

苯的衍生物的 HQSAR 研究

朱少敏

(大连交通大学环境科学与工程学院 ,116028 ,辽宁 ,大连)

[摘要] 应用分子全息结构描述子表征 69 种苯的衍生物建立相关 QSAR 模型 ,并且对分子全息产生过程中的碎片区分参数及碎片长度与模型质量的关系进行了探讨 . 在最佳全息条件下产生的模型相关系数 r^2 为 0.997、交叉验证相关系数 $q_{1.00}^2$ 为 0.841 ,研究结果表明所建立模型具有良好的拟合效果 ,同时具有较高的预测能力 .

[关键词] 苯的衍生物 ,分子全息 ,交叉验证 ,预测

[中图分类号] X132 ,O625 , [文献标识码] A , [文章编号] 1001-4616(2005)02-0067-04

Holographic QSAR of Benzene Derivatives

Zhu Shaomin

(College of Environmental Science and Engineering , Dalian Jiaotong University , 116028 , Dalian , China)

Abstract Based on molecular hologram QSAR model is derived to predict pLC_{50} of benzene derivatives in the fathead minnow (*Pimephales promelas*) , The factors that influence the quality of QSAR models are discussed. Under the best molecular hologram conditions , the model yields a $q_{1.00}^2$ of 0.841 and high predictive ability. The results show that HQSAR technique is suitable for use in screening large database of chemicals.

Key words benzene derivatives , holographic QSAR , variable selection , prediction

0 引言

定量活性相关技术(QSAR)在环境科学及药物合成中取得了广泛应用 ,一般而言 ,经典的 Hunsch 方法通常运用一些物理化学结构参数(Hammett 电子取代常数、正辛醇/水分配系数、摩尔折光率、轨道能量、原子电荷等) ,这些描述子之间经常存在明显的线性关系 ,其数值往往也需要实验测定或通过理论计算^[1] . 研究发为为了得到稳健的预测模型 ,参数的选择必须非常谨慎 ,因此应用该方法建立模型常需要花费较多的时间^[2] . 近年来以比较分子场分析(CoMFA)为代表的三维 QSAR 得到了迅速发展 ,并且在探讨作用机理、分子的优化设计等方面取得了巨大成功 ,然而在建立模型过程中必须确定分子的三维结构、生物活性构象及分子叠合规则 ,这些因素又制约了该技术的广泛应用 . 分子全息是新型的分子结构表征技术 ,其主要原理是将化合物的生物活性与其分子亚结构碎片类型所描述的分子结构之间建立相关模型 ,从而预测化合物的生物活性 . 由于分子全息描述子能自动、迅速地产生 ,并且建立模型过程中不需要分子的三维结构 ,目前在环境科学领域取得了广泛应用^[3,4] . 本文应用该类描述子表征苯的衍生物的分子结构 ,研究了 69 种化合物对黑头呆鱼(*Pimephales promelas*)的急性毒性(半致死浓度 LC_{50}) ,并建立了相应预测模型 ,结果表明所建模型具有良好的拟合效果 ,同时具有较高的预测能力 .

1 材料与方法

1.1 化合物及活性数据

69 种苯的衍生物结构见表 1 ,其半致死浓度(LC_{50})引自文献^[5] .

收稿日期 :2005-03-02.

作者简介 :朱少敏 ,女 ,1974— ,讲师 ,博士研究生 ,主要从事环境化学的教学与研究 ,E-mail :smzhu@dicp. ac. cn

万方数据

1.2 结构活性相关分析

分子模拟及统计分析采用 SYBYL 分子模拟软件在 SGI INDIG O2 工作站上进行 ,利用 Sketch 选项构建分子二维结构 ,选择 Gasteiger-Hucker 法计算部分原子电荷. 使用的碎片区分参数包括原子类型(A)、化学键类型(B)、连接性(Co)、氢原子(H)和手性(Ch) ,同时在不同的碎片长度(默认值为 4 - 7)范围内产生分子全息 ,分子全息长度 L 为 53 ~ 401 范围内 12 个预定的质数.

