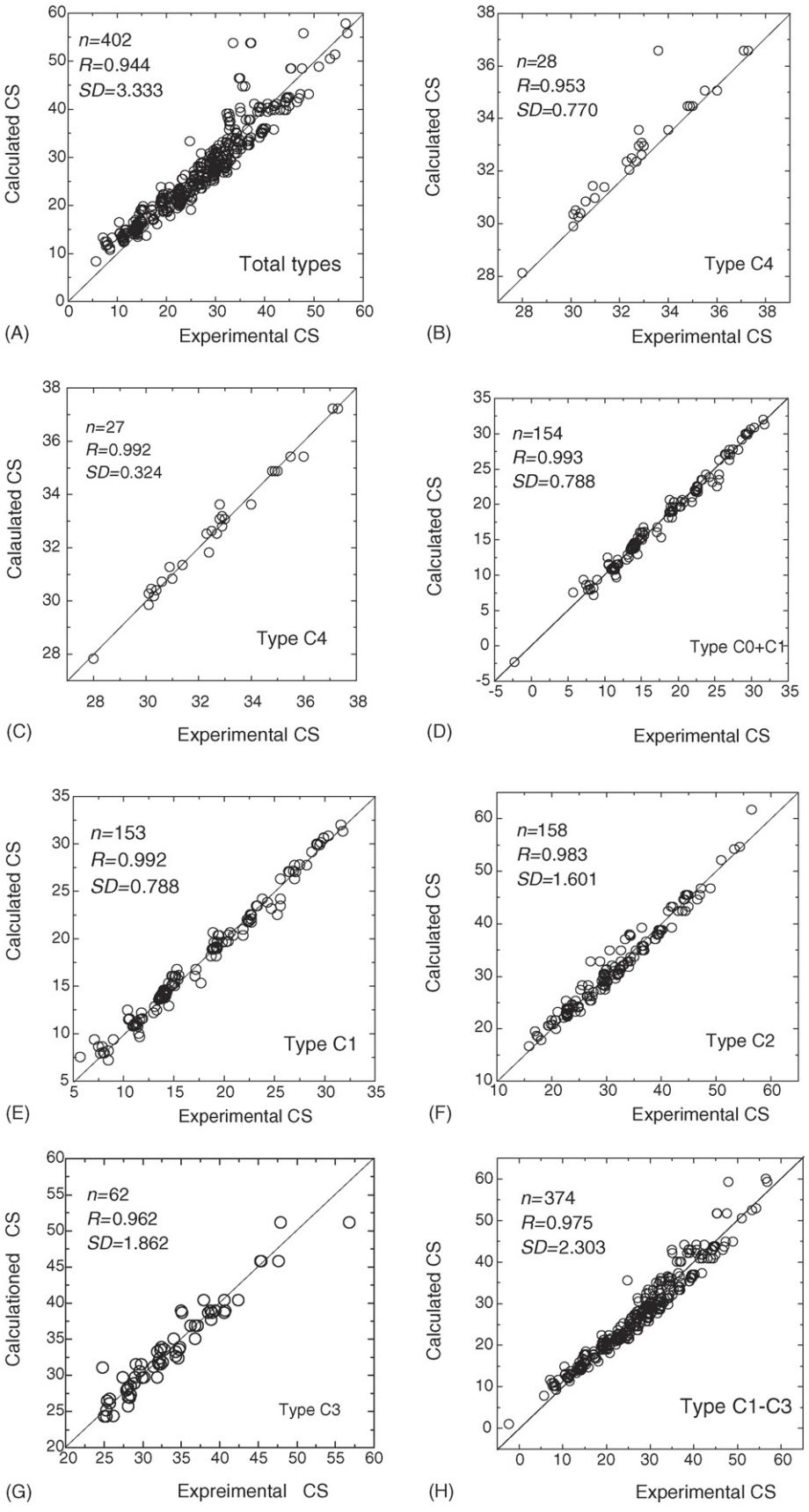

Fig. 2. Plot of the experimentally observed CS vs. the predicted CS for all types of carbon atom.

