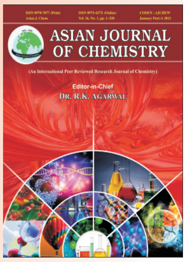
Asian Journal of Chemistry; Vol. 26, No. 12 (2014), 3507-3510

ASIAN JOURNAL OF CHEMISTRY

http://dx.doi.org/10.14233/ajchem.2014.16126



Studies on Quantitative Structure-Toxicity Relationship of Alcohols, Phenols, Ethers, Ketones and Esters

Zuoping Lan[1], Limin Liao[2], Yu Yu[1,*] and Zhaojing Zhu[1]

[1]College of Pharmacy, Chongqing Medical University, Chongqing 400016, P.R. China
[2]College of Resource and Environment Science, Neijiang Normal University, Sichuan 641300, P.R. China

*Corresponding author: Fax: +86 23 67195346; E-mail:yuyu3915@163.com

Received: 31 July 2013; *Accepted*: 8 October 2013; *Published online*: 5 June 2014; AJC-15283

In order to find out the quantitative relationship between toxicity and structures of alcohols, phenols, ethers, ketones, esters, some structures of compounds were characterized by the values of molecular vertexes and their interaction. Two models of the quantitative structure-toxicity relationship (QSTR) were established by the methods multiple linear regression (MLR) and stepwise regression (SMR). A comparison of two models indicated that the model 2 (M2) showed better simulation results and the multiple correlation coefficient (R) was 0.952 and the value of standard deviation (SD) was 0.325. Jacknife method used for testing its stability indicated that the regression model 2 had an acceptable stability and a good predictive ability. In addition, the model was tested by the cross-validation with the leave-one-out (LOO) procedure. And the multiple correlation coefficient in cross-validation (RCV) was 0.927 and the value of standard deviation (SDCV) was 0.396, which showed the stability and predictive ability of the model were desirable.

Keywords: Toxicity, Molecular vertex, Quantitative structure-toxicity relationship.

INTRODUCTION

In the industrial and agricultural production, alcohols, phenols, ethers, ketones, esters (APEKE) are widely used as solvents, additives, pesticides and refrigerants but they are harmful not only to the workers contacting them directly but also to the environment. The health of other people contacting them indirectly, as well as the growth of plants or animals, is endangered. With the rapid development of chemical industry and people's living standard, a large number of synthetic APEKE have entered into the environment and become a major environmental pollutant. Therefore, studying on the environmental behavior of the APEKE is of great significance, and now the evaluation on environmental risk of chemical substances has been focused. Studying on quantitative structure-toxicity relationship (QSTR) in various organic pollutants[1-3], the researchers have established mathematical models with the function of predictive toxicity and the model has been successfully implemented for organic toxicity prediction and evaluation[4]. In the paper, 37 APEKE (Table-1) were selected for establishing the model on study the quantitative relationship between toxicities and structures of the compounds. When the researchers characterizing the structures of the compounds, the non-hydrogen atoms (framework atoms) were looked as the vertexes of molecules (ignored the impacts of non-framework hydrogen atoms) and the molecular structures were characterized by the hybridization state of vertex atoms and the interactions between them (this method is called as "values of molecular vertexes and their interaction" in this paper). The quantitative toxicities of APEKE were studied relatively by the methods of multiple linear regression (MLR) and stepwise regression. The researchers found that there was a good linear relationship between the toxicities (log1/C) and the molecular structures of APEKE.

EXPERIMENTAL

Generally, the framework atoms (atoms of non-hydrogen) in a molecule are closely related to the main properties of the organic compounds, while the impact of the non- framework atoms (atoms of hydrogen) are ignored. In the special molecular whose atoms of hydrogen are ignored, each framework atom is regarded as the molecular vertex and both its state of the molecular vertex and the interaction between the vertices of the organic molecule influence the properties of the organic compounds greatly. The vertexes in different connection states and the interaction between them may influence the properties of molecular differently, while the vertexes in similar connection states and the interaction between them have similar effects on the properties of molecular and these similar effects can be assembled. In order to construct the model for studying values

