人工神经网络及分子拓扑参数在酚类有机物 QSBR 研究中的应用*

瞿福平 杨义燕 冯旭东 戴猷元
(清华大学化学工程系, 北京 100084 Email: fpqu@tsinghua.edu.cn)

摘要 利用分子拓扑参数作为输入参数, 探索了人工神经网络对 27 种酚类有机物的定量结构-生物降解性能关系(QSBR). 结果表明, 将人工神经网络运用于有机物的生物降解性能建模是可行的. 所建模型预测结果和文献数据十分接近, 预测能力优于已有文献报道, 且能够较好区分同分异构体.

关键词 人工神经网络, QSBR, 酚类有机物.

An Application of Artificial Neural Networks and Molecular Topological Index for the QSBR of Phenolic Organics*

Qu Fuping Yang Yiyan Feng Xudong Dai Youyuan
(Department of Chemical Engineering, Tsinghua University, Beijing 100084 China Email: fpqu@tsinghua.edu.cn)

Abstract A quantitative structure-biodegradability relationships (QSBR)-type model using artificial neural networks (ANN) was established for the 27 phenolic compounds, in which molecular topological index are calculated and taken as the input parameters. The results show that the model developed can make a better agreement between predicted and observed values for the biodegradability of the tested compounds than ever before.

Keywords ANN, QSBR, phenolic organics.

随着人们对有机物生物降解机理的认识、测试方法及评价标准的不断完善以及部分可比性有机物生物降解性能参数的获得, 定量结构-生物降解性能关系(quantitative structure biodegradability relationship, QSBR)研究逐渐活跃起来[1~5]. 然而, 在过去的 QSBR 研究中, 结构参数大多采用有机物的理化参数, 数学方法也多为线性回归、逐步回归、模式识别等, 其预测准确度往往不尽人意. 因此, 探索新的建模方法已成为 QSBR 的研究热点. 基于生物降解过程的复杂性, 本文以环境中广泛存在的酚类有机物为研究对象, 应用目前最为系统的生物降解性能资料[6], 以分子拓扑参数为结构参数和输入参数, 将人工神经网络模型用于酚类有机物 QSBR 研究中, 并建立相应酚类有机物生物降解性能的预测模型.

1 神经网络建模原理与方法

人工神经网络建模是近年来发展起来的一门新兴的信息处理技术[7~9], 它不同于传统的数学处理方法, 可以实现 $n \rightarrow m$ 维空间的非线性映射, 处理非常复杂的问题[10]. 神经元是神经网络的基本组成部分. 图 1 是一个人工神经单元的示意图, 图中 $x_1, \cdots, x_n$ 表示其他神经单元的轴突输出, $w_1, \cdots, w_n$ 为其它神经元与第 i 个神经元突触的联接, $w_1, \cdots, w_n$ 可以是正, 也可以是负, 分别表示兴奋性突触和抑制性突触, 数值的大小根据突触不同的化学变化而各不相

* 国家"九五"科技攻关课题(The National Key Science and Technology Project during the Ninth Five-year Plan Period): 96-909-05-01
瞿福平: 男, 32 岁, 博士后, 副教授
收稿日期: 1998-12-07

同. 每个人工神经元满足:

$$y_i = f(u_i) = f\left(\sum_{j=1}^{n} w_j x_j - \theta_i\right) \quad (1)$$

式中, u_i 为上一层神经元输出的加权累加值, θ 为阈值, $f(u_i)$ 是一个单调上升的有限值函数, 称为传递函数. $f(u_i)$ 通常取如下非线性形式:

$$f(u_i) = \frac{1}{1 + e^{-u_i/\gamma}} \quad (2)$$

式中, γ 为非线性因子.

按照不同模型建立起的具有各种拓扑结构的人工神经网络已有几十种, 其中的反向传播模型(BP)是应用最为广泛的模型之一, 如图2所示.

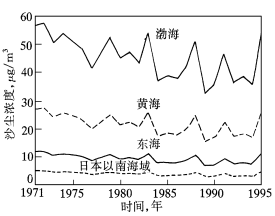

图1　神经元结构

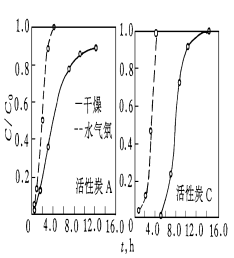

图2　三层神经网络

利用BP网络的学习功能, 给定一系列的输入输出学习模式令其学习, 并依据一定的学习规则调节各层节点之间的连接权重, 最终可使网络的实际输出与期望输出相比较达到一定的精度要求.

