Available online at www.sciencedirect.com
SCIENCE DIRECT®

Talanta 68 (2005) 31–39

Talanta

www.elsevier.com/locate/talanta

QSPR prediction of GC retention indices for nitrogen-containing polycyclic aromatic compounds from heuristically computed molecular descriptors

Rong-Jing Hu[a], Huan-Xiang Liu[a], Rui-Sheng Zhang[a,b,*], Chun-Xia Xue[a], Xiao-Jun Yao[a], Man-Cang Liu[a], Zhi-De Hu[a], Bo-Tao Fan[c]

[a] *Department of Chemistry, Lanzhou University, Lanzhou 730000, China*
[b] *Department of Computer Science, Lanzhou University, Lanzhou 730000, China*
[c] *Université Paris 7-Denis Diderot, ITODYS 1, Rue Guy de la Brosse, 75005 Paris, France*

Received 19 January 2005; received in revised form 22 April 2005; accepted 23 April 2005
Available online 31 May 2005

Abstract

Gas chromatographic retention indices of nitrogen-containing polycyclic aromatic compounds (N-PACs) have been predicted by quantitative structure–property relationship (QSPR) analysis based on heuristic method (HM) implemented in CODESSA. In order to indicate the influence of different molecular descriptors on retention indices and well understand the important structural factors affecting the experimental values, three multivariable linear models derived from three groups of different molecular descriptors were built. Moreover, each molecular descriptor in these models was discussed to well understand the relationship between molecular structures and their retention indices. The proposed models gave the following results: the square of correlation coefficient, R^2, for the models with one, two and three molecular descriptors was 0.9571, 0.9776 and 0.9846, respectively.

© 2005 Published by Elsevier B.V.

Keywords: Quantitative structure–property relationships; Retention indices; Heuristic method; Nitrogen-containing polycyclic aromatic compounds

1. Introduction

Nitrogen-containing polycyclic aromatic compounds (N-PACs) are derivatives of polycyclic aromatic hydrocarbons (PAHs) containing two or more fused aromatic rings that are composed of C- and H-atoms. N-PACs usually include cyano (CN), amino (NH_2), imino (NH), nitro (NO_2) and replacement of a CH group in the benzene rings by a nitrogen atom. They are produced mainly by the incomplete combustion of coal, petroleum and industrial processes, e.g. carbon anode and graphite production as well as the use of coal tar. PAHs and N-PACs are carcinogenic mutagenic and toxic [1–8]. Although N-PACs exist usually with much smaller quantity than PAHs, previous studies proved that N-PACs were more toxic than their parent PAHs [9]. These compounds are ubiquitous in the environment as priority contaminants. Some of them such as nitro-compounds can form adducts with DNA and interact with proteins [10]. Because of their toxicity, bioaccumulation and persistence in the environment and their potential adverse effects on human and wildlife, N-PACs have received considerable attention in recent decades.

Gas chromatography (GC), as one of the first chromatographic separation techniques, has been used to environmental analysis for many years. It was first applied to analyse PAHs in the early 1960s and then progressed rapidly and widely. Nowadays, GC continues to play an important role in the identification and quantification of these ubiquitous pollutants, such as volatile organic compounds (VOCs), pesticides, halogenated compounds and polycyclic aromatic hydrocarbons, in the environment [11]. It will continue to be a promising method in the environmental analysis in the future.

Quantitative structure–property relationship (QSPR) provides a promising method for the estimation of retention in-

[*] Corresponding author. Tel.: +86 931 891 2578; fax: +86 931 891 2582.
E-mail address: ruison@public.lz.gs.cn (R.-S. Zhang).

0039-9140/$ – see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.talanta.2005.04.034

dices of organic compounds based on the descriptors derived solely from the molecular structure to fit experimental data [12,13]. QSPR study cannot only develop a method for the prediction of the property of interests but also can identify and describe important structural features of molecules that are responsible for variations in molecular properties. 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 method has become very useful in the prediction of physicochemical properties. The main steps in this method includes: data collection, molecular geometry optimization, molecular descriptor generation, descriptor selection, model development and finally model performance evaluation [14,15].

To develop a QSPR model, molecular structures are often represented using molecular descriptors, which encode much structural information. In recent years, there has been a shift from empirical parameters to purely calculated descriptors, such as topological indices and quantum chemical descriptors. The advantage of these calculated descriptors over other empirical descriptors is the possibility to calculate descriptors solely from molecular structure and apply them to sets of structurally diverse compounds. After the calculation of molecular descriptors, linear methods, such as multiple linear regression (MLR), principal component regression (PCR), partial least squares (PLS) and heuristic method (HM) implemented in the software CODESSA and or non-linear methods can be used in the development of a mathematical relationship between the structural descriptors and the property [15,16]. QSPR has been investigated to describe and predict physicochemical property of PAHs and their derivatives. Ribeiro [12] and Ferreira [13] established QSPR models of boiling point, octanol–water partition coefficient and retention time index of some PAHs.

Quantitative structure–retention relationship (QSRR) has been studied widely of these compounds [17,18]. However, only a few papers have mentioned about the QSRR studies of N-PACs [19,20]. The purpose of the present study was to investigate the relationship between gas chromatographic retention indices of 117 N-PACs and their molecular parameters. Moreover, molecular descriptors were discussed to explore the influence of structural features on the values of RI. This paper provided a simple and straightforward way to predict the retention indices of N-PACs from their structures and gave some insight into structural features related to the retention of the compounds. The prediction results are satisfactory in all the three groups.

