ЖУРНАЛ СТРУКТУРНОЙ ХИМИИ

2017. Том 58, № 3 *Март – апрель* *С. 516 – 524*

UDC 541.6:541.021

STRUCTURAL CHARACTERIZATION AND CHROMATOGRAPHIC RETENTION TIME SIMULATION FOR SOME ALIPHATIC CARBOXYLIC ACIDS

L.-M. Liao[1,2], X. Huang[1], G.-D. Lei[1]

[1] *College of Chemistry and Chemical Engineering, Neijiang Normal University, Neijiang, Sichuan, P. R. China*
E-mail: leigdnjtc@126.com
[2] *College of Chemistry and Chemical Engineering, Chongqing University, Chongqing, P. R. China*

Received September, 8, 2015 *Revised — October, 14, 2015*

Molecular structural descriptors (MVEIs) are developed from the molecular vertex electronegativity interaction, and the molecular structures of 43 organic acids are characterized. Two quantitative structure-retention relationship models are built up by the multiple linear regression and the partial least squares regression. The correlation coefficients (R) of the two models are 0.990 and 0.988, and the standard deviations of them are 2.935 and 3.024, respectively. Then the two models are evaluated by the leave-one-out cross-validation and the correlation coefficients (R_{CV}) are 0.985 and 0.976, the standard deviations are 3.527 and 3.628, respectively. It is confirmed that MVEIs are largely dependent on the properties of the organic molecules.

DOI: 10.15372/JSC20170309

Keywords: organic acids, structural descriptor, retention time, structure and properties, molecular vertices electronegativity interaction (MVEI).

INTRODUCTION

With the development of science and technology, computers have been playing an increasingly important role in researching the properties of compounds [1]. In terms of forecasting the properties of organic compounds, researchers have made meaningful works and studies boiling points [2], solubility [3], distribution [4], biological activity [5], chromatographic retention behavior [6], etc. For the structural characterization, there are mainly two-dimensional (2D) description methods [7, 8] and three-dimensional (3D) description methods [9, 10]. However, 2D descriptors cannot determine the actual molecular spatial structure and distinguish *cis* and *trans* isomers, etc. 3D structural description methods, such as CoMFA or CoMSIA, have an intrinsic unfavorable drawback, i.e. the molecular conformation alignment before performing a QSAR study. Moreover, some other issues are also inevitable, such as the spatial lattice partition, control of variable amounts, and selection of appropriate probes in potential fields. Therefore, the development of new simple 3D molecular structural descriptors is very necessary. Organic acids are widely occurring substances in plant tissues. For example, organic acids are the main substance in tobacco, play an important role in the growth process of metabolism, and bring direct impacts on the quality of tobacco. In this paper, some of organic acids from tobacco leaves are adopted as research samples. New molecular structural descriptors (molecular vertex electronegativity interaction, MVEI) constructed from the previous studies [11—14] are used to characterize the structures of the research samples. Through multiple linear regression (MLR) and partial least squares (PLS) regression, two structure retention relationship (QSRR) models are constructed.

Table 1

43 *organic acids of tobacco leaves and their chromatographic retention time* (t_R)

