J Mol Model (2006) 12: 749–756
DOI 10.1007/s00894-005-0089-4

ORIGINAL PAPER

Chunhui Lu · Weimin Guo · Yang Wang · Chunsheng Yin

Novel distance-based atom-type topological indices *DAI* for QSPR/QSAR studies of alcohols

Received: 20 July 2005 / Accepted: 25 October 2005 / Published online: 13 January 2006

Abstract In this work, we propose a distance-based atom-type topological index (*DAI*) for quantitative structure-property/activity relationship (QSPR/QSAR) studies. The newly constructed index, which codes the structural environment of each atom type in a molecule, can be calculated simply. These atom-type topological indices, along with our recently proposed *Lu* index, were used to construct QSPR/QSAR models for several representative physical properties and biological activities of several data sets of alcohols with a range of non-hydrogen atoms by using multiple linear regression (MLR) analysis. The efficiency of these indices is verified by high quality QSPR models. The results indicate that the combined use of *Lu* and *DAI* indices promises to be a useful method for QSPR/QSAR analysis of complex compounds.

Keywords Topological index · Multiple linear regression · QSPR/QSAR · Alcohols

Introduction

The application of graph theory to chemistry and to structure-property-activity relationships (QSPR/QSAR) has led to the emergence of many topological indices, including the Randic index [1], Hosoya index [2], Balaban index [3], Bochev index [4], Wiener index [5], Xu index [6] and our recently introduced *Lu* index [7]. These topological indices have been applied widely in QSPR/QSAR studies. Although intensive work has been done on heteroatoms and multiple bonds [8–11], these simple indices only reflect overall properties, but cannot give a single value for a certain bond-type/group in a molecular graph. Many properties/activities of compounds, however, are determined by the overall features as well as certain multiple bonds and/or heteratoms. Because of the lack of information about a certain multiple bond and/or heteratom in molecular graphs, most of the existing simple topological indices are limited in their field of application.

In order to differentiate between the multiple bonds and/or heteratoms in molecular graphs and reflect the function of the special bond-type/group, the use of atom-type topological indices in QSAR/QSPR has received considerable attention due to their significant advantages over simple topological indices. Atom-type topological indices, unlike conventional simple topological indices that characterize a molecule as a whole, code the structural environment of each atom type in a molecule and further describe the structural information of a molecule at the atomic level. Therefore, implementations of the atom-type topological index may provide a breakthrough in QSAR/QSPR studies of complex compounds. One of the most important atom-type topological indices is the atom-type electrotopological state (E-state) index proposed by Hall et al. [12]. Recently, Ren [13–17] proposed a novel type of atom-type topological index, which has been used successfully in many QSAR/QSPR studies. Despite the significant achievements in this field, existing atom-type topological-index approaches to QSAR/QSPR may need further improvement.

In this work, we define a novel atom-type index based on the distance matrix of the molecular topological graph. Here, the distance between two vertices is the shortest distance between vertices i and j and is calculated by summing the relative bond length (take the C–C bond length 0.154 nm as 1) between two adjacent vertices in the shortest path. Therefore, these indices can differentiate between different multiple bonds and heteratoms. The novel atom-type index *DAI* can be calculated easily and shows a good correlation with the properties/activities of compounds under study by combined use with our recently proposed *Lu* index [7]. To illustrate the potential of these indices in QSPR/QSAR studies, two series of application examples were analyzed. First, several representative properties such as boiling point and water solubility of alcohols with a wide range of non-hydrogen atoms were selected for this case. The other series

C. Lu · W. Guo · Y. Wang · C. Yin (✉)
School of Environmental Science and Engineering,
Shanghai Jiao Tong University,
Shanghai, 20040, People's Republic of China
e-mail: csyin@sjtu.edu.cn
Tel.: +86-2154740825-608

of examples was related to biological activity and toxicity of alcohols.

Materials and methods

The Lu index is defined as follows: [7]

$$Lu = n^{1/2} \log \left[\frac{1}{2} \left(\sum_{i}^{n} \sum_{j}^{n} D_{ij} + \sum_{i}^{n} \sum_{j}^{n} D_{ij}^{2} \right) \right] \quad (1)$$

where n is the number of vertices in a molecular topological graph. D_{ij} is the shortest distance between vertices i and j and is calculated by summing the relative bond length [18] (take the C–C bond length 0.154 nm as 1) between two adjacent vertices in the shortest path.

For any atom i that belongs to the jth atom-type (take into account both the atomic nature and its connectivity) in a graph, the novel distance-based atom-type topological index $DAI_i(j)$ is expressed as follows:

$$DAI_{\mathrm{i}}(j) = 1 + \Phi_i(j) \quad (2)$$

$$\Phi_i(j) = n \bullet \frac{\sum_{j}^{n} D_{ij}}{\sum_{i}^{n} \sum_{j}^{n} D_{ij}} \quad (3)$$

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 the molecular topological graph. D_{ij} is defined and calculated as above.