2. Theory

2.1. Molecular descriptors

In QSPR studies, molecular descriptors of the chemical structures are important factors affecting the quality of the models. Various structural attributes of the molecule are used as the descriptors. They contain topological connectivity indices, properties depending on the charge distribution in the molecule and various thermodynamic functions at different temperatures and solvent characteristics. Their corresponding molecular descriptors include constitutional, topological, electrostatic and quantum-chemical, geometrical, thermodynamic descriptors, etc. Constitutional descriptors reflect only the molecular composition of the compound without using the geometry or electronic structure of the molecule, which related to the number of atoms, rings and bonds, for example, absolute and relative numbers of C, H, O, S, N, F, Cl, Br, I, P atoms; absolute and relative numbers of single, double, triple and aromatic bonds; molecular weight and average atomic weight; number of benzene rings, number of benzene rings divided by the number of atoms. Topological indices are two-dimensional (2D) descriptors based on graph theory concepts [21–23]. These indices are widely used in QSPR and QSAR studies. They help to differentiate the models according to their size, degree of branching, flexibility and overall shape. Electrostatic descriptors reflect characteristics of the charge distribution of the molecule such as total molecular surface area (TMSA) and partial positive surface area (PPSA). Quantum-chemical descriptors include information about binding and formation energies, partial atom charge, dipole moment and molecular orbital energy levels.

2.2. The heuristic method

The heuristic method (HM) (also called the heuristic multi-linear regression) performed in CODESSA was a procedure applied to pre-select the descriptors. It can perform a complete search for the best multi-linear correlations with a multitude of descriptors at a high speed in order to build the best multi-linear QSAR/QSPR model. The proceeding of the descriptor selection through heuristic method is shown as follows. First of all, all descriptors were pre-selected by eliminating: (i) those descriptors that are not available for each structure and (ii) descriptors having a small variation in magnitude for all structures. Descriptors for which values are not available for every structure in the data in question and which have a constant value for all structures in the data set are discarded. Thereafter, the one-parameter correlation equations for each descriptor are calculated. Then, to reduce further the number in the “starting set” of descriptors the following criteria are applied and a descriptor is eliminated if: (i) descriptors that give a F-test’s value below 1.0 in the one-parameter correlation and (ii) descriptors whose t-values are less than the user-specified value, etc. The left descriptor has a higher squared correlation coefficient in the one-parameter equations based on these descriptors. Next, starting with the top descriptor from the pre-selected list of descriptors the two-parameter correlations are calculated using the following pairs: the first descriptor with each of the remaining descriptors and the second descriptor with each of the remaining descriptors, etc. The best pairs as evidenced by the highest F-values in the two-parameter correlations are chosen and used for further inclusion of descriptors in a similar manner. A stepwise addition of further descriptor scales is performed to find the best multi-parameter regression models with the optimum values of statistical criteria (highest values of R^2, the R^2_{CV}, the F-value and the lowest s^2). R^2 is the correlation coefficient and s^2 is the squared standard error. R^2_{CV} is the cross-validated coefficient that describes the stability of a regression model obtained by focusing on the sensitivity of the model to the elimination of any single data point. The obtained regression is used to predict the value of this point, and the set-off estimated values calculated in this way is correlated with the experimental values.

The heuristic method usually produces correlations two to five times faster than other methods with comparable quality [24]. HM, as a good estimation method about what quality of correlation to expect from the data and a good tool to build models, has been applied to model and predict the retention indices of N-PACs in this paper. The result has proved the superiority of this method.

3. Experiment and methodology

3.1. Retention indices

The chromatographic data used were obtained from the paper by Vassilaros et al. [25] and consisted of gas chromatographic RI of 117 N-PACs. A complete list of the names and corresponding experimental RI values of compounds were given in Table 1. The equipment and procedures were described in Ref. [23]: the SE-52 coated fused-silica columns varied in length from 15 to 20 m and were either 0.3 mm i.d. or 0.2 mm i.d.; hydrogen was used as the carrier gas with a linear velocity of 100 cm/s; GC was performed using Hewlett-Packard 5880A gas chromatographs equipped with capillary systems at 40–265 °C at 4 °C/min with a 2-min initial isothermal period; the I values were generated from the raw retention data by use of a BASIC program written for the HP 5880GC and based on the equation of Van den Dool and Kratz [26].

3.2. Descriptor calculation and model developing

To obtain a QSRR model, compounds must be represented using molecular descriptors. Descriptors are generated solely from the molecular structures and aimed to numerically encode meaningful features of each molecule. The calculation process of the molecular descriptors is described as below: all the two-dimensional structures of the molecules were drawn using ISIS/Draw. Then the structures were pre-optimized using MM+ molecular mechanics force field and precisely optimized with semi-empirical AM1 method implemented in Hyperchem software package [27]. The final geometries were obtained with semi-empirical AM1 method in MOPAC program [28]. All calculations were carried out at restricted Hartree Fock level with no configuration interaction. The molecular structures were optimized using the Polak–Ribiere algorithm until the root mean square gradient was 0.01. The output files exported from MOPAC were transferred into software CODESSA, developed by Katritzky et al. [29,30], to calculate descriptors. CODESSA has been successfully used in various QSPR researches. In this program, a large number (>400) of molecular descriptors can be calculated on the basis of the geometrical and electronic structure of the molecules, which can be sorted into five classes: topological (Wiener index, Randic indices, Kier–Hall shape indices, etc.); constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.); geometrical (moments of inertia, molecular volume, molecular surface area, etc.); electrostatic (minimum and maximum partial charges, polarity parameter, charged partial surface area descriptors, etc.); quantum chemical (reactivity indices, dipole moment, HOMO and LUMO energies, etc.) [29,30]. Four descriptors have been totally calculated in present investigation: Randic index (order 3), Kier–Hall index (order 2), average valency of H-atom and number of benzene rings.

