

可变连接性指数预测卤代苯溶解度及分配系数

陈强, 孙敬敏, 胡亮

兰州大学大气科学学院, 兰州 730000

摘要 分子片段是因成键环境不同而具有不同特性的原子或官能团,也可以认为是影响化合物性质的一个分子结构单元。分子片段可变连接性指数(mfVCI)是定量结构-性质相关(QSPR)研究中的一种拓扑描述符。为了检验 mfVCI 对同组化合物不同性质预测的适用性,本文建立了 mfVCI 与 36 个卤代苯的水溶解度(S_w)和正辛醇/水分配系数(K_{ow})的线性 QSPR 模型,用规划求解方法得到最优解,并采用拟合优度诊断和交互验证法(内部交互验证和外部交互验证)对模型进行检验。结果表明,所建立的模型均具有较好的拟合能力($R^2>0.97$)和稳健性($q^2>0.89$),且预测能力强($q_{ext}^2>0.95$)。mfVCI 除具有一般可变分子连接性指数的优势外,还能更好地区分不同成键环境中原子或官能团对同组化合物不同性质的影响,能较好地应用于卤代苯的水溶解度及分配系数的预测。

关键词 分子片段可变连接性指数;QSPR;水溶解度;正辛醇/水分配系数;卤代苯

中图分类号 X131

文献标识码 A

doi 10.3981/j.issn.1000-7857.2011.15.006

Prediction of Aqueous Solubility and Partition Coefficients of Halogenated Benzenes Using Molecular Fragments Variable Connectivity Index

CHEN Qiang, SUN Jingmin, HU Liang

College of Atmospheric Sciences, Lanzhou University, Lanzhou 730000, China

Abstract For the molecular fragments Variable Connectivity Index (mfVCI), the molecular fragments are defined as the atoms or functional groups and are also regarded as the molecular structure unit for the major influence of the property. Different molecular fragments were endowed with different fragments weights. In order to test whether or not the mfVCI could be used for predicting different properties of the same molecules, the Quantitative Structure-Property Relationship (QSPR) models of aqueous solubility (S_w) and n -octanol/water partition coefficients (K_{ow}) for 36 halogenated benzenes based on the mfVCI were investigated. The fitting goodness of the QSPR models are mainly characterized by determination coefficient (R^2), standard deviation (SD), Fischer variance ratio (F), and significant lever (P). The internal cross-validation including leave-one-out and leave-many-out methods with cross-validated correlation coefficients (q_{lo}^2 and q_{mo}^2) was used to evaluate the model's robustness. The external cross-validation was used to evaluate the predictive power of the models developed from the training set by correlation coefficients (q_{ext}^2). The linear relationship was assumed between mfVCI and S_w or K_{ow} in here. The mfVCI was optimized by using Solver. The target function is the root mean calculated residual sum of squares for training set. The QSPR models with R^2 value of more than 0.97, q^2 value of more than 0.89, and q_{ext}^2 value of more than 0.95 have the good robustness and predictive ability as well as fitting ability. The mfVCI could distinguish the different role of same atoms or functional groups among different chemical bonding. The results show that the mfVCI could be well used in the prediction of both of aqueous solubility and n -octanol/water partition coefficients for halogenated benzenes. It could describe different properties equally well. The mfVCI maybe play an important role in the development of molecular descriptors on QSPR researches.

Keywords mfVCI; QSPR; aqueous solubility; octanol/water partition coefficient; halogenated benzene

收稿日期: 2011-01-10;修回日期:2011-03-25

基金项目: 兰州大学中央高校基本科研费自由探索面上项目(1zujbky-2009-7)