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

$$DAI(j) = \sum_{i=1}^{m} DAI_i(j) = m + \sum_{i=1}^{m} \Phi_i(j) \quad (4)$$

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

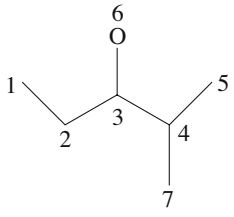

Fig. 1 The labeled molecular graph of 2-methyl-3-pentanol

As an illustration, Fig. 1 depicts the labeled molecular graph of 2-methyl-3-pentanol. The shortest distance matrix is expressed as follows:

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

For such a compound, there are two bond types containing C–C and C–O bonds in the molecular structure. Therefore, the relative bond lengths of these two bond types are 1 and 0.928, relatively. According to the definition above, the Lu index is calculated as

$$Lu = 7^{1/2} \log \left[\frac{1}{2} \left(\sum_{i=1}^{7} \sum_{j=1}^{7} D_{ij} + \sum_{i=1}^{7} \sum_{j=1}^{7} D_{ij}^{2} \right) \right]$$

$$= 7^{1/2} \log (13.6020) = 5.8571$$

The DAI indices are calculated as

$$DAI(\mathrm{CH_3}-) = DAI(1) + DAI(5) + DAI(7)$$

$$= \left(1 + 7 \times \frac{16.9286}{91.1428}\right) + 2\left(1 + 7 \times \frac{14.9286}{91.1428}\right)$$

$$= 6.5933$$

$$DAI(-\mathrm{CH_2}-) = DAI(2) = 1 + 7 \times \frac{11.9286}{91.1428} = 1.9161$$

$$DAI(-\mathrm{CH} <) = DAI(3) + DAI(4)$$

$$= \left(1 + 7 \times \frac{8.9286}{91.1428}\right) + \left(1 + 7 \times \frac{9.9286}{91.1428}\right)$$

$$= 3.4483$$

$$DAI(-\mathrm{OH}) = DAI(6) = 1 + 7 \times \frac{13.5714}{91.1428} = 2.0423$$

Regression analysis

Multiple linear regression

For each property, multiple linear regression using the Lu index and several DAI indices is used to develop the final models correlating the properties and activities of alcohols. The final model is obtained in the form of Eq. (5).

$$\text{property (activity)} = a_0 + a_1 Lu + \sum b_j DAI(j) \tag{5}$$

where a_0 is a constant, a_1 is the contribution coefficient of the Lu index, and b_j is the contribution coefficient of the jth group (atom type). Each coefficient describes the sensitivity of a property to each of the individual indices, so the constant coefficient of these parameters would reflect the relative importance of each index. As indices are added and removed, changes in the statistics can be monitored from model to model. Therefore, the significance of each index is evaluated by monitoring the statistics (t and F values) to choose a high quality subset of indices [19–21]. The standard error is used to evaluate the quality of the model constructed.

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 properties and activities, the leave-one-out method is used to do the cross-validation. In the present study, $n-1$ samples from a total dataset were used to construct a calibration set and to build a QSPR/QSAR model between descriptors and the property or activity examined using MLR. The property or activity of the sample is then predicted using one sample that was left out of the dataset. The procedure above is repeated until every sample in the total data set has been 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, which are defined as: [22]

$$r_{cv} = \sqrt{1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n} (y - \overline{y_i}) H^2}} \tag{6}$$

where y_i and $\hat{y}_i$ are the experimental and predictive value, respectively. $\bar{y}$ is the mean value of y_i.

$$s_{cv} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{N - M - 1}} \tag{7}$$

where N is the number of samples used for model building. M is the number of descriptors. For a reliable model, the r_{cv}^2 value should be >0.6. The model is considered to be excellent if r_{cv}^2 >0.9 [23].

Results and discussion

Correlations to *normal* boiling points of 138 alcohols (BP)

The boiling point at normal pressure of a compound is predetermined by the intermolecular interactions in the liquid and by the difference in the molecular internal partition function in the gas phase and in the liquid at the boiling temperature. Therefore, it is expected to be directly related to the chemical structure of the molecule, and indeed numerous methods have been developed over the years for estimating the normal boiling point of a compound from its structure [24]. As a starting point, we consider a data set of 138 alcohols [25–28] to develop the structure–boiling point model. The observed BP values are listed in Table 1. The best boiling point model is generated by the Lu index and all the DAI indices and is expressed in Eq. (8).

$$\begin{aligned} \text{BP} = & -163.1301 + 27.3869 Lu \\ & + 20.7888 DAI(\text{CH}_3-) \\ & - 6.1200 DAI(-\text{CH}_2-) \\ & - 37.4693 DAI(-\text{CH}) \\ & - 72.3142 DAI(>\text{C}<) \\ & + 56.9727 DAI(-\text{OH}) \\ N = 138, & \; r = 0.9963, r_{cv} = 0.9957, s = 3.270, \\ s_{cv} = 3.532, & \; F = 2970, P < 0.0001 \end{aligned} \tag{8}$$