Table 2
Analyses of the orthogonality and the signification of the independent variables

Variables	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
Mean	1.91	2.022	1.590	1.015	0.537	0.261	0.085	0.015	0.007
VIF	1.378	1.292	1.189	1.237	1.397	1.519	1.405	2.124	1.988
Tolerance	0.726	0.774	0.841	0.808	0.716	0.659	0.712	0.471	0.503
t-score	75.048	74.073	75.044	76.964	79.137	80.740	81.521	81.756	81.815

Table 1 shows part of the estimated ^{13}C NMR chemical shifts of the known 402 atoms and the other unknown 562 ^{13}C NMR chemical shifts predicted from Eq. (2). The sequence is following the rule generally used in compound naming. Though MLR, we can see the obtained results with regression coefficient being $R = 0.944$, standard deviation S.D. = 3.333 and F-statistic value $F = 358.343$, which are quite good relative to the 402 carbon atom data set. The plots of the calculated values against the observed values are shown in Fig. 2A. From this figure, we can see a linear relationship between the estimated chemical shift values of ^{13}C NMR spectra from the rooted path vector and the experimental chemical shift values, which can be fitted reasonably well on a straight line though origin. Variance inflation factors (VIF) were calculated to measure the degree of linear independence of each descriptor with respect to the other descriptors in the correlation. They were calculated using the following formula: $\mathrm{VIF} = 1/(1 - R^2)$ where R^2 corresponds to a correlation developed by setting one set of descriptor values as a property and performing a multiple linear regression (MLR) using the other sets of descriptors. This provides a quantitative basis for evaluating how the values of each descriptor correlate to the values of the remaining descriptors. We considered a VIF < 5 to be sufficient to reject linear dependence within the correlation descriptor set. The values of VIF of all correlations were all less than 2 (See Table 2), indicating that these descriptors showed no inter-correlation. On the other hand, the *t*-test was done to check the significance of the variables, and from the results we can find the minimum of the *t*-scores is 74.073, showing that the variables have obvious significance to the chemical shift of ^{13}C in alkanes that we studied.

On the other hand, according to ^{13}C NMR principle, the γ-position carbon atom which is apart three C–C bonds from the centered carbon has, in general, a negative effect on the CS value of the examined central atom, which is so-called γ-effect and reflects the negative contribution of P_3 to CS in this model. In view of quaternary carbon atom without C–H bond have not γ-effect, various carbon atoms in all alkanes were classified as four types, type 1, 2, 3 and 4, for primary (CH_3-), secondary ($CH_2<$), ternary (–CH<), and quaternary (>C<) carbons which represent the carbon atoms connecting one, two, three, and four carbon atoms through the carbon–carbon bonds, respectively. Of these 964 chemically non-equivalent carbon atoms, these mainly are four types non-equivalent atom as well as methane carbon atom (0 °C) and numbers for these four types of atoms are 153, 158, 62 and 27, respectively. As to quaternary carbon atoms, $R = 0.953$ (see Table 3). And the CS_{exp} plotted versus the CS_{obs} are shown in Fig. 2B. It can be seen that a dot deviating the gained line and obviously is outlier (no. 382 atom, the only one quaternary carbon atom in 2,3,3,4-

Table 3
Results of ^{13}C NMR chemical shift using estimated and predicted by MLR and cross-validation for alkanes

