Correlative Study On Chromatographic Retention Indices Of Polycyclic Aromatic Sulfur Heterocycles

Hui Zhang\textsuperscript{1,a}, Yan Chen\textsuperscript{1,b,*} Jing Li\textsuperscript{1,c}

\textsuperscript{1}School of Chemistry and Chemical Engineering, Xuzhou Institute of Technology, Xuzhou, Jiangsu, 221111, China

\textsuperscript{a} 653597093@qq.com, \textsuperscript{b}chenyan681110@126.com, \textsuperscript{c}ljingxz111@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 ($R_I$) 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 ($R_I$) 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

Chromatographic retention index of PASHs. In this paper, 114 kinds of PASHs were selected as research objects. The experimental values of their Chromatographic retention index ($R_I$) 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-hydrogen atom 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:

\[
I_n = \frac{4\delta_n^x + N_n}{N_n \delta_n} \tag{1}
\]

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

\[
\delta_n = \sigma_n - h_n \tag{2}
\]

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

\[
\delta_n^x = \sigma_n + \pi_n + k_n - h_n \tag{3}
\]

Here \( \pi_n \) and \( k_n \) are the number of electrons in \( \pi \) 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 (\( \Delta I \)). \( \Delta I \) of atom n is defined as:

\[
\Delta I_n = \sum_{j \neq n} \frac{I_n - I_j}{r_{n,j}^2} \tag{4}
\]

Here \( r_{n,j} \) 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; \( \Sigma \) 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 \( \Delta I_n \) of the atom n:

\[
E_n = \sum_i (I_i + \Delta I_i) \tag{5}
\]

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

<table>
<thead>
<tr>
<th>No.</th>
<th>Structural formula</th>
<th>Symbol</th>
<th>I</th>
<th>E-index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-CH\textsubscript{3}</td>
<td>sCH\textsubscript{3}</td>
<td>2.0000</td>
<td>E\textsubscript{1}</td>
</tr>
<tr>
<td>2</td>
<td>-CH\textsubscript{2}</td>
<td>ssCH\textsubscript{2}</td>
<td>1.5000</td>
<td>E\textsubscript{2}</td>
</tr>
<tr>
<td>3</td>
<td>&gt;CH\textsubscript{1}</td>
<td>sssCH</td>
<td>1.3333</td>
<td>E\textsubscript{3}</td>
</tr>
<tr>
<td>4</td>
<td>&gt;CH=</td>
<td>sdCH</td>
<td>2.0000</td>
<td>E\textsubscript{4}</td>
</tr>
<tr>
<td>5</td>
<td>&gt;CH=</td>
<td>aaCH</td>
<td>2.0000</td>
<td>E\textsubscript{7}</td>
</tr>
<tr>
<td>6</td>
<td>&gt;CH=</td>
<td>ssdC</td>
<td>1.6667</td>
<td>E\textsubscript{8}</td>
</tr>
<tr>
<td>7</td>
<td>&gt;CH=</td>
<td>aaCa</td>
<td>1.6667</td>
<td>E\textsubscript{9}</td>
</tr>
<tr>
<td>8</td>
<td>-S-</td>
<td>ssS</td>
<td>1.8333</td>
<td>E\textsubscript{31}</td>
</tr>
<tr>
<td>9</td>
<td>-S-</td>
<td>aSa</td>
<td>1.8333</td>
<td>E\textsubscript{32}</td>
</tr>
</tbody>
</table>

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

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:

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

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:

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

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

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].

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

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

$$M = (I + 1)H + (H + 1)O$$

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 ($n = \text{number of samples}$), 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 is very 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**