Correlative Study On Chromatographic Retention Indices Of Polycyclic Aromatic Sulfur Heterocycles

Hui Zhang[1,a], Yan Chen[1,b,*] Jing Li[1,c]

[1]School of Chemistry and Chemical Engineering, Xuzhou Institute of Technology, Xuzhou, Jiangsu, 221111, China

[a] 653597093@qq.com, [b]chenyan681110@126.com, [c]lijingxz111@163.com

Keywords: Polycyclic aromatic sulfur heterocycle; Chromatographic retention indices, the electrical topology state index; Neural network; Quantitative structure -retention relationship

Abstract. Based on the topological theory and MATLAB program, electrotopological state indexes (E_n) were calculated for 114 polycyclic aromatic sulfur heterocycles. A six-element regression model of quantitative structure-retention relationship (QSRR) for retention index (*RI*) as a function of E_n was constructed using leaps-and-bounds regression (LBR). The traditional correlation coefficient (R)，determination coefficient(R^2) and the cross-validation correlation coefficient (Q^2) were 0.995, 0.989 and 0.982 respectively. The model is highly reliable and has good predictive ability. The six structural parameters were used as the input neurons of artificial neural network, and a 6:8:1 network architecture was employed． A satisfied model was constructed with the back-propagation algorithm，the correlation coefficient (R^2) was 0.996. It can be concluded that the prediction results of BP-ANN model are better than MLR-QSRR model.

Introduction

Polycyclic aromatic sulfur compounds heterocycles (PASHs) are the main existing forms of sulfur in crude oil [1]，the existence of organic sulfur can greatly reduce the quality of petroleum products, and it will cause air pollution in the process of combustion. Some studies have indicated that PASHs compounds have a greater mutagenic and carcinogenic properties than those of polycyclic aromatic hydrocarbons.

Chromatographic retention index（*RI*）is an important parameter to study the environmental behavior of organic compounds. However, due to the PASHs exist a lot of isomers and the lack of PASHs standard, it is very difficult to test the gas chromatographic retention index of all possible PASHs.

In recent years, the quantitative structure-property/activity/retention relationship QSPR/QSAR /QSRR method was widely used in the prediction of physical and chemical properties, bioactivity and chromatographic properties of organic pollutants[2-5]. Among them, QSRR research has become a simple and effective method for the chromatographic research field, the results of the research work have also been fully recognized. On the basis of previous work[6-7], electrotopological status indexes *(E_n)* of 114 kinds of PASHs were calculated using the MATLAB programs[8-9],and by optimal subset regression method the six best variables were screened to establish the QSRR prediction model. According to the model to forecast the gas chromatographic retention index of PASHs, we found the predicted values were consistent with the experimental values. Combined with the artificial neural network (ANN)，the accuracy of prediction was further improved.

Materials and methods

Chomatographic retention index of PASHs. In this paper, 114 kinds of PASHs were selected as research objects. The experimental values of their Chromatographic retention index (*RI*) were obtained from the literature[10].

The Calculation of Electrical Topology Status Index. Electrical topological index was presented by the Kier and Hall[11], which is a descriptor reflecting some characters of molecular structure constructed by the topological environment and bonding-electron information of every non-hydrogenatom in molecule hydrogen-suppressed graph, together with some particular mathematical disposal.

Intrinsic value of a non-hydrogen atom is based on the Kier-Hall electronegativity and derived from the ratio of that electronegativity to the number of skeletal sigma bonds for that atom. The intrinsic value (I) of a non-hydrogen atom is defined as follows:

$$I_n = \frac{4\delta_n^v + N_n^2}{N_n^2 \delta_n} \quad (1)$$

where N_n is the principal quantum number for the valence shell of atom I, and δ_n is the number of connections(edges) in the skeleton (graph):

$$\delta_n = \sigma_n - h_n \quad (2)$$

in which σ_n is the number of electrons in the σ orbital, h_n is the number of hydrogen atoms bonded to atom i. δ_n^v in formula (1) Previously applied the definition in Kier and others' connectivity index，but Kier's δ_n^v can not clearly distinguish partial topological environment of atom n, Hall, together with others, revised δ_n^v in 1995 and gave the definition formula as below:

$$\delta_n^v = \sigma_n + \pi_n + k_n - h_n \quad (3)$$

Here π_n and k_n are the number of electrons in π orbital and in lone pairs, respectively.

