Chinese Chemical Letters Vol. 15, No. 4, pp 467-470, 2004
http://www.imm.ac.cn/journal/ccl.html

Prediction of the Aqueous Solubilities of Polychlorinated Biphenyls

Shu Shen LIU[1,2,*], Shi Hai CUI[1], Lian Sheng WANG[1]

[1]State Key Laboratory of Pollution Control and Resources Reuse, School of the Environment, Nanjing University, Nanjing 210093
[2]Department of Applied Chemistry, Guilin Institute of Technology, Guilin 541004

Abstract: Using the molecular electronegativity distance vector descriptors derived directly from the molecular topological structures, the aqueous solubilities of polychlorinated biphenyls (PCBs) were predicted. A three-variable regression equation with correlation coefficient of 0.9739 and the root mean square errors of 0.26 was developed. The descriptors included in the equation represent three interactions between three pairs of atomic types, *i.e.*, atom –C= and >C=, –C= and –Cl, and –Cl and –Cl. It has been proved that the aqueous solubilities of 137 PCB congeners can be accurately predicted as long as there are more than 65 calibration compounds.

Keywords: Polychlorinated biphenyls (PCBs), aqueous solubility, molecular electronegativity distance vector (MEDV).

Polychlorinated biphenyls (PCBs) are among the most widespread pollutants in the global ecosystem[1,2]. Because of the lipophilic nature of molecules, PCBs bioaccumulate in the food chain, and residues have been detected in fish and wildlife, and in human adipose tissue, milk and serum[3]. To explain the behaviour of these pollutants in the environment it is very useful to have a good knowledge of their physico-chemical properties, like aqueous solubility, activity coefficient, partition coefficient, that are all related to the hydrophobicity. These properties are particularly important because many PCBs are highly hydrophobic, their concentration in water remaining small and their accumulation in sediments[4] and aquatic organisms[5] significant. As for the prediction of the physico-chemical properties for organic compounds, many reliable methods have been developed by various investigators: quantitative structure activity relationship (QSAR) such as molecular connective indices, quantitative structure property relationship (QSPR), linear solvation energy relationship. Recently, Gramatica[6] used WHIM descriptors to develop some quantitative models between physico-chemical properties including aqueous solubility and structures of PCBs.

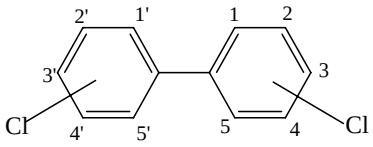
In the present study, we investigated the possibility of predicting the aqueous solubilities (S_W) of PCBs by using the molecular electronegativity distance vector[7-8] (MEDV) derived directly from their two-dimensional topological molecular structures and modified electrotopological state index. With the help of our program VSMP (variable selection and modeling based on the prediction)[9], a three-variable QSAR

* E-mail: ssliu@nju.edu or ssliuhl@263.net

equation with high prediction power has been developed.

137 PCBs (skeleton structure shown in **Figure 1**) and their S_W values (unit is mol/L) observed are directly taken from the literature[6] (see supporting materials). The original MEDV descriptors are calculated according to the literature[8]. Because only exist as three atomic types of nos. 2 (=C–), 3 (>C=), and 13 (–Cl), there are 6 interactions between these types and then only 6 nonzero MEDV descriptors for all 137 PCBs under study. The nonzero descriptors are $\boldsymbol{x}_{14}$ (interaction between types of nos. 2 and 2), $\boldsymbol{x}_{15}$ (nos. 2 and 3), $\boldsymbol{x}_{25}$ (nos. 2 and 13), $\boldsymbol{x}_{26}$ (nos. 3 and 3), $\boldsymbol{x}_{36}$ (nos. 3 and 13), and $\boldsymbol{x}_{91}$ (nos. 13 and 13), respectively. The six descriptors characterize well the molecular structures of the PCBs and the resolution ratio is 100%.

Figure 1 Three atomic types of PCBs as their MEDV descriptors



Three atomic types in PCB molecule are type of nos. 2 (=C–), 3(>C=), and 13(–Cl). Six interactions between them are 2–2, 2–3, 2–13, 3–3, 3–13, and 13–13, respectively, related to six MEDV descriptors, $\boldsymbol{x}_{14}$, $\boldsymbol{x}_{15}$, $\boldsymbol{x}_{25}$, $\boldsymbol{x}_{26}$, $\boldsymbol{x}_{36}$, and $\boldsymbol{x}_{91}$.