连接权重的调节规则为:

$$\Delta W_{ij}(n) = \epsilon\delta_j O_{ik} + \alpha\Delta W_{ij}(n-1) \quad (3)$$

式中, ΔW_{ji} 为隐含层中节点 j 与输入层中节点 i 之间连接权重的变化值; ϵ 为学习速率; α 为动量因子, O_{ik} 为学习模式 k 在节点 i 的输出, δ_k 为一中间参数.

2　网络输入参数的选取及处理

2.1　分子连接性指数的计算方法

从物质本身的结构出发, 影响生物降解的因素应该有溶解性、分子大小、疏水性能、电荷分布、空间排列等参数[11]. 网络的输入参数则应从这些参数中选取影响最大的因素. 在实际操作中, 有机物某些参数的缺乏常常成为制约QSBR研究的一大障碍. 如何用最简单的方法获得有机物的结构描述符成为人们研究的课题. 近年来, 根据拓扑理论发展起来的分子拓扑结构参数实现了这一需求. 它们具有简单、方便、不依赖于实验条件、在计算过程中不需考虑复杂电效应等优点. 在这类参数中, 影响最大的为Kier提出的分子连接性指数(MCI)[12,13]. 近年来, MCI在QSBR的研究中也逐渐得到应用[14].

分子连接性指数计算步骤为: ①写出化合物的结构式; ②写出隐氢图; ③标出原子点价; ④代入如下计算公式:

$$^0X = \sum \delta^{-1/2} \quad (4)$$

$$^1X = \sum (\delta_i \delta_j)^{-1/2} \quad (5)$$

……

$$^nX = \sum (\delta_i \delta_j \delta_k \cdots \delta_n)^{-1/2} \quad (6)$$

式中, δ_i 为成键原子的点价, 对于碳原子 $\delta_i = Z - h_i$, h_i 表示与该原子成键的 H 原子数. $i, j, k \cdots\cdots n$ 分别表示分子中依次排列的碳原子.

对于杂原子, $\delta_i = Z_v - h_i$, Z_v 表示杂原子价电子数, 相应的连接性指数用 $^nX^v$ 表示.

根据上述计算步骤得到的分子连接性指数, 可对分子结构差异实现定量化描述, 并具有以下特点: ①0阶、1阶和2阶路径项指数(^0X、^1X、^2X)可以反映分子的整体属性, 如分子大小、分子表面积和分子体积等. ②高阶路径项指数对于分子局部特征的描述十分有效. ③价指数可较好地反映带杂原子的杂环化合物或带杂原子取代基的物质的分子结构特征, 并可对杂原子进行校正.

本文计算了从0至4阶共14个参数, 它们是 ^0X, $^0X^v$, ^1X_p, $^1X^v_p$, ^2X_p, $^2X^v_p$, ^3X_p, $^3X^v_p$, ^3X_c, $^3X^v_c$, ^4X_p, $^4X^v_p$, $^4X_{pc}$, $^4X^v_{pc}$(指数中p为路径项, c为簇

项). 然后根据上述指数的特点, 选取了 7 个从 0 至 4 阶的价指数作为网络的输入参数, 它们是$^{0}X^{v}$, $^{1}X^{v}_{p}$, $^{2}X^{v}_{p}$, $^{3}X^{v}_{p}$, $^{3}X^{v}_{c}$, $^{4}X^{v}_{p}$, $^{4}X^{v}_{pc}$.

2.2 数据的预处理

在人工神经网络的建模过程中, 由于不同参数之间在数量级和变化幅度上的差异, 很可能使一些变化幅度小或绝对值小的数据被其他因素所掩盖, 从而影响神经网络的建模效果. 为了克服这一缺点, 目前已有多种方法用于数据的标准化处理, 比较而言. Z-变换法具有更为严谨的统计意义[15], 其处理方法如下:

$$Z_{ij} = (X_{ij} - \overline{X_j}) / S_j$$

式中, $\overline{X_j} = \frac{1}{m}\sum_{i=1}^{m} X_{ij}$, $S_j^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_{ij} - \overline{X_j})^2$

由于节点转换函数 $f(u)$ 的输出取值只能在 0～1 之间, 而实际的生物降解性能数据 ($\log q_{max}$) 变化范围在 0～2. 为了使网络在实际训练时易于收敛, 对学习模式中输出的生物降解性能数据进行了如下变换:

$$y_p = (y - 0.45) \times 0.55$$

从而可使网络的期望输出值处于 0.15～0.85 之间.

