分子电性距离矢量预测多氯萘的正辛醇/空气分配系数*

李正华 徐 盼 夏之宁**

(重庆大学化学化工学院, 重庆, 400044)

摘 要 应用分子电性距离矢量(MEDV)对多氯萘(PCNs)的 76种同系物进行结构表征, 通过多元线性回归方法建立了 PCNs的正辛醇/空气分配系数($\lg K_{OA}$)与 MEDV 之间的定量结构-性质关系(QSPR), 该模型具有显著的相关性(n= 24, R= 0.997, SD= 0.088, F= 417.546). 采用逐步回归的方法(SMR)从原模型参数中选取 2个参数建立了新模型, 其模型相关系数 R= 0.996, 继以留一法进行交互检验, 相关系数 R_{CV} = 0.995, 说明此定量构效相关模型具有很好的稳定性和预测能力. 运用模型对未有试验值的 52种 PCNs的 $\lg K_{OA}$进行了预测.

关键词 多氯萘, 分子电性距离矢量, 正辛醇/空气分配系数, 定量结构-性质关系.

多氯萘(polychlorinated naphthalenes, PCNs)是一类物理化学性质与多氯联苯(PCBs)相近的持久性有机污染物(POPs)[1]. 有机污染物的正辛醇/空气分配系数($\lg K_{OA}$)是描述污染物在空气和环境有机相间分配行为的重要参数[2-3]. Harner和 Bidleman等[4]采用产生柱法测定了 24种多氯萘的正辛醇/空气分配系数, 但是由于 PCNs纯品通常难以购置或制备, 通过实验获得所有 PCNs的 $\lg K_{OA}$的实验值有很大难度. 鉴于此, 通过定量构效关系方法预测其它未有实验值的 PCNs的 $\lg K_{OA}$是十分有意义的.

目前, 对 PCNs的正辛醇/空气分配系数的定量构效研究[5-8]主要采用量化参数作为定量构效模型的参数, 这类模型的建立需要大量的量化计算工作, 而分子电性距离矢量(MEDV)的计算易程序化, 只需输入分子的原子种类及连接关系, 在实际应用中更简便, 且所得模型具有良好的稳定性和预测能力.

本文采用分子电性距离矢量作为结构表征参数, 建立了高精度的 $\lg K_{OA}$与结构性质之间的相关方程, 并对其它未有实验值的 52种 PCNs化合物的 $\lg K_{OA}$值进行了预测.

1 分子电性距离矢量的计算

MEDV 是一种模拟分子中各非氢原子之间相互作用的新型矢量描述子, 它获得简便, 且具有很好的性质相关性[9], 其已经应用于有机化合物正辛醇/水分配系数、保留指数、生物活性等多种性质的定量构效关系的相关研究中[9-13]. MEDV 矢量中各个元素的计算借鉴库伦定理的形式, 将已经发现和合成的绝大多数有机物分子中各种非氢原子分为 4种类型, 这 4种非氢原子发生相互作用成以下几种方式组合: M_{kl}(其 k= 1, 2, 3, 4; $k \leqslant l \leqslant 4$), 表示第 k类原子和第 l类原子的作用项. M_{kl}可由下式计算:

$$M_{kl} = \sum_{i \in k,\ j \in l} \frac{q_i q_j}{d_{ij}^2} \quad (k = 1, 2, 3, 4;\ k \leqslant l \leqslant 4)$$

式中, k或 l为原子类型, 原子 i和 j分别属于第 k类原子和第 l类原子; q_i和 q_j为原子 i和 j的相对电性; d_{ij}表示原子 i和 j之间的距离(以相对键长表示), 是从原子通过一个或多个化学键连接到其它原子的所有路径中各个相对键长加和的最小值. 这样就得 10个变量: M_{11}, M_{12}, M_{13}, M_{14}, M_{22}, M_{23}, M_{24}, M_{33}, M_{34}, M_{44}, 即为 MEDV 描述子. 统写为: V_1, V_2, V_3, V_4, V_5, V_6, V_7, V_8, V_9, V_{10}, 具体计算方法参见文献[9-13]. 由于 PCNs分子不含第 4原子类型的非氢原子, 相应第 4类原子类型的 MEDV 元素实际为零, 因此 MEDV 只有 6个非零元素, 其元素对应为 V_1, V_2, V_3, V_5, V_6, V_8.