Because the value of aqueous solubility (S_W) for PCBs is very low, a negative logarithm transformation of S_W has to be performed before modeling a QSSR. The range of the values is from $-logS_W$ = 5.26 to 10.18 and the distribution is widespread and homogeneous (see **Figure 2**).

The *RRT* values of 30 compounds display below 0.4, 34 values between 0.4 and 0.5, 38 values between 0.5 and 0.6, 46 values between 0.6 and 0.7, 32 values between 0.7 and 0.8, and 30 show *RRT* values above 0.8.

To develop a stable and predicable quantitative structure-solubility relationship (QSSR) model between the MEDV descriptors and aqueous solubilities of 137 PCB congeners, it is essential to optimize the combinations of six MEDV descriptors entering into the final QSSR model. Here, the VSMP program developed in house is employed to select the best subset of descriptors. It has been found that the QSSR model including three MEDV descriptors, $\boldsymbol{x}_{15}$, $\boldsymbol{x}_{25}$, and $\boldsymbol{x}_{91}$, has the best model quality. The best model is as follows.

$$\begin{aligned} -logS_W &= (4.0173 \pm 0.1581) + (0.05168 \pm 0.00707)\cdot\boldsymbol{x}_{15} \\ &\quad + (0.05904 \pm 0.01015)\cdot\boldsymbol{x}_{25} + (0.5385 \pm 0.0122)\cdot\boldsymbol{x}_{91} \qquad (1) \\ n &= 137,\ m = 3,\ r = 0.9739,\ RMSEE = 0.26,\ F = 816.39 \text{ (Estimation)} \\ n &= 137,\ m = 3,\ q = 0.9724,\ RMSEP = 026 \qquad \text{(LOO prediction)} \end{aligned}$$

where n and m are the number of samples and the nonzero MEDV–13 descriptors, respectively. The r, *RMSEE* and F are the correlation coefficient, the root mean square

error, and Fischer statistic of estimations, respectively. The value after the symbol "±" in eq. 1 is the standard derivation related to the regression coefficient. A good QSSR model should have not only an excellent estimation ability for the internal example but also a good prediction ability for the external example. So, a leave–one–out (LOO) cross validation procedure is used to test the prediction ability of the model built. The q and *RMSEP* refer to the correlation coefficient (q) and the root mean square error (*RMSEP*) of predictions obtained in the LOO procedure. From equation 1, the 3–variable QSSR model has high estimation statistics and predictive ability. Obviously, the solubility of PCB compound in water is closely related to the molecular structure. To explain the effect of each atomic type on the S_W of PCB, the standard regression coefficients (b_0) of three MEDV descriptors are also calculated by our VSMP program and the b_0 values of three descriptors, $\boldsymbol{x}_{15}$, $\boldsymbol{x}_{25}$, and $\boldsymbol{x}_{91}$, are 0.216, 0.151, and 1.042, respectively. This shows that the most important descriptor affecting the S_W is the interaction between chloride atoms ($\boldsymbol{x}_{91}$). The second important descriptors are $\boldsymbol{x}_{15}$ and $\boldsymbol{x}_{25}$. These descriptors reflect the interactions between atom segment –C= and >C= as well as –C= and –Cl.

The values of $-\log S_W$ estimated by equation 1 and observed experimentally are listed in **Table 1** of supporting material together with the values of three optimal MEDV descriptors. The relationship graph between $-\log S_W$ estimated and observed is shown in **Figure 3**.

Figure 2 Distribution of $-\log S_W$.

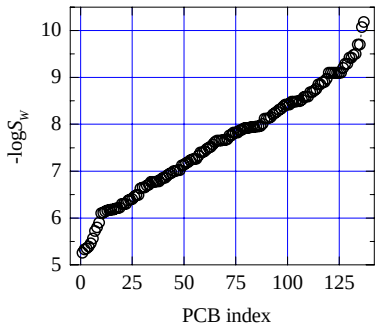
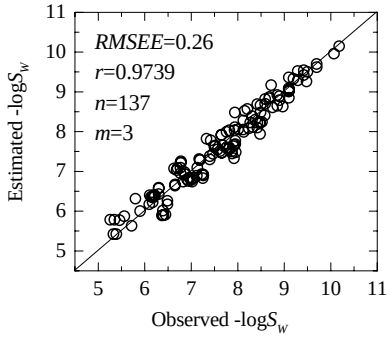


Figure 3 Plot of $-\log S_W$ estimated *vs* observed.