After the generation of descriptors, the heuristic method was used to select the sets of descriptors that are most relevant to the retention indices of these compounds. These descriptors can give some information on the affecting degree for retention of different descriptors and well understand the correlation between the experimental and calculated values. Then several sets of multi-linear models were automatically built by the same way in CODESSA.

4. Results and discussion

A total of 420 descriptors were calculated by the CODDESSA program for each of the compounds. After the heuristic reduction, the pool of descriptors was reduced to 209. To determine the optimum number of descriptors in a model, a variety of subset sizes were investigated. To select the sets of descriptors that are most relevant to retention indices and effectively show the relation between descriptors and retention indices of these compounds, three subsets with the descriptors from one to three were determined to establish the QSPR models. The predicted results for the three sets were listed in Table 1. The one to three parameter models are listed as follows:

One-parameter model

$$\begin{aligned}&\mathrm{RI} = 5.7952 \times 10^{1} + 4.8072 \times 10^{1}\mathrm{RI3}\\ &R^2 = 0.9571, \quad R^2_{\mathrm{CV}} = 0.9556, \quad F = 2565.88,\\ &\quad s^2 = 288.1697\end{aligned} \tag{1}$$

Two-parameter model

$$\begin{aligned}&\mathrm{RI} = 5.0772 \times 10^{3} + 8.1656 \times 10^{1}\mathrm{KHI2}\\ &\qquad -5.7892 \times 10^{3}\mathrm{AVH}\\ &R^2 = 0.9776, \quad R^2_{\mathrm{CV}} = 0.9761, \quad F = 2492.57,\\ &\quad s^2 = 151.5056\end{aligned} \tag{2}$$

Table 1
Experimental and calculated retention indices for N-PACs

No.	Compounds	Experimental	Predicted		
			One-para	Two-para	Three-para
1	1-Aminoindan	207.63	235.41	229.37	235.14
2	Quinoline	210.26	224.59	214.57	207.34
3	Isoquinoline	214.14	224.59	219.44	211.38
4	1-Methylindole	216.90	235.41	211.62	217.63
5	Indole	222.66	212.57	235.41	231.49
6	7-Azaindole	223.70	212.57	252.27	233.47
7	2-Methylquinoline	224.13	240.72	225.24	220.37
8	8-Methylquinoline	225.18	247.03	226.87	231.15
9	1-Methylisoquinoline	229.21	247.03	225.53	220.29
10	7-Methylquinoline	231.37	240.72	231.85	226.05
11	5-Aminoindole	232.12	228.70	239.29	243.22
12	3-Methylquinoline	232.47	240.72	230.98	234.92
13	7-Methylindole	235.49	235.01	240.18	240.00
14	4-Methylquinoline	235.77	247.03	228.14	222.60
15	3-Methylindole	239.20	235.41	235.58	236.45
16	2-Methylindole	240.10	225.09	238.88	239.24
17	2,7-Dimethylquinoline	244.04	256.85	251.60	246.10
18	2,6-Dimethylquinoline	244.19	256.85	251.42	245.95
19	1,2-Dimethylindole	244.42	260.60	227.36	234.24
20	2,2′-Bipyridyl	247.15	248.62	250.92	237.89
21	2,4-Dimethylquinoline	247.96	258.80	248.13	242.87
22	4-Azabiphenyl	252.35	248.62	255.66	241.99
23	2,4-Dimethylquinoline	256.65	241.22	260.59	261.38
24	1-Cyanonaphthalene	256.75	262.14	246.75	255.85
25	2,3-Dimethylindole	257.32	260.59	246.25	249.00
26	2-Cyanonaphthalene	260.88	261.09	249.43	258.30
27	5-Nitroindan	261.55	263.37	257.67	259.35
28	1-Aminonaphthalene	262.98	247.03	248.66	256.57
29	2-Aminonaphthalene	265.53	240.72	251.62	259.27
30	2,3,5-Trimethylindole	273.61	277.05	274.58	276.25
31	2-Aminobiphenyl	273.63	271.06	281.82	286.28
32	1-Nitronaphthalene	274.95	273.63	277.39	280.75
33	4-Azafluorene	279.85	302.70	317.72	298.02
34	2-Nitronaphthalene	280.63	275.39	277.44	281.14
35	3-Methyl-2-aminonaphthalene	283.73	268.84	267.64	276.43
36	2-Nitrobiphenyl	290.25	297.66	296.51	299.62
37	Phenazine	294.37	314.85	303.30	296.03
38	4-Aminobiphenyl	298.05	268.36	293.02	295.23
39	Benzo[*h*]quinoline	301.94	317.25	307.84	301.26
40	Acridine	304.04	314.85	305.38	299.48
41	Acridan (9,10-dihydroacridine)	304.11	314.85	321.70	324.14
42	Benzo[*f*]quinoline	307.94	317.25	306.98	300.54
43	Phenanthridine	307.94	317.25	305.72	309.10
44	3-Nitrobiphenyl	310.09	299.02	301.27	303.64
45	Carbazole	311.71	302.70	308.09	309.36
46	4-Nitrobiphenyl	314.59	301.96	301.14	303.51
47	3-Methylbenzo[*f*]quinoline	320.26	333.38	322.58	317.37
48	2-Methylbenzo[*f*]quinoline	320.50	333.71	328.36	331.96
49	2-Methylacridine	324.34	330.98	327.87	322.22
50	1-Methylcarbazole	324.45	325.74	326.33	328.27
51	4-Aminofluorene	325.11	322.52	334.42	335.03
52	1-Aminofluorene	327.21	325.74	332.02	333.09
53	3-Methylcarbazole	328.81	319.16	326.98	329.33
54	3-Aminofluorene	329.08	319.16	336.47	336.98
55	2-Methylcarbazole	329.61	318.83	327.69	329.87
56	9-Methylacridine	331.15	341.19	320.10	315.09
57	4-Methylcarbazole	331.88	322.52	323.63	326.19
58	2-Aminofluorene	331.91	318.83	337.89	338.08
59	6-Phenylquinoline	340.84	341.42	337.85	338.33
60	1,4-Dimethylcarbazole	343.16	346.24	344.56	347.18
61	2-Phenylindole	346.18	326.87	342.42	340.35
62	1,2-Dimethylcarbazole	347.31	354.60	342.55	345.42

