

有机污染物生物降解性预测模型

李彦莹, 李雪花*, 杨先海, 陈广超, 陈景文

大连理工大学环境学院工业生态与环境工程教育部重点实验室, 大连 116024

摘要: 生物降解性是评估污染物环境持久性的重要依据,也是化学品是否获准生产及进入市场的评价指标。采用 17 位生物降解领域专家评估的生物降解等级数据,通过功能树(FT)算法建立了包含 15 个分子结构参数的初级生物降解和最终生物降解预测模型。外部验证结果表明,模型具有较好的预测准确性,初级生物降解性加权准确度(weighted accuracy,WA): 训练集 WA=84.1%,验证集 WA=78.9%;最终生物降解性 WA: 训练集 WA=91.0%;验证集 WA=83.6%。预测正确性对化合物的杠杆值作图,表征了生物降解性模型的应用域。

关键词: 有机污染物; 初级生物降解性; 最终生物降解性; 预测模型; 功能树

文章编号: 1673-5897(2012)5-549-07 中图分类号: X171.5 文献标识码: A

Predicting Models for Biodegradability of Organic Pollutants

Li Yanying, Li Xuehua*, Yang Xianhai, Chen Guangchao, Chen Jingwen

Liaoning Key Laboratory of Industrial Ecology and Environmental Engineering, Ministry of Education, School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China

Received 29 June 2012 accepted 29 September 2012

Abstract: Biodegradability is significant for the evaluation of environmental persistence of pollutants, which has become a criterion of evaluation on a chemical allowing to be manufactured and market. The primary biodegradation and ultimate biodegradation models which contained 15 molecular structure descriptors were developed employing 17 biodegradation experts evaluating rating data and functional tree algorithm. The external validation results indicated that the models have a good predictive ability. The weighted accuracy (WA) of primary biodegradation is 84.1% and 78.9% for training set and validation set, respectively. WA for ultimate biodegradation is 91.0% and 83.6% for training set and validation set, respectively. Moreover, the applicability domain of the biodegradation models were characterized by the plots of prediction accuracy versus leverage values of chemicals.

Keywords: organic pollutants; primary biodegradability; ultimate biodegradability; prediction model; functional tree

生物降解是水环境中有机污染物最重要的降解途径^[1],它影响水环境中污染物的归趋、持久性和毒性^[2]。根据降解程度,生物降解可分为 3 个阶段^[3]: (1)初级生物降解: 有机污染物在微生物作用下,改

变母体化合物的结构,并使原化合物分子的完整性发生改变; (2)可接受生物降解: 除去有机污染物的毒性或人们不需要的特性; (3)最终生物降解: 有机污染物通过生物降解向无机物转化,完全被降解成

收稿日期: 2012-06-29 录用日期: 2012-09-29

基金项目: 国家 863 计划重点项目(2010AA065105)和工业生态与环境工程教育部重点实验室开放课题

作者简介: 李彦莹(1989-),女,学士,研究方向: 计算毒理学; E-mail: yyli@mail.dlut.edu.cn;

* 通讯作者(Corresponding author), E-mail: lixuehua@dlut.edu.cn

CO₂、水和其他无机物并被同化成微生物的一部分。根据经济合作与发展组织(OECD)的化学药品生物降解测试导则,我国制定了生物降解性测试的国家标准,包括溶解性有机碳(DOC)消减实验和 CO₂ 产生实验等 6 种测试方法^[4,9]。由于生物降解测试耗时长(测试周期为 28 d)、测试成本高^[10],目前仍有大量人工合成化学药品未经生物降解性测试^[11-12],此外,化学药品“环境友好设计”的理念要求在化学药品生命周期的早期,甚至在合成前的设计阶段就需要利用生物降解性信息指导产品设计^[2,13]。因此,预测生物降解性的理论模型扮演着日益重要的作用,欧美发达国家及 OECD 大力倡导定量结构-生物降解性相关关系(QSBR)预测化学药品生物降解性。

