

分子距边矢量结合神经网络法预测二噁英类 化合物 PCDFs 的 $\log K_{ow}$ 值

黄俊, 余刚, 张彭义

(清华大学 环境科学与工程系 环境模拟与污染控制国家重点实验室, 北京 100084)

摘要: 采用以基团间距离、边数为基准的分子距边矢量作为二噁英类化合物多氯代二苯并呋喃 (PCDFs) 的分子结构描述符, 结合反向传播人工神经网络方法建立了 PCDFs 的正辛醇/水分配系数 ($\log K_{ow}$) 与分子结构描述符之间的定量关系模型。结果表明, 所采用的结构描述符对于分子结构具有很好的区分能力, 所建立的模型对于校验样本的预测精度较高。利用所得模型, 对文献尚未报道 $\log K_{ow}$ 实验值的其他所有 85 种 PCDFs 给出了预测值。

关键词: 多氯代二苯并呋喃 (PCDFs); 持久性有机污染物 (POPs); 定量构效关系; 正辛醇/水分配系数; 神经网络; 分子距边矢量

中图分类号: O 625.2, X 820.6

文献标识码: A

文章编号: 1001-4160 (2002) 01-103-107

Predicting the $\log K_{ow}$ Using Molecular Distance – edge Vector Combined with Artificial Neural Network (ANN) Method

HUANG Jun, YU Gang, ZHANG Peng-yi

(Department of Environmental Science and Engineering, the Environmental Simulation and Pollution Control State

Key Laboratory, Tsinghua University, Beijing 100084, China)

Abstract: Polychlorinated dibenzofurans (PCDFs) are highly concerned as persistent organic pollutants and suspected endocrine disruptors, whose property $\log K_{ow}$ plays an important role for its environmental risk assessment. In this paper, a novel molecular distance – edge vector (V_{MDE} , μ in short) was introduced as structure descriptors, then a quantitative relationship of high accuracy was established with back – propagation artificial neural network method. An additional sample set was used to test the model with quite good results. With the established model, the $\log K_{ow}$ values of the other 85 PCDFs not belonging to the modeling sample set were given, these values has not been reported in literatures.

Key words: PCDFs; persistent organic pollutants (POPs); QSAR, $\log K_{ow}$; artificial neural network; molecular distance – edge vector

1 引言

多氯代二苯并呋喃 (polychlorinated dibenzofurans, 简称 PCDFs) 与多氯代二苯并 – 对 – 二噁英 (polychlorinated dibenzo – p – dioxins, 简称 PCDDs) 合称为氯代二噁英或二噁英类 (Dioxins)。目前, 二噁英类已经成为全球关注的环境问题焦点。在 2001 年 5 月 23 日 127 个国家和地区的代表正式签署的《关于持久性有机污染物 (POPs) 的斯德哥尔摩公约》(下文简称《POPs 公约》) 中, 二噁英类被定为首批采取全球控制行动的 12 种 POPs 化合物之一^[1]。按照《POPs 公约》的规定, 正辛醇/水分配系数 ($\log K_{ow}$) 是用于判断某种化合物是否属于持久性有机物的重要依据^[2]。对于 PCDFs 而言, 因氯原子的取代数目及位置的不同, PCDFs 共有 135 种可能的分子结构^[3]。而单一的 PCDF 纯物质通常难以获得, 并且一些在环境中共存

收稿日期: 2001-07-10; 修回日期: 2001-09-08

基金资助: 国家重点基础研究专项经费资助项目 (G1999045711)

作者简介: 黄俊 (1976—), 男, 江苏江阴人, 博士研究生, 环境科学专业, 研究方向: 有机污染物化学, 导师: 余刚。

联系人: 余刚。

的 PCDFs 即使用气质联用 (GC/MS) 方法仍难以分开, 因此目前还有相当数量的 PCDF_s 化合物难以获得 $\log K_{ow}$ 实验数据。在这种情况下, 可以采用定量构效关系 (QSAR) 来预测。但是, 以前的文献往往对精细结构的区分能力不够好, 从而导致模型的精度不如人意^[4], 或者采用的结构描述符数目过多而较复杂^[5]。

本文引入计算简便、对 PCDF_s 结构区分能力强的以基团间距离、边数为基准的分子距边矢量 (VMDE) 来作为分子结构描述符, 并采用误差反向传播人工神经网络 (BP-ANN), 建立了 PCDF_s 的 $\log K_{ow}$ 值的定量关系模型, 并进行了模型校验和未知样本预测。