Parameters	All (0–4 °C)	0–3 °C	(0 + 1) °C	1 °C	2 °C	3 °C	4 °C (28)	4 °C (27)
n	402	374	154	153	158	62	28	27
m	9	9	9	8	8	5	5	5
L	559	515	225	225	170	111	44	45
R	0.944	0.975	0.993	0.992	0.983	0.962	0.953	0.992
S.D.	3.333	2.303	0.788	0.788	1.601	1.862	0.77	0.324
F	358.343	773.372	1073.675	1129.078	523.595	138.976	43.475	265.418
U	35833.211	36912.109	6006.885	5615.006	10741.641	2409.396	28.891	138.891
Q	4355.422	1930.363	89.515	89.516	382.095	194.172	13.045	2.198
R^2_{CV}	0.886	0.948	0.919	0.982	0.956	0.897	0.828	0.936
R_{CV}	0.9413	0.9737	0.9586	0.991	0.9778	0.9471	0.9099	0.9674
b_0	3.022	0.983	−2.300	7.536	16.701	24.309	28.123	27.819
P_1	5.336	6.811	9.836					
P_2	7.356	7.584	8.146	8.146	7.503	5.373	2.117	2.351
P_3	−1.648	−2.029	−2.837	−2.837	−2.138	−1.775	0.600	0.549
P_4	0.839	0.809	0.727	0.727	0.847	0.322	−0.475	−0.440
P_5	0.210	0.106	0.044	0.044	0.018	−1.254	0.138	0.170
P_6	−0.138	0.043	0.067	0.067	−0.234	1.184	−0.104	−0.050
P_7	−0.506	−0.124	0.174	0.174	1.246			
P_8	2.486	1.715	−0.058	−0.058	−0.145			
P_9	−1.669	−1.101	0.300	0.300	−0.798			

L is the number of external prediction set.

Table 4
Statistics of CA model

ΔCS	All (0–4 °C)		0–3 °C		(0 + 1) °C		1 °C		2 °C		3 °C		4 °C (28)		4 °C (27)	
	N	*P* (%)	*N*	*P* (%)	*N*	*P* (%)	*N*	*P* (%)	*N*	*P* (%)	*N*	*P* (%)	*N*	*P* (%)	*N*	*P* (%)
≤1	92	22.89	144	38.50	132	85.71	131	85.62	79	50.00	31	50.00	27	96.43	27	100.0
≤2	199	49.50	240	64.17	149	96.75	148	96.73	135	85.44	51	82.26	27	96.43	27	100.0
≤3	283	70.40	303	81.02	154	100.0	153	100.0	150	94.94	57	91.94	28	100.0	27	100.0
≤4	343	85.32	358	95.72	154	100.0	153	100.0	155	98.10	60	96.77	28	100.0	27	100.0
≤5	377	93.78	367	98.13	154	100.0	153	100.0	156	98.73	60	96.77	28	100.0	27	100.0
≤6	387	96.27	367	98.13	154	100.0	153	100.0	158	100.0	61	98.39	28	100.0	27	100.0
≤8	393	97.76	372	99.47	154	100.0	153	100.0	158	100.0	62	100.00	28	100.0	27	100.0
≤10	396	98.51	372	99.47	154	100.0	153	100.0	158	100.0	62	100.00	28	100.0	27	100.0
>10	6	1.493	2	0.53	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00

tetramethylpentane). When delete the experiment value the result become much better (see Table 3 and Fig. 2C for detail). In addition, for 154 primary (including methane, Fig. 2D), 153 primary (not including methane, Fig. 2E),158 secondary (Fig. 2F), 62 ternary (Fig. 2G) and for 374 primary, secondly, tertiary carbon atoms (Fig. 2H), the correlation coefficients (R) are all exceed 0.97. Part of the estimated chemical shifts are listed in Table 1. Among it, the estimated chemical shifts of 0–3 and 4 °C were calculated by the model which was established through the corresponding experimental values of 0–3 and 4 °C, respectively. The two sets of values were listed in one column. Similarly, there are four sets of values in the column signed as 1234(27). Some elements in the rooted path vector are all zero in data set, so the number of non-zero elements for various types of carbon atoms are different. Among the primary carbon atoms, without and adding the one (0 °C) in methane molecule, the modeling equation and the statistic result are quite resemble. For both cases, the former intercept $b_0 = 7.536$ is equal to the latter intercept plus the first term $b_0 + b_1P_1 = -2.300 + 9.836 = 7.536$, which show the rationality that merging methane into the primary carbon set. The number of the estimated error within a given value range, e.g. $\Delta CS \leq 5$ ppm, and the corresponding percentage of the total atoms are listed for various types of atoms in Table 4. From Table 4, one can see that the atoms with absolute errors less than 5 ppm account for 93.78% in the case all 402 atoms were taken into consideration as the calibration set. It is in general illuminated that most of ^{13}C NMR chemical shift can be estimated and predicted by Eq. (2) with good results.

