文章编号:1000 - 2367(2017)01 - 0027 - 04 DOI:10.16366/j.cnki.1000 - 2367.2017.01.005

氯及溴代苯化合物生物富集因子预测

饶凡, 黄茜, 廖立敏

(内江师范学院 化学化工学院,四川 内江 641100)

摘 要:将有机化合物分子中的非氢原子分为四类,将不同非氢原子自身及非氢原子之间的关系参数化作为结构描述符,对部分氯苯及溴苯化合物分子结构进行了参数化表达.采用偏最小二乘回归(PLS)方法构建了化合物结构与其在鱼体内的富集因子(lgB_{CF})之间的关系模型,模型的建模相关系数(R^2)为 0.943,"留一法"交互检验的相关系数(Q^2)为 0.871.结果表明结构描述符能较好地表征化合物分子结构特征,所建模型稳定性好、预测能力强.

关键词:氯苯;溴苯;富集因子(lgB_{CF});结构与性质的关系(QSPR)

中图分类号:O625.2 **文献标志码**:A

氯代苯及溴代苯在化工、制药、制革等行业被广泛应用,从而成为环境中重要的污染物.该类化合物性质稳定,难以降解,在环境中可以长期滞留.氯代苯及溴代苯大多有毒,部分化合物还具有强的致癌、致畸及致突变的作用,严重威胁动植物生长、繁殖,通过食物链的积累最终威胁人类身体健康.生物富集因子(lgB_{CF})反映污染物在生物体内富集的程度,是评价有机污染物环境危害的常用参数之一.实验测定生物富集因子(lgB_{CF})成本高、周期长,通常采用 QSPR 方法来估算以获取相应的参数.在生物富集因子(lgB_{CF})的 QSPR 研究中,研究者们已经做过一些探索并取得优良的结果[1-3].在 QSPR 研究中化合物分子结构参数化表达是关键工作之一,目前研究者们在化合物分子结构参数化方面已经做过许多工作[4-7].本文将有机化合物分子中的非氢原子分为四类,将不同非氢原子自身及非氢原子之间的关系参数化作为结构描述符,对部分氯苯及溴苯化合物分子结构进行了参数化表达,进而采用偏最小二乘(PLS)回归方法构建了化合物生物富集因子(lgB_{CF})预测模型.模型拟合效果、稳定性及预测能力良好,可为芳烃类有机污染物的 QSPR 研究提供参考.

1 材料和方法

1.1 实验材料

选取 21 个含氯、溴苯类化合物为研究样本,化合物在鱼体内生物富集因子以 lgB_{CF} 表示,lgB_{CF} 实验值取自文献[8],按 lgB_{CF} 值的大小排序列于表 1.

1.2 实验方法

1.2.1 化合物分子结构表征

有机化合物的富集因子(lgB_{CF})除了与测量因素有关外,还与分子的结构相关,构成化合物原子种类、数目、原子之间的连接方式等都会影响其富集因子(lgB_{CF}).认为在分子结构隐氢图中,不同非氢原子及非氢原子之间的关系对化合物性质产生重要影响,而氢原子仅影响与其直接相连的非氢原子的染色值.首先将非氢原子按文献[6,7]方法分为四类,与 1、2、3、4 个其他非氢原子直接相连的非氢原子分别规定为第一、二、三、四类非氢原子,如与 3 个非氢原子相连的叔碳原子属于第三类非氢原子.然后采用文献[9]方法将非氢原子进行参数化染色,采用(1)式计算:

收稿日期:2016-08-23;**修回日期**:2016-11-05.

基金项目:四川省教育厅青年基金项目(13ZB0003);四川省科技厅应用基础项目(2015JY0077).

作者简介(通信作者):廖立敏(1981 -),男,湖南祁阳人,内江师范学院副教授,研究方向为化合物结构与性质的关系,E-mail: liaolimin523@126.com.

$$Z_i = [m_i(n_i - 1) - h_i]^{1/2}. \quad (1)$$

式中 i 为原子在分子中的编码,n_i 为非氢原子 i 的主量子数,m_i 为最外层电子数,h_i 为与其直接连接的氢原子数.

