Internet Electronic Journal of Molecular Design

February 2006, Volume 5, Number 2, Pages 102–115

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Danail Bonchev on the occasion of the 65th birthday

Prediction of Gas Chromatographic Retention Indices of Methylalkanes Produced by Insects

Fengping Liu,[1,2] Yizeng Liang,[1] and Chenzhong Cao [2]

[1] School of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, People’s Republic of China
[2] School of Chemistry and Chemical Engineering, Hunan University of Science and Technology, Xiangtan, 411201, People’s Republic of China

Received: September 19, 2005; Revised: December 12, 2005; Accepted: January 9, 2006; Published: February 28, 2006

Citation of the article:

F. Liu, Y. Liang, and C. Cao, Prediction of Gas Chromatographic Retention Indices of Methylalkanes Produced by Insects, *Internet Electron. J. Mol. Des.* **2006**, *5*, 102–115, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Prediction of Gas Chromatographic Retention Indices of Methylalkanes Produced by Insects [#]

Fengping Liu,[1,2] Yizeng Liang,[1] and Chenzhong Cao [2,*]

[1] School of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, People's Republic of China
[2] School of Chemistry and Chemical Engineering, Hunan University of Science and Technology, Xiangtan, 411201, People's Republic of China

Received: September 19, 2005; Revised: December 12, 2005; Accepted: January 9, 2006; Published: February 28, 2006

Internet Electron. J. Mol. Des. **2006**, *5* (2), 102–115

Abstract

Motivation. The methylalkanes studied in this work are produced by insects and are usually considered to be waterproofing agents present on the cuticle. A quantitative structure–retention relationships (QSRR) study has been carried out on a set of 177 methylalkanes by using molecular descriptors.

Method. A small number of molecular descriptors proposed by our team were used to establish a QSRR model. Multiple Linear Regression (MLR) analysis has been carried out to derive the best QSRR model. The model was supported by leave–one–out cross validation. Additional validation was performed on an external data set consisting of 30 methylalkanes.

Results. The best QSRR models for 177 methylalkanes are obtained with five structural descriptors, with $R^2 = 0.9999$ and SEC = 4.6. The QSRR model contains the molecular tightness index (MTI), the polarizability effect index (PEI), the number of carbon atom in the molecule backbone (N_C), the number of the 2–methyl group (N_{2-CH3}) and the number of the methyl group attached to the carbon backbone (N_{CH3}). Good results are obtained for the external data set with $R^2 = 0.9999$ and SEP = 3.7.

Conclusions. Compared with an earlier model for the prediction of these compounds, our model exhibits slightly improved performance, and the generated molecular descriptors have explicit physical meaning and easy to calculate. The model equations developed by present paper can be used to predict the chromatographic retention index of alkanes and support the identification of substances in cases the retention data for candidate structures are not available.

Keywords. Retention indices; methylalkanes; molecular descriptors; QSRR; quantitative structure–retention relationships; QSAR; quantitative structure–activity relationships; multiple linear regressions.

1 INTRODUCTION

The retention indices in gas chromatography have a long history since its introduction in 1958 by Kovats. Investigations and developments of mathematical models that are able to predict gas chromatographic retention data from chemical structures have found wide interest in studies on

[#] Dedicated on the occasion of the 65th birthday to Danail Bonchev.

[*] Correspondence author; phone: +867328291336; fax: +867328291001; E–mail: czcao@hnust.edu.cn.

quantitative structure–property relationships (QSPRs) [1]. QSPRs have been used to obtain simple model to explain and predict the chromatographic behavior of various classes of compounds.

Typical works in this field deal with 50–200 organic compounds, often belonging to a strictly defined class of substances. Aim is usually to create a model by using a small number of well interpretable molecular descriptors, although a great variety of much more than 1000 descriptors including structural, topological, geometrical, electrostatic and quantum–chemical index have been described and suggested for QSPR [2–3]. Recently published papers on relationships between molecular descriptors and the property of compounds, for instance, deal with sets of 149 alkanes [4], 130 methylalkanes [5], 400 alkenes [6], 150 alkyl benzenes [7–8], 200 polycyclic aromatic hydrocarbons [9], 60 polychlorinated naphthalenes [10], up to 100 esters, alcohols, aldehydes and ketones [11–14], 50 terpenes [15], 400 diverse organic compounds [16–17], 207 halogenated compounds [18], 13 acidic drugs [19], 28 organophosphonat esters [20], 846 toxicologically relevant compounds [21] and volatile organic compounds [22]. Typically, 20–300 molecular descriptors are tested and the final models contain less than 10 selected ones. Most used multivariate methods are multiple linear regression (MLR), partial least squares regression (PLS), principal component regression (PCR) and artificial neural networks (ANN). Despite the amount of literature available on the subject of QSPR of GC retention indices, many existing structural parameterization schemes need further improvement, and it is tedious and time–consuming to select structural descriptors from a pool composed of so many variables by many kinds of methods and programs.

The studied methylalkanes in this work produced by insects are usually considered to be waterproofing agents present on the cuticle. These components may also contain specific attractive chemical compounds used as lures. It is important to determine their chemical structures to make more effective lures. GC and GC–MS, the principal methods used to identify these alkanes, is problematic, because the interpretation of the spectra is difficult [5]. QSPR have been demonstrated to be a powerful tool to predict the retention indices (RI) of various compounds. In a previous work, Katritzky used AM1 parameterization within the semi–empirical quantum–chemical program MOPAC 6.0 and CODESSA program to calculate five types of molecular descriptors and established the QSPR model to predict the RI of the methylalkanes [5]. A number of 302 descriptors were calculated for each of 178 compounds studied. Finally, 4 descriptors were used to obtain a prediction model with a squared correlation coefficient of 0.9585 and a standard error of 5.8. This method needed a time–consuming selection and calculation of the descriptors. It is important to propose a simple and accurate model to identify these compounds. Many topological indices have been developed based on the molecular graph theory and have been proved useful in quantitative structure–property relationship (QSPR) studies. Due to the simplicity and efficiency of graph theoretical approaches, this paper also developed five topological indices to quantify the retention indices of the methylalkanes.

