ARTICLE

Modeling Gas Chromatographic Retention Indices of Oxygen-containing Compounds by Novel Atom-type Topological Indices

Xiao-fang Hu, Chun-hui Lu, Chun-sheng Yin*

School of Environmental Science and Engineering, Shanghai Jiaotong University, Shanghai 200240, China

(Dated: Received on September 2, 2005; Accepted on December 2, 2005)

Quantitative structure-retention relationship (QSRR) model for the estimation of retention indices (RIs) of 39 oxygen-containing compounds containing ketones and esters was established by our newly introduced distance-based atom-type indices DAI. The useful application of the novel DAI indices has been demonstrated by developing accurate predictive equations for gas chromatographic retention indices. The statistical results of the multiple linear regression for the final model are r=0.9973 and s=8.23. Furthermore, an external test set of 10 oxo-containing compounds can be accurately predicted with the final equation giving the following statistical results: r_{pred}=0.9966 and s_{pred}=8.56.

Key words: Topological indices, Retention indices, Oxygen-containing compounds

I. INTRODUCTION

For the purpose of prediction, quantitative structure-retention relationships (QSRR) between chromatographic retention indices (RIs) and different molecular descriptors have been studied for different compounds [1-6]. Using the QSRR method, structure parameters such as topological, geometric, electronic, and physical descriptors can be generated for molecules with similar structural features, and a subset can be selected that best describes the gas chromatographic retention data [7]. These descriptors derived from the molecular structure are believed to encode all the interactions including directional force, induction force, dispersion force, hydrogen bond, and so on. If the mobile phase and the stationary phase are the same for every solute, then only the differences in the structures of the solute molecules need to be encoded. Thus, all the numerical descriptors could be based on chemical structures of the solute molecules. In this study, we defined a novel atom-type index based on the shortest distance matrix of a molecular topological graph. The novel distance-based atom-type indices DAI can be easily calculated and shows a good correlation with the RIs of oxygen-containing compounds under study. In addition, the final model is validated to be predictive using an external test set.

II. METHODS

A. The definition of the DAI index

For any atom i that belongs to the kth atom-type (the definition is the same as that of Kier *et al.* [8]) in a graph, the novel distance-based atom-type topological index $DAI_i(k)$ is expressed as follows:

$$DAI_i(k) = 1 + \phi_i(k) \tag{1}$$

$$\phi_i(k) = n \cdot \frac{\sum_{j}^{n} D_{ij}}{\sum_{i}^{n}\sum_{j}^{n} D_{ij}} \tag{2}$$

where the parameter ϕ is considered as a perturbing term of the ith atom reflecting the effects of its structural environment, n is the number of total vertices in molecular topological graph, D_{ij} is the shortest distance between vertices i and j, and is calculated by summing the relative bond length [9] (take C−C bond length 0.154 nm as 1) between two adjacent vertices in the shortest path.

According to this definition, for kth atom-type in a molecular graph, the corresponding distance-based atom-type topological index, $DAI(k)$, is the sum of all $DAI_i(k)$ values of the same atom type in a molecular graph:

$$\begin{aligned} DAI(k) &= \sum_{i=1}^{m} DAI_i(k) \\ &= m + \sum_{i=1}^{m} \phi_i(k) \end{aligned} \tag{3}$$

where m is the count of atoms of the same type. Therefore, the value of $DAI(k)$ is equal to the number of sth atom-type plus total perturbation terms and is closely related to its structural environment.

As an illustration, Figure 1 depicts the labeled molecular graph of 2-methyl-3-pentanol. The shortest dis-

*Author to whom correspondence should be addressed. E-mail: csyin@sjtu.edu.cn; Fax: 0086-21-54740825-608.

tance matrix is expressed as follows:

$$D = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 2.928 & 4 \\ 1 & 0 & 1 & 2 & 3 & 1.928 & 3 \\ 2 & 1 & 0 & 1 & 2 & 0.928 & 2 \\ 3 & 2 & 1 & 0 & 1 & 1.928 & 1 \\ 4 & 3 & 2 & 1 & 0 & 2.928 & 2 \\ 2.928 & 1.928 & 0.928 & 1.928 & 2.928 & 0 & 2.928 \\ 4 & 3 & 2 & 1 & 2 & 2.928 & 0 \end{bmatrix}$$

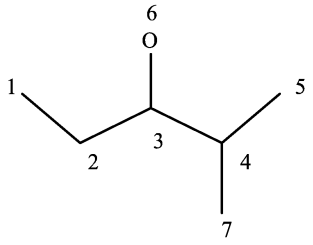

FIG. 1 The labeled molecular graph of 2-methyl-3-pentanol.

