

文章编号: 1007- 2985(2008) 01- 0095- 06

预测取代芳烃生物降解性的分子形状及连接性模型^{*}

关丽娜, 朱 南, 周 军, 冯长君

(徐州师范大学化学系, 江苏 徐州 221116)

摘 要: 应用分子连接性指数(${}^mX^v$)及分子形状指数(mK)分析影响 42 种取代芳烃在活性污泥中的生化需氧量(BOD), 建构了一个 5 变量的 QSBR 模型, 其可决系数(R^2)为 0.482, 估算标准误差为 10.706, 具有良好的稳健性与预测能力. 该模型比片段常数模型($R^2=0.223$)及文献[3]的人工神经网络(ANN)模型($R^2=0.427$)更为精确.

关键词: 取代芳烃; 生化需氧量; 连接性指数; 分子形状指数; 定量结构- 生物降解性模型

中图分类号: TQ413.21⁺3

文献标识码: A

取代芳烃大多为异生物合成物(xenobiotics), 在化工、医药、农药等领域具有广泛应用. 这些化合物可以通过多种途径进入环境, 直接造成生态环境的破坏; 甚至通过生物富集进入食物链, 经生物放大作用, 最终严重危害人类健康. 生物降解主要是指在微生物作用下使外源性物质的结构发生改变的过程, 它是有机污染物最重要的降解过程之一. 通常将生物降解分为 2 类^[1]: 其一是初始生物降解, 即改变母体化合物分子完整性的任何生物诱导的结构转化过程; 其二是最终生物降解或矿化过程, 即通过生物催化由有机物到无机物的代谢过程. 有机物在好氧条件下的最终生物降解能力可用其在一定时间内的生化需氧量(BOD)予以表征^[2].

陆光华等^[3]对含有烷基、羟基、羧基、硝基、羰基的 42 种取代芳烃的生化需氧量(BOD)进行了构效关系研究, 其生物降解数据摘自文献[4]. 他们使用量子化学 MOPAC6.0-AM1 法计算出这些化合物的 EHO-MO(分子最高占有轨道能)、 H_f (摩尔生成热)、TSA(分子表面积)、 μ (分子的偶极矩)以及 M_w (分子量)和 $\log P$ (P 为正辛醇/水分配系数), 将这 6 种参数与 BOD 进行多元线性回归及人工神经网络(ANN)研究, 获得有意义的研究结果. 其二元线性回归方程及 ANN 模型的估算值与实验值的相关系数(R^2)分别为 0.127, 0.427(笔者拟合). 笔者利用 Kier 等^[5]的分子连接性指数(${}^mX^v$)及其分子形状指数(mK)^[6]与这 42 个取代芳烃的 BOD 拟合, 经 spss13.0 统计软件系统给出的五元线性回归模型, 其 R^2 为 0.482. 该模型经 Jackknife 法^[7]检验, 具有良好的稳健性及预测能力. 研究结果不仅好于基团贡献法(片段常数法)^[8]($R^2=0.223$), 也优于文献[3]的 ANN 结果.

1 拓扑指数的建立与 BOD 数据来源

拓扑指数是对分子结构的数值化表征^[9-10], 因此它源于分子隐氢图的各种数学矩阵: 距离矩阵、邻接矩阵等, 经一定的矩阵运算便可衍生出分子结构的描述符, 即揭示分子结构某种特征的拓扑指数.

* 收稿日期: 2007- 11- 16

基金项目: (清华大学)环境模拟与污染控制国家重点联合实验室开放基金项目(KJ2007001); 徐州师范大学培育课题(05PYL04); 徐州师范大学学生科研立项课题

作者简介: 关丽娜(1986-), 江苏徐州人, 徐州师范大学化学系学生

通讯作者: 冯长君(1954-), 男, 教授, 主要从事物质构效学研究. E-mail: fengcixznu@xznznu.edu.cn.