2010年 9月 5日收稿.

* 国家自然科学基金(No 20775096); 科技部国际合作项目(No 2010DFA32680)资助.

** 通讯联系人, E-mail: chem-lalr.cqu@yahoo.com.cn

2 结果与讨论

应用 MEDV 对 PCNs进行结构表征, 删除 10个矢量中全为零的 4个变量, 采用多元线性回归方法建立了 PCNs化合物的分子电性距离矢量与 $\lg K_{OA}$的线性模型, 建立 6参数的模型 (Model 1), 所得结果如表 2; 由 Model 1所得的预测值及交互检验值见表 1, 它们与实验值相关图见图 1. 为了说明模型的稳定性和对外部样本的预测能力, 对模型进行了留一法交互检验 (LOO-CV).

表 1　PCNs的 $\lg K_{OA}$实验值和各模型预测值和交互检验值

Table 1　The experimental, model predicted and interactive-tested $\lg K_{OA}$ values of PCNs

序号	化合物	Exp[a]	Model 1		Model 2		序号	化合物	Exp[a]	Model 1		Model 2	
			Pred[b]	CV[c]	Pred[b]	CV[c]				Pred[b]	CV[c]	Pred[b]	CV[c]
01	Naphthalene		5.25		5.44		39	1,2,5,8-tetraCN	8.40	8.42	8.42	8.39	8.39
02	1-CN		6.08		6.17		40	1,2,6,7-tetraCN		7.95		7.90	
03	2-CN		5.68		5.81		41	1,2,6,8-tetraCN		8.04		8.05	
04	1,2-diCN		6.81		6.84		42	1,2,7,8-tetraCN		8.33		8.31	
05	1,3-diCN		6.50		6.59		43	1,3,5,7-tetraCN		7.72		7.77	
06	1,4-diCN	6.93	6.89	6.82	6.92	6.91	44	1,3,5,8-tetraCN		8.11		8.13	
07	1,5-diCN		6.89		6.92		45	1,3,6,7-tetraCN		7.64		7.64	
08	1,6-diCN		6.50		6.55		46	1,3,6,8-tetraCN		7.73		7.79	
09	1,7-diCN		6.50		6.56		47	1,4,5,8-tetraCN	8.45	8.50	8.52	8.51	8.51
10	1,8-diCN		6.90		6.95		48	1,4,6,7-tetraCN	8.13	8.02	7.98	7.98	7.96
11	2,3-diCN		6.41		6.48		49	2,3,6,7-tetraCN		7.56		7.52	
12	2,6-diCN		6.11		6.18		50	1,2,3,4,5-pentaCN		9.43		9.44	
13	2,7-diCN		6.11		6.18		51	1,2,3,4,6-pentaCN	8.91	9.05	9.12	9.02	9.03
14	1,2,3-triCN		7.53		7.55		52	1,2,3,5,6-pentaCN	9.15	9.05	9.05	8.98	8.97
15	1,2,4-triCN		7.62		7.63		53	1,2,3,5,7-pentaCN	8.73	8.74	8.74	8.74	8.74
16	1,2,5-triCN		7.62		7.59		54	1,2,3,5,8-pentaCN	9.13	9.12	9.12	9.11	9.11
17	1,2,6-triCN		7.23		7.22		55	1,2,3,6,7-pentaCN		8.67		8.61	
18	1,2,7-triCN		7.22		7.23		56	1,2,3,6,8-pentaCN		8.75		8.76	
19	1,2,8-triCN		7.62		7.63		57	1,2,3,7,8-pentaCN		9.05		9.02	
20	1,3,5-triCN	7.32	7.31	7.30	7.34	7.35	58	1,2,4,5,6-pentaCN		9.13		9.10	
21	1,3,6-triCN		6.92		6.96		59	1,2,4,5,7-pentaCN	8.86	8.82	8.81	8.85	8.85
22	1,3,7-triCN		6.92		6.97		60	1,2,4,5,8-pentaCN	9.18	9.21	9.21	9.23	9.23
23	1,3,8-triCN		7.32		7.37		61	1,2,4,6,7-pentaCN		8.74		8.70	
24	1,4,5-PCN	7.56	7.70	7.74	7.71	7.73	62	1,2,4,6,8-pentaCN	8.78	8.82	8.81	8.85	8.85
25	1,4,6-triCN	7.27	7.31	7.32	7.30	7.31	63	1,2,4,7,8-pentaCN	9.06	9.12	9.13	9.11	9.11
26	1,6,7-triCN		7.22		7.23		64	1,2,3,4,5,6-hexaCN	10.11	10.14	10.15	10.12	10.12
27	2,3,6-triCN		6.84		6.85		65	1,2,3,4,5,7-hexaCN	9.80	9.83	9.84	9.87	9.88
28	1,2,3,4-tetraCN		8.64		8.64		66	1,2,3,4,5,8-hexaCN	10.37	10.21	10.14	10.25	10.21
29	1,2,3,5-tetraCN		8.33		8.31		67	1,2,3,4,6,7-hexaCN	9.70	9.76	9.79	9.70	9.71
30	1,2,3,6-tetraCN		7.95		7.93		68	1,2,3,5,6,7-hexaCN		9.76		9.70	
31	1,2,3,7-tetraCN		7.95		7.94		69	1,2,3,5,6,8-hexaCN		9.84		9.83	
32	1,2,3,8-tetraCN		8.34		8.34		70	1,2,3,5,7,8-hexaCN	9.83	9.83	9.83	9.83	9.83
33	1,2,4,5-tetraCN	8.58	8.42	8.39	8.42	8.41	71	1,2,3,6,7,8-hexaCN		9.76		9.74	
34	1,2,4,6-tetraCN	8.08	8.03	5.02	8.01	8.01	72	1,2,4,5,6,8-hexaCN	9.89	9.91	9.92	9.95	9.96
35	1,2,4,7-tetraCN		8.03		8.02		73	1,2,4,5,7,8-hexaCN		9.91		9.95	
36	1,2,4,8-tetraCN	8.41	8.41	8.42	8.43	8.43	74	1,2,3,4,5,6,7-heptaCN		10.84		10.84	
37	1,2,5,6-tetraCN		8.35		8.26		75	1,2,3,4,5,6,8-heptaCN		10.91		10.97	
38	1,2,5,7-tetraCN		8.03		8.02		76	1,2,3,4,5,6,7,8-octaCN		11.91		11.99	