2 基本原理

2.1 结构表征

引用林治华等^[6-7]最近在研究 PCDF_s 气相色谱保留行为时发展的以基团间距离、边数为基准的分子距边矢量 (VMDE) 作为本研究的分子结构描述符, 即: 将 PCDF 分子中的苯环以拟原子方式处理, 从而将 PCDF 分子中的原子 (基团) 分成 3 类: 氯原子 (记为 1)、氧原子 (记为 2)、拟原子 (记为 3); 三者之间的相互作用记为 M_{11} 、 M_{12} 、 M_{13} 、 M_{22} 、 M_{23} 、 M_{33} , 其计算公式为:

$$M_{ij} = \sum_{j \geq i}^{n_j} \frac{1}{d_{ik,jl}^2} \quad (i, j = 1, 2, 3) \quad (1)$$

式中: $d_{ik,jl}$ 表示第 i 类中第 k 个原子与第 j 类中第 l 个原子之间的距离 (即最近相距键数), n_j 是 $d_{ik,jl}$ 的项数, 称为相互作用项数。

对于 PCDF_s 分子而言, 由于只有一个 O 原子, 因此 M_{22} 为零; 而对于所有 PCDF_s, 氧原子与苯环以及苯环与苯环之间的相互作用情况是完全相同的, 因此 M_{23} 、 M_{33} 可以不作考虑。从而定义分子距边矢量 (V_{MDE}) 为 (M_{11} , M_{12} , M_{13}), 简记为 $\mu = (\mu_1, \mu_2, \mu_3)$ 。文献 [6] 中已列出了所有 135 种 PCDF_s 的 μ 矢量, 可以发现, 没有 2 个 PCDF 分子的 μ 矢量是一样的, 表明该矢量对于 PCDF_s 具有较好的分子结构区分能力。

2.2 建模方法

我们采用目前应用最广泛、理论最完善的误差反向传播人工神经网络 (BP-ANN) 算法, 其具体的算法及所使用的计算机程序参见文献 [8] 和 [9]。

3 样本与方法

3.1 样本集的组织

Rulle 等^[4]收集整理了前人报道的共计 50 种 PCDF_s 分子的 $\log K_{ow}$ 实验数据, 本研究以此作为基本数据集。由于好的 QSAR 模型不仅要求建模误差小, 同时还要求有较强的预测能力, 即对于不包含在训练集中的样本也应具有较强的预测能力。基于这种考虑, 我们从中随机抽取出 5 种分子构成模型校验样本集, 而

以另外 45 种分子构成建模所需的训练样本集, 所有原始数据均列于表 1 中。

3.2 神经网络的结构、参数选择及训练

采用 3 层误差反向传播人工神经网络 (BP-ANN), 其中为带一个偏置节点的 3-2-1 体系, 网络的结构可以用骨架图直观地表示出来, 如图 1 所示。上面 BP-ANN 网络结构参数的选择依据如下:

(1) 样本容量 (设为 N) 与变量个数 (设为 M) 之比符合 $N/M \geq 5$ ^[10], 本研究中 $N/M = 45/3 = 15$;

(2) 输入层用于加载 μ 矢量, 输出层加载 $\log K_{ow}$ 数据, 故分别选择节点数分别为 3 和 1;

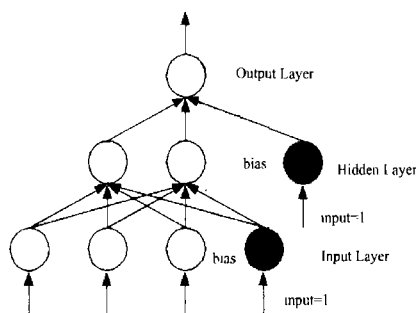


图 1 带一个偏置节点的 3-2-1 BP 神经网络示意图

Fig.1 3-2-1 BP-ANN with a bias

(3) 引入一个偏置节点的目的是调节传递函数, 提高网络性能;

(4) 隐层节点数的选择遵循了两点: 一是网络的规模在满足问题本身要求的前提下尽可能小; 二是网络中所有权值个数不大于训练样本数的一半, 在本研究中所有权值个数为 11 个 (含偏置), 远小于训练样本数的一半 22.5。