The t-value is a statistical parameter used to test the significant of each regression coefficient. The t-values are −9.50, 8.43, 8.53, −3.49, −7.45, −8.32 and 14.21, respectively. All indices in the model are statistically significant according to the t-values at the level of P<0.0001. More than 99.3% (r^2) of the variance in the experimental BP values are accounted for by this equation. The relative standard error is 3.3/170.2=1.9% of the mean values of the BP, which is approaching the experimental errors of boiling point measurements. On the other hand, the model is further validated using the leave-one-out cross-validation. The r_{cv} and s_{cv} are determined to be 0.9957 and 3.532 (°C), which are very close to the statistics of Eq. (8). The pairwise correlations between every pair of these indices were also performed. Cross-correlation analysis shows that the indices in the model are not highly correlated with each other (the pairwise correlation

Table 1 Calculated and experimental boiling points for 138 alcohols

		BP (°C)					BP (°C)		
No.	Compound	Experimental	Calculated	Residuals	No.	Compound	Experimental	Calculated	Residuals
1	1-butanol	117.6	109.0	8.6	70	3,4-dimethyl-2-hexanol	165.5	166.5	−1.0
2	2-methyl-1-propanol	107.9	102.9	5.0	71	2,5-dimethyl-2-hexanol	154.4	156.0	−1.6
3	2-butanol	99.5	97.9	1.6	72	4-methyl-4-heptanol	161.0	162.0	−1.0
4	2-methyl-2-propanol	82.4	87.6	−5.2	73	2,4,4-trimethyl-1-heptanol	168.5	166.0	2.5
5	1-pentanol	137.5	132.8	4.7	74	3-ethyl-3-hexanol	160.5	162.5	−2.0
6	3-methyl-1-butanol	131.0	126.9	4.1	75	2,3-dimethyl-2-hexanol	160.0	157.7	2.3
7	2-pentanol	119.3	120.2	−0.9	76	3,5-dimethyl-3-hexanol	158.0	155.5	2.5
8	2-methyl-1-butanol	128.0	125.6	2.4	77	2,3-dimethyl-3-hexanol	158.1	155.8	2.3
9	3-pentanol	116.2	118.6	−2.4	78	2-methyl-3-ethyl-2-pentanol	156.0	158.5	−2.5
10	3-methyl-2-butanol	112.9	114.7	−1.8	79	2,4,4-trimethyl-2-pentanol	147.5	145.6	1.9
11	2,2-dimethyl-1-propanol	113.1	116.1	−3.0	80	2,2,4-trimethyl-3-pentanol	150.5	149.8	0.7
12	2-methyl-2-butanol	102.3	108.3	−6.0	81	2,2-methyl-3-hexanol	156.0	156.0	0.0
13	1-hexanol	157.0	154.7	2.3	82	2,5-methyl-3-hexanol	157.5	159.9	−2.4
14	4-methyl-1-pentanol	151.9	148.4	3.5	83	4,4-methyl-3-hexanol	160.4	159.1	1.3
15	2-hexanol	140.0	141.1	−1.1	84	3,4-methyl-2-hexanol	165.5	166.5	−1.0
16	3-methyl-1-pentanol	153.0	148.6	4.4	85	6-methyl-2-heptanol	174.0	172.4	1.6
17	2-methyl-1-pentanol	148.0	146.1	1.9	86	3-methyl-1-heptanol	186.0	186.2	−0.2
18	3-hexanol	135.0	138.5	−3.5	87	2-methyl-3-ethyl-3-pentanol	158.0	156.7	1.3
19	2-ethyl-1-butanol	146.5	145.8	0.7	88	2,3,4-trimethyl-3-pentanol	156.5	150.1	6.4
20	4-methyl-2-pentanol	132.0	134.8	−2.8	89	1-nonanol	213.3	212.6	0.7
21	3,3-dimethyl-1-butanol	143.0	138.4	4.6	90	7-methyl-1-octanol	206.0	204.8	1.2
22	2,3-dimethyl-1-butanol	144.5	140.8	3.7	91	2-nonanol	198.5	198.0	0.5
23	2-methyl-2-pentanol	121.5	127.4	−5.9	92	3-nonanol	195.0	194.0	1.0
24	3-methyl-2-pentanol	134.3	135.3	−1.0	93	4-nonanol	192.5	191.4	1.1
25	2-methyl-3-pentanol	129.5	132.5	−3.0	94	5-nonanol	193.0	190.6	2.4
26	2,2-dimethyl-1-butanol	136.5	136.2	0.3	95	2-methyl-2-octanol	178.0	180.5	−2.5
27	3-methyl-3-pentanol	123.0	127.6	−4.6	96	2,6-dimethyl-2-heptanol	173.0	172.5	0.5
28	3,3-dimethyl-2-butanol	120.4	125.4	−5.0	97	2,6-dimethyl-3-heptanol	175.0	176.0	−1.0
29	2,3-dimethyl-2-butanol	118.4	122.4	−4.0	98	2,6-dimethyl-4-heptanol	174.5	174.9	−0.4
30	1-heptanol	176.4	175.2	1.2	99	3,6-dimethyl-3-heptanol	173.0	169.3	3.7
31	5-methyl-1-hexanol	170.0	168.4	1.6	100	2,2,3-trimethyl-3-hexanol	156.0	160.9	−4.9
32	2-heptanol	160.4	161.0	−0.6	101	3,5-dimethyl-4-heptanol	171.0	177.5	−6.5
33	4-methyl-1-hexanol	173.3	169.1	4.2	102	2,3-dimethyl-3-heptanol	173.0	171.4	1.6
34	2-methyl-1-hexanol	164.0	165.3	−1.3	103	2,4-dimethyl-4-heptanol	171.0	170.8	0.2
35	3-heptanol	157.0	157.7	−0.7	104	2-methyl-3-ethyl-3-heptanol	177.5	172.1	5.4
36	3-methyl-1-hexanol	169.0	168.0	1.0	105	2-methyl-3-ethyl-1-heptanol	193.0	194.7	−1.7
37	4-heptanol	156.0	156.5	−0.5	106	5-methyl-3-ethyl-3-heptanol	172.0	171.4	0.6
38	5-methyl-2-hexanol	151.0	154.0	−3.0	107	2,4,4-trimethyl-3-hexanol	170.0	166.8	3.2
39	2-methyl-3-hexanol	145.5	149.9	−4.4	108	3,4,4-trimethyl-3-hexanol	165.5	163.4	2.1
40	2-methyl-2-hexanol	143.0	145.7	−2.7	109	4-methyl-4-octanol	180.0	178.2	1.8
41	2,4-dimethyl-1-penanol	159.0	158.9	0.1	110	4-ethyl-4-heptanol	182.0	178.4	3.6
42	5-methyl-3-hexanol	148.0	150.8	−2.8	111	2-methyl-2-octanol	178.0	180.5	−2.5
43	3-methyl-3-hexanol	143.0	145.3	−2.3	112	1-decanol	231.1	229.9	1.2
44	2,4-dimethyl-2-penanol	133.1	139.2	−6.1	113	8-methyl-1-nonanol	219.9	221.8	−1.9
45	2,4-dimethyl-3-penanol	140.0	143.8	−3.8	114	2-decanol	211.0	215.2	−4.2
46	3-ethyl-3-penanol	142.0	145.8	−3.8	115	4-decanol	210.5	208.2	2.3
47	2,3-dimethyl-2-penanol	139.7	140.8	−1.1	116	3,7-dimethyl-1-octanol	212.5	210.6	1.9
48	2,3-dimethyl-3-penanol	139.0	139.8	−0.8	117	2,7-dimethyl-3-octanol	193.5	191.9	1.6
49	2,3,3-trimethyl-2-butanol	130.5	130.9	−0.4	118	2,6-dimethyl-4-octanol	195.0	191.1	3.9
50	3-methyl-2-hexanol	151.0	154.1	−3.1	119	2,3-dimethyl-3-octanol	189.0	186.8	2.2
51	1-octanol	195.2	194.4	0.8	120	5-methyl-5-nonanol	202.0	193.5	8.5