表 1 苯的衍生物对黑头呆鱼的半致死浓度(以 pLC_{50} 计)的实验值、拟合值及预测值

编号	化合物	半致死浓度(pLC_{50})			编号	化合物	半致死浓度(pLC_{50})		
		实验值	拟合值	预测值			实验值	拟合值	预测值
1	间苯二酚	3.04	3.00	3.12	36	1 μ -二甲苯	4.21	4.00	3.71
2	1 μ -二甲氧基苯	3.07	3.00	3.40	37	1 β -三甲苯	4.21	3.98	3.83
3	3-羟基苯甲醚	3.21	3.02	3.08	38	4-甲基-2 β -二硝基苯胺	4.21	4.14	3.59
4	2-甲基-4-苯胺	3.24	3.46	3.78	39	3-甲基-2 μ -二硝基苯胺	4.26	4.12	4.38
5	3-甲基苯酚*	3.29	3.24	3.27	40	4-氯-3-甲基苯酚*	4.27	4.11	3.98
6	甲苯	3.32	3.54	3.65	41	1 β -二氯苯	4.30	4.25	4.15
7	2-甲基-5-硝基苯胺	3.35	3.39	3.64	42	2 μ -二氯苯酚	4.30	4.32	4.30
8	4-硝基苯酚	3.36	3.73	3.99	43	4-氯甲苯	4.33	4.27	4.17
9	苯	3.40	3.49	3.64	44	2 μ β -三氯苯酚	4.33	4.34	4.90
10	1 β -二甲苯*	3.48	3.46	3.49	45	1 β -二硝基苯*	4.38	4.35	4.13
11	2-甲基-1 β -硝基苯胺	3.48	3.53	3.69	46	1 β -二氯苯	4.40	4.42	4.35
12	苯酚	3.51	3.52	3.54	47	4-甲基-3 β -二硝基苯胺	4.46	4.46	4.20
13	2-硝基甲苯	3.57	3.84	4.00	48	2 μ -二氯甲苯	4.54	4.55	4.50
14	4-甲基苯酚	3.58	3.67	3.70	49	1 μ -二氯甲苯	4.62	4.65	4.22
15	3-Nitrotoluene*	3.63	3.63	3.67	50	2 μ β -三溴苯酚	4.70	4.65	3.96
16	4-羟基-3-硝基苯胺	3.65	3.64	3.71	51	1 β β -三氯苯	4.74	4.73	4.29
17	2 β -二甲苯基酚	3.75	3.68	3.80	52	3 μ -二氯甲苯	4.74	4.66	4.58
18	2 μ -二硝基甲苯	3.75	3.86	4.04	53	2 μ β -三硝基甲苯	4.88	4.97	4.95
19	4-硝基甲苯	3.76	4.11	4.29	54	1 β β -三氯苯	4.89	4.89	4.72
20	氯苯*	3.77	3.86	3.87	55	5-甲基-2 μ -二硝基苯胺	4.92	4.86	4.69
21	2-甲基苯酚	3.77	3.70	3.67	56	1 β β -三氯苯	5.00	4.88	4.77
22	4-甲基-3-硝基苯胺	3.77	3.71	3.79	57	2-甲基-4 β -二硝基苯酚	5.00	5.00	4.64
23	4-甲基-2-硝基苯胺	3.79	3.70	3.81	58	2 β -二硝基甲苯	5.01	5.09	5.10
24	2-甲基-6-硝基苯胺	3.80	3.83	3.96	59	3 μ -二硝基甲苯	5.08	5.26	5.28
25	3-甲基-6-硝基苯胺*	3.80	4.05	4.35	60	2 β -二硝基甲苯*	5.15	4.85	4.27
26	3-氯甲苯	3.84	3.79	3.78	61	1 μ -二硝基甲苯	5.22	5.16	4.33
27	2 μ -二甲苯基酚	3.86	3.98	4.01	62	1 β β -三硝基甲苯	5.29	5.08	4.68
28	溴苯	3.89	3.79	3.70	63	2-甲基-3 β -二硝基苯胺	5.34	5.36	5.27
29	3 μ -二甲苯基酚	3.90	3.74	3.67	64	1 β β μ -四氯苯	5.43	5.45	5.37
30	3 β -二硝基甲苯*	3.91	4.13	4.68	65	1 β -二硝基苯	5.45	5.38	4.71
31	2 β -二硝基甲苯	3.99	4.11	4.14	66	2 β μ β -四氯苯酚	5.72	5.54	5.24
32	2-氯苯	4.02	4.06	4.05	67	1 β μ β -四氯苯	5.85	5.91	5.67
33	2 μ -二硝基苯酚	4.04	3.88	3.83	68	2 β μ β β -五氯苯酚	6.06	6.17	5.83
34	2-甲基-3 β -二硝基苯胺	4.12	4.12	4.38	69	2 β β -三硝基甲苯	6.37	6.47	6.28
35	5-甲基-2 β -二硝基苯胺*	4.18	3.93	4.25					