表 1 建模样本及检验样本

Table 1 Sample sets for modeling and test

Substitution pattern	Observed log K_{ow}	Substitution pattern	Observed log K_{ow}	Substitution pattern	Observed log K_{ow}	Substitution pattern	Observed log K_{ow}
2,8-	5.44	1,3,6,8-	6.37	1,2,3,4,7-	6.53	a 1,3,4,6,9-	6.34
1,2,3,4-	6.17	1,3,7,8-	6.34	1,2,3,4,8-	6.79	1,3,4,7,9-	6.33
1,2,3,6-	6.15	a 1,3,7,9-	6.34	1,2,3,6,7-	6.26	1,3,4,8,9-	6.51
1,2,3,8-	6.15	1,4,6,7-	6.15	1,2,3,6,8-	6.33	2,3,4,6,7-	6.47
1,2,3,9-	6.21	1,4,6,9-	5.6	1,2,3,6,9-	6.65	2,3,4,6,8-	6.59
1,2,4,8-	6.31	2,3,4,6-	6.11	1,2,3,7,8-	6.79	a 2,3,4,6,9-	6.53
1,2,6,7-	6.25	2,3,4,7-	6.06	1,2,4,6,7-	6.27	2,3,4,7,8-	6.92
1,2,7,8-	6.23	2,3,4,9-	6.17	1,2,4,6,8-	6.34	2,3,4,8,9-	6.42
1,2,7,9-	6.25	2,3,6,7-	6.31	1,2,4,6,9-	6.59	1,2,3,4,7,8-	7
1,3,4,6-	6.31	2,3,6,8-	6.73	1,2,4,7,8-	6.26	1,2,3,4,6,7,8-	7.4
a 1,3,4,7-	6.23	2,3,7,8-	6.53	1,2,4,7,9-	6.19	1,2,3,4,6,7,8,9-	7.97
1,3,4,8-	6.13	2,4,6,7-	6.25	1,3,4,6,7-	6.19		
1,3,4,9-	5.89	a 2,4,6,8-	6.17	1,3,4,6,8-	6.24		

a. this symbol with bond font means this sample is extracted out to form test sample set.

为保证网络对样本具有足够的敏感性和良好的拟合性, 将输入变量线性归一化至 -3 - 3 之间, 而将输出变量线性归一化至 0 - 1 之间, 其公式为: $V_1 = a \cdot V_0 + b$, a 和 b 值根据样本的最大值和最小值由程序自动计算获取。另外利用随机数矩阵进行权值矩阵的初始化, 随机数生成器的最大、最小输出设置为 3 和 -3。

4 结果与讨论

4.1 训练建模结果

取学习步长为 0.1, 记忆常数为 0.5, 网络训练的目标误差为 0.2, 将 45 种 PCDF_s 组成的训练样本集加载到 2.4 中的 BP 神经网络。所用机型为 PC 兼容机, CPU 为 Celeron 466, 128M 内存, 操作系统为 Windows ME。经过 4238 次的迭代 (约 32s), 网络训练完毕, 权值矩阵固定下来, 得到所需要的 PCDF_s 结构与正辛醇/水分配系数之间的定量相关模型, 有关层之间的连接权值列于表 2 中。

表 2 隐层与输入、输出层之间的连接权值

Table 2 Weight matrix between the input layer and hidden layer

	Input node 1	Input node 2	Input node 3	Input layer bias	Output node
Hidden node 1	0. 101899	- 1. 170129	0. 374317	- 2. 423910	- 5. 477703
Hidden node 2	1. 271974	1. 725676	- 1. 421026	- 2. 648534	3. 650005
Hidden layer bias	—	—	—	—	0. 028569

4.2 模型的精度校验

首先利用得到的模型计算出训练样本集中 45 种 PCDF_s 的 log K_{ow} 值, 并将模型值与文献值进行比较, 计算其相对百分误差 (RE, 单位为 %), 并通过如图 2 所示的误差频率分布图直观表示。从图 2 来看, 误差的分布基本上符合正态分布, 并且绝大多数样本的误差均在 $\pm 5\%$ 之间, 少量样本误差在 $\pm 5 - 10\%$ 之间, 这说明上面建立的模型具有较高的精度。需要指出的是, 由于我们数据进行了归一化处理, 程序所控制的目标误差是对归一化后的数据而言, 而上面计算的 RE 是对原始数据而言的。

