

醛类化合物分子结构与对鼠急性毒性关系

唐桂刚, 白乃彬(中国科学院生态环境研究中心, 北京 100085)

摘要: 通过检索美国国立职业与卫生研究所化学物质毒性效应登录(RTECS)1998 年版光盘系统, 搜集了 88 种醛类化合物对大鼠急性毒性口服 LD₅₀数据, 并利用遗传神经网络建立了醛类化合物分子结构与对大鼠急性毒性关系的 BP 神经网络模式, 模式的交互检验相关系数达 0.83~0.87, 具有较强的预报能力. 遗传神经网络是由遗传算法和神经网络耦合而成, 文章详细讨论了遗传神经网络的构造以及利用遗传神经网络建立 QSAR 模型的方法.

关键词: 醛类化合物; 遗传算法; 神经网络; 构效关系; 急性毒性

中图分类号: X174 文献标识码: A 文章编号: 0250-3301(2000)05-0089-05

Structure-Activity Relationships for Rat LD₅₀ of Miscellaneous Aldehydes

Tang Guigang, Bai Naibin(Research center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China)

Abstract: Acute toxicity data of 88 aldehydes to rat LD₅₀ were collected from optical disk system of Registry of Toxic Effects of Chemical Substances (RTECS) developed by National Institute for Occupational Safety and Health (NIOSH). Some models of artificial neural networks based on BP algorithm for relationships between acute toxicity of aldehydes to rat LD₅₀ and their molecular structure descriptors got from structural formula were then built by genetic neural network approach. The cross-validation correlation coefficient of these models were 0.83~0.87. Genetic neural network was obtained by integrating genetic algorithm and neural networks. How a genetic neural network was constructed and how a QSAR model was built by using it was described in detail in this paper.

Keywords: aldehydes; genetic algorithm; neural networks; QSAR; acute toxicity

醛类化合物应用较广, 且毒性较大, 为环境主要的污染物. 现生产与流通的醛类化合物约数百种, 仅有近百种作了急性毒性试验. 醛类化合物急性毒性的 QSAR 研究有助于对未作试验的醛类的预报, 解释毒性作用机理^[1,2].

1 样本

醛类化合物对大鼠的急性毒性数据来自于美国国立职业与卫生研究所化学物质毒性效应登录(RTECS)1998 年版光盘系统, 本文使用的是大鼠口服 LD₅₀(mg/kg), 在进行 QSAR 建模时取其对数, 学习集样本共 88 种, 数据见表 1.

在以往有关醛类化合物急性毒性的 QSAR 研究中, 作为自变量的分子结构描述符一般用 LOGP(1-辛醇/水分配系数的对数)和量子化学参数(如分子最低未占用轨道能量 E_{LUMO}), 获得了较好的线性关系模型^[1,2]. 但以往的研究中所用样本数较少(不超过 20 个), 包含的醛化合物种类不全面. 本文所用的样本集包括饱和脂肪醛、不饱和脂肪醛和芳香醛, 这样就有可

能建立具有预报各种醛类化合物急性毒性能力的 QSAR 模型. 依据醛类化合物分子结构式, 本研究选出了 16 种有可能对醛急性毒性有贡献的分子结构描述符, 见表 2.

2 GNN 方法

2.1 GNN 概述

遗传神经网络 GNN 是通过将遗传算法(Genetic Algorithm, GA)和基于反向传播算法(Back Propagation, BP)的神经网络耦合获得. 获得一个具有较好预报能力的 QSAR 模式必须选好参与建模的参数和从参数到特征的映射方法. BP 网络具有良好的非线性模拟能力, 在 QSAR 建模中得到广泛应用, 但建立 BP 网络模型需要确定参与网络训练的结构描述符以及网络拓扑结构和学习参数. 因而需要有某种方法找出合理的结构描述符、网络拓扑结构及学习参数的最佳组

基金项目: 国家自然科学基金资助项目(29877032)

作者简介: 唐桂刚(1973~), 男, 硕士, 主要研究方向为环境污染物的构效关系.