生物降解性预测模型仅从分子结构信息即可筛选快速生物降解(F)和缓慢生物降解(S)的化合物,QSBR 具有可预测化合物种类多和预测准确率高等优点^[14]。Boethling 等^[15]根据 22 位生物降解领域专家评估的 46 种化合物的水中耗氧最终生物降解等级,采用拓扑指数和大分子性质初步建立了最终生物降解预测模型,该模型准确地预测了外部验证集所包含的 40 个化合物中的 36 个,虽然预测精度较高,但因数据集太小所涉及的分子结构种类有限,对于新化学药品预测准确率可能具有不确定性。之后,Boethling 等^[16]又采用 QSBR 的方法根据 17 位专家评估的 200 种化合物生物降解等级建立了初级生物降解和最终生物降解的预测模型,预测准确率分别为 82.5% 和 83.5%,模型中使用了 36 个分子碎片与分子量作为描述符,并未考虑到分子碎片之间的相互作用及其他理化性质对生物降解性的影响。因此,本文考虑多种分子结构参数,并加入表征分子整体特征的特征参数,采用功能树(FT)方法,建立了初级生物降解和最终生物降解的预测模型,探讨了化合物生物降解性与其结构特征的关系。

1 材料和方法(Materials and methods)

1.1 数据来源

本研究使用的有机化合物生物降解性数据来源于 EPI suiteTM (Version 4.1, <http://www.epa.gov/oppt/exposure>) Biowin3 和 Biowin4 模块。EPI suiteTM 原始数据集包括 200 种化合物初级生物降解等级(Primary Rating)数据及其最终生物降解等级(Ultimate Rating)数据。数据库中 12 个为离子化合物无法计算 Dragon 分子结构描述符,5 个化合物名称有误或与 CAS 不符,因此从数据集中删除这 17 个

化合物。本研究的数据集包括 183 个有机化合物初级生物降解的等级数据和 183 个最终生物降解等级数据。数据集覆盖了 C、H、O、N、S、P、X 等原子构成的不同类型的官能团,包含了烷烃、烯烃、炔烃和芳香烃及其衍生物、醇、醛、酮、醚、酸和脂等多种有机污染物。Primary Rating 和 Ultimate Rating 数据是由 17 位生物降解领域专家根据化合物在典型的(或是可评估的)水生环境中所需要的生物降解时间确定的经验值。

对于初级生物降解,当 Primary Rating ≥ 3.5 时,判定为快速降解(F),Primary Rating < 3.5 时为缓慢降解(S);对于最终生物降解,当 Ultimate Rating > 2.5 时,判定为快速降解(F),Ultimate Rating ≤ 2.5 时为缓慢降解(S)。根据此判断原则,初级生物降解数据集中含 115 个快速降解(F)化合物,68 个缓慢降解(S)化合物;最终生物降解数据集中含 103 个快速降解(F)化合物,70 个缓慢降解(S)化合物。

1.2 分子描述符的计算

应用 DRAGON 软件(Version 2.1)计算了 0 ~ 3 维 DRAGON 分子结构描述符,包括原子构成描述符、拓扑描述符、BCUT 描述符、电荷描述符、芳香性指数、几何描述符、RDF 描述符、3D-MoRSE 描述符、WHIM 描述符、官能团、原子中心碎片等 18 类共 1480 个分子描述符。应用 ChemOffice 2010 软件中的 MOPAC 模块,采用 PM6 算法,计算了最终生成热(HOF)、最高占据分子轨道能(E_{HOMO})、碳原子最大电荷(Q_{Cmax})、偶极矩(dipole)、氢原子最小电荷(Q_{Hmin})等 18 个量子化学描述符。

1.3 模型的建立

首先将原始数据集以 2 : 1 的比例随机拆分成训练集和验证集。其中,训练集用于构建预测模型;验证集作为外部验证,用于评估模型的预测能力。为平衡训练集中快速降解和缓慢降解化合物的比例,经调整初级生物降解(primary)训练集包括 126 个化合物(70 个快速降解化合物与 56 个缓慢降解化合物),验证集包括 57 个化合物(45 个快速降解化合物与 12 个缓慢降解化合物);最终生物降解(ultimate)训练集包括 122 个化合物(66 个快速降解化合物与 56 个缓慢降解化合物),验证集包括 61 个化合物(37 个快速降解化合物与 24 个缓慢降解化合物)。

生物降解性(F 和 S)作为预测变量,以分子结构描述符作为分类指示变量,采用功能树(functional tree, FT)方法建立预测模型,将数据平均分成 10