The primary aim of the present work is: (1) based on the molecular graph theory, to propose a small set of molecular descriptor to reflect the structure of the methylalkanes; (2) to establish QSPR model of retention indices for these compounds using the proposed molecular descriptors. The strategy applied in this study is in some aspects different from previous works on retention modeling. In this paper, a novel molecular descriptor, the molecular tightness index (MTI) was proposed at the first time in our team to develop the QSPR models, and the selection of subsets of descriptors was guided by chromatographic experience. Descriptors selected in our MLR models provide information related to the different molecular properties participating in the physicochemical process that occurs in the gas chromatography, and these descriptors reflect the length of the carbon backbone, the relative position of the methyl substituent, the number of the methyl groups attached to the carbon chain and the conformation of the compounds. The notable merit of the present method is that the structural parameters derived directly from the molecular structures are easy to calculate and apply. In the following sections, we describe the data set, the selection and calculation of molecular descriptors as well as the computational methods employed, and the results of our work.

2 MATERIALS AND METHODS

2.1 Data Set

In this work, a set of 177 alkanes including monomethylalkanes, dimethylmethylalkanes, trimethylalkanes and tetramethylalkanes were studied (based to Ref. [5], we could not obtain the corresponding molecular structure of the compound 8m22mC22, which was removed from the data set). The data sets of the Kovats retention indices were chosen from Ref. [5]. Additionally, we used an external data set of 30 compounds to test the prediction quality of the QSPR model as Katritzky did [5]. The retention indices of all compounds was determined by GC and GC–MS under a single set of condition, which are listed in Table 2 and Table 3 together with the molecular descriptors and the predicted values of the retention indices.

2.2 Molecular Descriptors

Intermolecular solute–solute and solute–stationary phase interactions depending on the conformation of the structure are known to play an important role in determining the GC retention. The physicochemical properties related to the retention behavior of the compound are multi–dimensional. According to the basic factors that influence the retention indices of the compound, such molecular descriptors: the molecular tightness index (MTI), the polarizability effect index (PEI), the number of carbon atom in the molecule backbone (N_C), the number of the 2–methyl group ($N_{2\text{–}CH3}$) and the number of the methyl group attached to the carbon backbone (N_{CH3}) have been chosen to build the QSPR model.

2.3 Calculation of Molecular Descriptors

2.3.1 Calculation of N_C, $N_{2–CH3}$ and N_{CH3}

The number of carbon atom in the molecule backbone (N_C), the number of the 2–methyl group ($N_{2–CH3}$) and the number of the methyl group attached to the carbon backbone (N_{CH3}) can be obtained directly from the molecule structure.

2.3.2 Calculation of polarizability effect index (PEI)

According to our previous work [23–24], polarizability effect index (PEI) was proposed on the basis of the principle of a molecule being polarized in an electric field. The stabilizing energy caused by polarizability effect for a substitutent *R* interacting with point charge *q* is:

$$E(R) = \frac{-q^2}{2Dl^4} \times \sum \frac{\alpha_i}{\left[N_i \frac{1+\cos\theta}{1-\cos\theta} - \frac{2\cos\theta\left(1-\cos^{N_i}\theta\right)}{(1-\cos\theta)^2} \right]^2} \qquad (1)$$

where α_i is the polarizability of the *i*–th essential unit in the substituent *R*, *D* is the effective dielectric constant, *l* is the length of C–C bond, N_i is the point charge (*q*) to the *i*–th essential unit, and θ is the supplementary angle of bond angle ∠CCC (that is θ= 180º –109.5º =70.5º for the sp^3 hybridization). For the alkyl substituent *R*, α_i is approximately equal to a constant and the Eq. (1) is:

$$E(R) = K \times \sum \frac{1}{\left[N_i \frac{1+\cos\theta}{1-\cos\theta} - \frac{2\cos\theta\left(1-\cos^{N_i}\theta\right)}{(1-\cos\theta)^2} \right]^2} = K \sum(\Delta PEI) \qquad (2)$$
$$= K(PEI)$$

Here, $K = -q^2\alpha_i / 2Dl^4$. PEI is called polarizability effect index. The PEI value of an alkyl substituent *R* is the term of $\Sigma\ (1/\ [\ ]^2)$ in Eq. (2). $\Delta PEI = 1/\ [\ ]^2$ is the PEI increments of the *i*–th essential unit. Some ΔPEI values are listed in Table 1.

Table 1. ΔPEI values of the *ith* Essential unit in Alkyl substituent

N_i	ΔPEI	N_i	ΔPEI	N_i	ΔPEI	N_i	ΔPEI
1	1.00000	6	0.009052	11	0.002375	16	0.001073
2	0.140526	7	0.006388	12	0.001972	17	0.000945
3	0.048132	8	0.004748	13	0.001628	18	0.000838
4	0.023503	9	0.003666	14	0.001421	19	0.000749
5	0.013800	10	0.002196	15	0.001229	20	0.000673

We consider the 2–methyl nonane for example to compute the PEI. Figure 1 is the hydrogen–depleted molecular graph of this molecule. Take the first carbon (according to the nomenclature rule) as the beginning atom to calculate the PEI as follows:

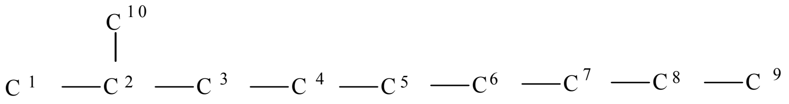

Figure 1. The hydrogen–depleted molecular graph of 2–methyl nonane.

$$\begin{aligned} PEI &= \sum PEI(R_i) \\ &= 1.0 + 0.1405 + 2 \times 0.04813 + 0.0235 + 0.0138 + 0.009052 + 0.004748 + 0.003666 \\ &= 1.2979 \end{aligned} \quad (3)$$

2.3.3 Calculation of the molecular tightness index (MTI)

Because the retention indices depend on the structure of a molecule, we defined the molecular tightness index (MTI) to reflect the branching and the shape of the molecule. Consider the 2–methyl nonane as the example to define and calculate the MTI. According to the hydrogen–depleted molecular graph of 2–methyl nonane (Figure 1), its distance matrix D is:

$$D=\begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 2 \\ 1 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 1 \\ 2 & 1 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 2 \\ 3 & 2 & 1 & 0 & 1 & 2 & 3 & 4 & 5 & 3 \\ 4 & 3 & 2 & 1 & 0 & 1 & 2 & 3 & 4 & 4 \\ 5 & 4 & 3 & 2 & 1 & 0 & 1 & 2 & 3 & 5 \\ 6 & 5 & 4 & 3 & 2 & 1 & 0 & 1 & 2 & 6 \\ 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 & 1 & 7 \\ 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 & 8 \\ 2 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 0 \end{bmatrix}$$