1.1 分子连接性指数

分子连接性指数(${}^m X_i^v$)是由 Kier, Hall 等在 Randic^[11] 的碳原子支化度指数(${}^m X_i$) 根据数学图在分子隐氢图的邻接矩阵基础上衍生出的分子描述符,用以揭示分子的大小、分支及刚性等结构特征.其核心概念之一是原子点价(δ_i^v):

$$\delta_i^v = (Z_i^v - h_i)(Z_i - Z_i^v - 1)^{-1}. \quad (1)$$

其中: Z_i^v , Z_i 分别是成键原子 i 的价电子数和电子总数; h_i 是与原子 i 直接键合的氢原子数,对于羟基中氧原子,其 $\delta_o^v = 5$.

根据非氢原子的 δ_i^v 定义 ${}^m X_i^v$ 为若干个相邻原子的点价乘积平方根的倒数:

$${}^m X_i^v = \sum_j^m S_j = \sum \prod (\delta_i^v)^{-0.5}. \quad (2)$$

其中: m 为 MCI 的阶数, $m = 0, 1, \dots$; t 代表子图的种类,即对不同分子根据需要可剖析成若干种局部拓扑图,常用有如图 1 所示的几种.

图 1 中的 a, b, c, d 分别称为链、星、星-链、环,依次对应 t 为 P, C, PC 及 CH , 笔者使用 Matlab 程序^[12] (经过修正) 计算标题化合物(见表 1) 的 8 个连接性指数: ${}^0 X_p^v, {}^1 X_p^v, {}^2 X_p^v, {}^3 X_p^v, {}^4 X_p^v, {}^5 X_p^v, {}^3 X_c^v, {}^4 X_{pc}^v$.

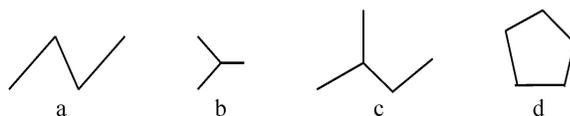


图 1 分子结构的基本子图的类型

1.2 分子形状指数(${}^m K$)

表征分子形状的参数常用分子体积(V)、表面积(S),但其本能只是反映分子的大小.对于实验测定 2 个分子的 V 或 S 相同,但其形态未必相同.目前建立描述分子形状参数侧重于量子化学法与拓扑指数法,后者以其简便性及在构效关系中常可获得良好的相关性,倍受世人青睐.表征分子形状的拓扑指数常用 Kier 的 Kappa 形状指数(${}^m K$).

1.2.1 烷烃的 Kappa 形状指数 对于烷烃及其环烷烃分子中碳原子均为 sp^3 杂化,其分子隐氢图中的每条边均是 C—C 单键.据此结构特征,在分子邻接矩阵基础上, Kier 定义 Kappa 形状指数(${}^m K$):

$${}^m K = [n_j - (m - 1)] \cdot [n_j - m]^{2 \cdot m} p_j^{-2}. \quad (3)$$

其中: m 是 Kappa 形状指数的阶数,本文只讨论 $m = 1, 2$ 情形; n_j 是分子 j 中非氢原子数; ${}^m P_j$ 为分子 j 中具有路径数为 m 的个数.由(6)式可得:

$${}^1 K = n_j(n_j - 1)^2 / ({}^1 P_j)^2, \quad (4)$$

$${}^2 K = (n_j - 1)(n_j - 2)^2 / ({}^2 P_j)^2. \quad (5)$$

例如甲基环己烷(A),其碳原子数 $n^v = 7$, ${}^1 P = 7$, ${}^2 P = 8$,代入(4), (5) 式得:

$${}^1 K = \frac{7 \times (7 - 1)^2}{7^2} = 5.143, \quad {}^2 K = \frac{(7 - 1)(7 - 2)^2}{8^2} = 2.344.$$

对于直链烷烃(不含支链), (4), (5) 式可还原为: ${}^1 K = n_j$, ${}^2 K = n_j - 1$.

1.2.2 含多重键及杂原子分子的 Kappa 形状指数 为了扩展烷烃的 Kappa 形状指数的适用范围, Kier 于 ${}^m K$ 公式中引入修正项 α , 分子 j 中原子 i 的修正值 α_i 定义为:

$$\alpha_i = r_i \cdot r^{-1} - 1. \quad (6)$$

其中: r 为 sp^3 杂化碳原子的原子半径($r = 77$ nm); r_i 为非 sp^3 杂化碳原子或其它杂原子的原子半径.例如羰基中的氧原子,其 sp^2 杂化的原子半径为 62 nm, 则 $\alpha_{(sp^2)} = \frac{62 \text{ nm}}{77 \text{ nm}} - 1 = -0.20$.