组进行交叉验证,具体运算由 Weka 软件(Version 3.6, <http://www.cs.waikato.ac.nz/ml/weka/>) 执行。

1.4 模型预测能力表征

模型的预测能力由训练集和验证集快速生物降解化合物与缓慢生物降解化合物的预测准确率及加权平均准确度表征。生物降解性模型可使用预测正确率来验证与评价,公式如下:

$$\text{预测准确度} = \frac{n_Y}{n_N + n_Y} \times 100\% \quad (1)$$

式中, n_Y 代表预测准确的化合物个数, n_N 代表预测错误的化合物个数。

$$\text{加权准确度 (WA)} = \frac{A_F \times n_F + A_S \times n_S}{n_F + n_S} \times 100\% \quad (2)$$

式中, n_F 代表快速降解化合物个数, n_S 代表缓慢降解化合物个数, A_F 代表快速降解预测准确度, A_S 代表缓慢降解预测准确度。

1.5 模型应用域表征

应用 Leverage 法表征分类模型的应用域。采用公式(3)计算化合物杠杆值(h)。

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (3)$$

式中, x_i 是每个化合物的描述符向量, x_i^T 是 x_i 的转置矩阵。警戒值(h^*)设定为 $3(m+1)/n$, 其中, m 值为描述符个数, n 为训练集化合物个数。矩阵运算在 MATLAB(7.0)软件中执行。

2 结果(Results)

2.1 初级生物降解分类预测模型

采用 FT 方法,筛选出 15 个分子结构描述符,建立了初级生物降解预测模型,结果见图 1。化合物进入 FT No.1 节点,经 Class0 和 Class1 两式计算概率值,以 0.462 为分界点,使化合物进入 FT 1 和 FT 2 两个节点继续计算。初级生物降解功能树模型表达式为:

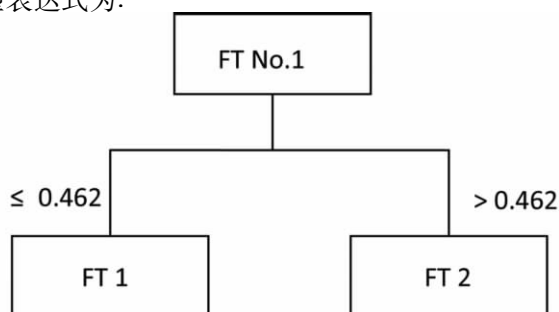


图1 初级生物降解性预测模型

Fig. 1 Prediction model of primary biodegradability

FT No.1:

$$\begin{aligned} \text{Class 0: } & -2.54 - 0.17 \times [n_N] - 0.23 \times [n_{Cl}] \\ & - 0.57 \times [n_{Cq}] - 0.14 \times [n_{CrH2}] - 0.14 [n_{CaR}] + 0.48 \\ & [n_{CONHR}] - 0.25 [n_{ROR}] - 0.31 [n_{PhX}] + 0.06 [C - \\ & 002] - 0.34 [C - 004] + 0.36 [C - 005] - 0.82 [C - \\ & 013] - 0.29 [E_{HOMO}] + 1.13 [Q_{Cmax}] + 0.08 [\text{dipole}] \end{aligned} \quad (4)$$

$$\begin{aligned} \text{Class 1: } & 2.54 + 0.17 [n_N] + 0.23 [n_{Cl}] + 0.57 \\ & [n_{Cq}] + 0.14 [n_{CrH2}] + 0.14 [n_{CaR}] - 0.48 [n_{CONHR}] + 0. \\ & 25 [n_{ROR}] + 0.31 [n_{PhX}] - 0.06 [C - 002] + 0.34 [C - \\ & 004] - 0.36 [C - 005] + 0.82 [C - 013] + 0.29 \\ & [E_{HOMO}] - 1.13 [Q_{Cmax}] - 0.08 [\text{dipole}] \end{aligned} \quad (5)$$