From the distance matrix D, we obtain the P_2 and P_3:

$$P_2 = \frac{1}{2} \sum N(d_{ij} = 2) \quad (4)$$

$$P_3 = \frac{1}{2} \sum N(d_{ij} = 3) \quad (5)$$

where $\sum N(d_{ij} = 2)$ is the number of the $d_{ij} = 2$ and $\sum N(d_{ij} = 3)$ is the number of the $d_{ij} = 3$ in the distance matrix D, d_{ij} is the length of the shortest path between vertex *i* and *j*. Then, the MTI index is defined as:

$$MTI = \frac{1}{2} \times \left[\left(\frac{P_2}{N-2} \right)^2 \right] + \left(\frac{P_3}{N-3} \right)^2 \quad (6)$$

where *N* is the number of vertex in molecular graph. For the 2–methyl nonane, P_2 = 9, P_3 =7, *N* = 10,

$$MTI = \frac{1}{2} \times \left[\left(\frac{9}{10-2} \right)^2 \right] + \left(\frac{7}{10-3} \right)^2 = 1.6328 \quad (7)$$

All the values of the molecular descriptors are listed in Table 2 and Table 3.

Table 2. Experimental and the calculated retention indices (RI) for 177 methylalkanes, with the values of the molecular descriptors

No	Compound	PEI	MTI	N_C	N_{CH3}	$N_{2–CH3}$	RI (Exp.)	RI (Cal.)	Δ
1	2mC9	1.2979	1.6328	9	1	1	966.5	951.8	14.7
2	3mC9	1.2733	1.9389	9	1	0	973.0	971.6	1.4
3	2mC11	1.3032	1.6050	11	1	1	1166.5	1158.7	7.8
4	3mC11	1.2786	1.8395	11	1	0	1172.5	1175.2	–2.7
5	2mC13	1.3068	1.5868	13	1	1	1366.5	1362.9	3.6
6	3mC13	1.2822	1.7769	13	1	0	1373.0	1377.4	–4.4
7	2mC15	1.3095	1.5740	15	1	1	1566.5	1565.6	0.9
8	3mC15	1.2848	1.7337	15	1	0	1573.7	1578.7	–5.0
9	2mC17	1.3115	1.5644	17	1	1	1765.8	1767.2	–1.4
10	3mC17	1.2869	1.7022	17	1	0	1774.0	1779.4	–5.4
11	2mC19	1.3131	1.5571	19	1	1	1966.0	1968.2	–2.2
12	3mC19	1.2884	1.6782	19	1	0	1974.3	1979.6	–5.3
13	10mC19	1.2673	1.6782	19	1	0	1943.0	1940.6	2.4
14	2mC21	1.3144	1.5512	21	1	1	2166.0	2168.6	–2.6
15	3mC21	1.2897	1.6593	21	1	0	2174.5	2179.4	–4.9
16	11mC21	1.2682	1.6593	21	1	0	2141.0	2139.7	1.3
17	2mC23	1.3154	1.5465	23	1	1	2364.0	2368.7	–4.7
18	3mC23	1.2908	1.6440	23	1	0	2374.5	2379.1	–4.6
19	12mC23	1.2689	1.6440	23	1	0	2337.0	2338.7	–1.7
20	2mC25	1.3163	1.5425	25	1	1	2563.0	2568.6	–5.6
21	3mC25	1.2917	1.6314	25	1	0	2574.4	2578.5	–4.1
22	13mC25	1.2696	1.6314	25	1	0	2534.5	2537.8	–3.3
23	2mC27	1.3171	1.5392	27	1	1	2763.0	2768.2	–5.2
24	3mC27	1.2924	1.6208	27	1	0	2774.4	2777.8	–3.4
25	14mC27	1.2702	1.6208	27	1	0	2733.0	2736.7	–3.7
26	2mC29	1.3177	1.5364	29	1	1	2962.2	2967.6	–5.4
27	3mC29	1.2931	1.6118	29	1	0	2974.0	2976.9	–2.9
28	15mC29	1.2706	1.6118	29	1	0	2931.5	2935.6	–4.1
29	2mC31	1.3183	1.5339	31	1	1	3161.5	3166.9	–5.4
30	3mC31	1.2936	1.6040	31	1	0	3174.1	3176.0	–1.9
31	4mC31	1.2839	1.6040	31	1	0	3157.5	3158.1	–0.6
32	5mC31	1.2792	1.6040	31	1	0	3150.0	3149.3	0.7
33	6mC31	1.2765	1.6040	31	1	0	3142.2	3144.4	–2.2
34	7mC31	1.2749	1.6040	31	1	0	3140.0	3141.4	–1.4
35	13mC31	1.2715	1.6040	31	1	0	3130.8	3135.3	–4.5
36	16mC31	1.2711	1.6040	31	1	0	3129.8	3134.4	–4.6
37	2mC33	1.3188	1.5317	33	1	1	3362.0	3366.1	–4.1
38	3mC33	1.2941	1.5973	33	1	0	3374.5	3375.0	–0.5
39	4mC33	1.2844	1.5973	33	1	0	3357.5	3357.1	0.4
40	5mC33	1.2797	1.5973	33	1	0	3350.0	3348.3	1.7
41	6mC33	1.2770	1.5973	33	1	0	3343.7	3343.4	0.3
42	13mC33	1.2721	1.5973	33	1	0	3328.5	3334.2	–5.7
43	17mC33	1.2715	1.5973	33	1	0	3328.5	3333.2	–4.7
44	2mC35	1.3192	1.5298	35	1	1	3562.0	3565.1	–3.1
45	3mC35	1.2946	1.5298	35	1	0	3574.3	3571.1	3.2
46	18mC35	1.2718	1.5914	35	1	0	3527.3	3531.9	–4.6
47	3m9mC23	1.2937	1.7808	23	2	0	2410.0	2410.7	–0.7
48	5m9mC24	1.2797	1.7683	24	2	0	2485.0	2483.6	1.4
49	3m11mC25	1.2936	1.7568	25	2	0	2609.0	2607.9	1.1
50	3m15mC25	1.2927	1.7568	25	2	0	2605.0	2606.2	–1.2
51	5m11mC25	1.2792	1.7568	25	2	0	2582.0	2581.2	0.8
52	5m17mC25	1.2781	1.7568	25	2	0	2585.0	2579.2	5.8
53	7m11mC25	1.2749	1.7568	25	2	0	2577.0	2573.3	3.7
54	2m6mC26	1.3231	1.6615	26	2	1	2704.0	2705.7	–1.7
55	4m8mC26	1.2860	1.7463	26	2	0	2695.0	2692.6	2.4