为了检验模型的预测能力, 我们利用 3.1 中所述的随机产生的检验样本来进行模型计算, 并将预测值与实验值进行比较, 其结果列于表 3 和图 3 中。

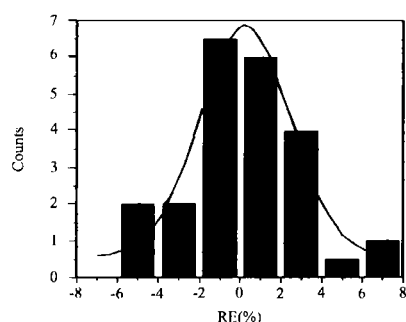


图2 模型的误差频率分布直方图

Fig.2 Relative error frequency plot

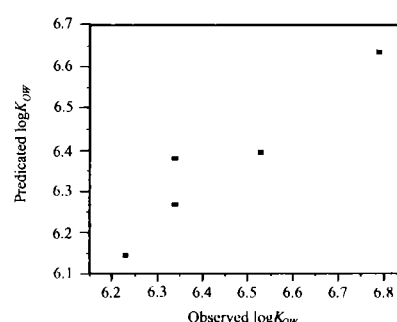


图3 检验样本的模型值与实验值的比较

Fig.3 Result for test samples prediction

表3 模型对于检验样本集的预测结果

Table 3 Predicting results for the test samples

Substitution pattern	Observed $\log K_{ow}$	Predicted $\log K_{ow}$	Error	RE (%)
1, 3, 4, 7-	6.23	6.14	-0.09	-1.4
1, 3, 7, 9-	6.34	6.27	-0.07	-1.2
2, 4, 6, 8-	6.79	6.64	-0.15	-2.3
1, 3, 4, 6, 9-	6.34	6.38	0.04	0.6
2, 3, 4, 6, 9-	6.53	6.39	-0.14	-2.1

由表3和图3可以看出,所建立的模型对于校验样本的预测精度较好,误差位于 $\pm 3\%$ 之间。因此,可以用于对未知样本的 $\log K_{ow}$ 值进行预测。

4.3 未知样本的预测

利用上述模型,我们对文献中尚未报道 $\log K_{ow}$ 值的其他85种PCDFs的 $\log K_{ow}$ 值进行了预测,其结果列于表4中。

表4 模型对85种尚未报道 $\log K_{ow}$ 实验值的FCDFs的预测结果Table 4 Predicted $\log K_{ow}$ values for 85 PCDFs without observed values reported

Substitution pattern	Predicted $\log K_{ow}$	Substitution pattern	Predicted $\log K_{ow}$	Substitution pattern	Predicted $\log K_{ow}$	Substitution pattern	Predicted $\log K_{ow}$
1-	5.51	1,2,7-	5.95	3,4,7-	5.79	2,3,4,7,9-	6.47
2-	5.51	1,2,8-	6.01	3,4,8-	5.84	1,2,3,4,6,7-	6.66
3-	5.50	1,2,9-	5.98	3,4,9-	5.81	1,2,3,4,6,8-	6.71
4-	5.49	1,3,4-	5.80	1,2,3,7-	6.28	1,2,3,4,6,9-	6.67
1,2-	5.66	1,3,6-	5.82	1,2,4,6-	6.11	1,2,3,4,7,9-	6.83
1,3-	5.63	1,3,7-	5.91	1,2,4,7-	6.21	1,2,3,4,8,9-	7.07
1,4-	5.59	1,3,8-	5.96	1,2,4,9-	6.22	1,2,3,6,7,8-	6.93
1,6-	5.59	1,3,9-	5.94	1,2,6,8-	6.25	1,2,3,6,7,9-	6.79
1,7-	5.63	1,4,6-	5.75	1,2,6,9-	6.23	1,2,3,6,8,9-	6.92
1,8-	5.67	1,4,7-	5.83	1,2,8,9-	6.36	1,2,3,7,8,9-	7.33
1,9-	5.65	1,4,8-	5.88	1,3,6,7-	6.15	1,2,4,6,7,8-	6.69
2,3-	5.64	1,4,9-	5.85	1,3,6,9-	6.19	1,2,4,6,7,9-	6.63
2,4-	5.59	2,3,4-	5.81	1,4,6,8-	6.12	1,2,4,6,8,9-	6.69
2,6-	5.60	2,3,6-	5.83	1,4,7,8-	6.20	1,3,4,6,7,8-	6.62
2,7-	5.65	2,3,7-	5.92	2,3,4,8-	6.21	1,3,4,6,7,9-	6.67
3,4-	5.57	2,3,8-	5.98	3,4,6,7-	6.02	2,3,4,6,7,8-	6.69
3,6-	5.57	2,3,9-	5.95	1,2,3,4,6-	6.39	1,2,3,4,6,7,9-	7.23
3,7-	5.62	2,4,6-	5.76	1,2,3,4,9-	6.51	1,2,3,4,6,8,9-	7.43
4,6-	5.54	2,4,7-	5.84	1,2,3,7,9-	6.58	1,2,3,4,6,7,8,9-	7.74
1,2,3-	5.93	2,4,8-	5.90	1,2,3,8,9-	6.67		
1,2,4-	5.86	2,4,9-	5.87	1,2,4,8,9-	6.54		
1,2,6-	5.86	3,4,6-	5.71	1,3,4,7,8-	6.46		