TABLE-1
37 APEKE COMPOUNDS AND THEIR log1/C

No	Compounds	x_2	x_3	x_6	x_8	x_{10}	log1/C (Exp)	log1/C (Cal)
1	Methanol	0.0000	0.0000	0.0000	0.0000	0.0000	0.24	-0.174
2	Ethanol	1.2500	0.0000	2.5644	0.0000	0.0000	0.54	0.793
3	Acetone	0.0000	1.6667	0.0000	0.0000	0.0000	0.54	0.718
4	Carbamate	0.0000	1.6667	2.6325	0.0000	1.5644	0.57	1.232
5	Isopropanol	2.5000	0.0000	3.1771	0.0000	0.0000	0.89	1.250
6	*tert*-Butyl alcohol	0.0000	0.0000	0.0000	4.5644	0.0000	0.89	0.787
7	Propanol	2.5000	0.0000	3.1771	0.0000	0.0000	0.96	1.250
8	Butanone	1.2500	1.6667	1.6700	0.0000	1.0000	1.04	1.341
9	Methyl acetate	0.0000	1.6667	2.5417	0.0000	1.5644	1.10	1.208
10	Ethyl formate	2.9167	0.0000	4.4414	0.0000	0.0000	1.15	1.679
11	*tert*-Amyl alcohol	1.2500	0.0000	1.8627	3.8144	0.0000	1.29	1.413
12	Isobutanol	1.2500	1.2500	1.6127	0.0000	1.0000	1.35	1.103
13	Urethane	1.2500	1.6667	2.7738	0.0000	1.8543	1.39	1.534
14	Butanol	3.7500	0.0000	3.4455	0.0000	0.0000	1.42	1.618
15	Ethyl acetate	1.2500	1.6667	2.6546	0.0000	1.8543	1.52	1.503
16	3-Pentanone	2.5000	1.6667	3.0622	0.0000	2.0000	1.54	1.891
17	Ether	2.5000	0.0000	2.9704	0.0000	0.0000	1.57	1.196
18	Isoamyl alcohol	2.5000	1.2500	2.6493	0.0000	1.2500	1.64	1.643
19	2-Pentanone	2.5000	1.6667	2.2041	0.0000	1.2500	1.72	1.750
20	1,3-Dichloro-2-propanol	2.5000	1.2500	2.1053	0.0000	1.2500	1.92	1.501
21	Ethyl propionate	2.5000	1.6667	3.8811	0.0000	2.8543	1.96	2.010
22	Propyl Acetate	2.5000	1.6667	2.8677	0.0000	1.9768	1.96	1.843
23	Acetal	2.5000	1.2500	3.9740	0.0000	3.7087	1.98	1.716
24	Ethyl isobutyrate	0.0000	2.9167	2.2417	0.0000	1.9271	2.24	1.759
25	Isobutyl acetate	1.2500	2.9167	2.1065	0.0000	3.2170	2.24	1.878
26	Butyl acetate	3.7500	1.6667	3.0137	0.0000	2.0441	2.3	2.171
27	Ethyl butyrate	2.5000	1.6667	3.8153	0.0000	2.8144	2.37	1.997
28	Aethylis valerianas	5.0000	1.6667	4.4408	0.0000	3.2155	2.72	2.711
29	Amyl acetate	5.0000	1.6667	3.1158	0.0000	2.0864	2.72	2.491
30	Trimethoxyphenol	8.3335	1.6667	2.0241	0.0000	4.7702	2.82	2.702
31	Acetophenone	8.3335	3.3334	1.1127	0.0000	4.1398	3.03	3.426
32	1,4-Dimethoxybenzene	6.6668	3.3334	4.0417	0.0000	9.4716	3.05	3.202
33	Phenylcarbamate	9.5835	3.3334	1.8763	0.0000	6.6886	3.19	3.640
34	1,3-Dimethoxybenzene	6.6668	3.3334	4.0272	0.0000	9.2876	3.35	3.219
35	Octanol	8.7500	0.0000	3.8056	0.0000	0.0000	3.40	2.903
36	Butyl valerate	7.5000	1.6667	4.6557	0.0000	3.4052	3.60	3.341
37	2-Methyl-6-isopropyl phenol	5.0001	6.2501	1.4050	0.0000	4.5686	4.26	4.223

of molecular vertexes and their interaction, all vertexes need to be classified based on their respective link style in the molecular. These vertexes are classified as four atomic types (A_1, A_2, A_3 and A_4) in the paper according to the number of each vertex linked in the chemical bond/bonds. For example, if a vertex is linked with 2 vertexes in the chemical bonds, the atomic type belongs to A_2.