Table 2. (Continued)

No	Compound	PEI	MTI	N_C	N_{CH3}	$N_{2\text{–}CH3}$	RI (Exp.)	RI (Cal.)	Δ
56	5m11mC26	1.2796	1.7463	26	2	0	2682.0	2680.7	1.3
57	6m10mC26	1.2773	1.7463	26	2	0	2678.0	2676.5	1.5
58	7m11mC26	1.2753	1.7463	26	2	0	2675.0	2672.8	2.2
59	3m7mC27	1.2972	1.7366	27	2	0	2809.0	2811.8	–2.8
60	3m15mC27	1.2935	1.7366	27	2	0	2805.0	2805.1	–0.1
61	5m11mC27	1.2799	1.7366	27	2	0	2782.0	2780.1	1.9
62	5m17mC27	1.2788	1.7366	27	2	0	2786.0	2778.0	8.0
63	7m23mC27	1.2741	1.7366	27	2	0	2774.0	2769.4	4.6
64	9m19mC27	1.2725	1.7366	27	2	0	2765.0	2766.4	–1.4
65	2m6mC28	1.3238	1.6494	28	2	1	2905.0	2904.8	0.2
66	2m10mC28	1.3198	1.6494	28	2	1	2899.0	2897.4	1.6
67	4m10mC28	1.2854	1.7276	28	2	0	2895.0	2889.0	6.0
68	5m15mC28	1.2794	1.7276	28	2	0	2882.0	2877.8	4.2
69	7m13mC28	1.2754	1.7276	28	2	0	2873.0	2870.5	2.5
70	3m7mC29	1.2978	1.7193	29	2	0	3008.0	3010.6	–2.6
71	3m13mC29	1.2945	1.7193	29	2	0	3004.0	3004.5	–0.5
72	5m13mC29	1.2800	1.7193	29	2	0	2982.0	2977.9	4.1
73	5m19mC29	1.2793	1.7193	29	2	0	2983.0	2976.5	6.5
74	7m17mC29	1.2752	1.7193	29	2	0	2973.0	2968.9	4.1
75	2m6mC30	1.3244	1.6390	30	2	1	3105.0	3103.8	1.2
76	2m10mC30	1.3204	1.6390	30	2	1	3099.0	3096.4	2.6
77	2m12mC30	1.3196	1.6390	30	2	1	3095.0	3095.0	0
78	3m7mC30	1.2981	1.7116	30	2	0	3108.0	3110.0	–2
79	4m10mC30	1.2860	1.7116	30	2	0	3094.0	3087.8	6.2
80	6m10mC30	1.2786	1.7116	30	2	0	3075.0	3074.1	0.9
81	3m7mC31	1.2984	1.7044	31	2	0	3209.0	3209.4	–0.4
82	3m13mC31	1.2951	1.7044	31	2	0	3203.5	3203.2	0.3
83	3m15mC31	1.2947	1.7044	31	2	0	3209.0	3202.6	6.4
84	5m13mC31	1.2806	1.7044	31	2	0	3180.5	3176.6	3.9
85	5m17mC31	1.2800	1.7044	31	2	0	3182.0	3175.5	6.5
86	7m11mC31	1.2769	1.7044	31	2	0	3170.2	3169.7	0.5
87	11m21mC31	1.2727	1.7044	31	2	0	3162.9	3161.9	1.0
88	2m8mC32	1.3222	1.6300	32	2	1	3297.0	3297.7	–0.7
89	4m8mC32	1.2879	1.6976	32	2	0	3292.0	3288.8	3.2
90	6m10mC32	1.27917	1.6976	32	2	0	3273.5	3272.8	0.7
91	8m12mC32	1.2757	1.6976	32	2	0	3266.0	3266.4	–0.4
92	9m21mC32	1.2739	1.6976	32	2	0	3262.0	3263.0	–1.0
93	14m18mC32	1.2724	1.6976	32	2	0	3257.5	3260.3	–2.8
94	3m9mC33	1.2971	1.6913	33	2	0	3403.0	3404.7	–1.7
95	3m15mC33	1.2952	1.6913	33	2	0	3409.0	3401.3	7.7
96	5m17mC33	1.2805	1.6913	33	2	0	3380.0	3374.2	5.8
97	5m19mC33	1.2804	1.6913	33	2	0	3382.0	3373.9	8.1
98	7m17mC33	1.2762	1.6913	33	2	0	3370.0	3366.3	3.7
99	11m23mC33	1.2731	1.6913	33	2	0	3362.4	3360.5	1.9
100	2m10mC34	1.3214	1.6221	34	2	1	3494.0	3494.2	–0.2
101	4m16mC34	1.2856	1.6854	34	2	0	3489.0	3482.5	6.5
102	6m10mC34	1.2796	1.6854	34	2	0	3473.8	3471.5	2.3
103	8m12mC34	1.2762	1.6854	34	2	0	3465.0	3465.1	–0.1
104	12m22mC34	1.2730	1.6854	34	2	0	3461.4	3459.2	2.2
105	13m17mC34	1.2731	1.6854	34	2	0	3455.0	3459.5	–4.5
106	3m7mC35	1.2993	1.6799	35	2	0	3609.5	3606.7	2.8
107	3m15mC35	1.2956	1.6799	35	2	0	3601.0	3600.0	1.0
108	5m9mC35	1.2830	1.6799	35	2	0	3580.0	3576.7	3.3
109	5m19mC35	1.2808	1.6799	35	2	0	3580.5	3572.6	7.9
110	7m17mC35	1.2767	1.6799	35	2	0	3569.7	3564.9	4.8
111	9m21mC35	1.2745	1.6799	35	2	0	3561.0	3561.0	0

Table 2. (Continued)