收稿日期: 1999-11-09

续表 1

化合物名称	分子结构描述符(描述符 A~ P 对应于表 2 中的 A~ P)																	LOG
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	(LD ₅₀)	
3-己烯醛	6	10	1	1	1	1	1	0	1	0	0	0	0	0	0	0	3.699	
顺式 3-己烯醛	6	10	1	1	1	1	1	0	1	0	0	0	0	0	1	0	3.193	
三聚乙醛	6	12	3	1	0	0	0	0	0	0	0	0	0	0	0	0	3.185	
环己烯-1-甲醛	7	10	1	1	2	2	1	1	0	0	0	0	0	0	0	0	3.021	
2,4-庚二烯醛	7	10	1	1	1	1	1	1	0	1	0	0	0	0	0	0	3.061	
2-庚烯醛	7	12	1	1	1	1	1	1	0	0	0	0	0	0	0	0	3.114	
2,3-二甲基-4-戊烯醛	7	12	1	1	2	2	2	0	0	1	0	0	0	0	0	0	3.753	
四聚乙醛	8	16	4	1	0	0	0	0	0	0	0	0	0	0	0	0	2.356	
二甲基-3-环己烯-1-甲醛	9	14	1	1	2	3	2	0	0	1	0	0	0	0	0	0	3.556	
2,6-壬二烯醛	9	14	1	1	1	1	1	1	0	0	0	0	0	0	1	0	3.699	
顺式-2,6-壬二烯醛	9	14	1	1	1	1	1	1	0	0	0	0	0	0	0	0	3.699	
2-壬烯醛	9	16	1	1	1	1	1	1	0	0	0	0	0	0	0	0	3.699	
2,6-二甲基-5-庚烯醛	9	16	1	1	2	1	1	0	0	0	1	0	0	0	0	0	3.699	
4-异丙烯基-1-环己烯-1-甲醛	10	14	1	1	2	2	1	1	0	0	0	0	0	0	0	0	3.236	
3,7-二甲基-2,6-辛二烯醛	10	16	1	1	1	2	1	1	0	0	0	0	0	0	0	0	3.695	
反式-3,7-二甲基-2,6-辛二烯醛	10	16	1	1	1	2	1	1	0	0	0	0	0	0	0	1	2.699	
2,4,6-三甲基-3-环己烯-1-甲醛	10	16	1	1	2	4	1	0	0	1	0	0	0	0	0	0	3.653	
2-癸烯醛	10	18	1	1	1	1	1	1	0	0	0	0	0	0	0	0	3.699	
10-十一碳烯-1-醛	11	20	1	1	1	1	1	0	0	0	0	0	0	0	0	0	3.699	
四氢苯甲醛	7	10	1	1	2	2	1	1	0	0	0	0	0	0	0	0	3.021	
1,2,3,6-四氢苯甲醛	7	10	1	1	2	2	1	0	0	1	0	0	0	0	0	0	3.391	
2,4-庚烯醛	7	10	1	1	1	1	1	1	0	1	0	0	0	0	0	0	3.061	
苯甲醛	7	6	1	1	1	1	1	0	0	0	0	1	0	0	0	0	3.114	
苯乙醛	8	8	1	1	1	1	2	0	0	0	0	0	1	0	0	0	3.190	
甲基苯甲醛	8	8	1	1	2	2	2	0	0	0	0	0	1	0	0	0	3.352	
alpha-甲基苯乙醛	9	10	1	1	2	2	2	0	0	0	0	0	1	0	0	0	3.447	
3-苯基-1-丙醛	9	10	1	1	1	1	2	0	0	0	0	0	0	1	0	0	3.699	
3-苯基-2-丙烯醛	9	8	1	1	1	1	2	1	0	0	0	0	0	1	0	0	3.346	
对-丙基-苯甲醛	10	12	1	1	2	2	1	0	0	0	0	1	0	0	0	0	3.204	
对-异丙基-苯甲苯	10	12	1	1	2	2	1	0	0	0	0	1	0	0	0	0	3.143	
2-(对-甲基)-丙醛	10	12	1	1	2	2	2	0	0	0	0	0	1	0	0	0	3.544	
3-苯基-alpha-甲基-2-丙烯醛	10	10	1	1	1	2	2	1	0	0	0	0	0	1	0	0	3.312	
对-异丙基苯基乙醛	10	14	1	1	1	2	2	0	0	0	0	0	1	0	0	0	3.613	
3-苯基-alpha-丁基-2-丙烯醛	13	16	1	1	1	2	3	1	0	0	0	0	0	1	0	0	3.643	
2,2-二甲基-3-(对-乙基苯基) 丙醛	13	18	1	1	3	1	2	0	0	0	0	0	0	1	0	0	3.699	
4-(1,1-二甲ethyl基)-苯丙醛	13	18	1	1	1	2	2	0	0	0	0	0	0	1	0	0	3.431	
2-甲基-3-(对-异丙基苯基) 丙醛	13	18	1	1	2	1	2	0	0	0	0	0	0	1	0	0	3.581	
2-苯亚甲基-庚醛	14	18	1	1	2	3	3	0	0	0	0	0	1	0	0	0	3.572	
3-(1,1-二甲ethyl基)-alpha-甲基-苯丙醛	14	20	1	1	1	2	2	0	0	0	0	0	0	1	0	0	3.568	
2-苯亚甲基-辛醛	15	20	1	1	2	2	3	0	0	0	0	0	1	0	0	0	3.491	