注: a. Hamer和 Bidleman[4]实验测定的 PCNs的 $\lg K_{OA}$值; b. 相应模型的预测值; c. 相应模型的交互检验值.

从建模 Model 1的结果看出, 所建模型的交互检验结果 R_{CV} = 0.993, 表明模型的稳定性较好. 源于样本体系中 lgK_{OA}的数目为 24, 而自变量为 6, 达到定量构效关系研究中普遍要求的样本数大于参数个数 3倍的条件. 但自变量间关系复杂, 有的变量对因变量的贡献大, 有的则不然, 甚至使方程的稳定性下降. 为了了解各个变量对因变量的影响程度, 减少上述多元线性回归模型中可能存在的过拟合现象, 得到更稳定的模型, 需对模型进行逐步回归分析, 进行变量筛选, 得到结果见表 2

表 2 逐步回归变量分析 (n= 24)

Table 2 Analysis of variables by SMR (n= 24)

m	a_0	a_1	a_2	a_3	a_5	a_6	a_8	R
01	5.604						0.333	0.994
02	5.011		0.226				0.349	0.996
03	5.062		0.212	0.020			0.330	0.996
04	16.139		2.277	1.731		−1.112	−1.719	0.997
05	22.303	3.362	4.015	2.086		−1.735	−2.713	0.997
06	−58.960	3.294	3.967	2.068	4.054	2.339	1.374	0.997

m	SD	F	U	Q	R_{CV}	SD_{CV}	F_{CV}	U_{CV}
01	0.101	1911.37	19.483	0.244	0.993	0.111	1581.867	19.437
02	0.086	1334.228	19.553	0.154	0.995	0.095	1072.411	19.516
03	0.087	853.903	19.554	0.153	0.994	0.102	622.325	19.498
04	0.084	688.368	19.572	0.135	0.994	0.109	408.776	19.481
05	0.086	530.989	19.574	0.133	0.993	0.118	278.591	19.456
06	0.088	417.063	19.574	0.133	0.993	0.122	219.306	19.456

注: a_1, a_2, a_3, a_5, a_6, a_8分别是变量 V_1, V_2, V_3, V_5, V_6, V_8的系数, a_0是常数项; 下标 CV代表交互检验的相应数值.