作者简介: 陈强,副教授,研究方向为环境化学,电子邮箱:chenqhqh@163.com

0 引言

定量结构-性质相关(QSPR)是研究物质结构与其性质之间相关关系,预测化合物性质的一种有效方法。分子描述符的建立是QSPR研究的核心任务,目前常用的分子描述符有经验理化参数、量子化学参数和拓扑指数等^[1]。基于化合物拓扑结构的拓扑指数计算简单,且应用效果好。Randić的简单分子连接性指数^[2]、Kier和Hall的价连接性指数^[3]及改进指数等被广泛应用于化合物性质的预测,但分子连接性指数的值是固定的,对化合物不同性质的预测结果不是很理想。为此,Randić提出的可变连接性指数既能区分处于不同成键环境的相同原子,又能根据化合物的不同性质得到原子对其性质贡献的权重^[4-5],能较好地用于化合物不同性质的预测。考虑官能团对有机物性质的重要作用及官能团和原子之间存在相互影响提出的分子片段可变连接性指数(mfVCI),既可区分不同原子或官能团,又可区分相同原子/官能团因成键环境不同而对化合物性质的不同影响^[6]。利用该指数建立的卤代苯发光菌毒性(pT_{50})及醇类沸点模型具有较好的拟合和预测能力,且稳健性强^[6-7]。

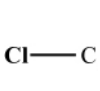
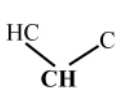
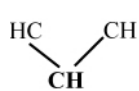
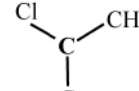
辛醇/水分配系数(K_{ow})和水溶解度(S_w)是表征有机污染物在土壤、水体和生物体内的富集、迁移和转化的重要参数^[8]。本研究将mfVCI应用于卤代苯的 S_w 和 K_{ow} 预测的QSPR研究中,考查模型的稳健性和预测能力,检验mfVCI对化合物不同性质的预测能力,为相关研究提供有效参考。

1 数据来源和方法

36个卤代苯的 S_w 和 K_{ow} 数据来源于文献[9]。

mfVCI是基于分子拓扑结构,考虑了原子/官能团不同成键环境的一种拓扑指数。分子片段定义为因成键环境不同而具有不同特性的原子或官能团,是影响化合物性质的分子结构功能单元^[6]。以邻二氯代苯为例,其包含的分子片段如表1所示。

表1 邻二氯代苯的分子片段
Table 1 Molecular fragments of ortho dichlorobenzene

名称	图示			
分子片段				
表达式	(C)Cl	(CH)CH(C)	(CH)CH(CH)	(CH)(C)(Cl)C
权重	w_1	w_2	w_3	w_4

mfVCI由增广矩阵计算得出。每一个分子片段被赋予不同的权重(w_i),分子片段的扩展点价(δ_i^m)为权重与该片段的Randić点价(δ_i)之和,即 $\delta_i^m=w_i+\delta_i$ 。mfVCI的计算与Randić的分子连接性指数类似,其一阶指数表达为

$$\chi^{mf}=\sum_{p=1}^m 1\chi_p^{mf}=\sum_{p=1}^m \prod_{i=1}^2 (\delta_i^{mf})_p^{-0.5} \tag{1}$$

其中, p 为化合物拓扑结构中所有边的数目; χ_p^{mf} 为子一阶路径mfVCI,即由顶点 i,j 构成的边的指数,

$$\chi_p^{mf}=\frac{1}{\sqrt{\delta_i^{mf}\delta_j^{mf}}}=\frac{1}{\sqrt{(\delta_i+w_i)(\delta_j+w_j)}} \tag{2}$$

且要求 $\delta_i^{mf}=w_i+\delta_i>0$,求解得到分子片段的最优权重。

2 mfVCI与 S_w 和 K_{ow} 的QSPR模型

假设mfVCI与化合物性质之间存在线性关系 $P=a\chi^{mf}+b$, a,b 为系数。以计算值与实验值的最小均方根误差为目标函数,采用规划求解方法获得方程参数。建立的QSPR模型见式(3)和式(4)。

$$\begin{aligned} -\lg S_w &= 0.0874 \times \chi^{mf} - 25.1017 \\ n &= 36, R^2 = 0.976, SD = 0.234, F = 1393.5, P < 0.001 \end{aligned} \tag{3}$$

$$\begin{aligned} -\lg K_{ow} &= 0.0083 \times \chi^{mf} + 1.1872 \\ n &= 36, R^2 = 0.998, SD = 0.054, F = 13848.3, P < 0.001 \end{aligned} \tag{4}$$