Table 1 (*Continued*)

No.	Compounds	Experimental	Predicted		
			One-para	Two-para	Three-para
63	2-Azafluoranthene	347.39	381.60	372.98	369.28
64	1-Azafluoranthene	348.17	381.60	369.69	366.45
65	1,3-Dimethylcarbazole	348.45	337.84	348.53	350.82
66	9-Cyanoanthracene	350.46	351.80	340.53	350.18
67	7-Azafluoranthene	350.50	381.60	368.95	365.88
68	9-Cyanophenanthrene	351.84	352.54	339.43	349.30
69	2-Nitrofluorene	353.06	353.50	359.20	356.47
70	4-Aminophenanthrene	353.97	337.07	346.05	353.65
71	9-Nitroanthracene	357.42	360.8	368.78	373.25
72	1-Azapyrene	357.73	381.86	372.04	368.59
73	4-Azapyrene	357.94	381.86	370.21	348.28
74	2-Azapyrene	362.43	381.86	375.23	371.39
75	1-Aminophenanthrene	362.62	339.96	340.59	349.41
76	1-Aminoanthracene	362.83	337.62	343.96	352.42
77	9-Aminophenanthrene	362.83	336.68	340.20	349.15
78	9-Aminoanthracen	363.91	341.19	354.44	360.14
79	Benzo[*def*]carbazole	363.92	367.31	366.58	372.34
80	3-Aminophenanthrene	365.60	333.71	358.50	363.64
81	2-Aminophenanthrene	365.80	333.38	343.92	352.39
82	2-Aminoanthracene	367.45	330.98	347.03	355.20
83	3,5-Diphenylpyridine	372.84	365.05	370.54	368.02
84	9-Phenylcarbazole	381.51	417.66	411.44	414.52
85	Benz[*c*]acridine	392.60	407.84	399.65	394.18
86	Benz[*a*]acridine	398.65	407.84	398.90	393.54
87	1-Azabenz[*a*]anthracene	400.00	407.84	404.05	397.81
88	4-Azachrysene	401.16	410.19	400.96	395.00
89	Benzo[*a*]carbazole	402.22	395.96	401.38	403.28
90	1-Azachrysene	407.18	410.19	399.78	394.04
91	Benzo[*b*]carbazole	409.63	393.29	400.19	402.74
92	3-Aminofluoranthene	409.97	404.99	405.89	407.79
93	2-Azachrysene	411.49	410.19	403.52	397.21
94	Benzo[c]carbazole	411.89	393.62	401.69	403.52
95	4-Aminopyrene	412.31	401.56	405.41	417.31
96	2-Aminopyrene	413.83	394.38	424.21	432.22
97	1-Aminopyrene	415.39	405.25	419.39	428.11
98	1-Nitropyrene	421.48	431.85	431.31	439.26
99	2,2′-Biquinoline	422.56	434.22	428.67	409.66
100	7,9-Dimethylbenz[*c*]acridine	438.32	450.92	446.81	440.20
101	5,7-Dimethylbenz[*a*]acridine	438.38	454.22	454.08	447.60
102	7,10-Dimethylbenz[*a*]acridine	439.46	450.59	457.95	451.16
103	2-Aminobenzo[*c*]phenanthrene	450.10	424.64	440.25	448.65
104	4-Aminobenzo[*c*]phenanthrene	451.51	430.89	436.57	445.40
105	10-Azabenzo[*a*]pyrene	455.40	472.53	465.73	453.39
106	6-Aminochrysene	463.19	429.95	445.91	452.62
107	9,10,12-Trimethylbenz[*a*]acridine	466.79	476.09	473.98	465.78
108	Dibenz[*a,c*]phenazine	474.08	498.57	484.79	479.64
109	5-Aminochrysene	487.88	427.00	438.37	446.83
110	Dibenz[*a,h*]acridine	488.55	500.84	493.43	488.44
111	Dibenzo[*a,i*]carbazole	490.57	489.23	492.05	495.17
112	Dibenz[*a,j*]acridine	490.66	500.84	492.72	487.84
113	6-Nitrobenzo[*a*]pyrene	501.71	519.29	513.04	514.46
114	Dibenzo[*a,g*]carbazole	502.30	486.89	492.25	495.33
115	Dibenzo[*c,g*]carbazole	502.92	484.82	496.11	498.31
116	7-Aminobenzo[*a*]pyrene	511.98	495.57	498.44	501.51
117	6-Aminobenzo[*a*]pyrene	515.66	499.68	496.98	500.01