2.3 人工网络的建模及预测

理论上, 3 层人工神经网络可以实现任一连续函数或映射. 本文采用了 3 层网络模型, 经过多次试探性运算得出了收敛速度较快的 7-10-1 三层结构, 相应的权重初始值为 0～1 之间随机取值, 学习初始速率取 0.05, 学习速率变动系数取 1%, 动量因子取 0.8, 非线性因子取 2. 采用逐次抽取 1 个样本校验的方法 (leave-one-out) 对 27 个酚类有机物的生物降解性能进行定量预测, 即当预测某化合物的生物降解性时, 其余 26 个有机物的输入输出学习模式对构成训练集提供网络学习, 直至收敛到一定精度, 然后用此网络对该有机物进行预测. 例如, 当预测第 1 种有机物时, 第 2～27 种依次

表 **1** 酚类有机物分子连接性指数(预处理后)和生物降解性能数据

名 称	$X_1(^{0}x^{v})$	$x_2(^{1}X^{v}_{p})$	$x_3(^{2}X^{v}_{p})$	$x_4(^{3}X^{v}_{p})$	$x_5(^{3}X^{v}_{c})$	$x_6(^{4}X^{v}_{p})$	$x_7(^{4}X^{v}_{pc})$	y(exp)	y(pred)	δy	相对误差/%
2-氨基-4-磺酸基酚	1.8209	2.3481	2.3353	1.8985	2.5366	1.7644	1.9884	0.2206	0.2257	0.0051	2.31
2-氨基酚	-0.9837	-0.9548	-0.9748	-0.7209	-0.9253	-1.0200	-0.5646	0.4807	0.4733	-0.0074	-1.54
3-氨基酚	-0.9837	-0.9748	-0.8541	-1.0256	-0.6700	-0.7095	-0.9368	0.3163	0.3259	0.0096	3.04
4-氨基酚	-0.9837	-0.9748	-0.8657	-0.9208	-0.6700	-1.1298	-0.8566	0.4246	0.4286	0.0040	0.94
2-氯酚	-0.1160	-0.0271	0.0253	0.3557	-0.0300	-0.1876	0.4347	0.3828	0.3800	-0.0028	-0.73
3-氯酚	-0.1160	-0.0471	0.2177	-0.2336	0.3636	0.4217	-0.3303	0.5775	0.5684	-0.0091	-1.58
4-氯酚	-0.1160	-0.0471	0.2061	-0.0721	0.3636	-0.3574	-0.2064	0.6325	0.6281	-0.0044	-0.70
2-氯-4 硝基酚	1.6731	1.4667	1.3479	1.4638	1.1084	1.2633	1.3788	0.1507	0.1593	0.0086	5.71
2, 4-二氯酚	1.5314	1.5650	1.9627	1.6157	2.0773	2.3288	1.5937	0.3141	0.3081	-0.0060	-1.91
2, 3-二甲基酚	1.1140	1.1386	1.2275	2.2037	1.1485	1.0452	2.5554	0.6017	0.6114	0.0097	1.61
2, 4-二甲基酚	1.1140	1.1186	1.4640	1.1802	1.6134	1.6262	1.2167	0.5500	0.5578	0.0078	1.42
2, 4-二硝基酚	1.8148	1.3687	0.8268	0.9560	0.3201	1.0119	0.6393	0.1804	0.1969	0.0165	9.15
2, 5-二羟基苯甲酸	0.8062	0.7900	0.5189	0.5314	0.2900	0.7098	0.4232	0.7992	0.7843	-0.0149	-1.86
2, 4-二氨基酚	-0.2041	-0.2901	-0.1092	-0.1956	0.1473	-0.3083	0.0255	0.3460	0.3376	-0.0084	-2.43
1, 2-二羟基苯	-1.1867	-1.1715	-1.2087	-0.9730	-1.1349	-1.2139	-0.7983	0.7117	0.7008	-0.0109	-1.53
1, 3-二羟基苯	-1.1867	-1.1915	-1.1046	-1.2109	-0.9119	-0.9742	-1.0788	0.7205	0.7205	0.0000	0.00
1, 4-二羟基苯	-1.1867	-1.1915	-1.1163	-1.1194	-0.9119	-1.3105	-1.0090	0.7062	0.6677	-0.0385	-5.45
2-羟基苯甲酸	0.2296	0.3423	-0.0958	0.1837	-0.5406	0.3701	-0.0303	0.8399	0.8493	0.0094	1.12
4-羟基苯甲酸	0.2296	0.3223	0.0032	0.0913	-0.3054	-0.0502	-0.1943	0.8525	0.8531	0.0006	0.007
2-羟基甲苯	-0.3248	-0.2501	-0.2153	0.0967	-0.2452	-0.3882	0.1943	0.7040	0.6900	-0.0140	-1.99
3-羟基甲苯	-0.3248	-0.2701	-0.0401	-0.4239	0.1150	0.1496	-0.4760	0.7095	0.7074	-0.0021	-0.30
4-羟基甲苯	-0.3248	-0.2701	-0.0518	-0.2761	0.1150	-0.5430	-0.3630	0.7095	0.7069	-0.0026	-0.37
2-硝基酚	0.0258	-0.1254	-0.4958	-0.2267	-0.8182	-0.0993	-0.3539	0.3828	0.3653	-0.0175	-4.57
3-硝基酚	0.0258	-0.1454	-0.3967	-0.3855	-0.6053	-0.1901	-0.5458	0.4362	0.4256	-0.0106	-2.43
4-硝基酚	0.0258	-0.1454	-0.4084	-0.2981	-0.6053	-0.4398	-0.4784	0.4147	0.4019	-0.0128	-3.09
酚	-1.7633	-1.6392	-1.7197	-1.5472	-1.7436	-1.5152	-1.5312	-0.7992	0.8232	0.0240	3.00
1, 3, 5-三羟基苯	-0.6099	-0.7438	-0.4785	-0.9473	-0.0813	-0.2542	-0.6964	0.4917	0.4728	-0.0189	3.84