分子 j 的修正值 α 则为所有的非 sp^3 杂化碳原子及杂原子的 α_i 之代数和: $\alpha_j = \sum \alpha_i$. 例如硝基苯的修正值 α :

$$\alpha = -0.13 \times 6 - (-0.20) + (-0.20) \times 2 = -1.38.$$

将 α 引入 ${}^m K$ 中的计算公式:

$${}^m K = (n_j + \alpha_j - (m - 1)(n_j + \alpha_j - 1)^2 \cdot ({}^1 P_j + \alpha_j)^{-2}. \quad (7)$$

同理

$$k = (n_j + \alpha)(n_j + \alpha - 1)^2 \cdot ({}^1P_j + \alpha)^{-2}, \quad (8)$$

$${}^2k = (n_j + \alpha - 1)(n_j + \alpha - 2)^2 \cdot ({}^2P_j + \alpha)^{-2}. \quad (9)$$

如硝基苯的 ${}^1K, {}^2K$:

$$K = (9 - 1.38)(9 - 1.38 - 1)^2 \cdot (14 - 1.38)^{-2} = 2.0968,$$

$${}^2K = (9 - 1.38 - 1)(9 - 1.38 - 2)^2 \cdot (11 - 1.38)^{-2} = 2.2593.$$

其它分子的 1K 及 2K 见表 1.

表 1 42 种化合物的部分拓扑指数

No.	Compounds	χ_p^v	χ_c^v	χ_{pc}^v	k	2k
1	对甲基异丙苯	6.89	0.55	0.83	4.17	2.87
2	α, α -三氯甲苯	6.23	0.47	0.76	3.85	3.09
3	α -氯甲苯	5.23	0.12	0.27	2.88	2.73
4	间苯二酚	4.20	0.15	0.16	2.62	1.94
5	邻苯二酚	4.20	0.13	0.21	2.62	1.94
6	邻甲氧基苯酚	5.16	0.12	0.25	3.34	2.62
7	对氨基苯甲醚	5.30	0.16	0.25	3.34	2.62
8	乙酸苯甲酯	6.41	0.20	0.24	4.02	3.88
9	苯乙醇	5.25	0.12	0.22	3.37	3.25
10	对异丙基苯甲醇	7.04	0.50	0.83	4.94	3.54
11	苯甲醛	4.37	0.10	0.15	2.02	2.26
12	苯乙酮	6.00	0.16	0.34	3.33	3.16
13	间苯二甲酸二甲酯	7.94	0.25	0.50	5.38	4.64
14	对苯二甲酸二甲酯	7.94	0.25	0.51	5.38	4.64
15	间苯二甲酸	6.02	0.26	0.43	3.93	3.25
16	对苯二甲酸	6.02	0.26	0.44	3.93	3.25
17	对异丙基苯甲酸	7.24	0.51	0.86	4.77	3.50
18	苯甲酸	4.74	0.13	0.22	2.62	2.42
19	<i>N</i> -甲基-2-氨基苯甲酸	6.16	0.19	0.42	3.99	3.27
20	对硝基苯甲酸	5.93	0.24	0.41	3.32	3.10
21	邻羟基苯甲酸	5.11	0.18	0.33	3.27	2.60
22	对甲基叔丁苯	7.81	1.50	1.40	4.97	2.69
23	乙酰替苯胺	5.80	0.19	0.19	3.30	3.13
24	<i>N</i> -乙酰乙酰基对甲苯胺	8.33	0.47	0.47	5.41	4.67
25	α -氨基甲苯	4.67	0.12	0.20	2.65	2.48
26	α -乙酰氨基甲苯	6.50	0.22	0.27	4.02	3.88
27	邻甲基苯酚	4.76	0.21	0.37	2.65	1.97
28	苯甲醚	4.80	0.07	0.15	2.65	2.48
29	对羟基苯甲醚	5.16	0.14	0.23	3.34	2.62
30	<i>N</i> -乙酰基-2-氨基苯酚	6.16	0.24	0.31	3.99	3.27
31	邻氨基苯甲醚	5.30	0.14	0.29	3.34	2.62
32	2,4,6-三氯苯酚	7.00	0.57	0.85	4.83	2.97
33	苯乙烯	4.67	0.10	0.18	2.06	2.31
34	对氨基苯甲醚	5.30	0.16	0.26	3.34	2.62
35	苯基乙胺	5.59	0.08	0.16	3.37	3.25
36	4-烯丙基-1,2-苯二甲醚	8.04	0.24	0.49	5.58	4.80