36个卤代苯的18个分子片段及对 S_w 和 K_{ow} 的分子片段权重 w_{si}, w_{ki} 见表2。模型预测值和实验值的拟合结果见表3。可以看出,由mfVCI建立的QSPR模型拟合能力好,具有显著的统计学意义。说明分子片段可变连接性指数能够很好地用于化合物结构与性质相关关系的研究。

表2 36个卤代苯的18个分子片段及权重系数
Table 2 18 molecular fragments and weights of the 36 halogenated benzenes

序号	分子片段	权重	
		w_{si}	w_{ki}
1	(C)Cl	-0.6959	-0.9979
2	(C)Br	-0.5683	-0.9986
3	(C)F	85.3419	-0.9990
4	(C)I	-0.9542	-0.9990
5	(2C)CH	-1.9926	24.1228
6	(CH)(C)CH	-1.9869	-1.6931
7	(2CH)CH	-1.9805	-1.9483
8	(2C)(Cl)C	-2.9798	-2.9380
9	(CH)(C)(Cl)C	-2.9655	-2.9553
10	(2CH)(Cl)C	-2.9497	-2.9665
11	(CH)(C)(Br)C	-2.9693	-2.9588
12	(2C)(Br)C	-2.9841	-2.9455
13	(2CH)(Br)C	-2.9547	-2.9666
14	(CH)(C)(F)C	-2.9693	-2.7394
15	(2CH)(F)C	-2.9509	-2.8075
16	(CH)(C)(I)C	-2.9594	-2.9644
17	(2C)(I)C	-2.9785	-2.9556
18	(2CH)(I)C	-2.9471	-2.9717

表3 $-\lg S_w$ 和 $-\lg K_{ow}$ 的实验值及预测值
Table 3 Experimental and predictive values of $-\lg S_w$ and $-\lg K_{ow}$

序号	化合物	$-\lg S_w$			$-\lg K_{ow}$		
		$\chi^2_{\text{总}}$	实验值	预测值	$\chi^2_{\text{总}}$	实验值	预测值
1	PhCl ₆	373.26	7.76	7.54	626.44	6.42	6.41
2	PhCl ₅	357.29	5.65	6.14	544.82	5.69	5.73
3	1,2,4,5-PhCl ₄	346.77	5.56	5.22	464.22	5.05	5.06
4	1,2,3,5-PhCl ₄	344.45	4.79	5.02	458.25	5.05	5.01
5	1,2,4-PhCl ₃	332.68	3.57	3.99	374.01	4.27	4.31
6	1,2,3-PhCl ₃	326.94	3.76	3.49	367.11	4.27	4.25
7	1,3,5-PhCl ₃	334.72	4.44	4.17	366.72	4.27	4.25
8	1,2-PhCl ₂	318.86	3.01	2.78	282.56	3.55	3.54
9	1,3-PhCl ₂	322.53	3.08	3.10	277.95	3.38	3.50
10	1,4-PhCl ₂	323.89	3.21	3.22	286.16	3.59	3.57
11	PhCl	313.78	2.35	2.34	194.41	2.81	2.81
12	1,2,4,5-PhBr ₄	364.92	6.98	6.81	580.84	6.01	6.03
13	1,2,3-PhBr ₃	344.58	5.03	5.03	454.86	4.98	4.98
14	1,2,4-PhBr ₃	344.69	4.50	5.04	459.19	4.98	5.02
15	1,3,5-PhBr ₃	348.42	5.60	5.37	446.32	4.98	4.91
16	1,2-PhBr ₂	326.05	3.50	3.41	341.52	4.07	4.03
17	1,3-PhBr ₂	330.29	3.38	3.78	331.05	3.75	3.95
18	1,4-PhBr ₂	330.29	4.07	3.78	339.30	4.07	4.02
19	PhBr	316.98	2.64	2.62	220.98	2.99	3.03
20	1,2-PhF ₂	309.93	2.00	2.00	168.23	2.59	2.59
21	1,3-PhF ₂	309.38	2.00	1.95	167.05	2.58	2.58
22	1,4-PhF ₂	310.46	1.97	2.05	165.02	2.58	2.56
23	PhF	307.06	1.79	1.75	133.84	2.27	2.30
24	1,2,3-PhI ₃	357.73	6.18	6.18	560.40	5.86	5.86
25	1,2,4-PhI ₃	356.70	5.93	6.09	566.16	5.85	5.91
26	1,3,5-PhI ₃	363.68	6.85	6.70	559.28	5.85	5.85
27	1,2-PhI ₂	334.12	4.24	4.12	410.67	4.65	4.61
28	1,3-PhI ₂	342.48	4.57	4.85	407.89	4.64	4.59
29	1,4-PhI ₂	344.49	5.25	5.02	417.67	4.64	4.67
30	PhI	324.07	2.95	3.24	260.16	3.36	3.36
31	1-Cl-2-Br-Ph	322.41	3.19	3.09	312.02	3.83	3.79
32	1-Cl-3-Br-Ph	326.41	3.21	3.44	304.50	3.83	3.73
33	1-Cl-4-Br-Ph	327.09	3.63	3.50	312.73	3.83	3.79
34	1-Cl-4-I-Ph	334.19	4.03	4.12	351.92	4.12	4.12
35	1-Br-2-I-Ph	329.81	3.83	3.74	376.03	4.36	4.32
36	1-Br-4-I-Ph	337.39	4.56	4.40	378.48	4.36	4.34