FT 1:

$$\begin{aligned} \text{Class 0: } & -2.71 - 2.08 [n_N] - 0.42 [n_{Cl}] - 0.18 \\ & [n_{Cq}] - 0.42 [n_{CrH2}] - 0.14 [n_{CaR}] + 0.48 [n_{CONHR}] + 0. \\ & 49 [n_{ROR}] - 0.31 [n_{PhX}] + 0.06 [C - 002] - 2.11 [C - \\ & 004] + 0.36 [C - 005] - 1.40 [C - 013] - 0.29 \\ & [E_{HOMO}] + 1.87 [Q_{Cmax}] + 0.08 [\text{dipole}] \end{aligned} \quad (6)$$

$$\begin{aligned} \text{Class 1: } & 2.71 + 2.08 [n_N] + 0.42 [n_{Cl}] + 0.18 \\ & [n_{Cq}] + 0.42 [n_{CrH2}] + 0.14 [n_{CaR}] - 0.48 [n_{CONHR}] - 0. \\ & 49 [n_{ROR}] + 0.31 [n_{PhX}] - 0.06 [C - 002] + 2.11 [C - \\ & 004] + 0.36 [C - 005] + 1.40 [C - 013] + 0.29 \\ & [E_{HOMO}] - 1.87 [Q_{Cmax}] - 0.08 [\text{dipole}] \end{aligned} \quad (7)$$

FT 2:

$$\begin{aligned} \text{Class 0: } & 0.29 - 0.17 [n_N] - 0.52 [n_{Cl}] + 0.27 \\ & [n_{Cq}] - 0.21 [n_{CrH2}] - 0.28 [n_{CaR}] + 0.03 [n_{CONHR}] - \\ & 0.51 [n_{ROR}] + 1.07 [n_{PhX}] + 0.06 [C - 002] - 0.34 [C - \\ & 004] + 1.26 [C - 005] - 0.82 [C - 013] - 0.04 \\ & [E_{HOMO}] + 1.55 [Q_{Cmax}] + 0.12 [\text{dipole}] \end{aligned} \quad (8)$$

$$\begin{aligned} \text{Class 1: } & -0.29 + 0.17 [n_N] + 0.52 [n_{Cl}] - 0.27 [n_{Cq}] \\ & + 0.21 [n_{CrH2}] + 0.28 [n_{CaR}] - 0.03 [n_{CONHR}] + 0.51 \times \\ & [n_{ROR}] - 1.07 \times [n_{PhX}] - 0.06 \times [C - 002] + 0.34 \times \\ & [C - 004] - 1.26 \times [C - 005] + 0.82 \times [C - 013] + 0. \\ & 04 \times [E_{HOMO}] - 1.55 \times [Q_{Cmax}] - 0.12 \times [\text{dipole}] \end{aligned} \quad (9)$$

初级生物降解性模型预测结果见表 1。结果表明,模型的预测能力较强,训练集准确度 $WA = 84.1\%$,验证集 $WA = 78.9\%$;此外,模型对快速生物降解化合物预测准确度略高于对缓慢生物降解化合物预测准确度(115 个快速降解化合物中正确预测 95 个,68 个缓慢降解化合物中准确预测 56 个)。初级生物降解预测模型包含 15 个分子描述符,其含义列于表 2。初级生物降解性受偶极矩、碳原子最大电荷、最高占据分子轨道能、碳原子中心碎片、杂原

子及醚键等因素的影响。

表1 初级生物降解性预测模型结果

Table 1 Performance of primary biodegradability prediction model

项 目	F 准确度 ¹⁾	S 准确度 ²⁾	WA ³⁾
训练集	84.3% (59/70)	83.9% (47/56)	84.1%
验证集	80.0% (36/45)	75.0% (9/12)	78.9%
汇总	95/115	56/68	—

注: ¹⁾ F 准确度: 快速降解化合物的预测准确度; ²⁾ S 准确度: 缓慢降解化合物的预测准确度; ³⁾ WA: 加权预测准确度。

表2 初级生物降解性预测模型中的15个分子描述符及其含义¹⁾

Table 2 Meanings of 15 molecular descriptors in primary biodegradability prediction model

描述符	含 义	最小值	最大值
E_{HOMO}	最高占据分子轨道能	-1.13×10	-6.11
Q_{Cmax}	碳最大电荷 -3.17×10^{-1}	8.16×10^{-1}	
dipole	分子偶极距	0	9.91
n_{N}	氮原子个数	0	5
n_{Cl}	氯原子个数	0	6
n_{Cq}	sp ³ 杂化季碳个数	0	3
n_{CH2}	环上 sp ³ 杂化仲碳个数	0	8
n_{CaR}	sp ² 杂化取代芳香碳个数	0	16
n_{CONHR}	脂肪族仲酰胺个数	0	2
n_{ROR}	脂肪族中的醚键个数	0	5
n_{PhX}	芳香环中卤-碳键个数	0	6
C-002	CH_2R_2	0	26
C-004	CR_4	0	3
C-005	CH_3X	0	4
C-013	CRX_3	0	1