* 为检验集内化合物.

2 结果与讨论

2.1 碎片区分参数对模型质量的影响

HQSAR 模型可以通过改变碎片区分因素和碎片长度来优化 ,碎片区分参数决定分子全息中所映射的拓扑结构信息^[6]. 本研究使用的碎片区分参数组合为{A , B }、{A , B , Co }、{A , B , Co , H } ,碎片长度为(1 ~ 7)与缺省值(4 ~ 7) ,不同碎片区分参数的 HQSAR 分析结果见表 2.

表 2 中结果表明在两种碎片长度(1 ~ 7 与 4 ~ 7)下 ,预测模型的质量均在碎片区分参数为原子、化学键及连接性时获得最佳结果 ,此时 q^2 分别为 0.841 和 0.810 ,交叉验证标准偏差为 0.334 与 0.362 ,两种模型均具有 — 68 — 万方数据

较强的预测能力.

表 2 碎片区分参数对模型质量的影响

Fragment distinction	1 ~ 7		4 ~ 7	
	q^2_{LOO}	SE	q^2_{LOO}	SE
{A ,B }	0. 745	0. 405	0. 710	0. 453
{A ,B ,Co }	0. 841	0. 334	0. 810	0. 362
{A ,B ,Co ,H }	0. 786	0. 378	0. 688	0. 512

2.2 碎片长度对模型质量的影响

碎片长度决定全息指纹中所含碎片的原子个数 ,能够代表不同的分子结构信息. 分子全息的产生可以在不同碎片长度下进行 ,一般包括碎片长度为 1、较短的碎片长度(1 ~ 3)、中等的碎片长度范围(4 ~ 7)和较大的碎片范围(8 ~ 10) ,其中碎片长度为 1 和(1 ~ 3)近似表征原子类型和官能团性质 ,4 ~ 7 碎片长度可以区分脂肪环的链长或芳香环的性质及取代基的位置和类型等结构特征^[7]. 为了得到最佳的 QSAR 模型 ,固定碎片区分参数为原子、化学键和连接性 ,扩大碎片长度的搜索范围 ,所得到的 HQSAR 模型分析结果见表 3.