A good fit alone does not guarantee that the model will be useful for prediction purposes. Some kind of validation is necessary to test how stable is the model and how well does it predict. The above LOO cross-validation experiment primarily tests the stability of the model (equation 1). However, it is not enough to prove the predictive ability of the model having merely a high value of q^2. To further test the predictive ability of the model for the external compounds without in the model, part of compounds are picked up from 137 PCBs to construct a calibration set which is used to develop a predict model and then predict the values of $-\log S_W$ in the remaining compounds. How to pick up the compounds in the calibration set is very important to the development of a predictive QSSR model. The procedure used in this study consists of two steps: (a)

ranking the $-\log S_W$ of 137 PCBs and (b) equidistantly picking up the compounds from the ranked PCB compound set. In this way, 25, 35, 45, 55, 65, 75, 85, 95, and 105 PCB compounds are respectively picked up to construct nine linear QSSR equation between three optimal MEDV descriptors and $-\log S_W$ of these PCBs. These QSSR models are respectively used to predict the values of $-\log S_W$ in remaining 112, 102, 92, 82, 72, 62, 52, 42, and 32 PCB compounds. The statistical results (see **Table 2**) together with various regression coefficients show that five QSSR models developed by the calibration sets including more than 65 PCB compounds have no significant differences with the model (eq. 1) derived from all 137 PCBs.

Table 2 Some calibration set models and their statistical parameters.

Calibration Set (m = 3)						Testing Set		
n	r	*RMSC* [a]	F	q	*RMSV* [a]	n	r_P	*RMSP* [a]
25	0.9756	0.24	138.30	0.9621	0.31	112	0.9721	0.27
35	0.9752	0.25	200.73	0.9688	0.28	102	0.9735	0.26
45	0.9671	0.29	197.33	0.9604	0.32	92	0.9773	0.24
55	0.9771	0.24	359.08	0.9737	0.26	82	0.9712	0.27
65	0.9723	0.26	351.24	0.9688	0.28	72	0.9753	0.25
75	0.9724	0.26	410.52	0.9692	0.28	62	0.9759	0.25
85	0.9712	0.27	449.07	0.9683	0.28	52	0.9783	0.24
95	0.9766	0.24	624.98	0.9745	0.25	42	0.9682	0.29
105	0.9731	0.26	601.62	0.9710	0.27	32	0.9772	0.25

[a] *RMSC*, *RMSV*, and *RMSP* are the root mean square errors in modeling, LOO validation, and prediction.

Acknowledgment

We are especially grateful to the China Postdoctoral Science Foundation and the Guangxi Natural Science Fund (No.0236063) for their financial supports.

References

1. S. Ayris, G.M. Currado, D. Smith, S. Harrad, *Chemosphere*, **1997**, *35*, 905.
2. S. Safe, *Crit. Rev. Toxicol.*, **1990**, *21*, 51.
3. S. Bandiera, S. Safe, A. B. Okey, *Chem. Biol. Interact.*, **1982**, *39*, 259.
4. M. Engwall, D. Broman, R. Ishag, C. Naf, Y. Zebuhr, *Environ. Toxicol. Chem*., **1996**, *15*, 213.
5. A. Borrell, A. Aguilar, A. Corsolini, S. Focardi, *Chemosphere*, **1996**, *32*, 2359
6. P. Gramatica, N. Navas, R. Todeschini, *Chemom. Intell. Lab. Syst*., **1998**, *40*, 53.
7. S. S**.** Liu, C. S. Yin, S. X. Cai, Z. L. Li, *J. Chem. Inf. Comput. Sci*., **2001**, *41*, 321.
8. S. S. Liu, S. H. Cui, Y. Y. Shi, L. S. Wang, *Internet Electron. J. Mol. Des*., **2002**, *1*, 610.
9. S. S**.** Liu, H. L. Liu, C. S. Yin, L. S. Wang, *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 964.

Received 3 April, 2003