Three-parameter model

$$\begin{aligned} \mathrm{RI} &= 4.4234 \times 10^3 - 7.3777 \times 10^1\mathrm{KHI2} \\ &\quad -4.4648 \times 10^3\mathrm{AVH} + 9.4760\mathrm{NBR} \\ R^2 &= 0.9846, \quad R^2_{\mathrm{CV}} = 0.9834, \quad F = 2401.06, \\ s^2 &= 105.5948 \end{aligned} \qquad (3)$$

Among these equations, Eq. (1) contains only one independent variable RI3 that means Randic index (order 3). The two parameters included in Eq. (2) are KHI2 and AVH that means Kier–Hall index (order 2) and average valency of a H-atom, respectively. In Eq. (3), number of benzene rings (NBR) was added in the three-parameter model.

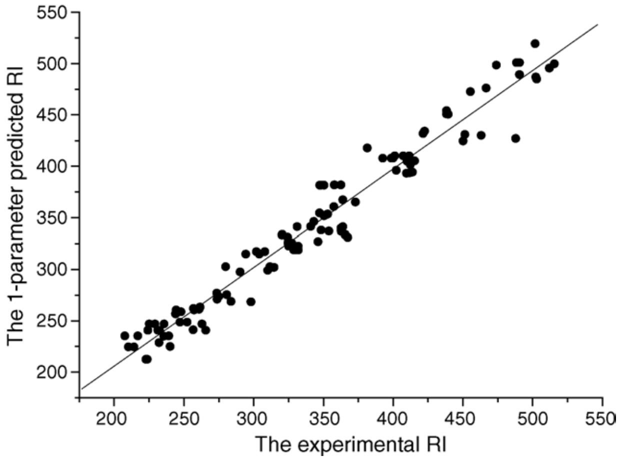

Fig. 1. The predicted RI vs. experimental values based on the one-parameter model by heuristic method.

These equations suggest that retention indices can be described as a sum of interactions of molecular topology, geometric and electronic properties, and quantum-chemical data. The values of the four descriptors were summarized in Table 2. The single-descriptor models were given in Table 3.

The one-parameter correlation equation obtained for the whole data set of 117 compounds is presented in detail in Table 1 and Fig. 1 with squared correlation coefficient $R^2 = 0.9571$. The cross-validated correlation coefficient $R^2_{CV} = 0.9556$, in comparison with correlation coefficient R^2, indicates the stability of the QSPR model. The descriptor in this model is Randic index (order 3) denoted as RI3, which also has the highest single parameter correlation $R^2 = 0.9571$. Randic index was first defined by Randic [21], whose primary purpose was to characterize the branch of hydrocarbon of methane series molecule quantitatively. The Randic index and its subsequent Kier index, which was improved by Kier and Hall [22,23] from the first were together called molecular connectivity index (χ). The general formula forχ is as follows in Eq. (4),

$$^{m}\chi = \sum(\delta_i\delta_j\delta_k\cdots)^{-1/2} \qquad (4)$$

where i, j and k correspond to the coordination numbers of atoms and m means the order of χ. In this study, m equals 3, so RI3 can be calculated by the following formula:

$$^{3}\chi = \sum(\delta_i\delta_j\delta_k\delta_l)^{-1/2} \qquad (5)$$

The descriptor was dependent on the connectivity of the atoms in a molecule. Being the derivatives of PAHs, the current set of compounds had similar structures that all the compounds had benzene rings, some of them had the same substituent –CH_3 and others had the similar substituents such as, –NH_2, –NO_2. Thus, all the structures of these compounds had the similar branch degree and molecular connectivity that the structure could be described by the same descriptor RI3.

Table 2
The value of the four descriptors of the compounds

No.	RI3	KHI2	AVH	NBR
1	3.6916	2.8440	0.9863	1
2	3.4663	2.1961	0.9797	0
3	3.4663	2.2228	0.9793	0
4	3.6916	2.4855	0.9843	1
5	3.2163	2.0649	0.9743	1
6	3.2163	1.9189	0.9693	0
7	3.8019	2.6453	0.9842	0
8	3.9332	2.6500	0.9840	1
9	3.9332	2.6175	0.9838	0
10	3.8019	2.6996	0.9839	0
11	3.5519	2.8817	0.9851	1
12	3.8019	2.7054	0.9841	1
13	3.6832	2.5168	0.9799	1
14	3.9332	2.6450	0.9837	0
15	3.6916	2.5173	0.9807	1
16	3.4767	2.5396	0.9804	1
17	4.1374	3.1488	0.9868	0
18	4.1374	3.1488	0.9868	0
19	4.2153	2.9036	0.9875	1
20	3.9663	2.4300	0.9768	0
21	4.1780	3.0977	0.9867	0
22	3.9663	2.4718	0.9765	0
23	3.8123	3.0431	0.9837	1
24	4.2476	2.6385	0.9804	2
25	4.2153	2.9206	0.9845	1
26	4.2256	2.6737	0.9805	2
27	4.2731	3.0637	0.9845	1
28	3.9332	2.5684	0.9791	2
29	3.8019	2.6066	0.9791	2
30	4.5577	3.4240	0.9867	1
31	4.4332	2.9533	0.9788	2
32	4.4865	2.7554	0.9768	2
33	5.0912	3.2321	0.9765	0
34	4.5231	2.7886	0.9772	2
35	4.3870	3.0559	0.9827	2
36	4.9865	3.1403	0.9789	2
37	5.3440	3.2010	0.9786	1
38	4.3770	2.9880	0.9774	2
39	5.3939	3.3618	0.9801	1
40	5.3440	3.3738	0.9807	1
41	5.3440	3.6161	0.9813	2
42	5.3939	3.3568	0.9802	1
43	5.3939	3.3626	0.9805	2
44	5.0146	3.1735	0.9786	2
45	5.0912	3.2162	0.9780	2
46	5.0758	3.1700	0.9785	2
47	5.7295	3.8060	0.9838	1
48	5.7364	3.8661	0.9836	2
49	5.6795	3.8772	0.9839	1
50	5.5705	3.6681	0.9812	2
51	5.5035	3.7153	0.9805	2
52	5.5705	3.7073	0.9808	2
53	5.4336	3.7196	0.9818	2
54	5.4336	3.7505	0.9806	2
55	5.4267	3.7196	0.9817	2
56	5.8919	3.7711	0.9837	1
57	5.5035	3.6681	0.9817	2
58	5.4267	3.7505	0.9804	2
59	5.8967	3.7769	0.9808	2
60	5.9969	4.1199	0.9844	2
61	5.5940	3.6363	0.9780	2
62	6.1709	4.1006	0.9845	2
63	6.7326	4.1363	0.9798	2