从表 2可以看出, 随变量个数 m 增加, 复相关系数 R 逐渐增加, 但在交互检验预测过程中, R_{CV}先增大而后减少. 从 MEDV 的 6个变量中筛选出了 4变量 (未引入变量 V_1, V_5), 此时回归系数 R 为 0.997 (与 6变量的 R 相等), 标准偏差 SD= 0.084(最小值), 而交互检验的回归系数 R_{CV}为 0.994, 交互检验的标准偏差 SD_{CV}为 0.109(均优于 6变量时的值), 说明 4变量的 QSAR 模型较 6变量的 QSAR 模型 (Model 1)具有更强的预测能力. 再从 MEDV的 6个变量中筛选出了 2变量 (引入变量 V_2, V_8), 此时回归系数 R 为 0.996, 标准偏差 SD= 0.086(均与最优值相近); 而交互检验的回归系数 R_{CV}为 0.995, 交互检验的标准偏差 SD_{CV}为 0.095(均达到最优值). 综合 R、SD、R_{CV}、SD_{CV}考虑, 可以认为 2个参数模型是相对最优模型, 统计意义更显著, 所得模型 (Model 2)见表 2, 由 Model 2所得的预测值及交互检验值见表 1, 它们与实验值相关图见图 2

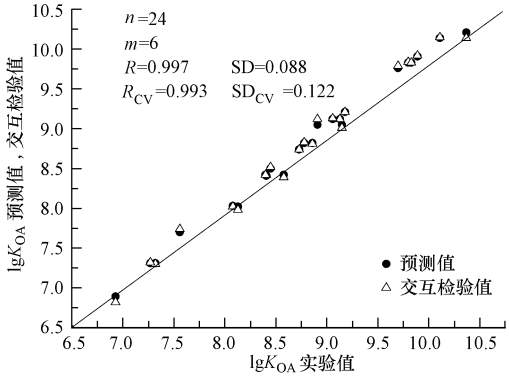

图 1 Model 1中 PCNs的 lgK_{OA}的实验值及交互检验值与预测值实验值对比图

Fig 1 Polt of predicted and interactive-tested *vs.* experimental lgK_{OA} values of 24 PCNs by model 1

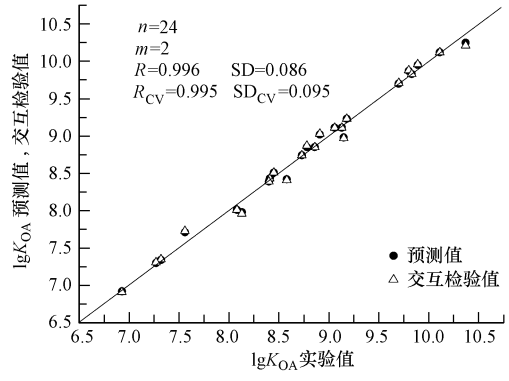

图 2 Model 2中 PCNs的 lgK_{OA}的实验值及交互检验值与预测值实验值对比图

Fig 2 Polt of predicted and interactive-tested *vs.* experimental lgK_{OA} values of 24 PCNs by model 2

3 结论

本文采用MEDV对76个PCNs进行结构表征,通过多元线性回归建立了MEDV描述子与PCNs的lgK_{OA}的相关模型,并通过各模型(m = 2, 6)对其余52种未知PCNs的lgK_{OA}进行预测,取得了较为满意的结果.本文所建立的模型中结构参数的取值完全来自分子本身的结构,不需加入任何经验性的性质参数或校正参数,较为客观.本研究对全面评估多氯萘的环境行为和安全性有一定的意义.