The DAI indices are calculated as

$$\begin{aligned} DAI_{(\mathrm{CH_3-})} &= DAI(1) + DAI(5) + DAI(7) \\ &= \left(1 + 7 \times \frac{16.9286}{91.1428}\right) + 2\left(1 + 7 \times \frac{14.9286}{91.1428}\right) \\ &= 6.5933 \\ DAI_{(\mathrm{-CH_2-})} &= DAI(2) \\ &= 1 + 7 \times \frac{11.9286}{91.1428} = 1.9161 \\ DAI_{(\mathrm{-CH<})} &= DAI(3) + DAI(4) \\ &= \left(1 + 7 \times \frac{8.9286}{91.1428}\right) + \left(1 + 7 \times \frac{9.9286}{91.1428}\right) \\ &= 3.4483 \\ DAI_{(\mathrm{-OH})} &= DAI(6) \\ &= 1 + 7 \times \frac{13.5714}{91.1428} = 2.0423 \end{aligned}$$

B. Multiple linear regression

For the retention indices, a multiple linear regression using several DAI indices is used to develop the final model correlating the retention indices of oxygen-containing compounds. The final model is obtained in the form of Eq.(4).

$$RI = a_0 + \sum b_j DAI(j) \tag{4}$$

where a_0 is a constant, and b_j is the contribution coefficient of jth group (atom type). When indices are added or removed, changes in the statistics from model to model can be monitored. Therefore, the significance of each index is evaluated by monitoring the statistics (t and F values) to choose a high quality subset of indices [10,11]. The standard error is used to evaluate the quality of the constructed model.

C. Model validation

In principle, cross-validation is a practical and reliable method for testing the significance of a model. Hence, to validate the final models generated individually for different stationary phase, the leave-one-out method is used to do the cross-validation. In the present work, $n-1$ samples from a total data set are used to construct a calibration set and to build a QSRR model between descriptors and retention indices using MLR. The retention indices of the sample are then predicted using one sample that was left out of the data set. The procedure above is repeated until every sample in the total data set is used for a prediction. The predictive ability of the model is quantified in terms of the corresponding leave-one-out cross-validated parameters, r_{cv} and s_{cv} values [12].

D. Data set

The data sets of Kováts retention indices of oxygen-containing compounds containing ketones and esters were taken from the literature [13]. The compounds range from 4 to 10 carbons, are linear and branched, and contain carbonyl and carboxyl. These RIs are measured on OV-1 at 333 K as shown in Table I.

III. RESULTS AND DISCUSSION

A. Regression analysis

The relationship between the experimental retention indices and DAI indices for 39 oxygen-containing compounds is of excellent quality. The QSRR obtained using five DAI indices shown in Table II, with its statistical parameters, is given as follows:

$$\begin{aligned} RI = {} & 19.0686 + 49.3519 DAI_{(\mathrm{CH_3-})} \\ & + 48.0110 DAI_{(\mathrm{-CH_2-})} + 21.8311 DAI_{(\mathrm{-CH<})} \\ & + 124.5146 DAI_{(\mathrm{O=})} + 10.6036 DAI_{(\mathrm{-O-})} \\ & r = 0.9973, \quad s = 8.23, \quad r_{\mathrm{cv}} = 0.9958, \\ & s_{\mathrm{cv}} = 10.32, \quad F = 1234, \quad P < 0.0001 \end{aligned} \tag{5}$$

All indices in the model are statistically significant according to the t-values at the level of P<0.0001. As it can be seen, this model produces a standard error of 8.23 and explains more than 99% of the variances in the experimental RIs for these compounds. According to Mihalić and Trinajstić's comments on the quality of