Table 2 (*Continued*)

No.	RI3	KHI2	AVH	NBR
64	6.7326	4.1088	0.9799	2
65	5.8223	4.1750	0.9845	2
66	6.1125	3.8056	0.9807	3
67	6.7326	4.1088	0.9801	2
68	6.1280	3.8027	0.9808	3
69	6.1479	3.9325	0.9793	2
70	5.8063	3.7321	0.9787	3
71	6.2998	3.9246	0.9775	3
72	6.7379	4.1386	0.9799	2
73	6.7379	4.1444	0.9803	0
74	6.7379	4.1711	0.9799	2
75	5.8664	3.7291	0.9796	3
76	5.8177	3.7677	0.9796	3
77	5.7981	3.7325	0.9797	3
78	5.8919	3.7325	0.9773	3
79	6.4352	3.9979	0.9789	3
80	5.7364	3.7673	0.9770	3
81	5.7295	3.7673	0.9796	3
82	5.6795	3.8059	0.9796	3
83	6.3883	4.1943	0.9810	2
84	7.4826	4.7053	0.9811	3
85	7.2785	4.5395	0.9808	2
86	7.2785	4.5345	0.9809	2
87	7.2785	4.5612	0.9804	2
88	7.3272	4.5225	0.9804	2
89	7.0313	4.3798	0.9783	3
90	7.3272	4.5175	0.9805	2
91	6.9757	4.4155	0.9790	3
92	7.2191	4.4761	0.9789	3
93	7.3272	4.5442	0.9802	2
94	6.9826	4.3798	0.9782	3
95	7.1477	4.5143	0.9795	4
96	6.9983	4.5525	0.9768	4
97	7.2244	4.5108	0.9770	4
98	7.7777	4.6978	0.9776	4
99	7.8272	4.7784	0.9792	1
100	8.1744	5.4403	0.9854	2
101	8.2431	5.6071	0.9865	2
102	8.1676	5.6612	0.9866	2
103	7.6278	4.9309	0.9793	4
104	7.7578	4.8927	0.9794	4
105	8.6240	5.3077	0.9803	2
106	7.7382	4.8932	0.9778	4
107	8.6982	5.8712	0.9868	2
108	9.1658	5.5007	0.9797	3
109	7.6768	4.8962	0.9792	4
110	9.2130	5.7002	0.9810	3
111	8.9714	5.5435	0.9790	4
112	9.2130	5.6951	0.9811	3
113	9.5967	5.8318	0.9795	4
114	8.9227	5.5435	0.9790	4
115	8.8796	5.5435	0.9783	4
116	9.1033	5.6750	0.9798	4
117	9.1888	5.6397	0.9795	4

Table 3
The single-descriptor models and their R^2 and s^2

Descriptor	Model	R^2	s^2
RI2	RI = 5.7952 + 4.8072RI2	0.9571	288.1697
KHI2	RI = 3.1981 + 8.1098KHI2	0.9332	448.6354
NBR	RI = 2.2819 + 5.6661NBR	0.6398	2419.5054
AVH	RI = 5.1922 − 4.9473AVH	0.0325	6499.6538

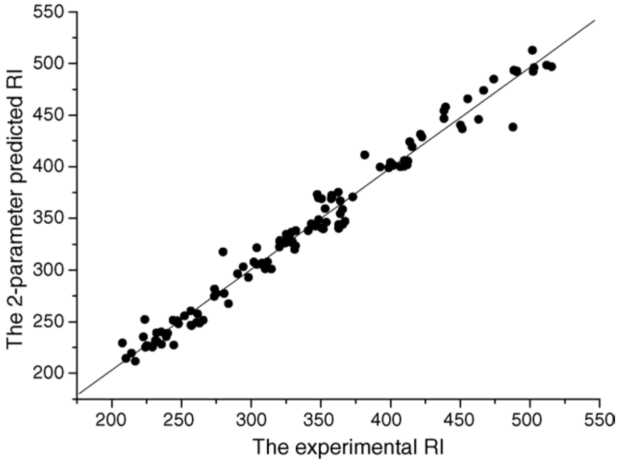

Fig. 2. The predicted RI vs. experimental values based on the two-parameter model by heuristic method.

