

人工神经网络用于异噻唑啉酮的定量结构-活性关系研究

何 琴¹, 黄保军², 李公春¹

(1. 许昌学院化学化工学院, 许昌 461000; 2. 许昌学院表面微纳米材料研究所, 许昌 461000)

摘 要: 采用误差反传前向人工神经网络 (artificial neural network, ANN) 建立了 21 种 2-(4-取代-苯基)-3-异噻唑啉酮类化合物的结构与其抗菌活性之间的定量关系模型 (ANN 模型), 以 21 种 3-异噻唑啉酮类化合物的量子化学参数和拓扑指数作为输入、抗菌活性作为输出, 所构建网络模型的交叉检验相关系数为 0.991 6、标准偏差为 0.080 1、残差绝对值 ≤ 0.221 , 应用于外部预测集, 预测集相关系数为 0.973 1; 而多元线性回归 (multiple linear regression, MLR) 法模型的相关系数为 0.841 8、标准偏差为 0.303 9、残差绝对值 ≤ 0.636 。结果表明, ANN 模型获得了比 MLR 模型更好的拟合效果。

关键词: 异噻唑啉酮; 定量结构-活性关系; 人工神经网络; 抗菌活性; 大肠杆菌

中图分类号: O622

文献标识码: A

文章编号: 1672-352X (2012)02-0323-04

QSAR of 3-isothiazolinone compounds using artificial neural network

HE Qin¹, HUANG Bao-jun², LI Gong -chun¹

(1. College of Chemistry and Chemical Engineering, Xuchang University, Xuchang 461000;

2. Institute of Surface Micro and Nano Materials, Xuchang University, Xuchang 461000)

Abstract: The study of the quantitative structure-activity relationship (QSAR) on 21 kinds of 2-(4-substituted-phenyl)-3-isothiazolinones was established by the artificial neural network based on the back propagation algorithm. For the artificial neural network method, the quantum chemical parameters about structure and the molecular topological index were used as the inputs of the neural network, and the antibacterial activities as the outputs of the neural network. As a result, the leave-one-out cross-validation regression coefficient was 0.991 6; the standard error was 0.080 1; the correlation coefficient of the test set was 0.973 1 and the absolute values of residual were less than 0.221. For comparison, the QSAR model was set up by multiple linear regressions (MLR) method. For the model built by MLR, the correlation coefficient was 0.841 8; the standard error was 0.303 9 and the absolute values of residual were less than 0.636. The results showed that the performance of neural network method is better than that of MLR method.

Key words: 3-isothiazolinone; quantitative structure-activity relationship; artificial neural network; antibacterial activity; *Escherichia coli*

异噻唑啉酮类化合物是一种广谱杀菌剂, 在工业生产、农林环保、海洋防污等领域有着广泛的应用^[1]。因其具有高效、低毒、环境友好等特点而引起生物、医药、化学等方面专家的广泛关注。近年来, 关于该类化合物的合成及抗菌活性的研究、定量构效关系 (quantitative structure-activity relationship, QSAR) 的研究虽然较多^[2-5], 但是其 QSAR 研究大多采用多元线性回归方法^[6-7], 采用人工神

神经网络方法的极少。为此, 作者采用人工神经网络方法对异噻唑啉酮类化合物的结构与其对大肠杆菌的抗菌活性进行建模, 并与多元线性回归法所建模型进行比较。

1 材料与方法

1.1 数据来源与预处理

论文所有数据均引自文献[7], 文献[7]在建模过

程中发现奇异值,并且剔除奇异值后重新建模。作者也将奇异值剔除选取 21 种 3-异噻唑啉酮类化合物。21 种 3-异噻唑啉酮类化合物的结构参数分别是量子化学参数(硫原子的亲核前线电子密度、S(1)-N(2)Wiberg 键极)和拓扑指数(3D-Balaban 指数、Schultz 分子价拓扑指数)。量子化学参数采用 Gaussian 03 量子化学程序包,在 B3YP/6-311G(d, p) 理论水平下计算得到,拓扑指数采用 Dragon 软件计算。