No.	Compound	Formula	t_R/min
01	2-Propenoic acid	$C_3H_4O_2$	8.16
02	*n*-Butanoic acid	$C_4H_8O_2$	11.36
03	Propanoic acid, 2,2-dimethyl-	$C_5H_{10}O_2$	12.42
04	Butanoic acid, 2-methyl-	$C_5H_{10}O_2$	14.82
05	*n*-Pentanoic acid	$C_5H_{10}O_2$	16.42
06	Propanoic acid, 3-chloro-	$C_3H_5ClO_2$	20.84
07	*n*-Hexanoic acid	$C_6H_{12}O_2$	21.88
08	Butanoic acid, 4-hydroxy-2-methylene-	C_5H_8O3	23.38
09	*n*-Heptanoic acid	$C_7H_{14}O_2$	27.34
10	Hexanoic acid, 2-ethyl-	$C_8H_{16}O_2$	29.86
11	Benzoic acid	$C_7H_6O_2$	31.75
12	2-Octenoic acid, *cis*-	$C_8H_{14}O_2$	32.40
13	*n*-Octanoic acid	$C_8H_{16}O_2$	32.53
14	Hexanoic acid, 2-methyl-	$C_7H_{14}O_2$	33.86
15	2-Propenoic acid, 3-(2-hydroxyphenyl)-, (*E*)-	$C_9H_8O_3$	34.69
16	*iso*-Nonanoic acid	$C_9H_{18}O_2$	35.99
17	*n*-Nonanoic acid	$C_9H_{18}O_2$	37.86
18	3-Chloroperbenzoic acid	$C_7H_5C_lO_3$	40.67
19	Decanynoic acid	$C_{10}H_{16}O_2$	41.31
20	Geranic acid	$C_{10}H_{16}O_2$	41.85
21	*n*-Decanoic acid	$C_{10}H_{20}O_2$	42.65
22	Benzoic acid, 4-hydroxy-3,5-dimethoxy	$C_9H_{10}O_5$	43.46
23	*iso*-Undecanoic acid	$C_{11}H_{22}O_2$	44.25
24	*n*-Undecanoic acid	$C_{11}H_{22}O_2$	47.04
25	Dibenzofuran-2-sulfonic acid	$C_{12}H_8O_4S$	49.97
26	*n*-Dodecanoic acid	$C_{12}H_{24}O_2$	51.43
27	Tridecanoic acid	$C_{13}H_{26}O_2$	54.09
28	*n*-Tridecanoic acid	$C_{13}H_{26}O_2$	55.42
29	*iso*-Tetradecanoic acid	$C_{14}H_{28}O_2$	58.09
30	*Z*-7-Tetradecenoic acid	$C_{14}H_{26}O_2$	58.75
31	*n*-Tetradecanoic acid	$C_{14}H_{28}O_2$	59.55
32	Pentadecanoic acid	$C_{15}H_{30}O_2$	62.21
33	*Z*-8-Methyl-9-tetradecenoic acid	$C_{15}H_{28}O_2$	62.62
34	*n*-Pentadecanoic acid	$C_{15}H_{30}O_2$	63.28
35	*iso*-Hexadecanoic acid	$C_{16}H_{32}O_2$	65.54
36	Hexadecenoic acid *Z*-11-	$C_{16}H_{30}O_2$	66.08
37	*n*-Hexadecanoic acid	$C_{16}H_{32}O_2$	67.14
38	*iso*-Heptadecanoic acid	$C_{17}H_{34}O_2$	69.27
39	*iso*-Octadecenoic acid	$C_{18}H_{34}O_2$	72.33
40	9,12-Octadecadienoic acid (*Z*,*Z*)-	$C_{18}H_{32}O_2$	72.47
41	Oleic acid	$C_{18}H_{34}O_2$	72.61
42	*iso*-Octadecenoic acid	$C_{18}H_{34}O_2$	72.74
43	*n*-Octadecanoic acid	$C_{18}H_{36}O_2$	73.40

Further, the robustness and prediction ability of the built models are tested by the leave-one-out (LOO) cross-validation, which made the results more reliable.

EXPERIMENTAL

Data sets. Li et al. [15] separated and analyzed the constituents of acidic compounds in tobacco leaves using comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry (GC×GC-TOFMS), and 43 organic acids were identified. The authors take these 43 compounds as research samples (Table 1). The retention time of the compounds was obtained with a non-polar DB-Petro column.

Structural characterization of compounds. Construction of molecular structural descriptors. For the molecular matrix skeleton, each non-hydrogen atom is a molecular vertex. The non-hydrogen atoms in the organic molecules are often C, O, N, P, S and halogen atoms (F, Cl, Br, and I). The vertex is the fundamental constitute of a molecule; molecular macroscopic properties are reflected from its compositive atomic levels. The properties of a molecule depend mainly on various interactions between the vertices in the molecule. These interactions vary with the electronegativity of the vertices and the distances between them. Without regard to non-framework hydrogen atoms, these vertices are classified as four vertex atomic types (A1, A2, A3, and A4) according to the number of each vertex connected through some chemical bond/bonds. If a vertex is linked to k ($k = 1, 2, 3, 4$) other vertices through chemical bonds, the vertex atomic type belongs to the kth one. For example, if a vertex atom is linked to two vertex atoms through two chemical bonds, the vertex atomic type belongs to the second one (A2). The interaction between four types of vertex atoms is defined as follows:

$$x_r = m_{nl} = \sum_{i=n,\ j=l} \frac{Z_i Z_j}{r_{ij}^2} \quad (n = 1, 2, 3, 4;\ \ n \le l \le 4), \tag{1}$$