No	Compound	PEI	MTI	N_C	N_{CH3}	$N_{2–CH3}$	RI (Exp.)	RI (Cal.)	Δ
112	2m12mC36	1.3210	1.615	36	2	1	3695.0	3691.6	3.4
113	5m17mC36	1.2812	1.6746	36	2	0	3680.0	3672.2	7.8
114	13m23mC36	1.2732	1.6746	36	2	0	3661.0	3657.4	3.6
115	3m15mC37	1.2834	1.6697	37	2	0	3779.0	3798.6	2.4
116	5m9mC 37	1.2960	1.6697	37	2	0	3801.0	3775.3	3.7
117	5m17mC37	1.2814	1.6697	37	2	0	3780.0	3771.5	8.5
118	13m23mC37	1.2733	1.6697	37	2	0	3759.0	3756.7	2.3
119	5m17mC38	1.2815	1.6650	38	2	0	3878.0	3870.8	7.2
120	4m8m12mC24	1.2868	1.8928	24	3	0	2520.0	2522.4	–2.4
121	5m9m13mC25	1.2816	1.8764	25	3	0	2610.0	2611.1	–1.1
122	4m8m12mC26	1.2877	1.8614	26	3	0	2719.0	2720.9	–1.9
123	3m7m11mC27	1.2995	1.8474	27	3	0	2838.0	2841.2	–3.2
124	3m8m12mC28	1.2981	1.8346	28	3	0	2918.0	2937.2	–19.2
125	3m7m11mC29	1.2998	1.8226	29	3	0	3037.0	3039.0	–2.0
126	5m13m17mC29	1.2809	1.8226	29	3	0	3007.0	3004.1	2.9
127	6m14m18mC30	1.2782	1.8114	30	3	0	3100.0	3097.9	2.1
128	3m7m11mC31	1.3004	1.8010	31	3	0	3236.5	3237.5	–1.0
129	5m13m17mC31	1.2814	1.80100	31	3	0	3205.4	3202.6	2.8
130	7m13m17mC31	1.2771	1.8010	31	3	0	3191.3	3194.7	–3.4
131	11m15m19mC31	1.2739	1.8010	31	3	0	3181.0	3188.6	–7.6
132	2m10m16mC32	1.3218	1.7239	32	3	1	3324.0	3321.4	2.6
133	4m12m16mC32	1.2868	1.7913	32	3	0	3316.0	3311.2	4.8
134	6m14m18mC32	1.2789	1.7913	32	3	0	3299.0	3296.4	2.6
135	12m16m20mC32	1.2736	1.7913	32	3	0	3281.0	3286.8	–5.8
136	3m7m15mC33	1.2999	1.7822	33	3	0	3436.5	3434.2	2.3
137	5m13m17mC33	1.2819	1.7822	33	3	0	3405.0	3401.0	4.0
138	7m11m15mC33	1.2784	1.7822	33	3	0	3389.0	3394.6	–5.6
139	11m15m19mC33	1.2744	1.7822	33	3	0	3379.0	3387.0	–8.0
140	2m10m16mC34	1.3223	1.7105	34	3	1	3524.0	3520.1	3.9
141	4m8m12mC34	1.2899	1.7736	34	3	0	3515.5	3514.6	0.9
142	6m14m18mC34	1.2792	1.7736	34	3	0	3497.0	3494.8	2.2
143	8m12m16mC34	1.2771	1.7736	34	3	0	3486.4	3490.9	–4.5
144	12m16m20mC34	1.2740	1.7736	34	3	0	3478	3485.2	–7.2
145	3m7m15mC35	1.3004	1.7656	35	3	0	3636.3	3632.7	3.6
146	5m9m13mC35	1.2845	1.7656	35	3	0	3605.0	3603.3	1.7
147	7m11m15mC35	1.2789	1.7656	35	3	0	3588.3	3592.9	–4.6
148	13m17m21mC35	1.2739	1.7656	35	3	0	3577.0	3583.8	–6.8
149	13m17m23mC35	1.2738	1.7656	35	3	0	3583.0	3583.6	–0.6
150	4m8m16mC36	1.2897	1.7580	36	3	0	3715.0	3711.8	3.2
151	8m12m16mC36	1.2775	1.7580	36	3	0	3685.0	3689.3	–4.3
152	14m18m22mC36	1.2738	1.7580	36	3	0	3676.0	3682.4	–6.4
153	3m7m15mC37	1.3008	1.7508	37	3	0	3835.0	3831.1	3.9
154	5m13m17mC37	1.2828	1.7508	37	3	0	3803.0	3797.9	5.1
155	7m13m19mC37	1.2783	1.7508	37	3	0	3784.0	3789.6	–5.6
156	15m19m23mC37	1.2737	1.7508	37	3	0	3775.0	3781.1	–6.1
157	16m20m24mC38	1.2736	1.7440	38	3	0	3873.5	3879.9	–6.4
158	5m13m17mC39	1.2831	1.7376	39	3	0	4001.0	3996.3	4.7
159	15m19m23mC39	1.2740	1.7376	39	3	0	3972.4	3979.5	–7.1
160	14m18m22mC40	1.2745	1.7315	40	3	0	4071.0	4079.2	–8.2
161	3m7m11m15mC29	1.3009	1.9218	29	4	0	3062.0	3065.6	–3.6
162	3m7m11m15mC31	1.3014	1.8942	31	4	0	3261.0	3263.8	–2.8
163	4m8m12m16mC31	1.2902	1.8942	31	4	0	3249.0	3243.0	6.0
164	3m7m11m15mC33	1.3014	1.8700	33	4	0	3459.0	3461.0	–2.0
165	4m8m12m16mC33	1.2907	1.8700	33	4	0	3448.0	3441.2	6.8
166	3m7m11m15mC35	1.3024	1.8485	35	4	0	3658.0	3660.2	–2.2
167	7m11m15m19mC35	1.2795	1.8485	35	4	0	3628.0	3618.1	9.9

Table 2. (Continued)

No	Compound	PEI	MTI	N_C	N_{CH3}	$N_{2–CH3}$	RI (Exp.)	RI (Cal.)	Δ
168	9m13m17m21mC35	1.2768	1.8485	35	4	0	3617.0	3613.0	4.0
169	11m15m19m24mC35	1.2752	1.8485	35	4	0	3605.0	3610.1	–5.1
170	6m10m12m16mC36	1.2826	1.8387	36	4	0	3723.0	3722.5	0.5
171	8m12m16m20mC36	1.2781	1.8387	36	4	0	3713.0	3714.2	–1.2
172	10m14m18m22mC36	1.2761	1.8387	36	4	0	3703.5	3710.5	–7.0
173	3m7m11m15mC37	1.3027	1.8294	37	4	0	3855.0	3858.4	–3.4
174	7m11m15m19mC37	1.2799	1.8294	37	4	0	3823.0	3816.3	6.7
175	9m13m17m21mC37	1.2772	1.8294	37	4	0	3813.0	3811.2	1.8
176	11m15m19m24mC37	1.2756	1.8294	37	4	0	3803.0	3808.3	–5.3
177	10m14m18m22mC38	1.2765	1.8206	38	4	0	3900.0	3908.7	–8.7