表 2 可能对醛类急性毒性有贡献的分子结构描述符

描述符类型	描述符			
原子数	A	碳原子总数		B 氢原子数
	C	氧原子数(醛基数)		
碳原子数	D	α 位碳原子数	E β 位碳原子数	F γ 位碳原子数
	G	δ 位碳原子数		
不饱和和脂肪醛	H	α 碳原子		I β 碳原子
	J	γ 碳原子		K δ 碳原子
双键的位置	L	醛基		M α 碳原子
	N	β 碳原子		
苯基位置	O	顺式结构		P 反式结构

合. 用遗传算法来进行特征选择和网络优化. 这种优化

-建模的组合方法近年来在 QSAR 中有较多的研究和应用. 在优化-建模组合方法中优化方法用得最多的是遗传算法, 最早有 Rogers 等^[3]和 Luke^[4]将遗传算法和线性回归组合, 近来又有将遗传算法和偏最小二乘法 (Partial least squares, PLS) 组合. 在遗传算法和神经网络的组合 (遗传神经网络, GNN) 中, 遗传算法的作用表现在 2 方面: 网络拓扑结构优化^[5]和特征筛选^[6]. Sung-Sau So 等^[6]对用遗传神经网络进行特征选择从而建立较好的网络模式作了深入细致的研究, 但仍然没有充分解决好如下问题: 网络拓扑结构; 遗传算法个体适应值, 搜索更多可能的特征组合.

2.2 遗传算法

本文使用遗传算法的 2 个变种: 遗传函数近似 (genetic function approximation, GFA)^[3]和进化程序 (evolutionary programming, EP)^[4]。2 种方法中, 遗传个体采用二进制编码, 结构见图 1, 适应值取神经网络对样本预报值与实验值之差的平方和的倒数。其第一步皆随机产生 N 个个体(二进制串), 每个个体包含参与建模的参数和网络结构参数(见图 1); 而第二步, GFA 和 EP 采用不同的策略产生下一代个体: GFA 途径是遗传算子复制、杂交和变异, 即首先按个体适应值的大小的比例将本代个体复制到下一代, 适应值大的个体在下一代有较多的表现机会, 小的个体被淘汰, 接着杂交算子交换个体遗传基因, 基于“基因块”假设, 较好的个体杂交后容易产生更好的个体, 最后变异算子按一定比例改变个体的二进制位; 而 EP 只采用变异的方法产生下一代个体, 对二进制串的位进行异或操作, 具体方法是先产生一 $0 \sim 1$ 的随机数 x , 若 x 在 $0 \sim 0.5$ 、 $0.5 \sim 0.8$ 、 $0.8 \sim 0.9$ 或 $0.9 \sim 1.0$, 则分别选择个体二进制串中的一位、二位、三位或四位作变异操作, 计算新产生 N 个个体适应值, 和原来的 N 个个体放在一起共 $2N$ 个个体, 按适应值大小排序, 前 N 个个体进入下一代, 这样, 前一代中较差的个体将被新产生的更好的个体所代替。