n and l represent the vertex atomic type; Z represents the relative electronegativity of the vertex atom relative to the C atom. For example, the relative electronegativity of the Cl atom is 3.16/2.55 = 1.1239; in MEDV [11—14], r_{ij} represents the ratio between the sum of the shortest bond length (between i and j atoms) and a length of the single C—C bond. In this paper, r_{ij} is improved to some extent. r_{ij} represents the ratio of the 3D distance (d) of vertex atoms under the molecular dominant conformation to the single C—C bond length ($BL_{C—C}$ = 0.1540 nm). According to Eq. (1), the interactions between the vertex atoms of a molecule can be assembled as follows: m_{11}, m_{12}, m_{13}, m_{14}, m_{22}, m_{23}, m_{24}, m_{33}, m_{34}, and m_{44}, shortened as x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9 and x_{10}. m_{11} (x_1) represents interactions between the vertex atoms whose vertex atomic type belongs to the first one (A1), and m_{12} (x_2) represents interactions between the vertex atoms whose vertex atomic type belongs to the first one (A1), and the vertex atoms whose vertex atomic type belongs to the second one (A2), and so on.

Generation of molecular descriptors. Three-dimensional molecular structures of 43 compounds are automatically generated by the Chemoffice 8.0 software, and then the semi-empirical quantum chemistry MOPAC 6.0 software contained in Chem3D is used to obtain the final optimized molecular structures at AM1 levels (cut-off value = 0.001 kJ/mol). For example, the final optimized molecular structure of No. 1 compound is shown in Fig. 1, and the spatial position coordinates for all vertex atoms in the molecule can be obtained (Table 2).

Then spatial distance (3D distance (d)) between the vertex atoms is calculated using their coordinates, and then r_{ij} is obtained with the 3D distance (d) and the single C—C bond length. The structural descriptors are obtained through the calculation of equation (1) with r_{ij} and the atomic relative electronegativity. Structural descriptors for other molecules are obtained in the same way. x_{10} is one full-zero vector, and thus, the other 9 variables (Table 3) are used for modeling.

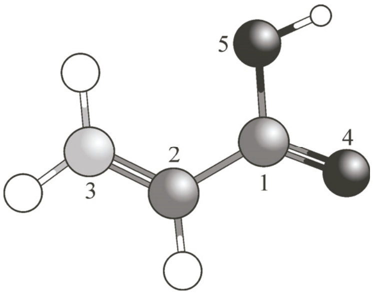

Fig. 1. 3D structure of 2-propenoic acid

Table 2

Spatial position coordinates of all vertex atoms of 2*-propenoic acid*

S. N.	Atom	Atomic relative electronegativity	x	y	z
1	C	1.0000	–3.3723	–0.0707	0.0000
2	C	1.0000	–4.6408	–0.8030	0.0000
3	C	1.0000	–5.8364	–0.2163	–0.0405
4	O	1.3490	–2.2261	–0.5311	0.0374
5	O	1.3490	–3.4737	1.2914	–0.0476

RESULTS AND DISCUSSION

MLR model. MLR is a classic modeling technique. Based on the past experience, a fine model shall comply with the empirical rule that the number of samples (N) / number of variables (n) is larger than 5. However, the number of samples in this study (N) is only 43, while the number of structural descriptors (i.e., independent) is 9. Therefore, the screening of variables is necessary. A stepwise multiple regression analysis contained in SPSS 13.0 is used to select the variables, and VIF [16] is calculated for each variable. VIF = $(1 - R^2)^{-1}$, where R represents the correlation between one independent variable and other variables. If VIF = 1.0, there is no correlation between the variables; when VIF = = 1.0 ~ 5.0, no obvious collinearity exists between the variables and the correlation equation can be accepted; if VIF > 5.0, there is obvious collinearity between the variables and the correlation equation cannot be accepted. Changes in correlation coefficients (R/R_{CV}) and standard deviations (SD/SD$_{\text{CV}}$) with the stepwise multiple regression (SMR) are shown in Fig. 2.

According to Fig. 2, *a* when four variables are introduced into the model, the multiple correlation coefficient (R) is 0.990 (close to the maximum value), and the cross-verification correlation coefficient (R_{CV}) reaches 0.985 (reaches the maximum value). When more variables are introduced, R increases not obviously, but R_{CV} decreases obviously. According to Fig. 2, *b*, when four variables are introduced into the model, the standard deviation (SD) is 2.935 (close to the minimum value), and cross-verification standard deviation (SD$_{\text{CV}}$) reaches 3.527 (reaches the minimum value). When more variables are introduced, SD decreases not obviously, but SD$_{\text{CV}}$ increases obviously. The results suggest that the 4-variable model is optimal, and the x_2, x_3, x_4 and x_5 are selected in the 4-variable model. VIF of the four variables are calculated and they are 4.587, 1.523, 1.279 and 4.795, respectively. All VIFs are in the range 1.0 ~ 5.0, suggesting that there is no significant collinearity between the variables and the equation is acceptable. The model complies with the empirical rule $N/n > 5$. The 4-variable model is shown in Eq. (2)