TABLE I Experimental and calculated RIs for 39 oxo-containing compounds

No.	Compound	RI				
		Exp	Calcd[a]	Res[a]	Calcd[b]	Res[b]
1	Ethyl acetate	600.00	610.07	−10.07	611.91	−11.91
2	Methyl propionate	615.20	604.21	10.99	601.86	13.34
3	Methyl isobutyrate	671.00	654.70	16.30	652.17	18.83
4	Propyl acetate	696.30	705.32	−9.02	706.32	−10.02
5	Methyl butyrate	705.60	696.83	8.77	695.76	9.84
6	Ethyl isobutyrate	744.60	751.57	−6.97	752.24	−7.64
7	Isobutyl acetate	757.70	755.21	2.49	754.81	2.89
8	Ethyl butyrate	784.00	794.87	−10.87	795.89	−11.89
9	Propyl propionate	792.60	798.15	−5.55	798.51	−5.91
10	Butyl acetate	796.20	799.81	−3.61	800.20	−4.00
11	Isopropyl butyrate	827.60	834.83	−7.23	836.87	−9.27
12	Ethyl isopentanoate	838.40	842.31	−3.91	842.64	−4.24
13	Isobutyl propionate	852.80	847.24	5.56	846.88	5.92
14	Propyl butyrate	881.50	891.25	−9.75	892.26	−10.76
15	1,3-Dimethylbutyl acetate	885.10	903.09	−17.99	912.74	−27.64
16	Pentyl acetate	896.40	894.15	2.25	893.77	2.63
17	Isobutyl isobutyrate	900.00	893.72	6.28	892.16	7.84
18	Methyl hexanoate	907.00	884.19	22.81	882.41	24.59
19	Isobuyl butyrate	940.30	939.74	0.56	939.68	0.62
20	Butyl butyrate	979.40	987.02	−7.62	988.12	−8.72
21	Ethyl hexanoate	982.90	981.98	0.92	981.85	1.05
22	Pentyl propionate	990.50	988.22	2.28	987.92	2.58
23	Hexyl acetate	996.50	988.55	7.95	986.04	10.46
24	3-Methyl-2-butanone	640.90	633.88	7.02	632.75	8.15
25	2-Pentanone	663.30	674.70	−11.40	676.48	−13.18
26	3,3-Dimethyl-2-butanone	693.10	698.48	−5.38	700.39	−7.29
27	4-Methyl-2-pentanone	721.20	726.86	−5.66	727.65	−6.45
28	4-Methyl-3-pentanone	733.00	725.66	7.34	724.93	8.07
29	3-Methyl-2-pentanone	734.80	729.83	4.97	729.30	5.50
30	3-Hexanone	764.80	767.69	−2.89	768.05	−3.25
31	2,4-Dimethyl-3-pentanone	779.00	774.92	4.08	774.06	4.94
32	5-Methyl-3-hexanone	816.70	818.15	−1.45	818.27	−1.57
33	2-Methyl-3-hexanone	820.00	818.29	1.71	818.13	1.87
34	4-Heptanone	853.40	861.64	−8.24	862.83	−9.43
35	3-Heptanone	865.80	862.58	3.22	862.19	3.61
36	2,2,4,4-Tetramethyl-3-pentanone	900.00	896.59	3.41	880.92	19.08
37	2,6-Dimethyl-4-heptanone	954.70	956.95	−2.25	957.67	-2.97
38	3-Octanone	966.00	960.95	5.05	959.95	6.05
39	2-Octanone	968.80	962.70	6.10	961.26	7.54

[a] From the data set; [b] From the cross-validation.

model [14], the constructed model represents excellent QSRR model.

On the other hand, the model is validated using the leave-one-out cross-validation. The r_{cv} and s_{cv} are determined to be 0.9966 and 10.32, which are very close to the statistics of Eq.(5). The cross-validation indicates the good stability of the QSRR model.