The values of molecular vertexes could be obtained through amending the calculation method of the atomic natural state (I) proposed by Hall and Kier[5] and the values were used for characterizing the impact of the properties of molecular in vertexes' own states. The following is the formula:

$$x_r = \sum_{i\in n}\sqrt{\nu_1/4}\cdot\left((2/n_i)^2\delta^i\sigma+\pi+1\right)/\delta^i\sigma \; (r = 1, 2, 3, 4) \quad (1)$$

In the formula, x_r is the value of all vertexes (i) whose atomic type belong to r; ν_i represents the electron numbers in electronic shell of atom of vertex (i); n_i is the main quantum numbers of vertex i; $\delta^i_{\sigma+\pi}$ represents the sum of electrons in bonds σ and π of the vertex; $d^i\sigma$ is the electron numbers in bonds σ of the vertex. For example, the values of a single vertex C_{SP}, C_{SP}^2 or C_{SP}^3 of the molecule are respectively 2.5000, 1.6667 or 1.2500. According to the classification of atom of vertex, one molecule consists of four types of vertexes at most so that four vertex values (shortened as x_1, x_2, x_3 and x_4) will eventually be obtained in each molecule.

The interaction between the vertexes is closely related to the electro-negativities of vertexes and the relative distance between vertexes. In general, the interaction increases with the increment of the electro negativities and the decrement of the relative distance. In this report, we learn from the literature[6-8] to characterize the impact of the properties of molecular influenced by the vertices' interaction. The interaction from four types of vertexes can be grouped in ten elements, M_{11}, M_{12}, M_{13}, M_{14}, M_{22}, M_{23}, M_{24}, M_{33}, M_{34} and M_{44}, shortened as x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13} and x_{14}. The formula was defined as followings:

$$x_r = m_{nl} = \sum_{i=n,j=1}\frac{Z_i.Z_j}{r_{ij}^2}$$

$$(n = 1, 2, 3, 4;\ n \le l \le 4;\ r = 5,6,....,14) \quad (2)$$

In the formula, n or *l* represents the types of vertexes i and j, i or j is an atom of vertexe in the molecular; Z_i and Z_j are the electro-negativities of atoms i and j relative to atom C (For example, the atoms of oxygen relative electro-negativity is 3.44/2.55 = 1.3490); r_{ij} represents the relative distance between the ith vertex and the jth vertex (*viz.* the proportion of sum of the experienced shortest path length relative to the C-C single bond length). Based on the above principle, there are 14 variables used for describing the structural information in each organic molecule compound. (In this study, x_{13} and x_{14} are described as zero for all the samples and the remaining 12 non-zero variables are applied to establishing models and next analysis.).

RESULTS AND DISCUSSION

In this work, 37 samples of APEKE selected. The level of toxicity is indicated by anesthetic activity (log1/C) of tadpoles. All the experimental values are taken from the literature[9]. All log1/C values of compounds are sorted from value size, which are shown in Table-1. Multiple linear regression (MLR), a classic method for models, is applied to linear fitting on independent variable and dependent variable. And then the least squares (LS) is applied to the results of linear fitting in order to obtain the best results models. Firstly, SPSS13.0 of the multiple linear regression (MLR) method was used for studying the relation between the structures and log1/C. At the same time, the model was evaluated by cross-validation with the leave-one-out (LOO) procedure and then the predictable model (M1) with 12 variables was obtained as followings:

$$\log 1/C = -0.456 - 0.044 \times x_1 - 0.338 \times x_2 + 1.017 \times x_3 + 0.291 \times x_4 + 0.562 \times x_5 + 0.425 \times x_6 + 0.119 \times x_7 + 0.075 \times x_8 + 0.481 \times x_9 - 0.325 \times x_{10} + 0.069 \times x_{11} - 1.014 \times x_{12} \quad (3)$$

N = 37, R = 0.966, SD = 0.309, F = 28.238; R_{CV} = 0.809, SD_{CV} = 0.707, F_{CV} = 3.774

N represents the number of samples, R represents multiple correlation coefficient, SD represents standard deviation, F represents the value of Fischer test; R_{CV} represents a cross-validated correlation coefficient, SD_{CV} represents the standard deviation of cross-validation, F_{CV} represents a Fischer test value of cross-validation.

The multiple correlation coefficient (R) of the model attained to 0.966, but the cross-validated correlation coefficient (R_{CV}) was only 0.809, that's to say, the results were quite different between R and R_{CV}. It showed the model with poor prediction as well as uncertainty. In addition, the standard deviation of the cross-validation (SD_{CV}) was also not ideal, which indicated considerable error to the prediction of the model. Usually a good model is consistent with the experience principles that should be "the number of samples / number of variables $\geqq$ 5". The number of variables was 12 in this model, however, with 37 samples which indicated that the number of variables were too much and over-linear-fitting might be happened. SPSS13. 0 was used for testing model (M1) and then t-statistics and variance inflation factor (VIF) of 12 variables were calculated. VIF was defined as: VIF $= (1-r^2)^{-1}$, r represents the degree of correlation of an independent variable with other variables (by correcting degree of freedom). If VIF is 1.0, which means no correlation between different variables; if VIF ranges from 1.0 to 10.0, which means no significant collinearity between variables so that the equation is acceptable; if VIF is larger than 10, which means that the equation is not acceptable. The analysis showed the model (M1) did have a certain degree of multi-collinearity (maximum VIF was 168.568), and not all variables had shown significant features (the values of part t-variables ranged in $-2 \leq t \leq 2$).