4.1. Cross-validation

In order to explain the stability and performance of the molecular modeling equations and their capacity for predicting chemical shifts of various atoms in external samples, a cross-validation technique called leave-one-out procedure, is employed to test the models above (see Table 3). In each time, only one sample drawn from working data set acts as a external prediction set and the remaining $n - 1$ samples construct a calibration set. The calibration set will be used to establish a new model between the observed CS value and the rooted path vector, and the new model is used to predict the chemical shift of the external set. In the same way, n times of modeling and prediction are executed for the working set containing n samples of each atom type. The average results of n times regression coefficients, correlation coefficients, and relative statistic parameters in the modeling stage are listed in Table 3. These models are used to predict CS of the external set sample. All of prediction results using the cross-validation technique with leave-one-out procedure are also listed in Table 3.

5. Conclusion

Chemical shift is one of the important characteristic parameters describing the local structure of any molecule that can reflect the simple but germane features of a compound. Rooted path vector is a good descriptor describing the microenvironment of a given carbon atom in an alkane and can easily be extended to the other molecules containing heteroatoms. Besides, the calculation of descriptors $\boldsymbol{P}$, is quite simple and easily done, only requiring to know the alkane primary structure and needing not to obtain any other property, parameter and perform require any complicated quantum chemical computation. It will be very useful in molecular structural characterization and chemical shift prediction of alkane molecules.

According to the regression coefficients, it was found that for primary, secondly, tertiary (including methane) carbon atoms both the first and the second variables have apparently positive contributions to chemical shifts of the 374 carbon atoms, but the third variables have obviously negative contributions either in the model of separation calculation or in the model of unification calculation. Therefore, we presumed from the results that the obviously negative contributions of the third variables to chemical shifts of the 374 carbon atoms possibly disclosed the γ-effect, which broadly existed in ^{13}C nuclear magnetic resonance spectra.

It demonstrated that the rooted path vector could generally be employed to predict the ^{13}C NMR chemical shift of alkane in a good agreement with the observed value except for few carbon atoms with relative great deviations. Especially the rooted path vector can be applied to work well in calculating the chemical shift if the carbon atom in methane, which show that the method have good applicability. In summary, this primary study has shown that the proposed new set of descriptors called the rooted path vector can differentiate the local characteristics and can be applied to describe the molecular structure of alkane. In order to explain the ability to characterize the

molecular structure of alkane, a further investigation is carried on modeling quantitative structure spectrum relationship for the other organic compounds.

Acknowledgements

All of the authors are especially grateful to the Fuk Yintung Educational Foundation (FYTF), the National New Drug Project of China, the National Chuihui Project Foundation (NCPF), Ministry of Mechanical Industries Fund (MMIF), the research initial funding of Chongqing University of Medical Sciences and Chongqing University Academic Fund (CQUF) for their partly financial support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2006.01.008.