Table 1 (continued)

No.	Compound	BP (°C)			No.	Compound	BP (°C)		
		Experimental	Calculated	Residuals			Experimental	Calculated	Residuals
52	6-methyl-1-heptanol	188.6	187.1	1.5	121	4-methyl-1-nonanol	216.0	221.7	−5.7
53	2-octanol	180.0	179.9	0.1	122	2-methyl-3-nonanol	200.0	200.4	−0.4
54	3-octanol	175.0	176.2	−1.2	123	2,2,5,5-tetramethyl-3-hexanol	170.0	166.0	4.0
55	4-methyl-1-heptanol	188.0	187.7	0.3	124	4-propyl-4-heptanol	191.0	193.6	−2.6
56	4-octanol	176.3	174.2	2.1	125	2,4,6-trimethyl-4-heptanol	181.0	178.1	2.9
57	2-ethyl-1-hexanol	184.6	181.9	2.7	126	3-ethyl-3-octanol	199.0	194.3	4.7
58	2-methyl-2-heptanol	156.0	163.4	−7.4	127	3-ethyl-2-methyl-3-heptanol	193.0	187.0	6.0
59	2,5-dimethyl-1-hexanol	179.5	176.4	3.1	128	1-undecanol	245.0	246.4	−1.4
60	5-methyl-2-heptanol	172.0	173.3	−1.3	129	2-undecanol	228.0	231.7	−3.7
61	6-methyl-3-heptanol	174.0	168.6	5.4	130	3-undecanol	229.0	227.6	1.4
62	3,5-dimethyl-1-hexanol	182.5	179.5	3.0	131	5-undecanol	229.0	222.6	6.4
63	3-methyl-2-heptanol	166.1	171.9	−5.8	132	6-undecanol	228.0	221.9	6.1
64	2-methyl-3-heptanol	167.5	167.1	0.4	133	1-dodecanol	261.9	262.2	−0.3
65	2-methyl-4-heptanol	164.0	166.7	−2.7	134	2-dodecanol	246.0	247.6	−1.6
66	5-methyl-3-heptanol	172.0	169.5	2.5	135	1-tridecanol	276.0	277.4	−1.4
67	3-methyl-3-heptanol	163.0	162.4	0.6	136	1-tetradecanol	289.0	292.0	−3.0
68	4-methyl-3-heptanol	170.0	169.4	0.6	137	1-pentadecanol	304.9	306.1	−1.2
69	3-methyl-4-heptanol	162.0	167.7	−5.7	138	1-hexadecanol	312.0	319.7	−1.0

coefficients | r |≤0.82). The calculated BP values and residuals for 138 compounds are shown in Table 1, and the plot of calculated versus observed BP values shown in Fig. 2 indicate no obvious deviation from linearity.