For the two-parameter model, Kier–Hall index (this is KHI2), as a topological descriptor, was substituted for Randic index in the one-parameter model. KHI2 also gave the single parameter correlation $R^2 = 0.9332$ compared with R^2 of RI3, which showed a great correlation between RI3 and KHI2. The Kier–Hall index, originally defined by Randic [21] and subsequently refined by Kier and Hall [22,23], was a series of numbers designated by order and subgraph type. Compared with Randic index, this descriptor could differentiate unsaturated molecules that included hetero-atoms and multiple bonds. In the current case, KHI2 gave the information on N hetero-atom with a similar connectivity pattern in the molecule, as well as the information on similar connectivity of unsaturated bonds in benzene rings.

Another descriptor in Eq. (2) was AVH, which belonged to the quantum-chemical descriptors. Table 1 show a poor correlation $R^2 = 0.0325$ between AVH and retention indices. But the correlation did not influence the final two-parameter result $R^2 = 0.9776$ (Fig. 2). This descriptor may be regarded as a correlative descriptor with the descriptor KHI2. The different position of N-atom as well as the different conjugated systems in each molecule led to the diverse average values of H-atom. From this descriptor, we can see that N-atom is also an influencing factor on the retention indices.

Compared with the two-parameter model, NBR was added in the three-parameter model. From Table 3 and Fig. 3, $R^2 = 0.6398$ indicated NBR have a great impact on RI. Observing Fig. 4, we can see that with the increasing of numbers of benzene rings, the calculated retention indices increase correspondingly. The GC retention indices of compounds were determined by the intermolecular interaction between stationary phase and N-PACs mainly. The numbers of benzene rings influence the molecular weight and the molecular weight increases along with the increasing of benzene rings. Thus, an increase in this descriptor enhances the van der Waals interaction between N-PACs and stationary phase and leads to an increase in the value of RI.

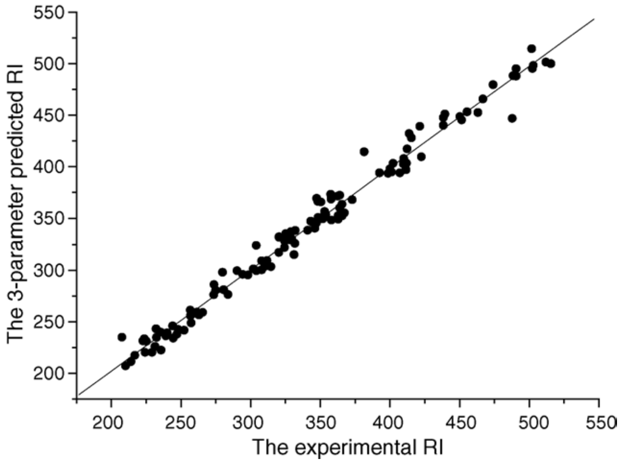

Fig. 3. The predicted RI vs. experimental values based on the three-parameter model by heuristic method.

From the above discussion, the three-parameter model, which shows the highest R^2, is obviously the best one. In order to evaluate its predictive ability, the whole data set was randomly divided into the test set and the training set and a leave-one-out cross-validation for the training set was performed. The test set contained 11 compounds 9, 19, 29, 39, 49, 59, 69, 79, 89, 99, 109 and the training set contained the others. The squared correlation coefficient (R^2) for the training set and the test set were 0.9863 and 0.9743, respectively, confirming the powerful predictive capability of the model. Fig. 5 shows the plot of the calculated versus experimental RI for the training set and the test set.

As can be seen from above discussion, the GC retention behavior of these compounds depended on the connection of the carbon backbone, the positions of N-atom and the conjugate bonds in benzene rings system. From the obtained results, we can see that the selected descriptors could account for these features and topological descriptor proved to be the most important factor influencing the GC retention index of N-PACs.

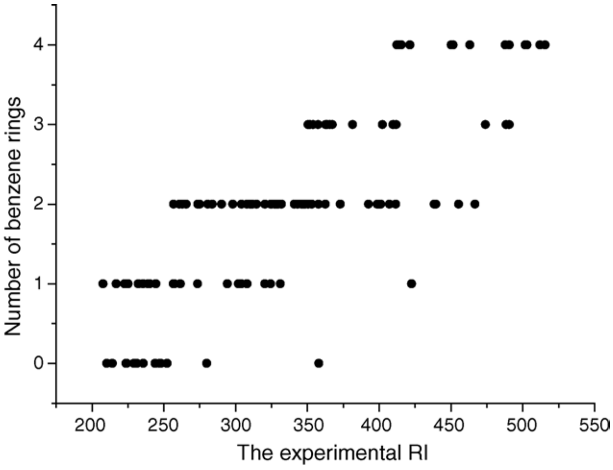

Fig. 4. The experimental RI vs. number of benzene rings.