References

[1] 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.
[2] L.P. Lindeman, J.Q. Adams, Carbon-13 nuclear magnetic resonance spectroscopy: chemical shifts for the paraffins though C9, Anal. Chem. 43 (1971) 1245–1252.
[3] G.W. Small, P.C. Jurs, Determination of topological similarity of carbon-atoms in the simulation of C-13 nuclear magnetic-resonance spectra, Anal. Chem. 56 (1984) 1314–1323.
[4] G.W. Small, T.R. Stouch, P.C. Jurs, Automated selection of models for the simulation of carbon-13 nuclear magnetic resonance spectra, Anal. Chem. 56 (1984) 2314–2319.
[5] H. Wiener, Structural determination of paraffin boiling points, J. Am. Chem. Soc. 69 (1947) 17–20.
[6] V. Kvasnicka, An application of neural networks in chemistry: prediction of C NMR chemical shifts, J. Math. Chem. 6 (1991) 63–76.
[7] 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.
[8] J.P. Doucet, A. Panaye, E. Feuilleaubois, P. Ladd, Neural networks and carbon-13 NMR shift prediction, J. Chem. Inf. Comput. Sci. 33 (1993) 320–324.
[9] 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.
[10] O. Ivanciuc, J.P. Rabine, D. Cabrol-Bass, A. Panaye, J.P. Doucet, ^{13}C NMR chemical shift prediction of the sp3 carbon atoms in the position relative to the double bond in acyclic alkenes, J. Chem. Inf. Comput. Sci. 37 (1997) 587–598.
[11] J. Meiler, R. Meusinger, M. Will, Neural network prediction of C-13 NMR chemical shifts of substituted benzenes, Monatsh. Chem. 130 (1999) 1089–1095.
[12] J. Meiler, R. Meusinger, M. Will, Fast determination of ^{13}C NMR chemical shifts using artificial neural networks, J. Chem. Inf. Comput. Sci. 40 (2000) 1169–1176.
[13] J. Meiler, W. Maier, M. Will, R. Meusinger, Using neural networks for ^{13}C NMR chemical shift prediction—comparison with traditional methods, J. Magn. Reson. 157 (2002) 242–252.
[14] J. Meiler, PROSHIFT: protein chemical shift prediction using artificial neural networks, J. Biomol. NMR 26 (2003) 25–37.
[15] D. Svozil, J. Pospichal, V. Kvasnicka, Neural network prediction of carbon-13 NMR chemical shifts of alkanes, J. Chem. Inf. Comput. Sci. 35 (1995) 924–928.
[16] 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.
[17] D. Nohair, D. Zakarya, Autocorrelation method adapted to generate new atomic environment: application for the prediction of 13-C chemical shift of alkanes, J. Chem. Inf. Comput. Sci. 42 (2002) 586–591.
[18] O. Ivanciuc, J.P. Rabine, D. Cabrol-Bass, C-13 NMR chemical shift sum prediction for alkanes using neural networks, Comput. Chem. 21 (1997) 437–443.
[19] S.S. Liu, H.L. 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. Chemometr. 15 (2001) 427–438.
[20] H. Hosoya, Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons, Bull. Chem. Soc. Jpn. 44 (1971) 2332–2339.
[21] M. Randic, On characterization of molecular branching, J. Am. Chem. Soc. 97 (1975) 6609–6615.
[22] L.B. Kier, L.H. Hall, W. Murray, Molecular connectivity. I. Relationship to local anesthesia, J. Pharm. Sci. 64 (1975) 1971–1974.
[23] A.T. Balaban, Topological indices based on topological distances in molecular graphs, Pure Appl. Chem. 55 (1983) 199–206.
[24] M. Randic, Molecular ID numbers: by design, J. Chem. Inf. Comput. Sci. 26 (1986) 134–136.
[25] H.P. Schultz, Topological organic chemistry. 1. Graph theory and topological indices of alkanes, J. Chem. Inf. Comput. Sci. 29 (1989) 227–228.
[26] Y. Miyashita, T. Okuyama, H. Ohsako, S. Sasaki, Graph theoretical approach to carbon-13 chemical shift sum in alkanes, J. Am. Chem. Soc. 111 (1989) 3469–3470.
[27] Y. Miyashita, Z. Li, S. Sasaki, Chemical pattern recognition and multivariate analysis for QSAR studies, Trends Anal. Chem. 12 (1993) 50–60.