However, for the same 138 compounds, the simple linear regression with the Lu index leads to a poorer correlation (r=0.9665 and s=9.797°C). Obviously, a single Lu index cannot give a simple and accurate correlation. As discussed in Section 1, the complement of additional atom-type indices seems to be necessary. All atom-type indices of these compounds were employed to model the BPs of 138 alcohols, in turn, and a high correlation model was constructed by combined use of the Lu index. The correlation coefficient and standard error of Eq. (8) can be compared to several results that have been reported. The model found by Yang et al. [29] using the extended adjacency matrix EAΣ and EAmax indices (EAΣ is the sum of the absolute eigenvalues of expanded adjacency matrix EA.) gave r=0.9837 and s=6.35 for 37 alcohols, but the molecular connectivity $^1\chi$ and $^1\chi^v$ indices provided a slightly superior model to those using EAΣ and EAmax indices. Analogously, a multiple linear model constructed by Yao et al. [30] using X_{m1}, X_{m2} and X_{m3} indices gave r=0.987 and s=7.4988 for the same series of 37 alcohols. Galvez et al. [31] reported that the three-parameter model using N (the number of the vertices) and two charge indices G_1 and J_2 gave r=0.979 and s=3.63 for 29 alcohols. Recently, Ren [15] developed a model by using AI and Xu indices and gave r=0.9957 and s=3.576 for the same data.

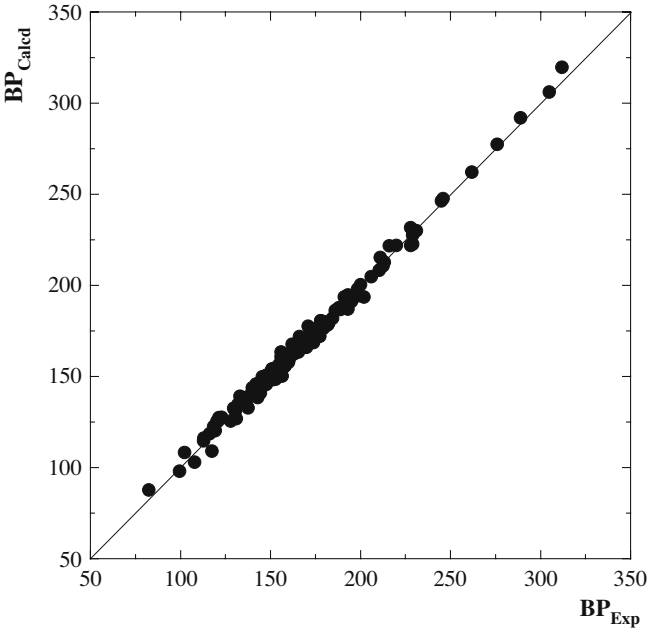

Fig. 2 Plot of observed vs. calculated BP for Eq. (8)

Correlations to water solubility of 63 alcohols (log(1/S))

As an extension of the above study, we selected a dataset of alcohols with their aqueous solubility to develop a structure–property model. Aqueous solubility is a particularly important property of organic compounds and widely applied in the field of pharmaceutical chemistry, biological

chemistry, and environmental science. The experimental water solubilities as log(1/*S*), where *S* is the solubility in moles per liter, are listed in Table 2 for 63 alcohols [32].

A model was developed using *Lu* and two *DAI* indices. The best three-parameter model is given below:

$$\begin{aligned} log(1/S) &= -3.7493 + 0.5196Lu - 0.0447DAI(\mathrm{CH_3}-) \\ &\quad + 0.6052DAI(-\mathrm{OH}) \\ N &= 63,\ r = 0.9876,\ r_{cv} = 0.9852, \\ s &= 0.1604,\ s_{cv} = 0.1760,\ F = 794,\ P < 0.0001 \end{aligned} \tag{9}$$

The *t*-values are −7.78, 42.82, −3.83 and 3.08, respectively. All indices in the model are statistically significant according to the *t*-values at the level of $P<0.0001$. This model produces a standard error of 0.1604 and explains more than 97.5% (r^2) of the variance in the experimental log(1/*S*) values. On the other hand, the model is further validated using the leave-one-out cross-validation. The r_{cv} and s_{cv} are determined to be 0.9852 and 0.1760, which are very close to the statistics of Eq. (9). The cross-validation demonstrates the model to be statistically significant. Cross-correlation analysis shows that the indices in the model are not highly correlated with each other (the pairwise correlation coefficients $|r| \leq 0.53$).