隐含层数及 各层节点数	学习 参数 α	动量 因子 η	初始 权值	参与 变量
----------------	-------------------	-----------------	----------	----------

图 1 进行特征选择与网络优化的 GNN 个体结构

2.3 BP 网络

在 GNN 中 BP 网络作为从特征到特性的映射方法, 它的输入描述符、拓扑结构及学习参数由 GNN 个体遗传因子确定, 在 Sung-Sau So 等^[6]的研究中, 网络结构作了限定(3-3-1), 这样就可能会漏掉一些好的网络模式, 比如 4 个输入描述符的模式。在本文中, 输入描述符及网络拓扑、学习参数都由 GNN 学习获得。规定 BP 网络输入描述符的个数小于样本数的 $1/5$, 网络最多包含 2 个隐含层数, 每个隐含层最多包含 16 个节点。网络学习参数包含学习参数 α 、动量因子 η 和网络的初始权值。

2.4 GNN 个体适应值

一般来说, 衡量一个模型质量和可信度标准是该模型的预报能力。为建立一个基于预测的模型, 个体适应值最好采用交互检验的相关系数(XR)或交互检验的误差(本文用样本预测值与实验值残差平方和, 记为 $XRmsE^2$)。获得它们的方法是: 每次从样本集中取一个样本作预测, 余下样本训练网络, 直至逐个样本皆被预测, 计算预测值和实验值的相关系数或残差平方和。但由于 BP 网络训练速度较慢, 用 XR 或 $XRmsE^2$ 作为个体适应值在样本数较大时将耗费大量机时。用表 1 的样本和表 2 的特征参数, 遗传算法群体数 60, 经过 40 代演化, 在 pentium II 333MHz 微机上大约需运行近 300h。为此, 本文对计算交互检验的相关系数或误差作了修正, 即每次随机取一定比例样本(如 10%)作预测, 剩下样本(如 90%)用于训练网络, 直至每个样本都被预测一遍。称之为部分交互检验相关系数(pXR)或部分交互检验误差(pXRmsE²)。每次取 10% 样本预测的记作 pXR_{10} 或 $pXRmsE_{10}^2$, 每次取 20% 样本预测的记作 pXR_{20} 或 $pXRmsE_{20}^2$ 。这样对于表 1 样本, 以 $pXRmsE_{10}^2$ 作为个体适应值, 用 pentium II 333MHz 微机, GNN 训练大约需要 30h。

3 结果与讨论

用 GNN 训练表 1 的样本, 其中 GFA 算法中的参数: 杂交率为 0.6, 变异率为 0.003, 群体规模 60, 最大演化代数 50; EP 算法参数: 全体规模 50, 最大演化代数 50。二者都以 $pXRmsE_{10}^2$ 作为个体适应值, GFAANN 算法结果如表 3, EP-ANN 算法结果如表 4。

由表 3 和表 4 可以看出, 经 GNN 训练后, 用最简单的描述符建立了具有较强预报能力的网络模型, 交互检验的相关系数达 0.83~0.87。表 3 和表 4 的所有模型中都包含 A、B、C、F、H 和 N, 说明它们是表达醛类大鼠 LD₅₀ 的核心描述符。一般认为, 化合物毒性显示取决于该化合物的穿透性和生物反应活性^[1,2]。描述符 A、B、C、F、N 与化合物形状和极性有关, 反映了化合物的穿透性。醛类化合物的生物反应活性常表现为碳-氧、碳-碳双键的断裂。描述符 C 为碳-氧双键的个数, C 越大化合物生物活性相对来说也越大。碳-碳双键中 GNN 只选用与醛基相邻的双键, 说明其它双键与

表 3 GF-ANN 训练结果

编号	学习参数 α	动量因子 η	隐含层节点数	初始权值	参与建模结构描述符	$XRmsE^2$	XR	$pXRmsE_{10}^2$
1	0.837	0.574	11	0.5	A B C F H J M N P	0.0335	0.840	0.0257
2	0.767	0.690	16	0.5	A B C F H J M N P	0.0337	0.839	0.0263
3	0.760	0.574	10	0.5	A B C F H J L M N P	0.0274	0.870	0.0265