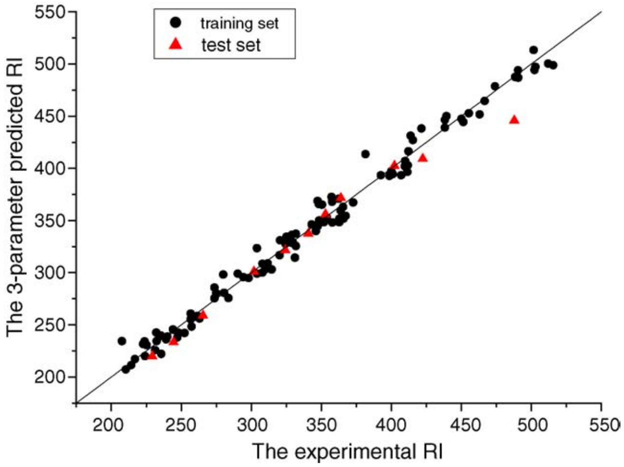

Fig. 5. The predicted RI of the training set and the test set vs. experimental values based on the three-parameter model by heuristic method.

5. Conclusion

A quantitative structure–property relationship model was derived to study the GC retention index of a diverse set of 117 N-PACs. Three QSPR models were developed with the squared correlation coefficient of 0.9571, 0.9776 and 0.9846. These models showed strong predictive ability. Among all the descriptors, topological descriptors were found to have high coding capabilities for the GC retention index and were selected to represent the chemical structures. The present work provides an effective method for the prediction of the GC retention indices for the N-PACs. This study also showed that the utility of the QSPR treatment involving descriptors derived solely from chemical structure and the correlation equation and descriptors can be used for the prediction of the retention index for unknown structures.

Acknowledgements

The authors thank the Association Franco-Chinoise pour la Recherche Scientifique & Technique (AFCRST) (Program PRA SI 02-03) for supporting this study.

References

[1] M. Kohler, T. Kunniger, P. Schmid, E. Gujer, R. Crockett, M. Wolfensberger, Environ. Sci. Technol. 34 (2004) 4766.
[2] C.L. Ritter, S.J. Culp, J.P. Freeman, M.M. Marques, F.A. Beland, D. Malejka-Giganti, Chem. Res. Toxicol. 15 (2002) 536.
[3] A.B. Ross, J.M. Jones, S. Chaiklangmuang, M. Pourkashanian, A. Williams, K. Kubica, J.T. Andersson, M. Kerst, P. Danihelka, K.D. Bartle, Fuel 81 (2002) 571.
[4] F.E. Speizer, Environ. Health Perspect. 70 (1986) 9.
[5] B. Fouillet, P. Chambon, M. Weill, Bull. Environ. Contam. Toxicol. 47 (1991) 1.
[6] J. Jocob, Pure Appl. Chem. 68 (1996) 301.
[7] A. Wernersson, G. Dave, Arch. Environ. Contam. Toxicol. 32 (1997) 268.

[8] P.S. Huovinen, M.R. Soimasuo, A.O. Oikari, Chemosphere 45 (2001) 683.

[9] M. Stefanova, S.P. Marinov, A.M. Mastral, M.S. Callen, T. Garcia, Fuel Process. Technol. 77–78 (2002) 89.

[10] C.L. Ritter, S.J. Culp, J.P. Freeman, M.M. Marques, F.A. Beland, D. Malejka-Giganti, Chem. Res. Toxicol. 15 (2002) 536.

[11] F.J. Santos, M.T. Galceran, TrAC, Trends Anal. Chem. 21 (2002) 672.

[12] F.A.L. Ribeiro, M.M.C. Ferreira, J. Mol. Struct. 663 (2003) 109.

[13] M.C.M. Ferreira, Chemosphere 44 (2001) 125.

[14] X.J. Yao, Y.W. Wang, X.Y. Zhang, R.S. Zhang, M.C. Liu, Z.D. Hu, B.T. Fan, Chemom. Intell. Lab. Syst. 62 (2002) 217.

[15] H.X. Liu, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, Anal. Chim. Acta 525 (2004) 31.

[16] H.X. Liu, C.X. Xue, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, J. Chem. Inf. Comput. Sci. 44 (2004) 1979.

[17] E.B. Ledesma, M.J. Wormat, Anal. Chem. 72 (2000) 5437.

[18] J. Kang, C. Cao, Z. Li, J. Chromatogr. A 799 (1998) 361.

[19] S. Liu, C. Yin, S. Cai, Z. Li, Chemom. Intell. Lab. Syst. 61 (2002) 3.

[20] B. Skrbic, N. Djurisic-Mladenovic, J. Cvejanov, Chemom. Intell. Lab. Syst. 72 (2004) 167.

[21] M. Randic, J. Am. Chem. Soc. 97 (1975) 6609.

[22] L.B. Kier, L.H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, 1976.

[23] L.B. Kier, L.H. Hall, Molecular Connectivity in Structure Activity Analysis, Research Studies Press, Letchworth, England, 1986.

[24] A.R. Katritzky, R. Petrukhin, R. Jain, M. Karelson, J. Chem. Inf. Comput. Sci. 41 (2001) 1521.

[25] D.L. Vassilaros, R.C. Kong, D.W. Later, M.L. Lee, J. Chromatogr. 252 (1982) 1.

[26] H. Van den Dool, P. Dec. Kratz, J. Chromatogr. 11 (1963) 463.

[27] HyperChem. 4.0, Hypercube (1994).

[28] J.P.P. Stewart, MOPAC 6.0, Quantum Chemistry Program Exchange, QCPE, No. 455, Indiana University, Bloomington, IN, 1989.

[29] A.R. Katritzky, V. S. Lobanov, M. Karelson, Comprehensive Descriptors for Structural and Statistical Analysis, Reference Manual, Version 2.0, 1994.

[30] A.R. Katritzky, V.S. Lobanov, M. Karelson, Chem. Soc. Rev. 24 (1995) 279.