The reciprocity between topology environment of the bonding atom n in a molecule and other atoms is expressed by the increment of inherent state value (ΔI). ΔI of atom n is defined as:

$$\Delta I_n = \sum_{j \neq n} \frac{I_n - I_j}{(r_{nj})^2} \quad \text{sum over all atoms } j \neq n \quad (4)$$

Here r_{nj} is the shortest route number between atom n and j in the molecule plus one; I_n is the n intrinsic value of atom j; Σ means to summarize all the other non-hydrogen atoms except for atom n in the molecule.

The electrotopological index (E_n) of atom n is defined as the sum of intrinsic value and ΔI_n of the atom n:

$$E_n = \sum_t (I_n + \Delta I_n)_t \quad (5)$$

The 9 atom types referred by the 114 kinds of compound molecules in this paper are listed in Table 1.

Table 1. Intrinsic Values of Atomic Types of Organic Molecules in This Work

No.	Structural formula	Symbol	I	E-index
1	-CH_3-	sCH_3	2.0000	E_1
2	-CH_2-	ssCH_2	1.5000	E_2
3	＞CH-	sssCH	1.3333	E_3
4	-CH=	sdCH	2.0000	E_6
5	-CH=	aaCH	2.0000	E_7
6	＞CH=	ssdC	1.6667	E_8
7	＞CH=	aaCa	1.6667	E_9
8	-S-	**ssS**	1.8333	E_{31}
9	-S-	aSa	1.8333	E_{32}

Here "s" means a single bonding, "d" a double bonding, "a" one bond in an aromatic ring

Analysis of Statistical Regression. Using the E_n of the 9 electrical state indices of molecules of PASHs as independent variables, and the chromatographic retention index as dependent variables. We had chosen the best variables of relativity with Chromatographic properties by MINITAB 14 software, and then built up the mathematical model between these indexes and RI. We had also used leave-one-out (LOO) to check all the models for the robustness and the prediction ability.

Table 2. The Results of En and RI with Regression of the Best Subset

No.	*R*	R^2	*S*	*F*	*FIT*	Variables
1	0.845	0.714	40.042	279.095	2.431	E_9
2	0.939	0.882	25.816	414.919	7.031	E_8E_9
3	0.986	0.971	12.765	1246.192	29.944	$E_7E_8E_9$
4	0.989	0.979	10.975	1274.275	39.088	$E_7E_8E_9E_2$
5	0.991	0.983	10.043	1221.693	41.633	$E_7E_8E_9E_2E_1$
6	0.995	0.989	7.888	1661.658	66.582	$E_7E_8E_9E_2E_1E_{32}$
7	0.995	0.990	7.339	1648.142	64.380	$E_7E_8E_9E_2E_1E_{32}E_{31}$

In Table 2, *R* is the correlation coefficient, *F* is the test value of Fischer, *S* is standard error, *FIT* is the function of Kubinyi [12-13], that the calculation formula is：

$$\mathrm{FIT} = \frac{R^2(y-b-1)}{(y+b^2)(1-R^2)} \tag{6}$$

In the formula, *y* is the sample size of the compounds, *b* is the number of variables. The bigger is *FIT*, the more stable is the model, and the better is the ability of prediction. When we chose the six variables of E_7, E_8, E_9, E_2, E_1 and E_{32}, we could obtain the best results by the data in Table 2.

Results and Discussion

Multiple linear regression model. Table 2 shows that the six-variable model has good stability and predictive ability, so the six-variable model was applied in this paper:

$RI = 71.776 + 12.722E_1 + 20.058\ E_2 + 13.842\ E_7 + 35.879\ E_8 + 12.191\ E_9 - 42.109\ E_{32}$ (7)

$n = 114$, $R = 0.995$, $R^2 = 0.989$, $S = 7.888$, $F = 1661.658$

Using QSRR Eq.(7), we could theoretically predict the *RI* values, The relative average error was 1.93%. Inspection of the model analysis used the software MINITAB by leave-one-out(LOO). The R_{cv}^2 values of QSRR model ($R_{cv}^2 = 0.982$) obtained are higher than 0.5, indicating high robustness of the models.