表 4 EP-ANN 训练结果

编号	学习参数 α	动量因子 η	隐含层节点数	初始权值	参与建模结构描述符	$XRmsE^2$	XR	$pXRmsE_{i0}^2$
4	0.822	0.636	11	0.25	A B C F H J M N P	0.0351	0.831	0.0239
5	0.922	0.767	11	1.0	A B C F H J N P	0.0391	0.812	0.0250
6	0.760	0.636	9	0.25	A B C F H J M N P	0.0318	0.849	0.0265
7	0.984	0.333	11	1.0	A B C F H N P	0.0312	0.851	0.0266
8	0.884	0.651	11	0.25	A B C F H J M N	0.0353	0.831	0.0272
9	0.930	0.654	11	0.25	A B C F H J L M N P	0.0335	0.843	0.0273

醛类化合物毒性没有显著关系。

GFA 为传统型遗传算法,侧重于基因的重组,与自然界的有性进化相似;EP 算法中个体由单个父代个体变异而来,相当于自然界的无性进化。由表 3 和表 4 可知,GFA 和 EP 都获得了较好的结果。

在表 3 和表 4 中, $pXRmsE_{i0}^2$ 与 $XRmsE^2$ 相比,前者略小,这是因为计算 $pXRmsE_{i0}^2$ 时,每次随机取 10% 的样本作预测,直至每个样本被预测一遍,这样即使使用相同的参数 $pXRmsE_{i0}^2$ 计算值也不定相同。因而 $pXRmsE_{i0}^2$ 与 $XRmsE^2$ 二者不完全成正比,用 XR 或 $XRmsE^2$ 衡量模型预报能力时,以 $pXRmsE_{i0}^2$ 作为遗传算法个体适应值不一定找到最优解。笔者用 $pXRmsE_{i0}^2$ 作为个体适应值是基于这样的假设: $XRmsE^2$ 值较小则 $pXRmsE_{i0}^2$ 也较小; $XRmsE^2$ 值较大则 $pXRmsE_{i0}^2$ 不会太小。从表 3 和表 4 可以看出,用 $pXRmsE_{i0}^2$ 作个体适应值不一定获得最优结果,但获得了满意的结果。

参考文献:

1 Mark T D Cronin, T Wayne Schultz. Structure-toxicity relationships for three mechanisms of action of toxicity to

vibrio fisheri, Ecotoxicology and environmental safety, 1998, **39**(1): 65~ 69.

2 Schultz T W, Bryant S E, D T Lin. Structure-toxicity relationships for tetrahymena: aliphatic aldehydes, Bull. Environ. Contam. Toxicol. , 1994, **52**(2), 279~ 285.

3 Brian T Luke. Evolutionary programming applied to the development of quantitative structure-activity relationships and quantitative structure-property relationships. J. Chem. Inf. Comput. Sci. , 1994, **34**(6): 1279~ 1287.

4 David Rogers, A J Hopfinger. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships, J. Chem. Inf. Comput. Sci. , 1994, **34**(4): 854 ~ 866.

5 Jian-hui Jiang, J rhong Wang, Xin-hua Song at al. Network training and architecture optimization by a recursive approach and a modified genetic algorithm, Journal of chem ometrics, 1996, **10**(3): 253~ 267.

6 Sung-Sao So, Martin Karplus. Evolutionary optimization in quantitative structure-activity relationship: An application of genetic neural networks, J. Med. Chem. , 1996, **39**(7): 1521 ~ 1530.

关于提供垃圾处理技术和产业政策专题文献调研服务的启事

近年来,垃圾处理和垃圾产业成为我国科技和产业相关单位及人员关注的热点,我中心对此进行了专题文献调研,可为用户提供专门的资料集一册,A4纸约300页,70余万字。内容包括我国环保状况与形势、环保产业状况与对策、环境管理的相关政策、国外与我国(包括港澳台地区)垃圾处理状况、对策与技术等五大部分,需要者请与中国科学院生态环境研究中心情报编辑室范秀英联系,电话:010-62932252,通讯地址:北京2871信箱,邮编100085。