非氢原子自身对化合物性质的影响按(2)式计算:

$$x_k = \sum_{i \in K} Z_i \quad (k = 1,2,3,4). \quad (2)$$

式中,k 表示非氢原子 i 的原子类型,Z_i 按式(1) 计算. 根据非氢原子的分类,对于一个有机化合物分子中最多含有四类非氢原子,因此最终可得到 4 个非氢原子自身对化合物性质贡献项,用 x_1,x_2,x_3 和 x_4 表示.

对于非氢原子之间的关系对分子性质的影响采用(3)式计算,这种关系反映出非氢原子之间的相关程度随距离增减呈反向变化以及随原子性质改变呈正向变化.

$$x_r = m_{nl} = \sum_{i \in n, j \in l} \frac{Z_i \cdot Z_j}{r_{ij}^2}, (n = 1,2,3,4; n \leq l \leq 4). \quad (3)$$

Z 按(1) 式计算;r_{ij} 是非氢原子 i、j 之间的相对距离(即键长之和与碳碳单键键长的比值,如果 i、j 之间有多条路径,则以最短的为准);n 和 l 为原子所属类型. 化合物分子中四类非氢原子可以组合出以下 10 种关系项:m_{11},m_{12},$\cdots$,m_{44},简写为 x_5,x_6,$\cdots$,x_{14}. 这样最多将有 14 个变量(结构描述符) 来描述化合物结构信息.

1.2.2 建模与评价

采用偏最小二乘回归(PLS)建模,PLS 是近几十年发展起来的多元统计方法,在定量构效关系中广泛应用. PLS 特别适合于变量数较多而样本数较少的情况下进行建模,它通过对 X 和 Y 矩阵同时做双线性分解,并将分解所得潜隐变量再做一次最小二乘拟合以得到最终模型,详细原理请参见文献[10]. 以建模相关系数(R^2)、"留一法"交互检验相关系数(Q^2)及标准偏差(SD)对模型质量进行评价. 一般认为,建模相关系数 R^2 在 0.64 ~ 1.0 之间,"留一法"交互检验相关系数(Q^2) ≥0.50,标准偏差(SD)与数值范围之比在 10% 范围内,表明模型具有良好的预测能力和稳定性[11].

2 结果与讨论

由于研究样本中不含有第四类非氢原子,因而得到的结构描述符中与第四类非氢原子相关的 x_4,x_8,x_{11},x_{13},x_{14}这 5 个变量为全"0"项,剩余 9 个变量(列于表 1)用于建模分析. 将化合物结构描述符作为自变量 X,化合物富集因子($\lg B_{CF}$)作为因变量 Y,建立偏最小二乘(PLS)模型,模型主成分数(A)为 5,建模相关系数(R^2)为 0.943,处在 0.64 - 1.0 之间;交互检验的相关系数(Q^2)为 0.871,大于 0.50;标准偏差(SD)为 0.164,与数值范围(4.26 - 1.70 = 2.56)之比为 6.41%,小于 10% 的标准.

图 1 为 21 个样本在 PLS 前两个主成分得分空间散点图,可以发现 95% 以上的样本点都落在 95% 置信度 Hotelling T^2 椭圆置信圈内,说明化合物结构描述符能够恰当表现研究样本化合物分子结构特征,并在统计模型中得到正确反映,模型总体质量良好,可以用于分析影响富集因子($\lg B_{CF}$)的结构因素.

模型对每个样本的富集因子($\lg B_{CF}$)拟合程度的好坏,还可以从样本的残差进行分析,图 2 为化合物的富集因子($\lg B_{CF}$)模型计算值的标准化残差累积概率分布图,样本标准化残差基本服从正态分布,所有样本的标准化残差均小于 ±2 倍标准偏差,进一步说明模型拟合能力优良.