1.2 建模方法

QSAR 研究中,多元线性回归(MLR)是经典的建模方法^[8-9],该方法可以获得因果模型,且模型直观明了、物理意义明确。但是,该方法也存在一定的不足,如要求模型变量参数相互正交、化合物或样本数大于描述变量等^[9-10]。而人工神经网络方法因其独特的结构与功能,能够模拟任何因果关系不明确、推理规则不确定的非线性问题^[11-13]。

BP 网络是应用最为广泛的人工神经网络之一,它是一个以有向图为拓扑结构的动态系统,模拟人脑神经系统对连续的和间断的数值输入做出反馈而完成线性的或非线性的学习及预测工作,完成从 i 维空间(输入节点数为 i)到 m 维空间(输出节点数为 m)的映射,输出公式为:

$$y_i = f\left(\sum_{i=1}^i w_{ij} x_i\right)$$

BP 网络由若干层神经元组成,神经元之间通过权重连接,但同层的神经元之间没有联系。其结构如图 1 所示,图中圆圈表示神经元。在 BP 网络中,网络是否停止学习由学习目标函数决定。含有 N 个样本的训练集的学习目标函数为:

$$E_{AV} = 1/2n \sum_{j=1}^n (d_j(n) - y_j(n))^2$$

其中, $d_j(n)$ 为节点 j 在第 n 次迭代中的期望输出, $y_j(n)$ 为实际输出。当输出达到目标函数要求时,网络停止学习,并构建好各网络参数。当输出未达到函数要求时,计算误差,并将误差信号沿各层反向传播,同时按梯度下降法修正权值,开始新一轮学习,直到误差达到要求。权值修正公式为:

$$\omega_{ji}(n+1) = \omega_{ji}(n) + \eta \delta_j(n) y_i(n)$$

其中 $w_{ji}(n)$ 为第 n 次迭代时节点 i, j 之间的连接权值, η 为学习步长(速率), $\delta_j(n)$ 为局部梯度, $y_i(n)$ 为来自 i 节点的输出(作为 j 节点的输入)。

采用 ANN 进行 QSAR 建模时,选择 2-(4-甲基-苯基)-3-异噻唑啉酮、5-氯-2-甲基-3-异噻唑啉酮、2-(4-甲基-苯基)-3-异噻唑啉酮和 5-氯-2-(4-

乙氧羰基-苯基)-3-异噻唑啉酮作为外部预测集,其它 17 个样本作为训练集。

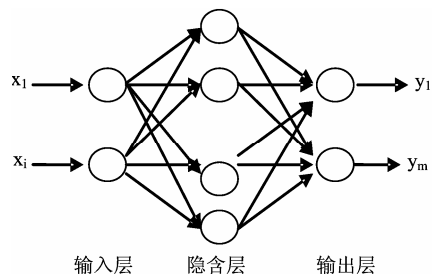


图 1 BP 网络结构图

Figure 1 BP neural network's structure

1.3 模型的检验

模型验证在 QSAR 建模中非常重要,其结果直接决定模型外部预测能力和真实有效性。交互检验(cross validation)是目前较为广泛使用的一种模型检验方法,然而有研究结果显示^[14]交互检验相关系数的大小与模型预测能力并没有明显的相关关系,对模型预测能力的评价只能通过外部样本集即预测集来进行。模型的外部预测能力可以用外部预测样本复相关系数和外部预测集交互检验来衡量。以 ANN 进行建模,模型有效性通过 R (相关系数)、 Rcv (LOO 法交互检验相关系数)、 Se (标准偏差)等加以判别。

2 结果与分析

2.1 QASR 模型的建立

选择硫原子的亲核前线电子密度、S(1)-N(2)Wiberg 键极、3D-Balaban 指数、Schultz 分子价拓扑指数作为输入参数、抗菌活性作为输出,构建 21 种 3-异噻唑啉酮类化合物的 3 层神经网络 QSAR 模型。神经网络参数如下:网络拓扑结构为 $4 \times 4 \times 1$, 输入层、隐含层和输出层的传递函数均为 logsig 函数,目标函数为 1×10^{-9} ,学习速度为 0.1,迭代次数为 10 000 次。