$$t_{\text{R}} = 1.701 + 2.175x_2 + 0.884x_3 + 1.460x_4 + 0.873x_5. \qquad (2)$$

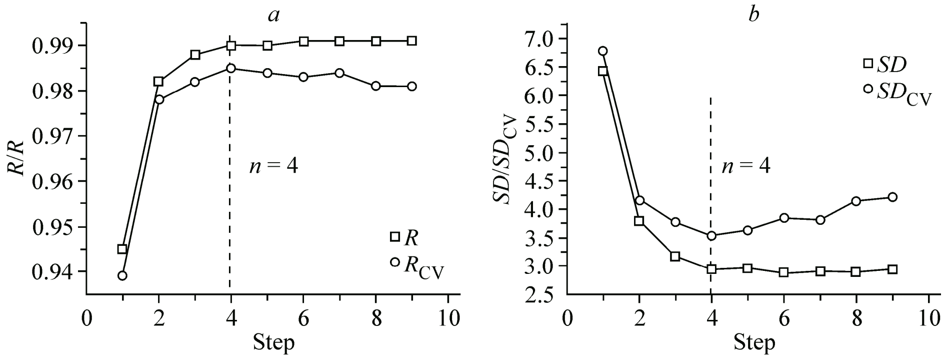

Fig. 2. Plot of R and R_{CV} (*a*) and plot of SD and SD$_{\text{CV}}$ (*b*) change with SMR

Table 3

Descriptor values of compounds

No.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	t_R/min
01	2.6495	2.9808	3.9339	0.0000	0.0000	1.1399	0.0000	0.0000	0.0000	8.16
02	2.1929	4.6982	3.6754	0.0000	1.0112	1.6425	0.0000	0.0000	0.0000	11.36
03	7.0809	0.0000	5.1387	4.7567	0.0000	0.0000	0.0000	0.0000	1.0205	12.42
04	4.3880	2.9753	7.7772	0.0000	0.0000	1.6331	0.0000	1.0206	0.0000	14.82
05	2.0210	6.0145	3.5828	0.0000	2.6414	2.0456	0.0000	0.0000	0.0000	16.42
06	2.3483	4.8297	3.7408	0.0000	1.0112	1.6425	0.0000	0.0000	0.0000	20.84
07	2.1925	6.3848	5.4275	0.0000	2.6414	3.5021	0.0000	0.2962	0.0000	21.88
08	4.1822	5.8118	7.9056	0.0000	1.0112	2.6892	0.0000	1.1399	0.0000	23.38
09	1.8945	8.6475	3.5065	0.0000	7.0149	2.8258	0.0000	0.0000	0.0000	27.34
10	3.6431	10.1485	6.8083	0.0000	4.2267	5.5612	0.0000	1.0206	0.0000	29.86
11	1.2705	5.5690	5.1014	0.0000	8.4193	6.7828	0.0000	1.1399	0.0000	31.75
12	1.8178	9.7678	3.4749	0.0000	9.7948	3.2529	0.0000	0.0000	0.0000	32.40
13	1.8143	9.4779	3.4747	0.0000	9.6016	3.0599	0.0000	0.0000	0.0000	32.53
14	3.4810	9.5828	4.1542	0.0000	4.6734	2.4010	0.0000	0.0000	0.0000	33.86
15	2.3399	9.8724	7.8936	0.0000	10.0447	12.0185	0.0000	1.9172	0.0000	34.69
16	3.0506	9.8773	6.3491	0.0000	7.0151	5.4125	0.0000	0.2341	0.0000	35.99
17	1.7320	10.1977	3.4442	0.0000	12.3945	3.2621	0.0000	0.0000	0.0000	37.86
18	1.8403	10.7363	7.3141	0.0000	7.1098	12.1894	0.0000	2.2386	0.0000	40.67
19	1.6597	10.3009	3.3548	0.0000	15.3665	3.3649	0.0000	0.0000	0.0000	41.31
20	4.8202	9.6391	10.1230	0.0000	4.2327	7.9254	0.0000	1.2175	0.0000	41.85
21	1.6832	10.8138	3.4254	0.0000	15.3638	3.4339	0.0000	0.0000	0.0000	42.65
22	5.0350	13.0006	15.8334	0.0000	3.9406	15.8416	0.0000	6.8671	0.0000	43.46
23	2.7608	14.2336	3.5947	0.0000	15.3639	3.4339	0.0000	0.0000	0.0000	44.25
24	1.6345	11.3638	3.4075	0.0000	18.4876	3.5869	0.0000	0.0000	0.0000	47.04
25	3.3886	9.7044	6.8512	4.1656	16.2079	26.6926	2.8294	7.5365	2.1684	49.97
26	1.6024	11.8516	3.3953	0.0000	21.7487	3.7220	0.0000	0.0000	0.0000	51.43
27	1.5704	12.2960	3.3837	0.0000	25.1336	3.8449	0.0000	0.0000	0.0000	54.09
28	1.5704	12.2960	3.3837	0.0000	25.1336	3.8449	0.0000	0.0000	0.0000	55.42
29	2.5165	13.7725	5.8593	0.0000	21.7488	7.1069	0.0000	0.1229	0.0000	58.09
30	1.5902	13.0746	3.3899	0.0000	29.6074	4.0675	0.0000	0.0000	0.0000	58.75
31	1.5478	12.6991	3.3751	0.0000	28.6308	3.9561	0.0000	0.0000	0.0000	59.55
32	1.5253	13.0714	3.3669	0.0000	32.2311	4.0587	0.0000	0.0000	0.0000	62.21
33	2.1802	15.4526	5.2214	0.0000	24.3773	9.1513	0.0000	0.1762	0.0000	62.62
34	1.5253	13.0714	3.3669	0.0000	32.2311	4.0587	0.0000	0.0000	0.0000	63.28
35	1.5253	13.9656	3.3668	0.0000	36.2753	4.1606	0.0000	0.0000	0.0000	65.54
36	1.5335	13.6250	3.3693	0.0000	36.8598	4.1844	0.0000	0.0000	0.0000	66.08
37	1.5085	13.4148	3.3606	0.0000	35.9266	4.1532	0.0000	0.0000	0.0000	67.14
38	2.3820	15.2767	5.7318	0.0000	32.2311	7.7542	0.0000	0.0944	0.0000	69.27
39	2.3479	15.9866	5.7017	0.0000	36.1290	8.1287	0.0000	0.0879	0.0000	72.33
40	1.5977	14.8957	3.3922	0.0000	46.3865	4.5090	0.0000	0.0000	0.0000	72.47
41	1.5083	14.3747	3.3599	0.0000	44.8469	4.4223	0.0000	0.0000	0.0000	72.61
42	2.3479	15.9866	5.7017	0.0000	36.1290	8.1287	0.0000	0.0879	0.0000	72.74
43	1.4789	14.0340	3.3496	0.0000	43.5768	4.3233	0.0000	0.0000	0.0000	73.40