对模型进行了 20 次 Y 随机排序验证. 以原始变量 Y 和排序后的变量 Y 的相关系数对模型的 R^2 和 Q^2 作图(图 3),并作线性回归. 根据文献[12]判断标准,好的模型要求 R^2 和 Q^2 的截距要分别小于 0.300 和 0.050. 图中可以看到本文所建模型 R^2 和 Q^2 回归线的截距分别为:0.156 和 -0.415,因此可以认为模型的良好结果并非偶然因素所致. 分析各变量的标准系数发现,x_9 系数较大并且与富集因子($\lg B_{CF}$)Y 呈负相关,x_9 对应于第二类非氢原子之间的关系项,说明第二类原子少的化合物可能具有较大的富集因子($\lg B_{CF}$);x_1 系数也较大并且与富集因子($\lg B_{CF}$)Y 呈正相关,x_1 对应于第一类非氢原子自身染色值,说明第一类原子多的化合物可能具有较大的富集因子($\lg B_{CF}$). 而研究样本中的第一类非氢原子基本上都是氯或溴取代基,第二类非氢原子为苯环上未被取代的碳原子,第一类非氢原子(取代基)越多,意味着第二类非氢原子(苯环上未被取代的碳原子)就会越少. 因而苯环上取代基越多的化合物,可能具有较大的富集因子($\lg B_{CF}$),例如 10

号化合物苯环上碳原子全部被取代,因而表现出最大的富集因子(lgB_{CF})值.

表 1　化合物及其在鱼体内生物富集因子

序号	化合物	x_1	x_2	x_3	x_5	x_6	x_7	x_9	x_{10}	x_{12}	lgB_{CF}	Cal.	Err.
1	溴苯	4.582 6	8.660 3	2.000 0	0.000 0	5.678 3	5.957 8	19.234 6	11.105 1	0.000 0	1.70	1.94	0.24
2	氯苯	3.741 7	8.660 3	2.000 0	0.000 0	4.972 2	5.601 8	19.234 6	11.105 1	0.000 0	1.85	1.65	-0.20
3	1,2-二氯苯	7.483 3	6.928 2	4.000 0	1.355 2	6.884 9	14.736 3	13.300 5	13.704 1	4.911 0	2.48	2.53	0.05
4	1,4-二氯苯	7.483 3	6.928 2	4.000 0	0.555 7	9.075 9	12.206 4	10.026 5	21.265 0	0.545 7	2.52	2.56	0.04
5	1,3-二氯苯	7.483 3	6.928 2	4.000 0	0.826 1	8.465 8	12.910 9	10.538 1	20.083 7	1.227 7	2.65	2.51	-0.14
6	1,3-二溴苯	9.165 2	6.928 2	4.000 0	1.143 4	9.644 9	13.892 2	10.538 1	20.083 7	1.227 7	2.80	2.94	0.14
7	1,4-二溴苯	9.165 2	6.928 2	4.000 0	0.780 2	10.338 1	13.091 7	10.026 5	21.265 0	0.545 7	2.83	3.02	0.19
8	六溴苯	27.495 5	0.000 0	12.000 0	20.209 8	0.000 0	75.087 4	0.000 0	0.000 0	38.469 1	3.04	3.15	0.11
9	1,2-二溴苯	9.165 2	6.928 2	4.000 0	1.834 8	7.899 4	15.907 8	13.300 5	13.704 1	4.911 0	3.10	2.88	-0.22
10	1,2,3-三氯苯	11.225 0	5.196 2	6.000 0	3.536 6	7.319 2	25.578 1	8.287 2	14.176 7	11.049 6	3.11	3.19	0.08
11	1,2,4-三氯苯	11.225 0	5.196 2	6.000 0	2.737 1	9.510 1	23.048 2	5.013 3	21.737 6	6.684 4	3.26	3.22	-0.04
12	1,3,5-三氯苯	11.225 0	5.196 2	6.000 0	2.478 4	10.480 9	21.927 3	2.762 4	26.935 7	3.683 2	3.28	3.16	-0.12
13	1,2,3,5-四氯苯	14.966 6	3.464 1	8.000 0	5.744 6	8.465 8	35.597 3	0.920 8	20.083 6	14.050 8	3.36	3.66	0.30
14	1,2,4-三溴苯	13.747 7	5.196 2	6.000 0	3.758 4	10.847 4	25.018 3	5.013 3	21.737 6	6.684 4	3.66	3.59	-0.07
15	1,2,4,5-四氯苯	14.966 6	3.464 1	8.000 0	5.474 2	9.075 9	34.892 8	0.409 2	21.265 0	13.368 7	3.76	3.71	-0.05
16	1,2,3,4-四氯苯	14.966 6	3.464 1	8.000 0	6.273 7	6.884 9	37.422 7	3.683 2	13.704 1	17.734 0	3.77	3.68	-0.09
17	1,2,4,5-四溴苯	18.330 3	3.464 1	8.000 0	7.516 8	10.338 1	38.120 9	0.409 2	21.265 0	13.368 7	3.79	3.89	0.10
18	1,3,5-三溴苯	13.747 7	5.196 2	6.000 0	3.430 2	11.899 7	23.803 1	2.762 4	26.935 7	3.683 2	3.85	3.56	-0.29
19	五氯苯	18.708 3	1.732 1	10.000 0	9.836 9	4.972 2	50.974 6	0.000 0	11.105 1	25.646 1	3.86	3.95	0.09
20	2,4,5-三氯甲苯	12.225 0	3.464 1	8.000 0	3.533 8	7.505 6	29.121 9	0.409 2	21.265 0	13.368 7	3.87	3.99	0.12
21	六氯苯	22.449 9	0.000 0	12.000 0	14.755 4	0.000 0	68.059 1	0.000 0	0.000 0	38.469 1	4.26	4.02	-0.24