续表

No.	Compounds	χ_p^v	χ_c^v	χ_{pc}^v	K	\bar{K}
37	苯甲醇	4.54	0.12	0.19	2.65	2.48
38	邻甲基苯甲酸	5.66	0.26	0.51	3.30	2.62
39	邻苯二甲醚	6.13	0.12	0.30	4.11	3.34
40	邻硝基苯甲酸	5.93	0.22	0.43	3.32	3.10
41	对苯二酚	4.20	0.15	0.17	2.62	1.94
42	苯酚	3.83	0.07	0.09	1.97	1.76

1.2.3 mK 中蕴含的分子形状信息 由表 1 中各种有机物分子的 1K 及 2K 可见:

(1) 1K 所显示的结构信息主要为分子结构的复杂度,确切地说是反映分子的环性.笔者提出分子环比(β)予以表征, (β) 定义为分子中所含的环数(g_j)与其原子数之比:

$$\beta = g_j \cdot n_j^{-2}. \quad (10)$$

如苯酚、萘酚的分别为 1/7, 2/11, 其 1K 依次为 1.967 7, 2.954 1.

(2) 2K 主要揭示分子中原子的空间密度.因为在原子数 n_j 相同下,分子越似球体,其 2P 值越大,该分子的体积越小,相应分子的原子空间密度越大,据(9)式,该分子的 2K 越小.

1.3 BOD 数据来源

这 42 个有机物的生物降解数据来源于文献[4].此是用驯化的活性污泥为降解微生物,测定它们的 5 日生化需氧量(BOD_5)及最后一天的生化需氧量(BOD_u),令

$$BOD = (BOD_5 / BOD_u) \times 100 \quad (11)$$

作为描述生物降解性的参数.对于一般化合物而言,其 BOD 值越大,相应的生物降解性越强.表 2 列出本文 42 个有机物的 BOD.

表 2 42 种化合物的 BOD 的实验值与计算值

No.	Compounds	BOD			No.	Compounds	BOD		
		Exp.	Cal.	Er./%			Exp.	Cal.	Er./%
1	对甲基异丙苯	88	66	- 25	22	对甲基叔丁苯	52	56	8
2	α, α - 三氯甲苯	90	87	- 4	23	乙酰替苯胺	69	64	- 7
3	α - 氯甲苯	71	71	0	24	N- 乙酰乙酰基对甲苯胺	74	72	- 3
4	间苯二酚	67	70	4	25	α - 氨基甲苯	64	74	16
5	邻苯二酚	83	73	- 12	26	α - 乙酰氨基甲苯	77	79	3
6	邻甲氧基苯酚	90	77	- 15	27	邻甲基苯酚	50	65	29
7	对氨基苯甲醚	65	71	9	28	苯甲醚	56	68	21
8	乙酸苯甲酯	95	81	- 14	29	对羟基苯甲醚	86	75	- 13
9	苯乙醇	87	91	5	30	N- 乙酰基- 2- 氨基苯酚	60	74	24
10	对异丙基苯甲醇	85	95	11	31	邻氨基苯甲醚	55	74	35
11	苯甲醛	66	65	- 2	32	2, 4, 6- 三氯苯酚	83	77	- 7
12	苯乙酮	67	69	4	33	苯乙烯	53	59	11
13	间苯二甲酸二甲酯	99	90	- 9	34	对氨基苯甲醚	65	71	10
14	对苯二甲酸二甲酯	84	91	8	35	苯基乙胺	90	76	- 15
15	间苯二甲酸	78	84	8	36	4- 烯丙基- 1, 2- 苯二甲醚	89	95	7
16	对苯二甲酸	75	85	13	37	苯甲醇	94	78	- 17
17	对异丙基苯甲酸	89	86	- 4	38	邻甲基苯甲酸	94	72	- 24
18	苯甲酸	85	70	- 17	39	邻苯二甲醚	96	82	- 15
19	N- 甲基- 2- 氨基苯甲酸	85	83	- 3	40	邻硝基苯甲酸	100	74	- 26
20	对硝基苯甲酸	62	72	16	41	对苯二酚	70	71	1
21	邻羟基苯甲酸	88	80	- 9	42	苯酚	85	63	- 26