基于弦和路径分析的环系最小环最小集识别方法

陈维明, 孙传涛, 朱翠娣, 郑崇直

(中国科学院上海有机化学研究所 计算机化学开放实验室, 上海 200032)

摘要: 最小环最小集 (SSSR) 是环系的重要特征之一。本文提出一种基于生成树、弦和路径长度的环系最小环最小集识别方法, 不需事先进行环系识别。由于采用以分枝节点和枝为基本单元的树图表示, 降低了树图的复杂度, 并且识别过程只需在图上搜索一次, 大大加快了识别过程。给出了2个包含复杂环系的化学结构作为最小环最小集识别的例子。

关键词: 环系分析; 最小环的最小集; SSSR

中图分类号: O6-051

文献标识码: A

文章编号: 1001-4160 (2002) 01-107-111

收稿日期: 2001-07-10; 修回日期: 2001-10-18

基金资助: 国家自然科学基金 (39970870); 国家科技部重点攻关计划专题 (96-920-37-05) 资助项目

作者简介: 陈维明 (1953—), 男, 上海人, 研究员, 博士, 计算机化学专业。

5 结论

(1) 引入以基团为基准的分子边距矢量作为 PCDF_s 的结构描述符, 利用 BP 人工神经网络建立了二噁噻类化合物多氯代二苯并呋喃 (PCDF_s) 的重要环境化学性质——正辛醇/水分配系数 ($\log K_{ow}$) 的定量结构性质关系模型。

(2) 所建立的模型对训练样本集的误差较小, 并且基本呈现正态分布; 对校验样本集的预测精度较高, 误差位于 $\pm 3\%$ 之间, 具有较好的预测能力。

(3) 利用所建立的模型, 对文献报道中尚未收集到实验值的其他 85 种 PCDF_s 的 $\log K_{ow}$ 值进行了预测, 这涵盖了本研究基础数据集 50 个样本以外的 PCDF_s 所有可能的结构, 对于在精细结构水平上正确评价 PCDF_s 的环境风险具有重要的参考价值。

参 考 文 献

- [1] 联合国环境规划署. 关于持久性有机污染物的斯德哥尔摩公约. 全权代表会议的最后文件 (中文版), 瑞典斯德哥尔摩全权代表会议筹备会议, 2000, 5:22-23.
- [2] 联合国环境规划署. 关于持久性有机污染物的斯德哥尔摩公约 (中文版), 瑞典斯德哥尔摩, 2001, 5:23.
- [3] 郑明辉, 刘鹏岩, 包志成, 徐晓白. 二噁噻的生成及降解研究进展. 科学通报, 1999, 44(5):455-463.
- [4] Rulle P. The n-octanol and n-hexane/water partition coefficient of environmentally relevant chemicals predicted from the mobile order and disorder (MOD) thermodynamics. Chemosphere, 2000, 40(5):457-512.
- [5] Kari T, Juhani R. Electronic eigenvalue (EEVA): a new QSAR/QSPR descriptor for electronic substituent effects based on molecular orbital energies. A QSAR approach to the Ah receptor binding affinity of polychlorinated biphenyls (PCBs), dibenzo-p-dioxins (PCDDs) and dibenzofurans (PCDFs), Chemosphere, 2000, 41(6):843-848.
- [6] 林治华, 刘树深, 李志良. 多氯代二苯并呋喃在不同色谱柱上的气相色谱保留行为——定量结构-色谱保留关系 (QSRR) 的研究. 色谱, 2001, 19(2):116-123.
- [7] 林治华, 徐红, 刘树深, 李志良. 多氯联苯 (PCBs) 分配系数的估算和预测. 计算机与应用化学, 2000, 17(2):189.
- [8] 黄俊, 周申范, 国振双, 许国旺. 人工神经网络预测二维色谱柱效. 色谱, 2000, 18(1):1-4.
- [9] 黄俊, 周申范, 唐婉莹. TNT 生化降解时间序列的人工神经网络预报模型. 环境科学研究, 2000, 13(2):3-5.
- [10] 许禄, 吴亚平, 胡昌玉, 李华. 苯胺类化合物结构——毒性定量构效关系研究. 中国科学 (B 辑), 2000, 30(1):1-7.