In order to further examine the impact of variables to the model and eliminate the over-fitting phenomenon in the model, then the researchers tried to find a better model and conducted a stepwise regression (SMR) analysis on variables. SMR, a classical variable selection method for linear models, is used for testing the significance levels of variables in turn so that the orders of variable introduction or removal are determined. This method reflects the principle of "in or out" orderly. In the research, introducing significant variables by SMR in turn and employing RCV as the objective function by cross-validated, a 5-parameters regressive model (M2) was obtained:

$$\log 1/C = -0.174 + 0.238 \times x_2 + 0.535 \times x_3 + 0.261 \times x_6 + 0.211 \times x_8 - 0.111 \times x_{10} \quad (4)$$

N = 37, R = 0.952, SD = 0.325, F = 59.369; R_{CV} = 0.927, SD_{CV} = 0.396, F_{CV} = 37.968

The model (M2) was effective and better than M1 and it conformed to the "samples/number of variables $\geqq$ 5" rule. Compared with M1, the multiple correlation coefficient (R) and the cross-validation multiple correlation coefficient (R_{CV}) of M2 was a little lower, the standard deviation (SD) was a little higher, but the cross-validation standard deviation (SD_{CV}) significantly decreased. Considering the multiple correlation coefficients (R and R_{CV}) and standard deviation (SD and SD_{CV}), the model (M2) was better than M1. In addition, variable numbers have decreased from 12 to 5, which lightened the complexity of the model significantly. Regression test again on the M2 showed that t-values, all the absolute values of variables, were larger than 2 (the minimum t was 2.540), while the VIF value were significantly lower (maximum VIF was 3.719), which confirmed the model with a high quality. In order to prove its good stability and reliability, robustness test on model (M2) was carried out by Jackknife[10] method. Each time, we removed some compounds whose single digits of molecules were 0, 1, 2, ... and 9 from 37 compounds, then the remaining compounds used as modeling group and then regression analysis was conducted based on eqn (4). After repeating them 10 times, the multiple correlation coefficients (R) of the models were shown by radar graph (Fig. 1) in which graph scale spacing was 0.02. Fig. 1 showed all multiple

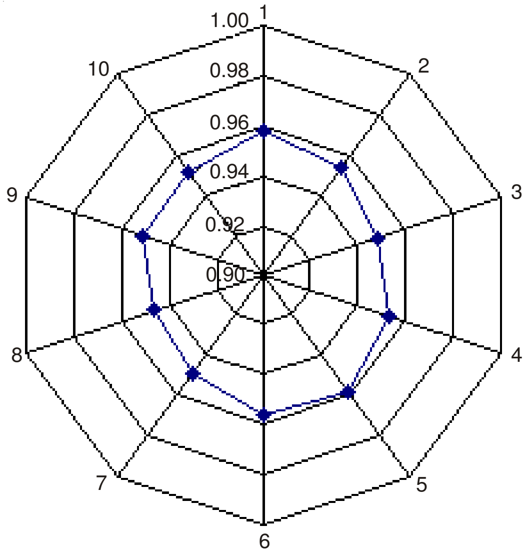


Fig. 1. Radar graph of R of M2 (10 times)

correlation coefficients (R) fell back into the range of 0.94-0.96, so the linear regression model had a good robustness. Meanwhile, the model (M2) evaluated by the cross validation with the leave-one-out (LOO) procedure, the correlation coefficients (R_{cv}) was 0.927, which was a bit lower than R value (0.952) of the original model. Additionally, the value of the cross-validation standard deviation (SD_{CV}) of the model was 0.396, which was a bit larger than the original value SD (0.325). All proved the stability and the predictability of the model (M2) were excellent.