参 考 文 献

[1] Jerzy Falandysz. Polychlorinated naphthalenes: an environmental update [J]. Environmental Pollution, 1998, 101: 77-90

[2] Tom Harner, Nicholas J L, Green Kevin C Jones. Measurements of octanol-air partition coefficients for PCDD/Fs: a tool in assessing air-soil equilibrium status [J]. Environ Sci Technol, 2000, 34: 3109-3114

[3] Terry F Bidleman, Paul A Helm, Birgit M Braune, et al. Polychlorinated naphthalenes in polar environments——A review [J]. Science of the Total Environment, 2010, 408: 2919-2935

[4] Tom Harner, Terry F Bidleman. Measurement of octanol-air partition coefficients for polycyclic aromatic hydrocarbons and polychlorinated naphthalenes [J]. J Chem Eng Data, 1998, 43: 40-46

[5] Qin L T, Liu H S, Liu H L, et al. A new predictive model for the bioconcentration factors of polychlorinated biphenyls (PCBs) based on the molecular electronegativity distance vector (MEDV) [J]. Chemosphere, 2008, 70: 1577-1587

[6] Chen Jingwen, Xue Xingya, Karl-Werner Schramm, et al. Quantitative structure-property relationships for octanol/air partition coefficients of polychlorinated naphthalenes, chlorobenzenes and p, p'-DDT [J]. Computational Biology and Chemistry, 2003, 27: 165-171

[7] M Staikova F, Wania b D J Donaldson. Molecular polarizability as a single-parameter predictor of vapour pressures and octanol-air partitioning coefficients of non-polar compounds: a priori approach and results [J]. Atmospheric Environment, 2004, 38: 213-225

[8] Tomasz Puzyn, Jerzy Falandysz. QSPR modeling of partition coefficients and henry's law constants for 75 chloronaphthalene congeners by means of six chemometric approaches—— a comparative study[J]. J Phys Chem Ref Dat, 2004, 36(1): 203-214

[9] 刘树深, 刘堰, 李志良, 等. 一个新的分子电性距离矢量(MEDV) [J]. 化学学报, 2000, 58(11): 353-1357

[10] 廖立敏, 梅虎, 郑怀礼, 等. 气中痕量挥发性有机物的结构表征和保留时间的估计与预测 [J]. 环境化学, 2007, 26(6): 838-840

[11] 仝建波, 李云飞, 刘淑玲, 等. 多氯代二苯并呋喃定量结构性质关系的研究 [J]. 计算机与应用化学, 2010, 27(2): 225-227

[12] 张亚辉, 刘征涛, 刘树深, 等. MEDV描述子预测取代芳烃类化合物的藻毒性 [J]. 环境科学研究, 2009, 22(7): 823-827

[13] 崔世海, 杨静, 刘树深, 等. 基于分子电性距离矢量预测有机污染物的生物富集因子 [J]. 中国科学 B, 2007, 37(3): 248-253

PREDICTING THE OCTANOL-AIR PARTITING COEFFICIENT OF PCNs USING MOLECULAR ELECTRONGATIVITY-DISTANCE VECTOR

LI Zhenghua XU Pan XIA Zhining

(College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400044, China)

ABSTRACT

Molecular electrongativity-distance vector (MEDV) was used to describe the chemical structures of 76 polychlorinated naphthalenes(PCNs). A reasonable quantitative relationship model between the lgK_{OA} of PCNs and the molecular electrongativity-distance vector (MEDV) was achieved by multiple linear regression (MLR). The results of significance test were quite satisfactory (n = 24, R = 0.997, SD = 0.088, F = 417.546). A more predictive model with high correlation coefficient (R = 0.996) was constructed by selecting two parameters form all the elements in the MEDV vectors of the former model through stepwise multiple regression (SMR). The performance of the two-parameter model was tested through cross-validation by the leave-one-out procedure (LOO) and satisfactory results were obtained (R_{CV} = 0.995). Then octanol-air partition coefficients (lgK_{OA}) of 52 PCNs with unknown experimental values were predicted by the models.

Keywords: polychlorinated naphthalenes(PCNs), molecular electrongativity-distance vector(MEDV), octanol-air partitioning coefficient(lgK_{OA}), quantitative structure-property relationship(QSPR).