Model fitting: $N = 43$, $R = 0.990$, SD = 2.935, $F = 451.061$; cross-validation: $R_{\text{CV}} = 0.985$, $\text{SD}_{\text{CV}} = 3.527$, $F_{\text{CV}} = 309.372$, where N represents the number of samples; R represents the multiple correlation coefficient; SD is the standard deviation of estimation; F is the Fischer test value. CV represents cross-validation.

The multiple correlation coefficient (R) and the CV correlation coefficient (R_{CV}) of the established model are desirable (greater than 0.95), indicating that the stability and predictability of the model are satisfactory. Another indicator used to evaluate the quality of the model is SD. The model is good and the prediction accuracy is acceptable when the ratio of SD to the value range is less than 10 % [17]. SD and SD_{CV} of the model are 2.935 and 3.527, and the retention time range between the maximum and minimum values is 65.24. The ratios of SD and SD_{CV} to the retention time range (65.24) are 4.50 and 5.41 %, respectively. They are less than 10 %, indicating that the accuracy of the model prediction is acceptable.

PLS regression model. The PLS regression is a widely used modeling method at present, which has an advantage of effectively overcoming multicollinearity issues and especially suits for the condition of a sample size smaller than the variable number. In addition, PLS has the desirable property that the precision of the model parameters is improved with increasing number of relevant variables and observations.

The Simca-P11.5 software is used to build the PLS model for the samples. To ensure the model has a good prediction capability; LOO cross-validation is used to test the model. A model is constructed from 9 descriptors (Table 3) of the samples. The number of principal components (A) is 3. The correlation coefficient (R) and the cross-validation correlation coefficient (R_{CV}) of the model are 0.988 and 0.976. SD and the cross-validation standard deviation (SD_{CV}) of the model are 3.024 and 3.682, respectively. R and R_{CV} are larger than 0.95, showing the strong predictive ability and stability of the model. The ratios of SD and SD_{CV} to the retention time range (65.24) are 4.64 and 5.56 %, respectively. They are less than 10 %, also indicating the acceptable accuracy of the model prediction.