In the equation (4), x_2 and x_3 represents their state values of the second and third type vertex atoms in compounds, x_6 represents the interaction value between the first and second type vertex, so the second and third type vertexes have important influence on toxicity of the compounds in this sample set. The toxicities of 37 compounds (log1/C) predicted by the model (M2) are listed in Table-1 (Cal.). To observing the results of model fitting easily, the above results are shown in Fig. 2. It shows that most of the sample points are near to or on diagonal line of the square in Fig. 2, as shows the calculated values and experimented values are almost similar and the models constructed in this work have good estimation of stability and qualify for favorable prediction. However, the prediction results of some samples in the model (M2) are inspection errors and the reasons may be due to their own particularity of the molecular structure. Moreover, experimental data itself may be a certain degree of errors and lead to the error results. And to some degree, the method of molecular structure characterization may be not perfect because this method is based on two-dimensional molecular structure, while the actual molecular structure is three-dimensional. We should take these problems into account in the future researches.

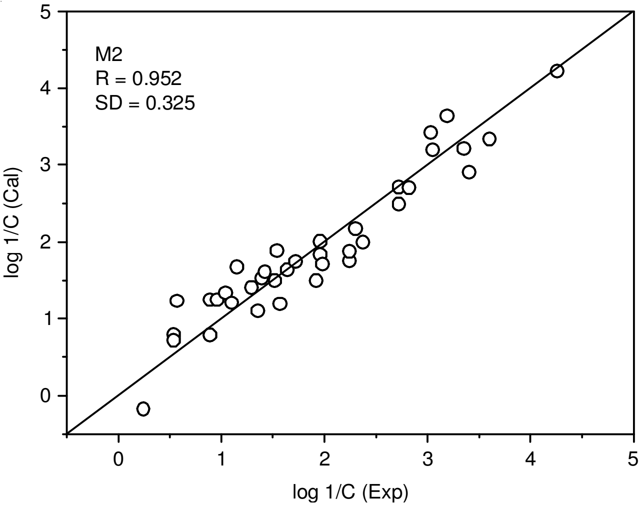


Fig. 2. Plot of calculated values *vs.* experimented values (log 1/C)

Conclusion

In the paper, the toxicity (log 1/C) of 37 APEKE were studied by practice the theory of "values of molecular vertexes and their interaction" and the researchers got very satisfied results as expected. All the parameters of molecular structure were derived from the molecules themselves. Compared with the popular three-dimensional molecular modeling method[11,12], our method has the advantage of simpler calculation. Model (M2) constructed in this study was tested by the interactive validation and Jackknife method. The results showed that it was acceptable with overall robustness and good predictive ability. This model was proposed to simulate the toxicity (log l/C) of APEKE organic compounds and the results were in little error as far as experimental results. Model (M2) has a certain predictive ability,especially to the toxicity (log l/C) of APEKE organic compounds. The prediction values by model are of great reference value in the case of lacking experimental data. At the same time, the results obtained in this paper have considerable reference value for the study of quantitative structure-toxicity relationship of organic pollutants.

ACKNOWLEDGEMENTS

The authors gratefully acknowledged the financial support of National Natural Science Foundation of China (No. 81172097).

REFERENCES

1. W.H. Li, X. Xu, Z.G. Xu and Y.F. Luo, *J. Math. Med.*, **16**, 168 (2001).
2. B. Wang, J.S. Zhao, Y.J. Yu, X.D. Wang and L.S. Wang, *Environ. Sci.*, **25**, 89 (2004).
3. J.J. Hu, C.L. Lu, H. Zhou, Y.Z. Zhang and Y.D. Jian, *Environ. Chem.*, **29**, 48 (2010).
4. S P. Bradbury, *Environ. Res.*, **2**, 89 (1994).
5. L.H. Hall and L.B. Kier, *J. Chem. Inf. Comput. Sci.*, **35**, 1039 (1995).
6. L.L. Sun, L.P. Zhou, Y. Yu, Y.K. Lan and Z.L. Li, *Chemosphere*, **66**, 1039 (2007).
7. L. Liao, H. Mei, J. Li and Z. Li, *J. Mol. Struct. Theochem.*, **850**, 1 (2008).
8. S.S. Liu, C.S. Yin, S.X. Cai and Z.L. Li, *Chemom. Intell. Lab. Syst.*, **61**, 3 (2002).
9. L.S. Wang and Z.L. Li, Molecular Connectivity and Molecular Structure-Activity, China Environmental Science Press, Beijing, p. 250 (1992).
10. W.S. Dietrich, N.D. Dreyer, C. Hansch and D.L. Bentley, *J. Med. Chem.*, **23**, 1201 (1980).
11. A.A. San Juan, *J. Mol. Graph. Model.*, **26**, 482 (2007).
12. J. Yoo, K.M. Thai, D.K. Kim, J.Y. Lee and H.J. Park, *Bioorg. Med. Chem. Lett.*, **17**, 4271 (2007).