2 结果与讨论

将文献[4]中42种取代芳烃的BOD及相应的8个分子连接性指数(${}^m X_i^v$)及2个分子形状指数(${}^m K$)输入SPSS统计软件,经多元线性回归得:

$$\text{BOD} = b_0 + \sum b_k E_k, n' = 42, Q^2 = 0.528, F = 2.54, s = 11.38. \quad (12)$$

其中: E_k 代表 ${}^m X_i^v, {}^m K$; n', Q^2, F, s 依次为样本容量、可决系数、Fischer 检验值及估计标准误差。为了使建立的BOD模型具有稳定性及可靠性,所选变量的个数应满足统计学规则:样本容量与变量个数之比应大于或等于5。在本实验中样本容量为42,据此规则选择5个变量的拓扑指数集(其数据见表1。限于篇幅,未能列出其余拓扑指数)比较合适。经向后逐步回归法建立的5元数学模型:

$$\text{BOD} = 100.733 - 15.670 {}^0 X_p^v - 51.734 {}^2 X_p^v - 67.413 {}^3 X_c^v + 132.618 {}^4 X_{pc}^v + 33.641 {}^2 K, \\ n' = 42, Q^2 = 0.365, F = 4.141, s = 11.999. \quad (13)$$

为了与文献[3]作比较,本文对前37个标题化合物进行逐步回归:

$$\text{BOD} = 100.733 - 33.363 {}^0 X_p^v - 24.650 {}^3 X_c^v + 63.244 {}^4 X_{pc}^v + 16.574 {}^1 K + 29.788 {}^2 K, \\ n' = 37, Q^2 = 0.482, F = 5.77, s = 10.706. \quad (14)$$

按(14)式给出的计算值见表2中的“Cal.”(其中“Er.”为相对误差),它与相应实验值基本吻合。该式给出的37个化合物的平均估算误差仅为10.73%,小于文献[3]ANN法的11.4%(本文按其计算值与相应实验值拟合得其 $R^2 = 0.436, S = 10.515$);但对余下5个化合物BOD的平均预测误差为18.40%,大于文献[3]ANN法的13.0%。

为了进一步检验模型(14)的稳健性,采用Jackknife的逐组剔除法(因 $n' > 30$):即每次分别剔除个数为1, 2, 3, ..., 0的化合物,用余下34个或33个化合物的数据建模。重复10次,结果列于表3。

表3 BOD的Jackknife法检验

剔除的化合物	1	2	3	4	5	6	7	8	9	0
Q^2	0.584	0.429	0.407	0.478	0.518	0.507	0.469	0.505	0.483	0.507
F	7.592	4.059	3.708	4.947	5.802	5.553	4.771	5.715	5.223	5.748
S	9.63	11.22	11.21	11.28	10.74	10.87	10.46	10.31	10.99	10.54

由表3可见: Q^2 的波动范围较小,基本呈正态分布。说明(14)式具有总体可接受的稳健性及良好的预测能力。

近年来,物质的定量结构——生物降解性的相关性(QSBR)研究已成为环境科学的一个重要分支。进行有机物的QSBR研究的关键是其生物降解性能数据的测定及选用合适的分子结构描述符。根据生物降解机理^[13],影响其因素有物质的溶解性、疏水性以及分子的大小、电荷分布、空间排列等^[14]。因此,能够表征这些因素的如电性参数、空间参数、疏水性参数等便成为进行QSBR研究中常用的分子结构描述符。

本文42种有机物所呈现出的疏水性、电性及空间因素显然是由其分子结构特征所确定的。这些因素在模型(14)中是以 ${}^0 X_p^v, {}^3 X_c^v, {}^4 X_{pc}^v, {}^1 K, {}^2 K$ 予以表征的,它们共同揭示了影响其生物降解性的本质因素。因此,该模型对这些有机物的BOD不仅呈现较高的相关性,而且具有良好的稳健性及预测能力,优于文献[3]的ANN法研究结果。