3 QSPR 模型的检验

为检验模型的预测能力,随机抽取 10 个样本作为测试集,其余作为训练集建立模型。采用文献[10]中交叉验证方法检验模型的稳健性和预测能力。由训练集得到的模型如式(5)和式(6)所示:

$$-\lg S_w=0.0462\times\chi^2_{\text{总}}+1.6166$$

$$n=26,R^2=0.974,SD=0.247,F=911.4,P<0.001, \quad (5)$$

$$q^2_{\text{loo}}=0.899,q^2_{\text{lno}}=0.897,q^2_{\text{ext}}=0.956$$

$$-\lg K_{ow}=0.0365\times\chi^2_{\text{总}}+2.2365$$

$$n=26,R^2=0.998,SD=0.051,F=10219.8,P<0.001, \quad (6)$$

$$q^2_{\text{loo}}=0.997,q^2_{\text{lno}}=0.997,q^2_{\text{ext}}=0.988$$

可见,模型 R^2 均大于 0.7,具有较好的拟合能力,且具有显著的统计学意义。 $q^2>0.89,q^2_{\text{ext}}>0.95,q^2$ 和 q^2_{ext} 均较接近于 1,模型具有较好的稳健性和预测能力^[10]。计算值与实验值的相关性如图 1 所示。

4 结论

本文利用分子片段可变连接性指数(mfVCI)对 36 个卤代苯水溶解度和正辛醇/水分配系数进行了 QSPR 研究。可以看出:① mfVCI 有良好的结构区分性,分子片段能降低 mfVCI 的简并度;② mfVCI 能够很好地预测卤代苯的水溶解度和正辛醇/水分配系数,能满足化合物不同性质的预测。