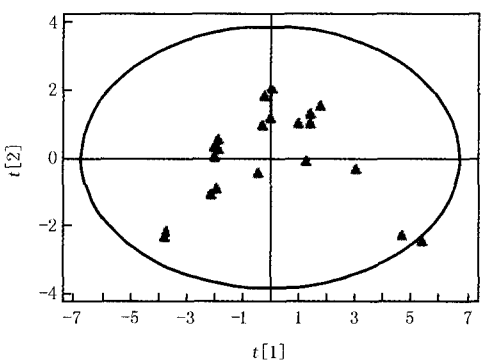

图1　样本在前2个主成分得分分布

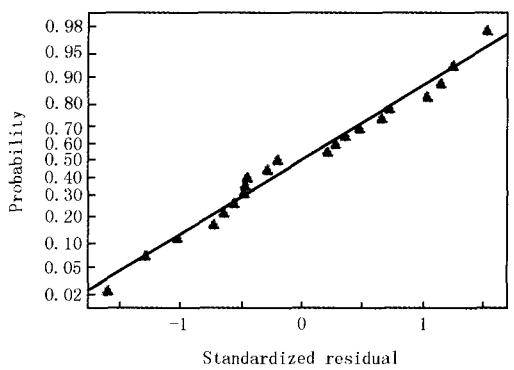

图2　标准化残差累积概率分布图

PLS 模型对全部化合物的富集因子(lgB_{CF})进行了预测,预测值列于表 1 的 Cal. 列,误差列入 Err. 列. 图 4为计算值与实验值的相关图,图中可以发现绝大部分样本点都分布在过原点的平分线附近,说明模型对化合物的富集因子(lgB_{CF})预测值准确性高、误差小,效果优良.

3　结　论

对部分含氯、溴的芳烃化合物结构进行了参数化表达,进而采用偏最小二乘回归(PLS)方法构建了该类化合物定量结构-生物富集因子(lgB_{CF})的 QSPR 模型. 模型经检验,具有可接受的预测能力与总体稳健性,

可以用于含芳烃类化合物生物富集因子(lgB_{CF})预测. 本文所采用的分子结构描述符与文献[3-5]相比,不需要考虑分子构象优化等,因而具有简单易懂、计算方便、计算工作量小等优点,由于其基于分子二位平面结构计算得到,因而也有所不足,如不能反映分子的立体结构特征、不能区分顺反异构体、对于一些特殊的立体结构特征还难以表达,这些将在今后的研究中予以克服. 本文可为环境中有机污染物的 QSPR/QSAR 研究提供一种新的方法,具有一定的参考价值.