注: ¹⁾ 最小值和最大值: 分别指正确预测的151个化合物描述符空间的最小值和最大值, 为预测模型描述符域。

2.2 最终生物降解分类预测模型

基于15个分子结构描述符, 使用功能树(FT)方法, 得到最终生物降解预测模型。在 FT No.1 节点, 经 Class0 和 Class1 两式计算概率值, 以 0.581 为分界点, 化合物被分为 Class=0 或再经 n_{CaR} 筛选, 大于3个即为 Class=1, 小于或等于3个则进入 FT 1 继续计算。最终生物降解功能树模型表达式为:

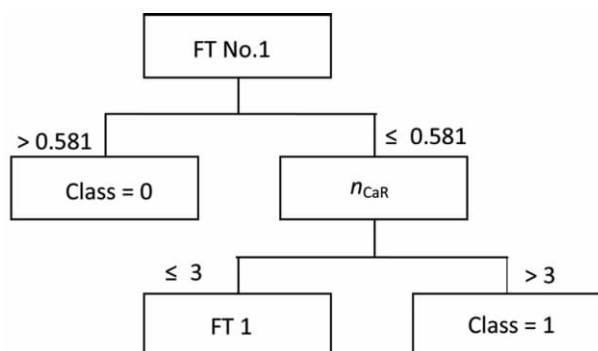


图2 最终生物降解性预测模型

Fig. 2 Prediction model of ultimate biodegradability

FT No.1:

$$\text{Class 0: } -3.18 - 0.31 \times [n_{\text{Cl}}] - 7.58 \times [\text{MWC}_{09}] - 1.78[\text{MATSi}_e] - 1.61[\text{MATSi}_p] + 0.52[\text{Mor}_{22u}] + 1.33[\text{Mor}_{29m}] + 2.73[\text{E}_{2u}] - 1.57[\text{E}_{2m}] - 8.51[\text{R}_{5v+}] - 0.18[n_{\text{CaR}}] + 1.52[n_{\text{NHRPh}}] - 0.20[\text{C}-006] + 0.41[\text{O}-056] + 0[\text{HOF}] - 0.37[\text{E}_{\text{HOMO}}] \quad (10)$$

$$\text{Class 1: } 3.18 + 0.31[n_{\text{Cl}}] + 7.58[\text{MWC}_{09}] + 1.78[\text{MATSi}_e] + 1.61[\text{MATSi}_p] - 0.52[\text{Mor}_{22u}] - 1.33[\text{Mor}_{29m}] - 2.73[\text{E}_{2u}] + 1.57[\text{E}_{2m}] + 8.51[\text{R}_{5v+}] + 0.18[n_{\text{CaR}}] - 1.52[n_{\text{NHRPh}}] + 0.20[\text{C}-006] - 0.41[\text{O}-056] + 0[\text{HOF}] + 0.37[\text{E}_{\text{HOMO}}] \quad (11)$$

FT1:

$$\text{Class 0: } -21.04 - 0.31[n_{\text{Cl}}] - 10.49[\text{MWC}_{09}] - 4.21[\text{MATSi}_e] - 0.08[\text{MATSi}_p] + 1.37[\text{Mor}_{22u}] + 3.54[\text{Mor}_{29m}] + 22.44[\text{E}_{2u}] - 3.35[\text{E}_{2m}] + 3.62[\text{R}_{5v+}] + 0.84[n_{\text{CaR}}] + 1.52[n_{\text{NHRPh}}] + 0.08[\text{C}-006] + 0.41[\text{O}-056] + 0[\text{HOF}] - 0.97[\text{E}_{\text{HOMO}}] \quad (12)$$

$$\text{Class 1: } 21.04 + 0.31[n_{\text{Cl}}] + 10.49[\text{MWC}_{09}] + 4.21[\text{MATSi}_e] + 0.08[\text{MATSi}_p] - 1.37[\text{Mor}_{22u}] - 3.54$$

$$[\text{Mor}_{29\text{m}}] - 22.44[\text{E}_{2\text{u}}] + 3.35[\text{E}_{2\text{m}}] - 3.62[\text{R}_{5\text{v}+}] - 0.84[\text{n}_{\text{CaR}}] - 1.52[\text{n}_{\text{NHRPh}}] - 0.08 \times [\text{C} - 006] - 0.41 \times [\text{O} - 056] + 0 \times [\text{HOF}] + 0.97 \times [\text{E}_{\text{HOMO}}] \quad (13)$$