Fig. 3 presents the scoring distribution scatter of 43 samples at the first and second PLS principal component spaces. It can be seen that most samples fall into the Hotelling T^2 confidence ellipse with 95 % confidence, only with exception of two samples (No. 3 and No. 25, <5 %). Compounds No. 3 and No. 25 contain the fourth type of vertex atoms, while other compounds do not contain this type of vertex atoms. Therefore, they have a certain degree of particularity. Statistical results indicate that the structural descriptors built in this article can successfully give structural characteristics of organic acid compounds.

The VIP (variable importance in the projection) can reflect the explanatory power of each variable on Y (Fig. 4). Usually, the variable whose VIP value is greater than 1, has a greater correlation and a larger explanatory power to Y. For this system, the VIP values of x_5 and x_2 are greater than 1,

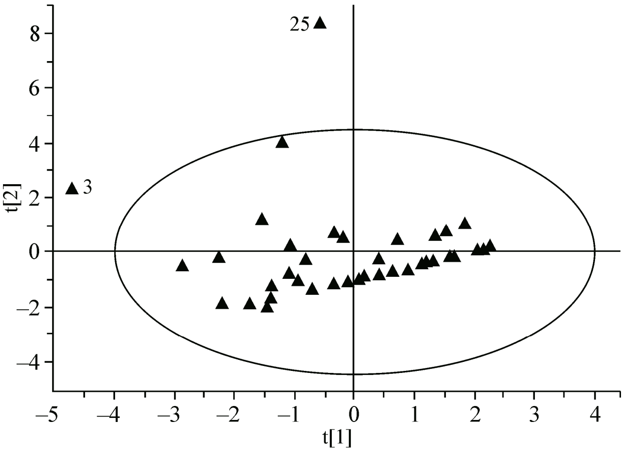

Fig. 3. Front two principal component score distribution

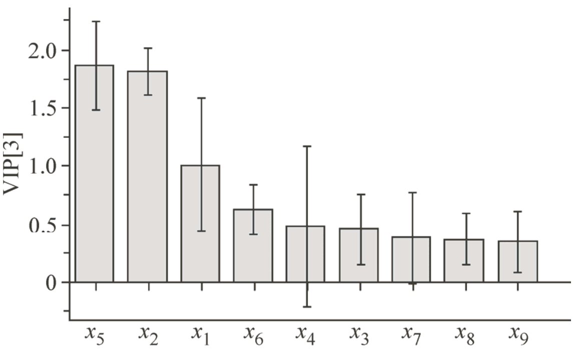

Fig. 4. Importance of variables

Table 4

Experimental values, calculated values and errors of compound retention time

No.	t_R/min	Cal$_{(MLR)}$	Err$_{(MLR)}$	Cal$_{(PLS)}$	Err$_{(PLS)}$	No.	t_R/min	Cal$_{(MLR)}$	Err$_{(MLR)}$	Cal$_{(PLS)}$	Err$_{(PLS)}$
1	8.16	11.661	3.501	10.745	2.585	23	44.25	49.252	5.002	50.929	6.679
2	11.36	16.051	4.691	15.639	4.279	24	47.04	45.571	–1.469	45.716	–1.324
3	12.42	13.186	0.766	11.027	–1.393	25	49.97	49.096	–0.874	48.431	–1.539
4	14.82	15.045	0.225	12.892	–1.928	26	51.43	49.469	–1.961	49.415	–2.015
5	16.42	20.256	3.836	20.170	3.75	27	54.09	53.380	–0.710	53.094	–0.996
6	20.84	16.395	–4.445	16.062	–4.778	28	55.42	53.380	–2.040	50.094	–5.326
7	21.88	22.691	0.811	22.490	0.61	29	58.09	55.824	–2.266	57.148	–0.942
8	23.38	22.211	–1.169	21.014	–2.366	30	58.75	58.985	0.235	58.493	–0.257
9	27.34	29.734	2.394	30.206	2.866	31	59.55	57.303	–2.247	56.757	–2.793
10	29.86	33.482	3.622	34.779	4.919	32	62.21	61.249	–0.961	60.419	–1.791
11	31.75	25.673	–6.077	25.392	–6.358	33	62.62	61.210	–1.410	63.761	1.141
12	32.40	34.570	2.170	35.204	2.804	34	63.28	61.249	–2.031	60.418	–2.862
13	32.53	33.770	1.240	34.254	1.724	35	65.54	66.725	1.185	65.709	0.169
14	33.86	30.296	–3.564	31.542	–2.318	36	66.08	66.496	0.416	65.329	–0.751
15	34.69	38.920	4.230	40.945	6.255	37	67.14	65.217	–1.923	64.080	–3.06
16	35.99	34.921	–1.069	36.083	0.093	38	69.27	68.135	–1.135	69.942	0.672
17	37.86	37.748	–0.112	38.170	0.31	39	72.33	73.056	0.726	73.773	1.443
18	40.67	37.725	–2.945	40.291	–0.379	40	72.47	77.598	5.128	75.819	3.349
19	41.31	40.488	–0.822	40.640	–0.67	41	72.61	75.092	2.482	73.284	0.674
20	41.85	35.308	–6.542	36.519	–5.331	42	72.74	73.056	0.316	73.773	1.033
21	42.65	41.663	–0.987	41.970	–0.68	43	73.40	73.233	–0.167	71.426	–1.974
22	43.46	47.411	3.951	47.930	4.47						