Mathematical Models of Neural Networks. In recent years, artificial neural networks in many areas had extensive researches and applications [14], based on its excellent nonlinear function approximation ability, which can achieve a high degree of match between the mapping of input and output in line with targets relationships. The merit-based selection of the six parameters E_1, E_2, E_7, E_8, E_9 and E_{32} in the electrical state indices as the input layer BP neural network, the chromatographic retention index（*RI*）as variables of output layer, and number of hidden layer unit in accordance with the proposed rule of Xu. et al [15] .

$2.2 > \rho\ (= N/M) \geq 1.4$ (8)

where *N*, and *M* are the total number of samples and the total weight of the network respectively. *M* value is calculated by the following formula

$M = (I + 1)H + (H + 1)Q$ (9)

In the formula, *I*, *H*, and *Q* are the number of input layer neurons, hidden layer and an output layer respectively. Here we took the hidden layer *H*=8, the 114 sample data into a training set (*n* = number 2,4, and 5 in the data respectively), the test set (*n* = 1 in the data respectively) and validation set (*n* =

3in the data respectively), the correlation coefficient of the models were obtained, in them, $r = 0.998$ of training set, $r = 0.998$ of the test set, $r = 0.998$ of the test set, $r = 0.998$ of the total correlation coefficient. The model is stable. Goodness of fit of the predicted values and experimental values isvery ideal, and relative average error is 0.85%. The neural network and multiple regression method were compared, the results showed that prediction model of using neural network method established better forecast ability than multiple regression method.

Summary

(1) Electrotopological state indices contain the electronic structure of atoms and molecular space structure topological information, so there is a good correlation between the E_n and chromatographic retention index(*RI*).

(2) The QSRR model is has a good robustness and prediction ability tested by leave-one-out (LOO) cross validation.

(3) The indexes of E_1, E_2, E_7, E_8, E_9 and E_{32} contained main factors of chromatographic retention, so the main structural fragments of the chromatographic retention index are $-CH_3$，$-CH_2-$,-CH=,＞CH=,-S-.

(4) Prediction ability of neural network method is superior to multiple regression prediction model. The nonlinear relationship between RI and E_1, E_2, E_7, E_8, E_9 and E_{32} is presented.

Acknowledgement

This work was supported by the National Natural Science Found of China (No. 21272095) and Natural Science Fund for colleges and Universities in Jiangsu Province (14KJB430022).

References

[1] X F Zeng, J Liu, J H Liu, et al.Chin J. Anal. Chem., Vol. 34 (2006) , p. 1546-1550.

[2] Y Chen, W Yue, B Wang. J. Wuhan Univ. (Nat. Sci. Ed.), Vol. 60 (2014), p.52-56.

[3] C J Feng, W H Yang.Chinese J. Struct. Chem., Vol. 33 (2014), p. 830-834.

[4] C J Feng, W H Yang, L L Mu. Chinese J. Struct. Chem., Vol. 27 (2008), p. 575-587.

[5] C Wang, W Wang, L Q Pan, et al. Chem. Vol. 76 (2013), p. 929-934.

[6] Y Chen .Bull Sci Technol, Vol. 29 (2013), p. 16-19.

[7]Y Chen. Food Sci. Vol. 32 (2011), p. 274-277.

[8] Q. N. Hu, Y. Z. Liang, Y. L. Wang Comp. App. Chem. Vol. 20 (2003), p. 386-390.

[9] T. Zhang, Y.Z. Liang, C.X. Zhao. Chin. J. Anal. Chem. Vol. 34 (2006), p. 1607-1610.

[10] Z H Li, F S Cheng, Z N Xia.Chin. J. Chromatogr. Vol. 29 (2011), p. 63-69.

[11] L. H. Hall, L. B.Kier. J. Chem. Inform. Mod. Vol. 35 (1995), p. 1039-1045.

[12] L. S.Urra, M. P.Gonzalez, M.Teijeira.Bio. Med. Chem., Vol. 15 (2007), p. 3565-3571.

[13] L. S. Urra, M. P.Gonzalez, M. Teijeira.Bio. Med. Chem., Vol.14 (2006) p. 7347-7358.

[14] C. J. Feng, L. L. Mu, W. H. Yang. Acta.Chim.Sin. Vol. 66 (2008), p. 2093-2098.

[15] L. Xu , X. G. Shao:*Methods of Chemometric,* Science Press, Beijing (2004).