表 3 不同碎片长度对 HQSAR 分析结果的影响

碎片长度	分子全息长度	交互验证(LOO)		非交互验证		PCs
		q^2_{LOO}	SE _{CV}	r^2	SE	
1	71	0. 693	0. 442	0. 746	0. 402	6
1 ~ 3	83	0. 662	0. 463	0. 778	0. 376	6
1 ~ 7	251	0. 841	0. 334	0. 997	0. 049	12
3 ~ 10	97	0. 827	0. 316	0. 987	0. 086	7
4 ~ 7	257	0. 810	0. 362	0. 995	0. 060	11
8 ~ 10	311	0. 553	0. 533	0. 928	0. 214	6

表 3 中数据显示最佳分子全息产生的条件为 :碎片长度为 1 ~ 7 ,全息长度为 251 ,碎片区分参数为原子、化学键和连接性 ,此时模型 LOO 交叉验证 q^2 为 0. 841 ,标准预测误差 SE_{CV} 为 0. 334 ,主成分数为 12 ,非交叉验证相关系数为 0. 997 ,标准误差为 0. 049 ,69 种苯的衍生物对黑头呆鱼的半致死浓度的负对数值(pLC₅₀)的实验值与拟合值、预测值的相关性分别见图 1 与图 2.

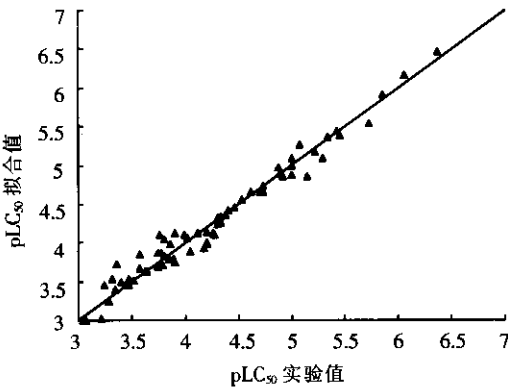


图 1 苯的衍生物的 pLC₅₀ 实验值与拟合值

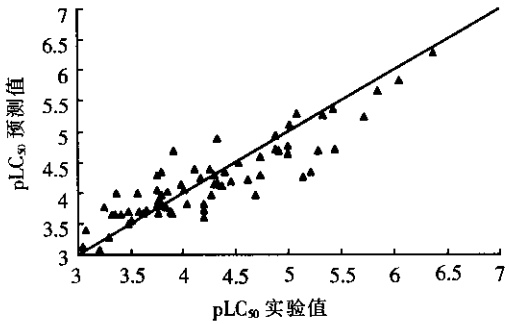


图 2 苯的衍生物的 pLC₅₀ 实验值与预测值

2.3 模型质量的评价

检验所建立 QSAR 模型的质量通常应用 LOO 交叉验证方法 ,然而该方法获得的较高 q^2 值并不一定表明所建立模型具有较强的预测能力与稳健度^[8]. 业已证明交叉验证 Leave-N-out(LNO)方法比常见的 LOO 方法更能检测模型的稳健度. 为了进一步验证 HQSAR 模型的预测能力和稳健度 ,将 69 种苯的衍生物的活性数据按大小排列 ,均匀抽取 13 种化合物构建检验集 ,其余 56 种化合物为训练集用于建立相关模型 ,从而预测检验集中化合物的活性. 分析结果表明 LNO 交叉验证相关系数 q^2_{LNO} 为 0. 837 ,标准误差为 0. 236 ,最佳主成分数为 12 ,非交叉相关系数为 0. 996 ,标准估计误差为 0. 049 ,由此可见基于训练集所建模型与全部数据集的 HQSAR 模型的万方数据

拟合和预测能力都很相近,这说明基于分子全息的 QSAR 模型具有很好的预测能力和稳健度.

3 结论

本文应用分子全息结构描述子表征 69 种苯的衍生物并建立相应 QSAR 模型,并应用 LOO 与 LNO 交叉验证方法检验模型质量.结果表明所建立模型具有良好的预测能力与稳健度.分子全息 QSAR 技术的特点决定其在环境科学领域研究构效关系方面必将拥有广阔的应用前景.