It should be mentioned that a single *Lu* index yields a slightly poorer correlation (r=0.9767 and s=0.2206). Here the results indicated that a single *Lu* index related to molecular size cannot model aqueous solubility satisfactorily. The two-variable regression based on the combined use of *Lu* and *DAI*(−OH) produces an obviously improved model with r=0.9844 and s=0.1807. Here the role of −OH groups seems to be much important to the aqueous solubility of alcohols possibly due to the hydrogen-

Table 2 Calculated and experimental log(1/*S*) for 63 alcohols

		log(1/*S*)					log(1/*S*)		
No.	Compound	Experimental	Calculated	Residuals	No.	Compound	Experimental	Calculated	Residuals
1	ethanol	−1.10	−1.41	0.31	33	5-methyl-2-hexanol	1.38	1.39	−0.01
2	1-propanol	−0.62	−0.69	0.07	34	2-methyl-3-hexanol	1.32	1.21	0.10
3	1-butanol	−0.03	−0.01	−0.01	35	2-methyl-2-hexanol	1.07	1.08	−0.01
4	2-methyl-1-propanol	−0.10	−0.22	0.12	36	2,2-dimethyl-1-pentanol	1.52	1.30	0.21
5	2-butanol	−0.47	−0.31	−0.15	37	4,4-dimethyl-1-pentanol	1.55	1.48	0.06
6	1-pentanol	0.59	0.62	−0.03	38	2,4-dimethyl-1-pentanol	1.60	1.40	0.19
7	3-methyl-butanol	0.51	0.44	0.06	39	3-methyl-3-hexanol	0.98	1.14	−0.16
8	2-pentanol	0.28	0.32	−0.04	40	2,4-dimethyl-2-pentanol	0.93	1.08	−0.15
9	2-methyl-1-butanol	0.46	0.37	0.08	41	2,4-dimethyl-3-pentanol	1.22	1.03	0.18
10	3-pentanol	0.21	0.24	−0.03	42	3-ethyl-3-pentanol	0.83	1.06	−0.23
11	3-methyl-2-butanol	0.18	0.12	0.05	43	2,3-dimethyl-2-pentanol	0.87	1.01	−0.14
12	2-methyl-2-butanol	−0.15	0.04	−0.19	44	2,3-dimethyl-3-pentanol	0.84	0.96	−0.12
13	1-hexanol	1.21	1.24	−0.03	45	2,2-dimethyl-3-pentanol	1.15	1.01	0.13
14	4-methyl-1-pentanol	1.14	1.08	0.05	46	2,2,3-trimethyl-3-butanol	1.27	0.83	0.43
15	2-hexanol	0.87	0.95	−0.08	47	2,3,3-trimethyl-2-butanol	0.71	0.83	−0.12
16	2-methyl-1-pentanol	1.11	0.97	0.13	48	1-octanol	2.35	2.41	−0.06
17	3-hexanol	0.80	0.83	−0.03	49	2-octanol	2.09	2.14	−0.05
18	2-ethyl-1-butanol	1.01	0.89	0.11	50	2-ethyl-1-hexanol	2.11	2.02	0.08
19	4-methyl-pentanol	0.79	0.77	0.01	51	2-methyl-2-heptanol	1.72	1.86	−0.14
20	3,3-dimethyl-1-butanol	0.50	0.85	−0.35	52	3-methyl-3-heptanol	1.60	1.73	−0.13
21	2,3-dimethyl-1-butanol	0.37	0.79	−0.42	53	1-nonanol	3.01	2.97	0.03
22	2-methyl-2-pentanol	0.49	0.66	−0.17	54	7-methyl-1-octanol	2.49	2.84	−0.35
23	3-methyl-2-pentanol	0.71	0.69	0.01	55	2-nonanol	2.74	2.71	0.02
24	2-methyl-3-pentanol	0.70	0.64	0.05	56	3-nonanol	2.66	2.58	0.07
25	2,2-dimethyl-1-butanol	0.91	0.73	0.17	57	4-nonanol	2.59	2.50	0.08
26	3-methyl-3-pentanol	0.36	0.57	−0.21	58	5-nonanol	2.49	2.48	0.00
27	3,3-dimethyl-2-butanol	0.61	0.51	0.09	59	2,6-dimethyl-4-heptanol	2.16	2.16	−0.00
28	2,3-dimethyl-2-butanol	0.37	0.46	−0.09	60	3,5-dimethyl-4-heptanol	2.51	2.04	0.46
29	1-heptanol	1.81	1.83	−0.02	61	2,2-diethyl-1-pentanol	2.42	2.45	−0.03
30	2-heptanol	1.55	1.55	0.00	62	1-decanol	3.63	3.52	0.10
31	3-heptanol	1.44	1.43	0.01	63	1-dodecanol	4.67	4.58	0.08
32	4-heptanol	1.40	1.38	0.02					