Table 3. Experimental and the calculated retention indices (RI) for external test set of 30 methylalkanes, with the values of the descriptors

No	Compound	PEI	MTI	N_C	N_{CH3}	$N_{2–CH3}$	RI (Exp)	RI (Cal.)	Δ
1	5mC27	1.2779	1.6208	27	1	0	2750.3	2750.1	0.2
2	7mC29	1.2743	1.6118	29	1	0	2939.8	2943.1	–3.3
3	7m11mC21	1.2729	1.8098	21	2	0	2172.0	2174.1	–2.1
4	3m11mC23	1.2927	1.7808	23	2	0	2405.0	2407.5	–2.5
5	3m7mC25	1.2964	1.7568	25	2	0	2608.5	2611.9	–3.4
6	5m9mC25	1.2801	1.7568	25	2	0	2586.0	2581.9	4.1
7	4m10mC26	1.2847	1.7463	26	2	0	2692.5	2689.4	3.1
8	6m13mC26	1.2764	1.7463	26	2	0	2681.0	2674.0	7.0
9	5m15mC27	1.2790	1.7366	27	2	0	2783.2	2777.9	5.3
10	7m11mC27	1.2756	1.7366	27	2	0	2767.2	2771.7	–4.5
11	9m11mC27	1.2738	1.7366	27	2	0	2765.0	2768.4	–3.4
12	4m8mC28	1.2867	1.7276	28	2	0	2895.0	2891.4	3.6
13	5m9mC29	1.2815	1.7193	29	2	0	2982.0	2981.1	0.9
14	7m19mC31	1.2756	1.7044	31	2	0	3166.0	3169.0	–3.0
15	9m19mC31	1.2737	1.7044	31	2	0	3165.0	3165.6	–0.6
16	2m10mC32	1.3209	1.6300	32	2	1	3291.0	3292.9	–1.9
17	2m12mC34	1.3206	1.6220	34	2	1	3494.0	3492.2	1.8
18	6m14mC34	1.2785	1.6854	34	2	0	3475.0	3473.2	1.8
19	3m7m13mC27	1.2986	1.8474	27	3	0	2840.0	2839.3	0.7
20	2m10m18mC28	1.3205	1.7568	28	3	1	2918.0	2917.8	0.2
21	9m13m17mC29	1.2747	1.8226	29	3	0	2995.0	2992.7	2.3
22	5m9m13mC31	1.2835	1.8010	31	3	0	3200.0	3206.8	–6.8
23	7m11m15mC31	1.2779	1.8010	31	3	0	3191.3	3196.5	–5.2
24	9m13m17mC31	1.2753	1.8010	31	3	0	3192.2	3191.7	0.5
25	5m9m23mC33	1.2831	1.7822	33	3	0	3409.0	3404.4	4.6
26	7m13m17mC33	1.2776	1.7822	33	3	0	3395.0	3394.2	0.8
27	9m13m17mC33	1.2758	1.7822	33	3	0	3391.9	3390.9	1.0
28	6m10m14mC34	1.2808	1.7736	34	3	0	3496.0	3499.4	–3.4
29	6m12m16mC34	1.2798	1.7736	34	3	0	3500.0	3497.6	2.4
30	10m14m18mC34	1.2752	1.7736	34	3	0	3489.0	3489.1	–0.1

2.4 Multiple Regression Analysis

Statistical evaluation of the data and multivariate data analysis has been performed mainly by the software products Origin and Bilin program packages [25]. Additional programs have been developed in Matlab 6.0 [26]. All work has been performed on personal computers running under operating system Microsoft Windows 2000. Correlation coefficient (R), adjusted (R^2_A), variance ratio (F) and standard error of estimate (SEE) were used to judge the statistical quality of the regression equations. The program also generated the predicted values of retention indices. The final equations had regression coefficients and variance ratio (*F*) significant to more than 95% level as revealed by the student *t*–statistic and *p*–values. Use of more than one variable in the multivariate equation was justified by autocorrelation study with the help of the program.

2.5 Validation of the QSRR Models

The predictive powers of the equations were validated by leave–one–out (LOO) cross–validation method, where one compound is deleted at once and prediction of the activity of the deleted compound is made based on the QSPR model. The process is repeated after elimination of another compound until all of the compounds have been deleted at once. For the validation of the models, predicted residual sum of square (*PRESS*), total sum of squares (*SSY*), cross–validated R^2 (R^2_{CV}), standard error of *PRESS* (S_{PRESS}) and predictive standard error or uncertainty factor (*PSE*) for the final equations were considered.

As a further test of the utility of the model, the retention indices of 30 methylalkanes not to be used for building the QSPR model were predicted. The compounds in the external test set were measured by using the same methodology as the training set. Then the appropriate descriptor values were inserted into the correlation equation, and the respective retention indices were calculated.

3 RESULTS AND DISCUSSION

The best five parameters correlation equation obtained for the whole set of 177 compounds is presented in detail in the following Eq. (8).

$$\begin{aligned} RI &= -2376.611(\pm 55.291) + 1844.268(\pm 40.247)PEI + 44.927(\pm 15.852)MTI \\ &+ 99.181(\pm 0.136)N_C + 20.124(\pm 1.559)N_{CH_3} - 51.398(\pm 2.462)N_{2-CH_3} \\ N &= 177 \ R = 0.9999 \ R^2 = 0.9999 \ R^2_A = 0.9999 \ F(5{,}171) = 627419 \ p<0.000 \\ SEE &= 4.6 \ PRESS = 3913.6 \ SSY = 3017.2 \ R^2_{CV} = 0.9999 \ S_{PRESS} = 5.0 \ PSE = 4.8 \end{aligned} \quad (8)$$

where *N* is the number of data points, *R* is correlation coefficient. R^2_A, *F*, *p*, *SEE*, *PRESS*, *SSY*, R^2_{CV}, S_{PRESS} and *PSE* are adjusted R^2, ratio between the variances of observed and calculated activities, probability factor related to *F*–ratio, standard error of estimate, predicted residual sum of squares, total sum of squares, cross validated R^2, standard error of PRESS and uncertainty factor respectively. The values within the parenthesis are confidence intervals of corresponding parameters.

The calculated retention indices are shown in Table 2 and plotted against the experimental values in Figure 2. The average error of the whole set of 177 compound is 3.7, which is lower than 4.6 in Katritzky’s paper. For the external set, a correlation coefficient of R^2 = 0.9999 and *SEP* = 3.7 was achieved. The calculated retention indices are shown in Table 3 and plotted against the experimental values in Figure 3.