最终生物降解模型预测结果见表 3。结果表明,模型的预测能力较强,训练集 $\text{WA} = 91.0\%$,验证集 $\text{WA} = 83.6\%$ 。最终生物降解模型包含 15 个分子描述符,其含义列于表 4。最终生物降解性取决于最终生成热、最高占据分子轨道能、电负性、极化率、氯原子个数等。

3 讨论 (Discussion)

3.1 应用域表征和离群点分析

采用 Leverage 方法表征了初级生物降解和最终生物降解预测模型的应用域,模型杠杆预警值分别为 0.381 和 0.393。模型应用域见图 3 和图 4, Y 为准确预测, N 为错误预测。初级生物降解预测模

型训练集中二十八烷醇、氨磺磷和曲拉通 X-100(聚氧乙烯-8-辛基苯基醚)及验证集中二碳酸二叔丁酯和十四烷酸十六烷基酯的 $h > h^*$,但仍被准确预测,可见该模型有一定的外推能力。另外,训练集中 20 个化合物及验证集中 10 个化合物虽然 $h < h^*$,但被错误预测。采用正确预测的 151 个化合物的描述符空间来表征初级生物降解预测模型的应用域,描述符域(最大值、最小值)见表 2。最终生物降解验证集中共有 8 个化合物 $h > h^*$,特别是阿特拉津和曲拉通 X-100(聚氧乙烯-8-辛基苯基醚)的 h 值远大于 h^* ,但仍被准确预测,可见该模型有一定的外推能力。训练集中 10 个化合物及验证集中 8 个化合物虽然 $h < h^*$ 但被错误预测。采用准确预测的 162 个化合物的描述符空间的范围来表征最终生物降解预测模型的应用域,描述符域(最大值、最小值)见表 4。

表 3 最终生物降解性预测模型结果

Table 3 Performance of ultimate biodegradability prediction model

项 目	F 准确度 ¹⁾	S 准确度 ²⁾	WA ³⁾
训练集	89.4% (59/66)	92.9% (52/56)	91.0%
验证集	86.5% (32/37)	79.2% (19/24)	83.6%
汇总	91/103	71/80	—

表 4 最终生物降解性预测模型中的 15 个分子描述符及其含义¹⁾

Table 4 Meanings of 15 molecular descriptors in ultimate biodegradability prediction model

描述符	含 义	最小值	最大值
n_{Cl}	氯原子个数	0	6
MWC_{09}	第 9 阶分子路径数	0	0.4
$\text{MAT}_{\text{S}_{1\text{e}}}$	原子拓扑距离为 1 的电负性性质	-5.98×10^{-1}	4.19×10^{-1}
$\text{MAT}_{\text{S}_{4\text{p}}}$	原子拓扑距离为 4 的极化率性质	-5.14×10^{-1}	5.50×10^{-1}
$\text{Mor}_{22\text{u}}$	无权重 3D 分子结构描述符	-1.13	3.77
$\text{Mor}_{29\text{m}}$	原子质量加权的 3D 分子结构描述符	-6.25×10^{-1}	1.32
$\text{E}_{2\text{u}}$	原子质量加权的原子在第二个主成分上的分布	2.65×10^{-1}	7.98×10^{-1}
$\text{E}_{2\text{m}}$	原子在第二个主成分上的密度	2.40×10^{-2}	1.41
$\text{R}_{5\text{v}+}$	范德华体积加权的 R 最大自相关指数	0	1.49×10^{-1}
n_{CaR}	sp^2 杂化取代芳香碳个数	0	16
n_{NHRPh}	芳香族仲胺个数	0	2
C - 006	CH_2RX	0	12
O - 056	醇羟基的氧原子中心碎片 0	4	
HOF	最终生成热	-1.75×10^3	4.75×10^2
E_{HOMO}	最高占据分子轨道能	-1.13×10	-6.11