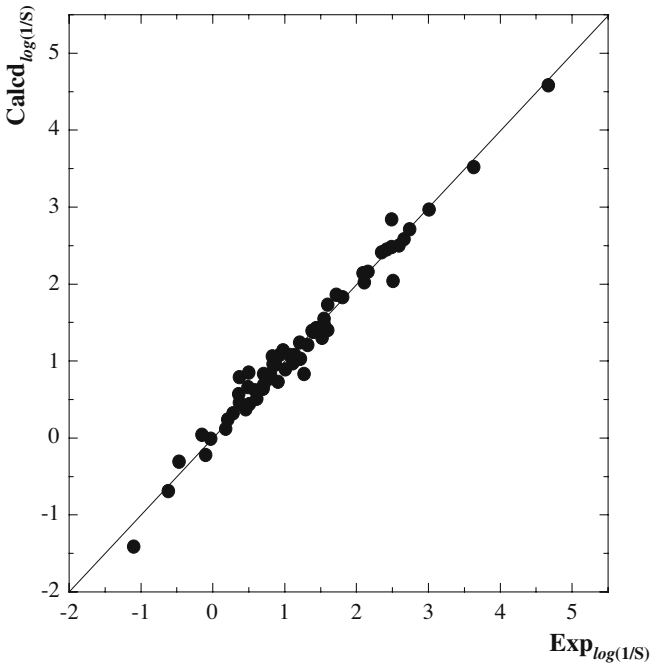

Fig. 3 Plot of observed vs. calculated log(1/*S*) for Eq. (9)

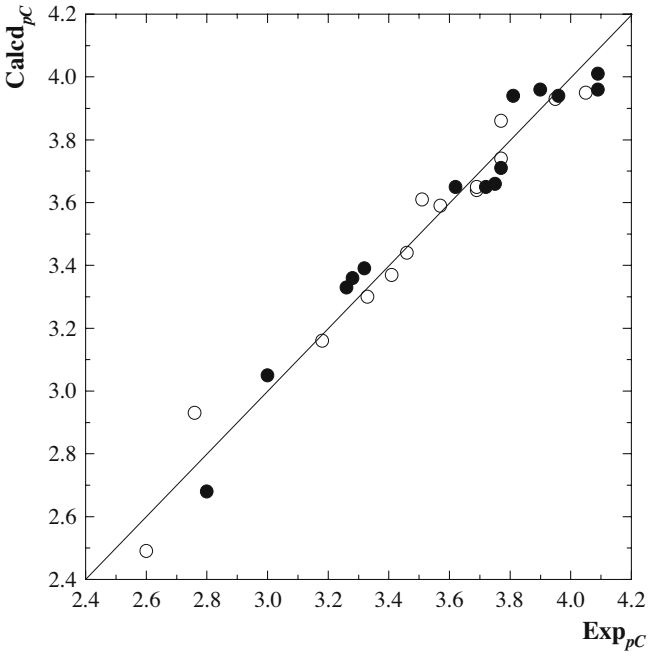

Fig. 4 Plot of observed vs. calculated pC for Eqs. (10) (○) and (11) (•)

bonding interactions. Finally, the best correlation is obtained in terms of up to three indices (Eq. 9). The calculated log(1/*S*) values and residuals for 63 compounds are listed in Table 2, and the plot of calculated versus observed log(1/*S*) values is shown in Fig. 3.

Correlations to biological activities of 14 alcohols

In this section, we will provide other examples of applications of these novel topological indices with the aim of further verifying their applicability to biological activities and toxicities.

Toxicity of organic compounds is one of the particular interesting biological activities in the scientific community due to its impact on environment and human health [23]. The toxicities of 14 alcohols on tomatoes and spiders are taken directly from [33] and shown in Table 3, where the toxicities (pC) are 50% inhibitory growth impairment concentration (−logLC50). First, the model for describing

Table 3 Toxicities (pC) of 14 alcohols on tomatoes and spiders

		pC[a]			pC[b]		
No.	Compound	Experimental	Calculated	Residuals	Experimental	Calculated	Residuals
1	methanol	2.60	2.49	0.11	2.80	2.68	0.12
2	ethanol	2.76	2.93	−0.17	3.00	3.05	−0.05
3	1-propanol	3.33	3.30	0.03	3.32	3.39	−0.07
4	2-propanol	3.18	3.16	0.02	3.26	3.33	−0.07
5	1-butanol	3.69	3.64	0.05	3.77	3.71	0.06
6	2-methyl-1-propanol	3.57	3.59	−0.02	3.72	3.65	0.07
7	2-butanol	3.46	3.44	0.02	3.62	3.65	−0.03
8	2-methyl-2-propanol	3.41	3.37	0.04	3.28	3.36	−0.08
9	1-pentanol	4.05	3.95	0.10	4.09	4.01	0.08
10	3-methyl-1-butanol	3.95	3.93	0.02	4.09	3.96	0.13
11	2-pentanol	3.77	3.74	0.03	3.90	3.96	−0.06
12	2-methyl-1-butanol	3.77	3.86	−0.09	3.96	3.94	0.02
13	3-pentanol	3.69	3.65	0.04	3.81	3.94	−0.13
14	2-methyl-2-butanol	3.51	3.61	−0.10	3.75	3.66	0.09

[a]Toxicities on tomatoes
[b]Toxicities on spiders

toxicity on tomatoes is generated. The best two-parameter model is given below.

$$pC = 0.2448 + 1.0360Lu + 0.2280DAI(-\mathrm{OH})$$
$$N = 14,\ r = 0.9827,\ r_{cv} = 0.9633,\ s = 0.0766, \quad (10)$$
$$s_{cv} = 0.1109,\ F = 160,\ P < 0.0001$$