therefore these two variables can explain *Y* with a relatively strong explanatory power. x_5 corresponds to the interactions between the second type of vertices (A2); x_2 corresponds to the interactions between the first (A1) and the second types of vertices (A2). The results show that the number of substituents, the size and complexity of the substituents significantly affect the retention behavior of the compound.

Comparison of models. The retention time of the samples is estimated by the MLR model (Eq. (2)) and the PLS model, and the results are listed in Table 4. Fig. 5 presents a plot of the observed retention time versus estimated and calculated ones, and Fig. 6 presents the residual distribution of calculated results.

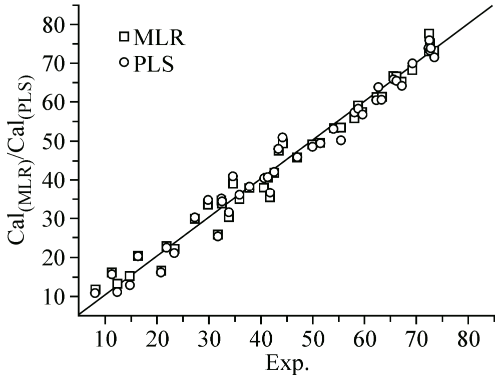

Fig. 5. Calculated vs. experimental values

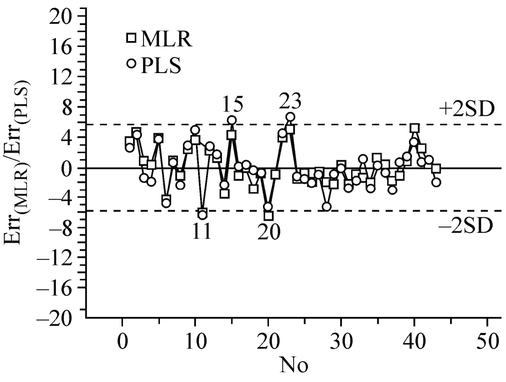

Fig. 6. Comparative residuals of the compounds

Table 5

Comparisons among different QSRR models

No.	Descriptors	N	Selected variables	n/A	Method	R	SD	F	R_{CV}	SD_{CV}	F_{CV}
1	MEDV [12]	53	x_1—x_{10}	10	MLR	0.906	∇	—	0.903	∇	—
2	MEDV [12]	53	x_1, x_3, x_5, x_6, x_1, x_7, x_{10}	7	MLR	0.906	∇	—	0.904	∇	—
3	MEDV [13]	63	x_1, x_2, x_3, x_5, x_6, x_9	6	MLR	0.982	∇	247.554	0.974	∇	—
4	MEDV [13]	63	x_3, x_5, x_6	3	MLR	0.980	∇	487.159	0.977	∇	415.315
5	MEDV [14]	45	x_1—x_{10}	10	MLR	0.952	∇	33.128	0.925	∇	20.030
6	MEDV [14]	45	x_1-x_4, x_5, x_6, x_9, x_{10}	6	MLR	0.947	∇	105.039	0.923	∇	25.843
7	MEDV*	43	x_2—x_5	4	MLR	0.982	4.096	113.990	0.976	4.422	193.342
8	MVEI*	43	x_2—x_5	4	MLR	0.990	2.935	451.061	0.985	3.527	309.372
9	MEDV*	43	x_1—x_9	3	PLS	0.978	3.891	—	0.951	4.451	—
10	MVEI*	43	x_1—x_9	3	PLS	0.988	3.024	—	0.976	3.628	—

"∇" — not listed here; "—" — not available; "*" — results of this work.