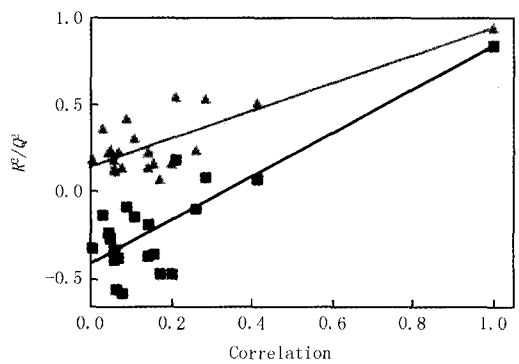

图3 Y随机排序验证结果

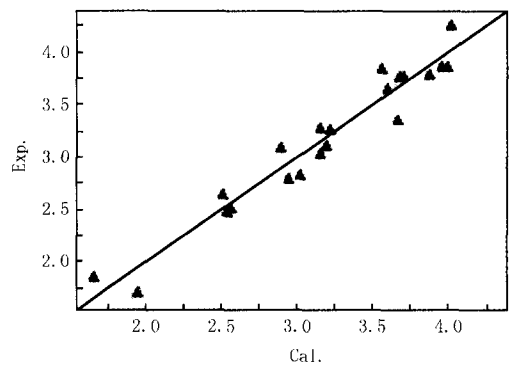

图4 实验值与计算值相关图

参 考 文 献

[1] 冯惠,李鸣建. 部分多氯联苯生物富集因子的 QSAR 研究[J]. 环境科学与技术,2013,36(11):49-53.

[2] 秦红,陈景文,王莹,等. 有机污染物生物富集因子定量预测模型的建立与评价[J]. 生态毒理学报,2009,54(1):27-32.

[3] 郑玉婷,乔显亮,杨先海,等. 卤代有机化合物生物富集因子的定量结构-活性关系模型[J]. 生态毒理学报,2013,8(5):772-777.

[4] 汪斌,常自超,舒茂,等. 氰基吡咯烷类 FAP 抑制剂的分子对接及 3D-QSAR 研究[J]. 化学研究与应用,2016,28(7):47-53.

[5] Shu M,Zhang Y R,Tian F F,et al. Molecular Docking and 3D-QSAR Research of Biphenyl Carboxylic Acid MMP-3 Inhibitors[J]. Chinese J. Struct. Chem. ,2012,31(3):443-451.

[6] 廖立敏. 醛酮化合物结构与保留指数关系的研究[J]. 化学研究与应用,2015,27(5):617-623.

[7] 李建凤,谢永红,雷光东. 部分聚合物结构与热容变化关系研究[J]. 计算机与应用化学,2016,33(7):833-837.

[8] Lu X X,Tao S,Hu H Y,et al. Estimation of bioconcentration factors of nonionic organic compounds in fish by molecular connectivity indices and polarity correction factors [J]. Chemosphere,2000,41(10) : 1675-1688.

[9] 王晓荣,堵锡华. mB 对气相色谱保留指数的 QSPR 研究[J]. 南京工业大学学报,2002,24(3):31-37.

[10] 梅虎,周原,孙立力,等. 氨基酸结构描述子矢量 VHSE 及其在肽 QSAR 中的应用[J]. 化学通报,2005,68(7):534-540.

[11] 顾云兰,陈鑫,简美玲. 苯胺类化合物结构与毒性的密度泛函理论研究[J]. 化学研究与应用,2015,27(8):1139-1144.

[12] Andersson P M,Sjöstrom M,Lundstedt T. Preprocessing peptide sequences for multivariate sequence-property analysis[J]. Chemometr Intell Lab Syst,1998,42:41-50.

Bioconcentration Factors Prediction for Chlorobenzene and Bromobenzene Compounds

Rao Fan, Huang Xi, Liao Limin

(College of Chemistry and Chemical Engineering, Neijiang Normal University, Neijiang 641100, China)

Abstract: The organic molecule non-hydrogen atoms were grouped into four categories, the different non-hydrogen atoms and the relationship between them were used as structural descriptors to parameterize the molecular structure of some chlorobenzene and bromobenzene compounds. The partial least squares regression (PLS) method was used to construct a model of relationship between the structures and bioconcentration factors (lgB_{CF}) of the compounds. The correlation coefficient (R^2) was 0.943, "leave one out" cross validation correlation coefficient (Q^2) was 0.871, respectively. The results showed that the structural descriptors could characterize the molecular structures of the compounds well, the stability of the model was good, and the predictive power was strong.

Keywords: chlorobenzene; bromobenzene; bioconcentration factors (lgB_{CF}); structure and property relationship (QSPR)