注: ¹⁾最小值和最大值: 分别指准确预测的 151 个化合物描述符空间的最小值和最大值, 为预测模型的描述符域。

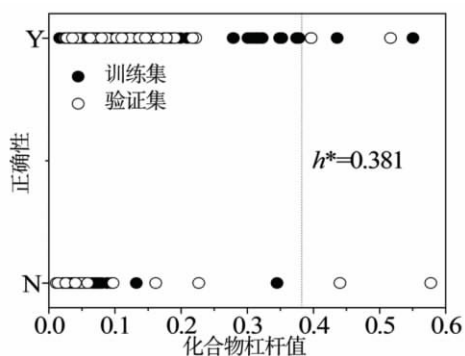


图3 初级生物降解模型的预测正确性与化合物杠杆值关系 (Y: 正确预测, N: 错误预测)

Fig. 3 Plots of prediction correctness versus h values of chemicals for primary biodegradation model

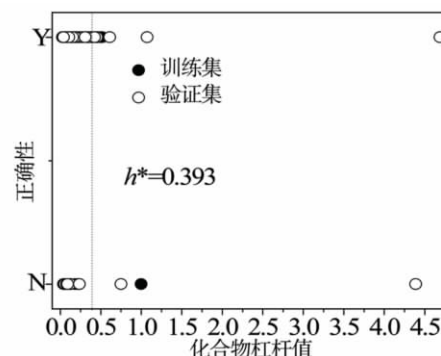


图4 最终生物降解模型预测正确性与化合物杠杆值关系 (Y: 正确预测, N: 错误预测)

Fig. 4 Plots of prediction correctness versus h values of chemicals for ultimate biodegradation model

3.2 生物降解预测模型的比较

本研究构建的预测模型与 EPI Suite 中 Biowin3 (最终生物降解模型)和 Biowin4 (初级生物降解模型)预测模型进行了比较(表 5-6)。EPI suite 中 Biowin3 和 Biowin4 预测模型是基于 36 个分子结构碎片及分子量建立。该模型考虑了化合物分子结构碎片对生物降解性的影响,但并没有考虑到分子碎片之间的相互作用对生物降解性的影响,而且新化学品可能未包含该模型的任何结构碎片,从而导致化合物的生物降解速率只能凭借分子量来评价^[16]。

本研究构建的生物降解预测模型,除分子结构碎片外还包含了描述分子整体特征的描述符:初级生物降解模型中包含了偶极矩、最高占据分子轨道能;最终生物降解模型中包含了最终生成热、最高占据分子轨道能、电负性性质、极化率性质等。经模型内外部验证结果表明,基于 15 个分子结构描述符所建立的预测模型具有较强的预测正确性。此外,本研究还采用 Leverage 方法和描述符空间的范围分别定义了模型的应用域。

表5 本研究的初级生物降解预测模型与 Biowin 4 模型的比较

Table 5 Performance comparison of the present model with EPI suite Biowin 4 model

模 型	算 法	化合物 个数 ¹⁾	描述符 个数	训练集准确度			验证集准确度		
				F(%) ²⁾	S(%) ³⁾	WA(%) ⁴⁾	F(%)	S(%)	WA(%)
本研究	多元线性回归	183	15	84.3 (59/70)	83.9 (47/56)	84.1 (106/126)	80.0 (36/45)	75.0 (9/12)	78.9 (45/57)
Biowin 4	多元线性回归	200	37	84.9 (101/119)	79.0 (64/81)	82.5 (165/200)	—	—	—

注: ¹⁾ 训练集和验证集中化合物的总个数; ²⁾ F(%): 训练集和验证集快速降解化合物预测准确度; ³⁾ S(%): 训练集和验证集缓慢降解化合物预测准确度; ⁴⁾ WA(%): 训练集和验证集总体加权预测准确度。

表6 本研究的最终生物降解模型与 Biowin 3 模型的比较

Table 6 Performance comparison of the present model with EPI suite Biowin 3 model