In order to obtain insights into the molecular mechanism of interactions between eluent and stationary phase, the relative importance of structural features in molecules was analyzed. Selection of the five descriptors was based on the structure of the molecule and the properties related to the retention data; therefore all descriptors can be well interpreted in terms of chromatography. One of them, PEI connecting with the polarizability shows the prominent positive effect on the retention indices. The results indicate that polarizability is a significant factor in these molecules. The MTI that bases on the molecular graph theory and distance matrix and characterizes the size and the shape of the molecule has a positive effect on the retention indices, which is in line with the experimental experience. N_C and N_{CH3} reflecting the length of the molecule backbone and the branching of the methylalkanes also has positive affect on the retention data. The magnitude of these descriptors increases with (1) in the number of atoms in the molecule and (2) in branching. Within the group of the methylalkanes, the 2–methylalkanes possess the different retention indices with change of length of the carbon chain, consequently the descriptor $N_{2–CH3}$ behaves as an indicator descriptor and shows different influence of the 2–methylalkanes to the retention data.

Table 4. Correlation coefficient matrixes for independent variables

	PEI	MTI	N_C	N_{CH3}	$N_{2–CH3}$
PEI	1				
MTI	–0.3089	1			
N_C	–0.1433	0.0944	1		
N_{CH3}	–0.1007	0.8181	0.5162	1	
$N_{2–CH3}$	0.8122	–0.5028	–0.2431	–0.2602	1

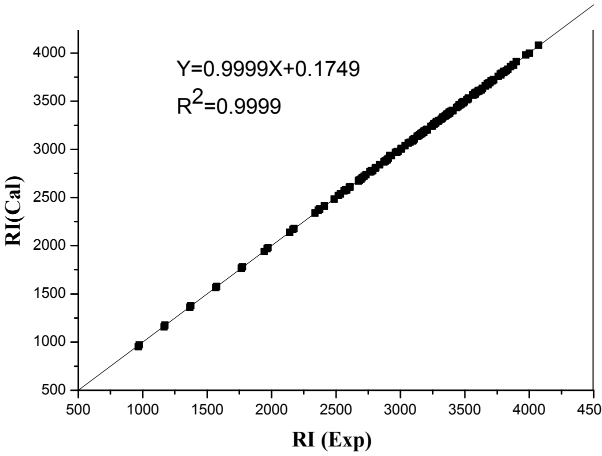

Figure 2. Plot of the calculated vs the experiment retention indices (RI) for the 177 methylalkanes.

To understand more clearly how the retention indices depend on the structure of the molecule, one can examine the property *vs*. descriptor relationship. Analyzing this relationship reveals some general trends. As already mentioned, the retention indices of the methyl–branched alkanes depend (1) on the polarizability of the molecule (2) on the length of the carbon backbone (3) on the branching and shape of the molecule (4) on the position of the methyl groups connected to the backbone.

It was to be expected that the GC retention indices of methylalkanes should modeled by molecular structural descriptors that reflect the relative position and the number of the methyl groups attached to the carbon backbone, the conformation of the compound, and the length of the carbon backbone. As our QSPR model shows, these molecular differences are best described by the selected descriptors.

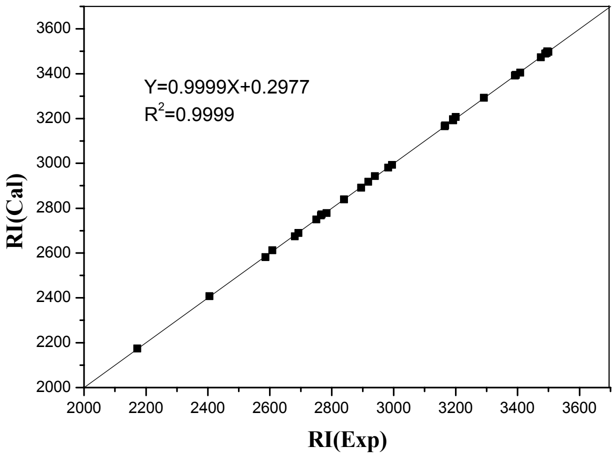

Figure 3. Plot of the calculated *vs*. the experiment retention indices (RI) for the external test set.

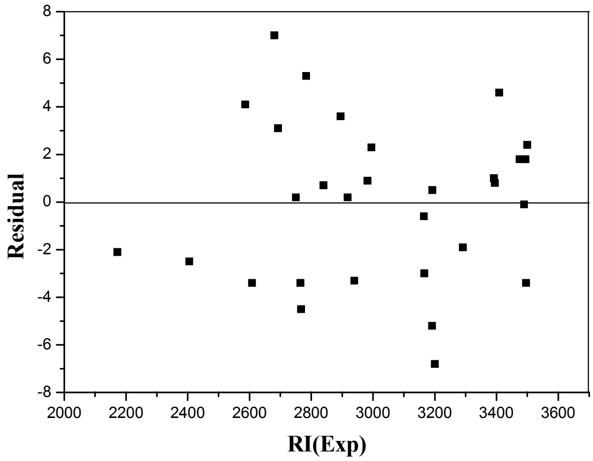

Figure 4. Plot of the residuals *vs*. the experiment retention indices (RI) for the test set.

It is well known that correlated descriptor variables can lead to unstable models. Therefore, we investigated the inter–correlations among our descriptor variables. The results in Table 4 show that no high linear correlation between them. In Figures 2 and 3 we present the retention indices as calculated by the MLR model compared with the experimental values from the database. The training set includes 177 methylalkanes and test set has 30 methylalkanes. The good correlation between the calculated and experimental values suggests that descriptors generated in the model are extremely sensitive to the retention indices. Figure 4 shows the plots of the residuals against the experimental values of the retention indices for the test set. The propagation of the residuals in both side of zero indicates that no systematic error exist in the development of the QSPR model.

4 CONCLUSIONS

A quantitative structure–property relationship model was derived to study the GC retention indices of methyl–branched alkanes for a diverse set of 177 compounds. A five descriptor equation was developed with a squared correlation coefficient of 0.9999 and a standard error of 4.6, which is close to the average experiment error of 4. Compared with the Katritzky's model for the prediction of these compounds, our model is simple and exhibits superior performance. The descriptors appeared in the model coding the chemical structure effectively and simply provide information related to the different molecular structure and molecular properties participating in the physicochemical process that occurs in the GC separated process. The correlation equation and descriptors can be used for the prediction of retention indices for similar compounds in cases where retention values were not readily available. The advantage of this approach over other methods lies in the fact that the descriptors used can be calculated from structure alone and are not dependent on any experiment properties. This paper provided a simple and straightforward way to predict the retention indices of the alkanes from their structures and gave some insight into structural features related to the retention of the compounds and the construction of structural descriptors.