The *t*-values are 0.44, 13.41 and 3.78, respectively. All indices in the model are statistically significant according to the *t*-values at the level of $P<0.0001$. The model can explain 96.6% of the variance in the experimental pC values for 14 alcohols. The correlation coefficient and standard error of the leave-one-out cross-validation procedure are 0.9633 and 0.1109. The cross-validation results demonstrate the model to be statistically significant. The correlation coefficient between Lu index and $DAI(-\mathrm{OH})$ index is 0.46 and shows the two variables to be relatively uncorrelated.For toxicity on spider, we obtain the following two- parameter models.

$$pC = 2.4851 + 0.2571Lu - 0.1392DAI(> \mathrm{C} <)$$
$$N = 14,\ r = 0.9771,\ r_{cv} = 0.9633,\ s = 0.0856, \quad (11)$$
$$s_{cv} = 0.1109,\ F = 122,\ P < 0.0001$$

The *t*-values are 31.48, 15.18 and −3.18, respectively. Each coefficient is also highly significant. This model explains more than 95.5% of the variance in the experimental values of pC for 14 alcohols with a standard error of 0.0856. The calculated pC values and residuals are shown in Table 3. A comparison of calculated and observed toxicities is shown in Fig. 4.

Conclusion

The atom-type topological index *DAI* based on the distance matrix of the molecular graph can describe the different structural environments of each atom-type in a molecule at the atomic level. Multiple linear regression using Lu and *DAI* indices can provide high-quality QSPR/QSAR models for properties and activities of alcohols. The final models were shown to be statistically significant and reliable by the leave-one-out cross-validation procedure. The results indicate that the combined use of Lu and *DAI* indices is useful and feasible for QSPR/QSAR analysis of complex compounds.

References

1. Randiá M (1975) J Am Chem Soc 97:6609–6615
2. Hosoya H (1971) Bull Chem Soc Jpn 44:2332–2339
3. Balaban AT (1995) Chem Phys Lett 89:399–404
4. Bonchev D, Trinajstic N (1977) J Chem Phys 67:4517–4533
5. Wiener H (1947) J Am Chem Soc 69:17–20
6. Ren B (1999) J Chem Inf Comput Sci 39:139–143
7. Lu C, Guo W, Hu X, Wang Y, Yin C (2005) J Math Chem (accepted)
8. Balaban AT, Bonchev D, Seitz WA (1993) J Mol Struct (THEOCHEM) 280:253–260
9. Balaban AT, Ivanciuc O (1989) FORTRAN-77 Computer program for calculating topological index *J* for molecules containing heteroatoms. In: Graovac A (ed) MATH/CHEM/COMP 1988 Studies in Physical and Theoretical Chemistry, No. 63. Elsevier, Amsterdam, pp 193–211
10. Ivanciuc O, Ivanciuc T, Balaban AT (1998) J Chem Inf Comput Sci 38:395–401
11. Balaban AT (1986) MATCH Commun Math Comput Chem 21:115–122
12. Hall LH, Mohney B, Kier LB (1991) J Chem Inf Comput Sci 31:76–82
13. Ren B (2002) Comput Chem 26:223–235
14. Ren B (2002) J Mol Struct (THEOCHEM) 586:137–148
15. Ren B (2002) J Chem Inf Comput Sci 42:858–868
16. Ren B (2003) J Chem Inf Comput Sci 43:161–169
17. Ren B (2003) J Chem Inf Comput Sci 43:1121–1131
18. Yang P, Gao XH (1987) Chemical bonding and structure–property relation (in Chinese). Higher Education, Being, China
19. Maw HH, Hall LH (2001) J Chem Inf Comput Sci 41:1248–1254
20. Rose K, Hall LH, Kier LB (2002) J Chem Inf Comput Sci 42:651–666
21. Ren B (2003) Chemometr Intell Lab 66:29–39
22. Xu L (1996) Chemometrical method (in Chinese). Scientific Press of China, Beijing, China
23. Ren B (2003) J Comput-Aided Mol Des 17:607–620
24. Katritzky AR, Maran U, Lobanov VS, Karelson M (2000) J Chem Inf Comput Sci 40:1–18
25. Lide D (2003–2004) CRC Handbook of Chemistry and Physics, 84th edn. CRC, Boca Raton, Florida
26. Lide DR, Milne GWA (1992) Handbook of data on common organic Compounds. CRC, Boca Raton, Florida
27. Dean JA (1999) Lange's handbook of chemistry, 15th edn. McGraw–Hill Book Company
28. Yaws CL (1999) Chemical properties handbook. McGraw-Hill, Beijing, China
29. Yang YQ, Xu L, Hu, CY (1994) J Chem Inf Comput Sci 34:1140–1145
30. Yao YY, Xu L, Yang, YQ, Yuan, YS (1993) J Chem Inf Compt Sci 33:590–594
31. Galvez J, Garcia R, Salabert MT, Soler R (1994) J Chem Inf Compt Sci 34:520–525
32. Nelson TM, Jurs PC (1994) J Chem Inf Compt Sci 34:601–609
33. Kier LB, Hall LH (1986) Molecular connectivity in structure–activity studies. Research Studies, Letchworth