模 型	算 法	化合物 个数 ¹⁾	描述符 个数	训练集准确度			验证集准确度		
				F(%) ²⁾	S(%) ³⁾	WA(%) ⁴⁾	F(%)	S(%)	WA(%)
本研究	多元线性回归	183	15	89.4 (59/66)	92.9 (52/56)	91.0 (111/122)	86.5 (32/37)	79.2 (19/24)	83.6 (51/61)
Biowin 3	多元线性回归	200	37	93.5 (101/108)	71.7 (66/92)	83.5 (167/200)	—	—	—

注: ¹⁾ 训练集和验证集中化合物的总个数; ²⁾ F(%): 训练集和验证集快速降解化合物预测准确度; ³⁾ S(%): 训练集和验证集缓慢降解化合物预测准确度; ⁴⁾ WA(%): 训练集和验证集总体加权预测准确度。

综上,本研究采用 FT 算法,建立了水中好氧初级生物降解和最终生物降解预测模型。外部验证结果表明,模型有较高的预测准确度,初级生物降解训练集 WA 为 84.1%,验证集为 78.9%;最终生物降解训练集 WA 为 91.0%,验证集为 83.6%。应用域表征结果显示,所构建的模型可用于预测烷烃、烯烃、炔烃和芳香烃及其衍生物、醇、醛、酮、醚、酸和脂等多种类有机污染物在水中好氧初级和最终生物降解性。

通讯作者简介: 李雪花(1980—),女,环境工程博士,讲师,主要研究方向为污染生态化学和计算毒理学,发表学术论文 10 余篇。

参考文献:

- [1] Peijnenburg W J G M. Structure - activity relationships for biodegradation: A critical review [J]. *Pure and Applied Chemistry*, 1994, 66(9): 1931 - 1941
- [2] Rücker C, Kümmerer K. Modeling and predicting aquatic aerobic biodegradation: A review from a user's perspective [J]. *Green Chemistry*, 2012, 14(4): 875 - 887
- [3] 金志刚, 张彤, 朱怀兰. 污染物生物降解[M]. 上海: 华东理工大学出版社, 1997. 3
- [4] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. GB/T 21803-2008 化学品快速生物降解性 DOC 消减试验[S]. 北京: 中国标准出版社, 2008
- [5] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. GB/T 21856-2008 化学品快速生物降解性二氧化碳产生试验[S]. 北京: 中国标准出版社, 2008
- [6] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. GB/T 21856-2008 化学品快速生物降解性改进的 MITI 试验[S]. 北京: 中国标准出版社, 2008
- [7] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. GB/T 21802-2008 化学品快速生物降解性密闭瓶法试验[S]. 北京: 中国标准出版社, 2008
- [8] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. GB/T 21857-2008 化学品快速生物降解性改进的 OECD 筛选试验[S]. 北京: 中国标准出版社, 2008
- [9] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. GB/T 21801-2008 化学品快速生物降解性呼吸计量法试验[S]. 北京: 中国标准出版社, 2008
- [10] Philipp B, Hoff M, Germa F. Biochemical interpretation of quantitative structure activity relationships (QSAR) for biodegradation of N-Heterocycles: A complementary approach to predict biodegradability [J]. *Environmental Science & Technology*, 2007, 41 (4), 1390 - 1398
- [11] Rud n C, Hansson S O. Registration, evaluation, and authorization of chemicals (REACH) is but the first step how far will it take us? Six further steps to improve the European chemicals legislation [J]. *Environmental Health Perspectives*, 2010, 118(1): 6 - 10
- [12] Christensen F M, Eisenreich S J, Rasmussen K, et al. European experience in chemicals management: integrating science into policy [J]. *Environmental Science & Technology*, 2011, 45(1): 80 - 89
- [13] Boethling R S, Sommer E, DiFiore D. Designing small molecules for biodegradability [J]. *Chemical Reviews*, 2007, 107(6): 2207 - 2227
- [14] Raymond J W, Rogers T N, Shonnard D R, et al. A review of structure based biodegradation estimation methods [J]. *Journal of Hazardous Materials*, 2001, 84(2-3): 189 - 215
- [15] Boethling R S, Sabljic A. Screening level model for aerobic biodegradability based on a survey of expert knowledge [J]. *Environmental Science & Technology*, 1989, 23(6): 672 - 679
- [16] Boethling R S, Howard P H, Meylan W, et al. Group contribution method for predicting probability and rate of aerobic biodegradation [J]. *Environmental Science & Technology*, 1994, 28(3): 459 - 465

◆