构成训练集,而预测第 2 种有机物时,则以第 3～27 种,第 1 种为序构成训练集,依次循环类推.用此方法,运行约 35000 次,可达到收敛精度,在 Pentium-166 微机上所需时间约 2.5h.

3 结果与讨论

计算结果如表 1 所示,图 3 直观地示出了实测值与预测值的比较.

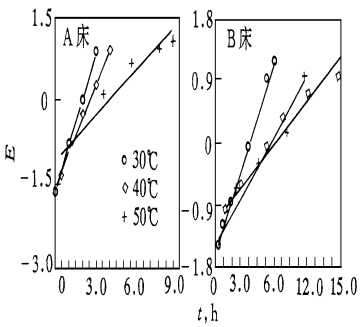

图 3 酚类有机物生物降解性能预测值与文献值的比较

从表 1 及图 3 可以看出,人工神网络模型计算结果和已有文献资料数据十分接近(均方根误差 σ= 0.013),且能够较好区分过去很难区分的同分异构体结构,该模型的预测能力优于目前居于领先水平的 Okey 等人[3]的报道.因此,采用分子连接性指数与人工神经网络的结合对于探索有机物生物降解性能规律是可行的.

参 考 文 献

1 Vaishnav D D, Boethling R S and Babeu L. Quantitative Structure-Biodegradability Relations for Alcohols, Ketones and Alicyclic Compounds. Chemosphere, 1987, **16**: 695～703

2 Okay R W. A QSBR Development Procedure for Aromatic Xenobiotic Degradation by Unacclimated Bacteria. Water Research, 1993, **65**: 772～780

3 Okay R W et al. A QSBR-based Biodegradability Model-A QSBR. Water Research, 1996, **30**(9): 2206～2214

4 何苗,张晓健,瞿福平等.杂环化合物好氧生物降解性能与其化学结构相关性的研究.中国环境科学, 1997, **17**(3): 199～202

5 瞿福平,张晓健,何苗等.氯代芳香化合物好氧生物降解性能与其化学结构相关性研究.环境科学, 1998, **19**(6): 26～28

6 Pitter P. Biodegradability of Organic Substances in the Aquatic Environment. USA: CRC Press, 1992. 65～83

7 孙益民,何鸣鸿,乔芝郁.神经网络建模计算稀土卤化物的熔化焓.自然科学进展, 1998, **8**(1): 108

8 宋新华,陈茁,俞汝勤等.人工神经网络用于对位取代苯酚定量构效关系的研究,中国科学(B 辑), 1993, **23**(3): 245

9 沈洲,韩朔睽,张爱茜等.利用人工神经网络研究含硫芳香族化合物结构与毒性的关系.环境化学, 1997, **16**(2): 138

10 焦李成.神经网络系统理论.西安:西安电子科技大学出版社. 1992. 1～16

11 刘次全等.量子生物学及其应用.北京:高等教育出版社, 1990. 380～389

12 Kier L B et al. Molecular Connectivity in Chemistry and Drug Research. Academic Press, 1986. 10～22

13 王连生.有机物定量结构-活性相关.北京:中国环境科学出版社, 1993. 12～24

14 王飞越等.有机物结构-活性定量关系及其在环境化学和环境毒理学中的应用.环境科学进展, 1992, 2(1): 26

15 萨特·L·考夫曼著,刘昆元译.聚类分析法解析分析化学数据.北京:化学工业出版社, 1990. 14