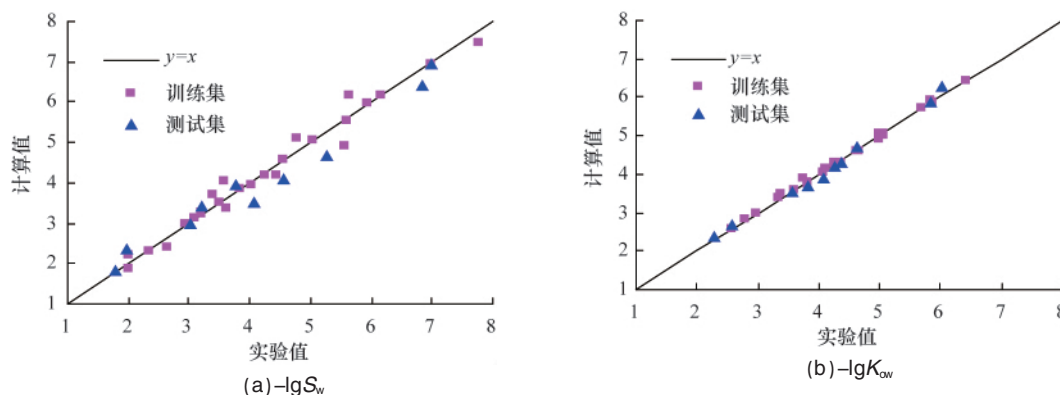


图 1 计算值与实验值的相关性

Fig. 1 Correlation between the experimental and the calculated values

参考文献 (References)

- [1] 王鹏. 定量构效关系及研究方法[M]. 哈尔滨: 哈尔滨工业大学出版社, 2004.
Wang Peng. Method of study quantitative structure activity relationships [M]. Harbin: Harbin Institute of Technology Press, 2004.
- [2] Randić M. On characterization of molecular branching [J]. *Journal of the American Chemical Society*, 1975, 97(23): 6609–6615.
- [3] Kier L B, Hall L H. Molecular connectivity. VII: Specific treatment of heteroatoms[J]. *Journal of Pharmaceutical Sciences*, 1976, 65(12): 1806–1809.
- [4] Randić M. Novel graph theoretical approach to heteroatoms in quantitative structure –activity relationships [J]. *Chemometrics and Intelligent Laboratory Systems*, 1991, 10(1–2): 213–227.
- [5] Randić M. The connectivity index 25 years after[J]. *Journal of Molecular Graphics & Modelling*, 2001, 20(1): 19–35.
- [6] Chen Q, Kou Y W, Wang Q, et al. A molecular fragments variable connectivity index for studying the toxicity (*Vibrio fischeri* pT₅₀) of substituted-benzenes [J]. *Journal of Environmental Science and Health*

- Part A-Toxic/Hazardous Substanc*, 2009, 44(3): 289–295.
- [7] Chen Q, Kou Y W, Pan F, et al. Application of molecular fragments variable connectivity index to predicting boiling points of alcohols[J]. *Journal of the Iranian Chemical Society*, 2010, 7(4): 1012–1020.
- [8] 余训爽, 余讯民. 氯代芳烃化合物水溶解性和分配系数的 QSAR 研究 [J]. 科技通报, 2006, 22(1): 4–9.
Yu Xunshuang, Yu Xunmin. *Bulletin of Science and Technology*, 2006, 22(1): 4–9.
- [9] 杨锋, 颜肖慈, 欧阳礼, 等. 改进了有价连接性指数及其应用实例—— δ_i 和 δ^* 在结构/性质相关中的比较研究[J]. 化学物理学报, 1998, 11(3): 221–226.
Yang Feng, Yan Xiaoci, Ouyang Li, et al. *Chinese Journal of Chemical Physics*, 1998, 11(3): 221–226.
- [10] Tropsha A, Gramatica P, Gombar V K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models[J]. *Qsar & Combinatorial Science*, 2003, 22(1): 69–77.

(责任编辑 岳臣)

·学术动态·



“第八届中国通信学会学术年会”征文

由中国通信学会主办的“第八届中国通信学会学术年会”将于 2011 年 12 月在北京召开。

征文范围: 智慧城市与绿色 IT; 通信网络技术与应用; 密码学与信息安全; 信号与信号处理以及其他与通信相关的研究内容。

征文截止日期: 2011 年 8 月 20 日。

联系电话: 010-62262770。

会议网址: <http://www.leaderstudio.net/nianhui2011/index.html>。