人工神经网络的 QSAR 模型样本模拟结果及实验值的残差分析见表 1 和图 2。从表 1 和图 3 可以看出,与多元线性回归法相比,神经网络的预测值与实验值之间的残差更小,残差各点在纵轴零对应的直线上下基本均匀分布,各残差值的绝对值均小于 0.221。这说明用神经网络预测异噻唑啉酮的抗菌活性更为准确,拟合能力比多元线性回归法更好。

2.2 QSAR 模型的检验

为了检验所建 QSAR 模型的稳定性,分别采用

了留一交互检验检验其稳定性、外部样本集检验其外部预测能力。所得 ANN 模型和 MLR 模型的相关系数 (R)、标准差 (Se)、留一交互检验相关系数 (Rcv) 分别见表 2、图 3 和图 4, 结果表明神经网络的稳定性更好。为了进一步验证模型的外部预测能力, 同时, 便于与 MLR 模型进行比较, 分别选用 2-(4-甲氧基苯基)-3-异噻唑啉酮和 5-氯-2-(4-甲基

苯基)-3-异噻唑啉酮作为外部预测集, 另外, 随机增加了 2-(4-甲硫基苯基)-3-异噻唑啉酮和 5-氯-2-(4-乙氧羰基苯基)-3-异噻唑啉酮为外部预测集, 当预测集样本数为 4、训练集样本数为 17 时, 所得结果见表 1, 训练集相关系数 R 为 0.991 6, 预测集相关系数 R 为 0.973 1, 模型预测能力较强、稳定性良好。

表 1 BP 网络和多元线性回归的预测结果

Table 1 Results of BP-ANN and MLR

化合物 Compound	BP-ANN预测结果			MLR预测结果 ^[7]		
	Prediction results of BP-ANN			Prediction results of MLR		
	实验值 Observed value	计算值 Calculated value	残差 Residual	实验值 Observed value	计算值 Calculated value	残差 Residual
2-(4-氯苯基)-3-异噻唑啉酮	4.449	4.282	-0.167	4.449	4.608	0.159
2-(4-硝基苯基)-3-异噻唑啉酮	3.869	3.857	-0.012	3.869	3.553	-0.316
2-(4-乙氧羰基苯基)-3-异噻唑啉酮	3.697	3.723	0.026	3.697	3.468	-0.229
2-(4-三氟甲基苯基)-3-异噻唑啉酮	4.088	4.105	0.017	4.088	3.728	-0.360
2-苯基-3-异噻唑啉酮	4.151	4.166	0.015	4.151	4.545	0.394
2-(4-甲基苯基)-3-异噻唑啉酮	3.980	3.978	-0.002	3.980	4.339	0.359
2-(4-甲氧基苯基)-3-异噻唑啉酮*	3.918	3.904	-0.014	3.918	4.378	0.460
2-(4-乙氧基苯基)-3-异噻唑啉酮	4.645	4.633	-0.012	4.645	4.201	-0.444
2-(4-丙氧基苯基)-3-异噻唑啉酮	4.070	4.088	0.018	4.070	3.974	-0.096
2-(4-二甲基氨基苯基)-3-异噻唑啉酮	4.041	4.028	-0.013	4.041	3.951	-0.090
2-(4-甲硫基苯基)-3-异噻唑啉酮*	4.047	4.142	0.095	4.047	4.529	0.482
5-氯-2-(4-氯苯基)-3-异噻唑啉酮	5.912	5.691	-0.221	5.912	5.276	-0.636
5-氯-2-(4-硝基苯基)-3-异噻唑啉酮	3.709	3.672	-0.037	3.709	4.033	0.324
5-氯-2-(4-乙氧羰基苯基)-3-异噻唑啉酮*	3.753	3.753	0.000	3.753	4.029	0.276
5-氯-2-(4-三氟甲基苯基)-3-异噻唑啉酮	4.048	4.053	0.005	4.048	4.258	0.210
5-氯-2-苯基-3-异噻唑啉酮	5.546	5.558	0.012	5.546	5.244	-0.302
5-氯-2-(4-甲基苯基)-3-异噻唑啉酮*	4.653	4.479	-0.174	4.653	5.050	0.397
5-氯-2-(4-甲氧基苯基)-3-异噻唑啉酮	5.382	5.455	0.073	5.382	5.167	-0.215
5-氯-2-(4-乙氧基苯基)-3-异噻唑啉酮	4.708	4.722	0.014	4.708	4.959	0.251
5-氯-2-(4-丙氧基苯基)-3-异噻唑啉酮	4.430	4.551	0.121	4.430	4.736	0.306
5-氯-2-(4-二甲基氨基-苯基)-3-异噻唑啉酮	5.007	4.955	-0.052	5.007	4.935	-0.072