The calculated RIs and residual for 39 compounds are shown in Table I. The plot of estimated RIs against the

TABLE II The values of descriptors of the compounds

No.	*DAI*				
	CH_3-	$-CH_2-$	$-CH<$	O=	$-O-$
1	4.4938	1.9371	0.0000	2.0694	1.7498
2	4.5551	1.8864	0.0000	2.0072	1.8716
3	6.6246	0.0000	1.7687	2.0063	1.9120
4	4.5397	3.8260	0.0000	2.0887	1.7397
5	4.6159	3.7758	0.0000	1.9953	1.9089
6	6.7154	2.0212	1.8067	1.9720	1.7981
7	6.6248	1.7560	1.8697	2.1319	1.7560
8	4.7056	5.8416	0.0000	1.9600	1.7983
9	4.6789	5.8530	0.0000	1.9979	1.7351
10	4.5722	5.7015	0.0000	2.1099	1.7589
11	6.8666	3.8825	1.9039	1.8483	1.7714
12	6.7970	3.8036	1.8920	1.9629	1.8362
13	6.7748	3.7777	1.9115	2.0266	1.7337
14	4.7356	7.8722	0.0000	1.9438	1.7442
15	8.7313	1.7747	3.6330	2.1667	1.7747
16	4.5984	7.5782	0.0000	2.1310	1.7890
17	8.9113	1.8040	3.8432	1.9755	1.7330
18	4.6679	7.5536	0.0000	2.0168	1.9787
19	6.8568	5.7721	1.9448	1.9623	1.7320
20	4.7711	9.8164	0.0000	1.9503	1.7312
21	4.7576	9.6922	0.0000	1.9519	1.8625
22	4.7148	9.6943	0.0000	2.0277	1.7502
23	4.6209	9.4620	0.0000	2.1510	1.8223
24	6.4738	0.0000	1.7288	2.0687	0.0000
25	4.4790	3.7039	0.0000	2.0621	0.0000
26	8.4804	0.0000	0.0000	2.0953	0.0000
27	6.5467	1.7331	1.8079	2.1044	0.0000
28	6.6270	1.9222	1.7658	1.9974	0.0000
29	6.5488	1.9222	1.6876	2.0756	0.0000
30	4.6264	5.6833	0.0000	1.9873	0.0000
31	8.6575	0.0000	3.6797	1.9938	0.0000
32	6.7027	3.7016	1.8508	2.0092	0.0000
33	6.6984	3.8431	1.8049	1.9655	0.0000
34	4.6968	7.6532	0.0000	1.9543	0.0000
35	4.6606	7.6042	0.0000	1.9951	0.0000
36	12.8154	0.0000	0.0000	1.9681	0.0000
37	8.9263	3.5618	3.8623	1.9438	0.0000
38	4.1994	10.1120	0.0000	2.0010	0.0000
39	4.0567	10.1284	0.0000	2.0653	0.0000

corresponding experimental RIs were shown in Fig.2.

B. The predictability of the QSRR model

To demonstrate the predictability of the Eq.(5), we randomly select an external data set [13], as shown in Table III. The predictive correlation coefficient r_{pred} and standard error s_{pred} are 0.9966 and 8.56 for the test set, showing a good predictive power of the constructed model. Table III lists calculated, experimental RIs and descriptors for the test compounds.

TABLE III Calculated, experimental RIs and descriptors for the test set of compounds

No.	Compound	RIs			*DAI*				
		Exp	Calcd	Res	CH_3-	$-CH_2-$	$-CH<$	O=	$-O-$
1	Ethyl propionate	694.20	702.18	−7.98	4.6397	3.8979	0.0000	1.9937	1.7688
2	Methyl isopentanoate	761.30	745.34	15.96	6.6968	1.7378	1.8552	2.0167	1.9557
3	Butyl propionate	891.40	893.33	−1.93	4.7004	7.7797	0.0000	2.0109	1.7347
4	2-Ethylbutyl actetae	957.00	949.35	7.65	6.7410	5.6286	1.7257	2.1737	1.7956
5	3-Pentanone	676.40	673.46	2.94	4.5618	3.7607	0.0000	1.9974	0.0000
6	2-Methyl-3-pentanone	733.00	725.66	7.34	6.6270	1.9222	1.7658	1.9974	0.0000
7	2-Hexanone	767.90	769.85	1.95	4.5229	5.5775	0.0000	2.0864	0.0000
8	5-Methyl-2-hexanone	836.50	820.46	16.04	6.6047	3.5209	1.8707	2.1328	0.0000
9	2-Heptanone	868.70	865.28	3.42	4.5577	7.4647	0.0000	2.1114	0.0000
10	2,2-Dimethyl-3-heptanone	964.70	973.09	−8.39	8.7284	5.8634	0.0000	1.9416	0.0000