Acknowledgment

The project is supported by a grant of the National Natural Science Foundation of China (No.20472019) and the Scientific Research Fund Provincial Education Department (No.04A015).

5 REFERENCES

[1] R. Kaliszan, *Quantitative Structure–Chromatographic Retention Relationships*, Wiley, New York, 1987.
[2] M. Karelson, *Molecular Descriptors in QSAR/QSPR*, Wiley, New York, 2000.
[3] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*,Wiley/Vch, Weinheim, 2000.
[4] Q. S. Xu, D. L. Massart, Y. Z. Liang, and K. T. Fang, Two–step Multivariate Adaptive Regresion Splines for Modeling A Quantitative Relationship between Gas Chromatography Retention Indices and Molecular Descriptors. *J. Chromatogr. A*. **2003**, *998*, 155–167.
[5] A. R. Katritzky and K. Chen, QSPR Correlation and Prediction of GC Retention Indexes for Methyl–branched Hydrocarbons Produced by Insects, *Anal. Chem*. **2000**, *72*, 101–109.
[6] Y. Du, Y. Liang, and D. Yun, Data Mining for Seeking Accurate Quantitative Relationship between Molecular Structureand Retention Indices of Alkenes by Projection Pursuit. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1283–1292.
[7] J. M. Sutter, T. A. Peterson, and P. C. Jurs, Prediction of Gas Chromatographic Retention Indices of Alkylbenzenes. *Anal. Chim. Acta*. **1997**, *342*. 113–122.
[8] A. Yan, G. Jiao, Z. Hu, and B. T. Fan, Use of Artificial Neural Networks to Predict the Gas Chromatographic Retention Index Data of Alkylbenzenes on Carbowax–20M. *Comp. Chem*. **2000**, *24*, 171–178.
[9] S. Liu, C. Yin, S. Cai, and Z. Li, Molecular Structural Vector Description and Retention Index of Polycyclic Aromatic Hydrocarbons. *Chemom. Intell. Lab. Syst*. **2002**, *61*, 3–15.
[10] J. Olivero and K. Kannan, Quantitative Structure–retention Relationships of Polychlorinated Naphthalenes in Gas Chromatography. *J. Chromatogr. A*. **1999**, *849*, 621–627.
[11] T. Körtvelyesi, M. Görgenyi, and K. Heberger, *Anal. Chim. Acta*. **2001**, *428*, 73–82.
[12] M. H. Fatemi, Simultaneous Modeling of the Kovats Retention Indices on OV–1and SE–54 Stationary Phases Using Artificial Neural Networks. *J. Chromatogr. A*. **2002**, *955*, 273–280.

[13] B. Ren, Atom–type–based AI Topological Descriptors for Quantitative Structure–retention Index Correlations of Aldehydes and Ketones. *Chemom. Intell. Lab. Syst.* **2003**, *66*, 29–39.

[14] B. S. Junkes, R. D. M. C. Amboni, R. A. Yunes, and V. E. F. Heinzen, Prediction of the Chromatographic Retention of Saturated Alcohols on Stationary Phases of Different Polarity Applying the Novel Semi–empirical Topological Index. *Anal. Chim. Acta.* **2003**, *477*, 29–39.

[15] M. Jalali–Heravi and M. H. Fatemi, Artificial Neural Network Modeling of Kovats Retention Indices for Noncylic and Monocylic Terpenes. *J. Chromatogr. A.* **2001**, *915*, 177–183.

[16] M. Pompe and M. Novic, Prediction of Gas–Chromatographic Retention Indices Using Topological Descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 59–67.

[17] M. Jalali–Heravi and Z. Garkani–Nejad, Use of Self–training Artificial Neural Networks in Modeling of Gas Chromatographic Ralative Retention Times of A variety of Organic Compounds. *J. Chromatogr. A.* **2002**, *945*, 185–194.

[18] T. Ivanciuc and O. Ivanciuc, Quantitative Structure–retention Relationship Study of Gas Chromatographic Retention Indices for Halogenated Compounds, *Internet Electron. J. Mol. Des.* **2002**, *1*, 94–107, http://www.biochempress.com/.

[19] T. Hanai, R. Miyazaki, E. Kamijima, H. Homma, and T. Kinoshita, Computational Prediction of Drug–Albumin Binding Affinity by Modeling Liquid Chromatographic Interactions, *Internet Electron. J. Mol. Des.* **2003**, *2*, 702–711, http://www.biochempress.com/.

[20] Y. S. Prabhakar, a Combinatorial Protocol in Multiple Linear Regression to Model Gas Chromatographic Response Factor, *Internet Electron. J. Mol. Des.* **2004**, *3*, 150–162, http://www.biochempress.com/.

[21] Z. Garkani–Nejad, M. Karlovits, W. Demuthb, T. Stimpfl, W. Vycudilik, M. Jalali–Heravi, and K. Varmuza, Prediction of Gas Chromatographic Retention Indices of a Diverse Set of Toxicologically Relevant Compounds, *J. Chromatogr. A.* **2004**,*1028*, 287–295.

[22] F. Luan, C. Xue, R. Zhang, C. Zhao, M. Liu, Z. Hu, and B. Fan, Prediction of Retention Time of a Variety of Volatile Organic Compounds Based on the Heuristic Method and Support Vector Machine, *Anal. Chim. Acta* **2005**, *537*, 101–110.

[23] C. Cao and Z. Li, Molecular Polarizability. 1. Relationship to Water Solubility of Alkanes and Alcohols, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1–7.

[24] C. Cao and H. Yuan, On Molecular Polarizability: 3. Relationship to the Ionization Potential of Haloalkanes,Amines, Alcohols, and Ethers, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1010–1014.

[25] H. Kabinyi, in: R. Mannhold, P. Krogs Gaad–Larsen, H. Timmerman (Eds.), *QSAR: Hansch Analysis and Related Approaches*, VCH, Weinheim, 1993.

[26] Mathworks Inc., Software Matlab, Natick MA, 2000.

Biographies

Fengping Liu is an associate professor of chemistry at the Hunan University of Science and Technology and a Ph.D. candidate in applied chemistry at the Central South University, People's Republic of China.