Fig. 5 shows that most sample points are near the 45° diagonal line and the residual errors of most samples are smaller than 2SD in Fig. 6. For the MLR model, the residual errors of only two compounds (No.11 and No.20) are little bigger than 2SD. For PLS model, the residual errors of three compounds (No.11, No.15, and No.23) are slightly larger than 2SD. In addition, the correlation coefficient (R and R_{CV}) and standard deviations (SD and SD_{CV}) of the two models also indicate that the predictive ability of the MLR model is slightly better than that of the PLS model. Overall, the simulation results of the two models are satisfactory.

Molecular vertex electronegativity interaction (MVEI) is obtained based on the improvement of MEDV [11—14]. In [12—14], MEDV was used to characterize the structures of volatile compounds of natural products and relevant QSRR models were established through MLR. The results are listed from No. 1 to No. 6 in Table 5. For a more fair comparison, MEDV is also used to characterize the structures of 43 samples of this research. The same variable combination and methods as in this research are used for modeling. The relevant results are given for No.7 (MLR) and No.9 (PLS) in Table 5.

According to Table 5, the fitting results and cross-validation results of MVEI models are obviously superior to those of MEDV models. It indicates that the characterization capability of MVEI for the compound structure is better than that of MEDV and the improvement of MEDV in this paper is successful.

CONCLUSIONS

Interactions between molecular vertices are calculated as structural descriptors (MVEIs) based on three-dimensional structures of molecules. 43 organic acids are characterized by the structural descriptors and two QSRR models are constructed through MLR and PLS regression. The results show that the descriptors constructed in this research can characterize well the difference in the molecular structures of organic compounds. The models can be used to predict the retention time of organic acids under the experimental conditions described in the literature. Predicted values have some reference value when there is no experimental value. Molecular structural descriptors are constructed entirely from the molecular calculation, and need no consideration of the conformational optimization, overlapping issues, etc. The calculation is relatively simple, fast and easy. Therefore, they are expected to promote the QSPR/QSAR study of organic compounds.

This work was supported by the Foundation of Education Bureau, Sichuan Province (13ZB0003) and Technology Bureau, Sichuan Province (2012j13-141).

REFERENCES

1. *Pang J., Ma Z., Shen B.S. et al.* // Chin. J. Struct. Chem. – 2014. – **33**. – P. 480 – 489.
2. *Qin S., Liao L.M.* // Comput. Appl. Chem. – 2012. – **29**. – P. 973 – 976.
3. *Eslam P., Reza A.H., Jamal S.A. et al.* // J. Mol. Liq. – 2015. – **204**. – P. 162 – 169.
4. *Liao L.M., Li J.F., Lei G.D. et al.* // J. Struct. Chem. – 2011. – **52**. – P. 1111 – 1114.
5. *Liao L.M., Li J.F., Wang B.* // Chin. J. Struct. Chem. – 2011. – **30**. – P. 1397 – 1402.
6. *Qin S., Li J.F., Liao L.M.* // Chin. J. Struct. Chem. – 2012. – **31**. – P. 665 – 672.
7. *Todeschini R., Gramatice P., Provenzani R.* // Chemom. Intell. Lab. Syst. – 1995. – **27**. – P. 221 – 229.
8. *Bravig R., Gancia E., Mascagni P.* // J. Comput.-Aided Mol. Des. – 1997. – **11**. – P. 79 – 92.
9. *Travis R.H., Richard J.S., Patricia L. et al.* // Bioorg. Med. Chem. Lett. – 2015. – **25**. – P. 327 – 332.
10. *Yu S.L., Yuan J.Y., Shi J.H. et al.* // Chemom. Intell. Lab. Syst. – 2015. – **146**. – P. 34 – 41.
11. *Sun L.L., Zhou L.P., Yu Y. et al.* // Chemosphere. – 2007. – **66**. – P. 1039 – 1051.
12. *Zhu W.P., Yang S.B., Liao L.M. et al.* // Chin. J. Struct. Chem. – 2009. – **28**. – P. 391 – 396.
13. *Wu J.H., Zhang S.W., Zhang C.J. et al.* // J. Shanxi Univ., Nat. Sci. Ed. – 2010. – **33**. – P. 425 – 429.
14. *Li J.F.* // Sci. Technol. Food Ind. – 2014. – **35**. – P. 292 – 295.
15. *Li H.F., Lu X., Lu H.L. et al.* // Chem. J. Chin. Univ. – 2006. – **27**. – P. 612 – 617.
16. *Zhu L.L.* // Food Sci. – 2011. – **32**. – P. 109 – 112.
17. *Sung-Sun S., Karplus M.* // J. Comput.-Aided Mol. Des. – 1999. – **13**. – P. 243 – 258.