综上所述,分子连接性指数(${}^m X_i^v$)及分子形状指数能对有机分子实现量化表征,它们较为全面地揭示了影响有机物BOD的主要因素,可用于其它取代芳烃生物降解性的估算与预测。目前,国内对分子形状指数的研究较少,有必要对其进一步开发与应用。

参考文献:

- [1] 陆光华,周燕洁. 苯甲酸、苯酚及苯胺类在自然水体中的生物降解速率[J]. 河海大学学报:自然科学版,2004,32(6):

611-613.

- [2] 赵元慧, 杨邵贵. 松花江中取代苯酚和苯胺类的生物降解性及 QSBR 研究 [J]. 环境科学学报, 2002, 22(1): 45-50.
- [3] 陆光华, 赵元慧, 程香菊, 等. 取代芳烃的生物降解性与结构相关性研究 [J]. 环境化学, 2002, 21(1): 62-67.
- [4] DEARDEN J C, NICBOLSON R M. The Prediction of Biodegradability by Use of Structure-Activity Relationships: Correlation of Biological Oxygen Demand with Atomic Charge Difference [J]. Pestic. Sci., 1986, 17: 305-310.
- [5] KIER L M, HALL L H. Molecular Connectivity in Chemistry and Drug Research [M]. New York: Academic Press, 1976: 82.
- [6] 许禄, 胡昌玉. 应用化学图论 [M]. 北京: 科学出版社, 2000: 247-258.
- [7] 沐来龙, 冯长君. 环价连接性指数与饱和烃沸点的 QSPR 研究 [J]. 化工学报, 2004, 55(10): 1702-1705.
- [8] 陈刚. 定量结构与生物降解性能关系的研究及应用 [J]. 武汉理工大学学报: 交通科学与工程版, 2004, 28(3): 459-462.
- [9] 李鸣建, 冯长君. 取代芳烃对剑尾鱼和稀有句鲫的生物活性模型 [J]. 吉首大学学报: 自然科学版, 2006, 27(5): 102-105.
- [10] 杨春峰, 冯长君, 沐来龙. 取代苯甲酸生物降解速率常数的拓扑计算 [J]. 吉首大学学报: 自然科学版, 2006, 27(5): 106-108.
- [11] RANDIC M. On Characterization of Molecular Branching [J]. J. Amer. Chem. Soc., 1975, 69(1): 17-21.
- [12] 胡黔南, 梁逸曾, 王亚丽. 直观队列命名法的基本原理及其在矩阵与拓扑指数计算中的应用 [J]. 计算机与应用化学, 2003, 20(4): 386-389.
- [13] 徐向阳, 任艳红, 黄绚, 等. 典型有机污染物微生物降解及其分子生物学机理的研究进展 [J]. 浙江大学学报: 农业与生命科学版, 2004, 30(6): 684-689.
- [14] 瞿福平, 杨义燕, 冯旭东, 等. 定量结构-生物降解性能关系研究原理及进展 [J]. 中国环境科学, 1999, 19(1): 18-21.

Molecular Shape and Connectivity Models for Predicting Biodegradability of Substituted Aromatic Compounds

GUAN Lina, ZHU Nan, ZHOU Jun, FENG Changjun

(School of Chemistry and Chemical Engineering, Xuzhou Normal University, Xuzhou 221116, Jiangsu China)

Abstract: Five variables quantitative structure-biodegradability relationship (QSBR) model was established applying Kier's molecular connectivity indices (${}^mX_i^v$) and shape indices (mK) simulation analysis of the biochemical oxygen demand (BOD) of 42 substituted aromatic compounds with (by) activated sludge. The model determined that coefficient (R^2) is 0.482, and standard deviation is 10.706, which possessed robustness and predictability. The model was superior to fragment constant model ($R^2 = 0.223$) and the neural network approach ($R^2 = 0.427$) of literature [3].

Key words: substituted aromatic compounds; biochemical oxygen demand (BOD); molecular connectivity indices; molecular shape indices; quantitative structure-biodegradability relationship (QSBR) model

(责任编辑 易必武)