TABLE IV Correlation coefficients of five *DAI* parameters

DAI	CH_3-	$-CH_2-$	$-CH<$	O=	$-O-$
CH_3-	1.0000				
$-CH_2-$	−0.6221	1.0000			
$-CH<$	0.6402	−0.5443	1.0000		
O=	−0.1093	−0.1232	−0.0761	1.0000	
$-O-$	0.2552	0.1761	−0.1043	0.0102	1.0000

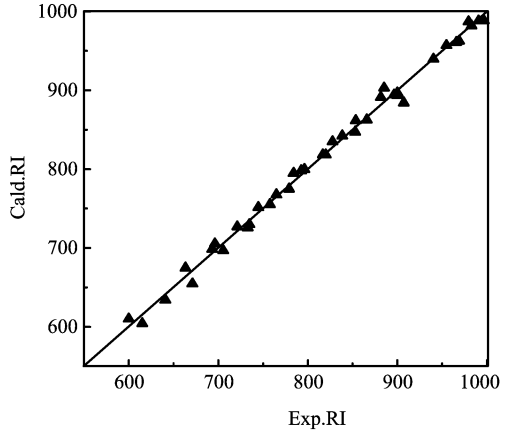

FIG. 2 Plot of observed vs. calculated RIs for the Eq.(7).

C. Variable correlation problems

According to the principle of statistics, a regression equation is of no relevance when the applied explanatory variables were mutually interrelated by simple or multiple correlations. Here, the bivariate correlation of two sets of variables, including the dependent variables (Table IV) was investigated. As it was shown by the correlations of the independent variables given in Table IV, the five-parameter models for modeling RIs have cleared up the possibility.

IV. CONCLUSIONS

Kováts retention indices of 39 oxygen-containing compounds on OV-1 stationary phases are well correlated with novel atom-type topological indices *DAI* using the MLR method. Excellent structure-retention index model show the efficiency of these indices in the QSRR studies. In addition, the final model is validated to be statistically reliable and predictive using the leave-one-out cross validation and an external test set.

[1] B. da S. Junkes, R. D. de M. C. Amboni, R. A. Yunes and V. E. F. Heinzen, Anal. Chim. Acta. **477**, 29 (2003).
[2] B. Ren, Chemometer. Intell. Lab. Syst. **66**, 29 (2003).
[3] E. Estrada and Y. Gutierrez, J. Chromatogr. **A 858**, 187 (1999).
[4] S. Liu, C. Yin, S. Cai and Z. Li, Chemometer. Intell. Lab. Syst. **61**, 3 (2002).
[5] S. S. Liu, H. L. Liu, Z. N. Xia, C. Z. Cao and Z, L. Li, J. Chin. Chem. Soc. **47**, 455 (2000).
[6] J. M. Sutter, T. A. Peterson and P. C. Jurs, Anal. Chim. Acta. **342**, 113 (1997).
[7] F. Yang, X. Yan, L. Ouyang, L. Wang, M. Luo and S. Qu, Chin. J. Chem. Phys. **11**, 221 (1998).
[8] L. H. Hall and L. B. Kier, J. Chem. Inf. Comput. Sci. **35**, 1039 (1995).
[9] S. Liu, Y. Liu, S. Cai and Z. Li, Acta Chimica Sinica **58**, 1353 (2000).
[10] H. H. Maw and L. H. Hall, J. Chem. Inf. Comput. Sci. **41**, 1248 (2001).
[11] K. Rose, L. H. Hall and L. B. Kier, J. Chem. Inf. Comput. Sci. **42**, 651 (2002).
[12] L. Xu, *Chemometrical Method*, Beijing: Science Press, China (1996).
[13] The Sadtler Standard Gas Chromatography Retention Index Library, Sadtler Research Laboratories, Philadelphia, PA, 1985.
[14] Z. Mihalić and N. J. Trinajstić, J. Chem. Educ. **69**, 701 (1992).