注: *为外部预测集; 文献[7]仅仅选用 2-(4-甲氧基-苯基)-3-异噻唑啉酮和 5-氯-2-甲基-3-异噻唑啉酮作为预测集。

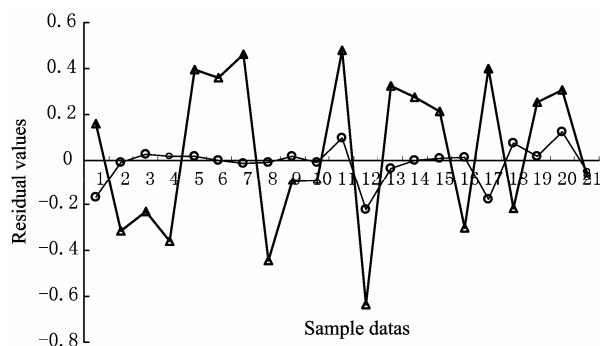


图 2 人工神经网络和多元线性回归预测残差散点图

Figure 2 Plots of residual values of ANN and MLR

通过文献[7]的多元线性回归可知, 硫原子的亲核前线电子密度、 $S(1)-N(2)$ Wiberg 键极、3D-Balaban 指数、Schultz 分子价拓扑指数、异噻唑啉酮环与苯环之间的二面角、 $S(1)$ 原子的净电荷、LUMO 轨道的组成等都能影响化合物的抗菌活性, 而硫原子的亲核前线电子密度、 $S(1)-N(2)$ Wiberg 键极、3D-Balaban 指数、Schultz 分子价拓扑指数是主要影响因素, 能为异噻唑啉酮类化合物的开发及改进提供理论依据。在此基础上, 采用神经网络建模可以显著提高模型预测能力和稳定性。可见, 2 种方法结合, 在试验设计、环境保护、药物设计研

究等方面将可发挥重要作用。

表 2 QSAR 模型效果分析

Table 2 Analysis of performance for QSAR models

Model	R	Se	Rcv
MLR	0.841 8	0.303 9	
ANN	0.991 6	0.080 1	0.973 1

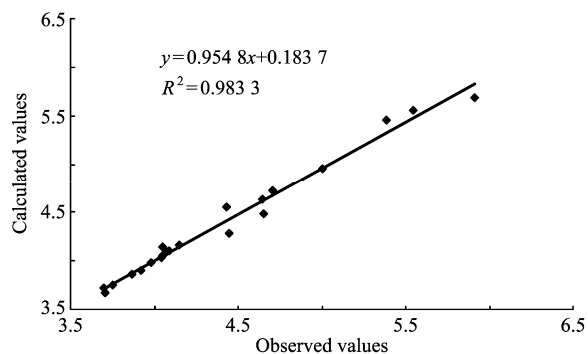


图 3 ANN 留—交叉检验计算值与实验值相关分析

Figure 3 Plots of experimental and calculated pMIC for the ANN analysis of the LOO