[参考文献]

- [1] Topliss J G, Edwards R P. Chance factors in studies of quantitative structure-activity relationships[J]. J Med Chem, 1979, 22(10): 1238—1246.
- [2] David A W, Frank R B. Holographic QSAR of benzodiazepines[J]. Quant Struct Act Relat, 1998, 17(3): 224—251.
- [3] 崔世海, 刘树深, 王晓栋, 等. 雌二醇衍生物的 HQSAR 研究[J]. 科学通报, 2003, 48(3): 239—242.
- [4] Cui S, Wang X, Liu S, *et al.* Predicting toxicity of benzene derivatives by molecular hologram derived quantitative structure-activity relationships (QSARS) [J]. SAR and QSAR in Environmental Research, 2003, 14(3): 223—231.
- [5] Subhash C, Basak Brian D Gute, *et al.* A comparative QSAR study of benzamides complement inhibitory activity and benzene derivatives acute toxicity [J]. Computers and Chemistry, 2000, 24(2): 181—191.
- [6] Leming M S, Hong F, Weida T, *et al.* QSAR models using a large diverse set of estrogen [J]. J Chem Inf Comput Sci, 2001, 41(1): 186—195.
- [7] 王连生. 有机污染物化学 [M]. 北京: 高等教育出版社, 2004.
- [8] Wolfgang S. Receptor based 3D QSAR analysis of estrogen receptor ligands-merging the accuracy of receptor-based alignment with the computational efficiency of ligand-based method [J]. Journal of Computer-Aided Molecular Design, 2000, 14(16): 559—572.

[责任编辑 孙德泉]

(上

接第41页)

[参考文献]

- [1] Chatwin P C, Allen C M. Mathematical models of dispersion in rivers and estuaries [J]. Ann Rev Fluid Mech, 1985, 17: 119—149.
- [2] Chaudhry M H, Cass D E, Edinger J E. Modelling of unsteady-flow water temperatures [J]. J Hydraul Eng, 1983, 109(5): 657—669.
- [3] Fattah Q N, Hoopes J A. Dispersion in anisotropic homogeneous porous media [J]. J Hydraul Eng, 1985, 111: 810—827.
- [4] Gane C R, Stephenson P L. An explicit numerical method for solving transient combined heat conduction and convection problems [J]. Int J Numer Meth Engrg, 1979, 14: 1141—1163.
- [5] Guvanasen V, Volker R E. Numerical solutions for solute transport in unconfined aquifers [J]. Int J Numer Meth Fluids, 1983, 3: 103—123.
- [6] Douglas J, Russell T F. Numerical methods for convection-dominated diffusion problem based on combining the method of characteristics with finite element or finite difference procedure [J]. SIAM J, Numer Anal, 1982, 19(5): 871—885.
- [7] Rigal A. Numerical analysis of two-level finite difference schemes for unsteady diffusion-convection problems [J]. Int J Numer Methods Engrg, 1989, 28(2): 1001—1021.
- [8] 陆金甫, 张宝琳, 徐涛. 求解对流扩散方程的交替分段显-隐式方法 [J]. 数值计算与计算机应用, 1998, 19(3): 161—167.
- [9] 张宝琳, 符鸿源. 一类交替块 Crank-Nicolson 方法的差分图 [J]. 科学通报, 1999, 40(11): 1148—1152.
- [10] 王文治. 求解扩散方程的一类交替分组显式方法 [J]. 山东大学学报(理学版), 2002, 37(3): 194—199.
- [11] 陆金甫, 杜正平, 刘晓遇. 对流占优扩散问题的一种特征差分方法 [J]. 清华大学学报(自然科学版), 2002, 42(8): 35—37.
- [12] 田振夫. 泊松方程的高精度三次样条差分方法 [J]. 西北师范大学学报(自然科学版), 1996, 32(2): 13—17.
- [13] Evans D J, Abdullah A R. A new explicit method for the diffusion-convection equations [J]. Comput Math Appl, 1985, 11: 145—154.

[责任编辑 陆炳新]