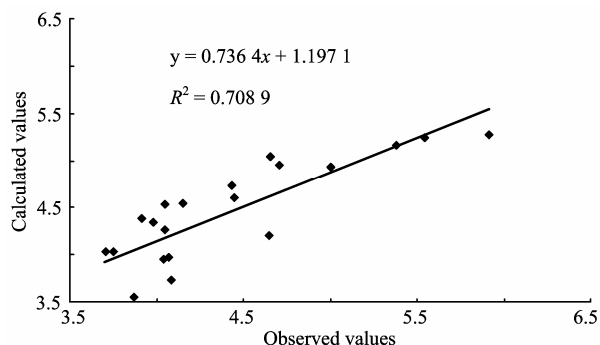


图 4 MLR 留—交叉检验计算值与实验值相关分析

Figure 4 Plots of experimental and calculated pMIC for the MLR analysis of the LOO

参考文献:

- [1] Critchley M, Bentham R. The efficacy of biocides and other chemical additives in cooling water systems in the control of amoebae [J]. *J Appl Microbiol*, 2009, 106 (3): 784-789.
- [2] 方永勤,张雁. N-甲基-1, 2-苯并异噻唑啉-3-酮的合成研究[J]. *化学世界*, 2009, (6): 361-363.
- [3] 祝德义,李彦春,靳丽强,等. 一株嗜盐菌的筛选及异噻唑啉酮对其抑制作用[J]. *中国皮革*, 2009, 38 (21): 12-15; 25.
- [4] 姜晓辉,于良民,董磊,等. 新型防污剂异噻唑啉酮衍生物的合成、生物毒性与防污性能研究[J]. *精细化工*, 2007, 24(2): 125-128; 189.
- [5] 于良民,魏长峰,姜晓辉,等. 4,5-二氯-2-甲氧丙基-4-异噻唑啉-3-酮的合成及杀藻性能研究[J]. *涂料工业*, 2008, 38 (6): 1-4.
- [6] Morley J O, Oliver A J, Charlton M H. Structure-activity relationships in 3-isothiazolones [J]. *Org Biomol Chem*, 2005, 3 (20): 3713-3719.
- [7] 夏树伟,孙玮,于良民,等. 2-(4-取代-苯基)-3-异噻唑啉酮抗菌活性的定量构效关系研究[J]. *化学学报*, 2007, 65(23): 2707-2714.
- [8] Livingstone D J, Salt D W. Judging the significance of multiple linear regression models [J]. *J Med Chem*, 2005, 48(3): 661-663.
- [9] 陈景文,李雪花,于海瀛,等. 面向毒害有机物生态风险评价的(Q)SAR 技术: 进展与展望[J]. *中国科学(B 辑): 化学*, 2008, 38 (6): 461-474.
- [10] Camargo A B, Marchevsky E, Luco J M. QSAR study for the soybean 15-lipoxygenase inhibitory activity of organosulfur compounds derived from the essential oil of garlic [J]. *J Agric Food Chem*, 2007, 55 (8): 3096-3103.
- [11] Francisco J P P, Xerardo G M, Humberto G D. Multi-target spectral moment QSAR versus ANN for antiparasitic drugs against different parasite species [J]. *Bioorg Med Chem*, 2010, 18(6): 2225-2231.
- [12] 杜雨静,范英芳. 人工神经网络用于三苯基丙烯腈衍生物的定量结构-活性关系模型[J]. *化工进展*, 2010, 29(1): 25-28.
- [13] 肖方竹,邓健,彭国文,等. QSAR 结合人工神经网络预测取代氯苯酚生物毒性[J]. *化学研究与应用*, 2009, 21(6): 888-810.
- [14] Liu H X, Papa E, Gramatica P. QSAR prediction of estrogen activity for a large set of diverse chemicals under the guidance of OECD principles [J]. *Chem Res Toxicol*, 2006, 19(11): 1540-1548.