电性距离矢量和神经网络用于三唑并嘧啶磺酰胺类除草剂的QSAR研究

陈 艳，冯长君，堵锡华
（徐州工程学院 化学化工学院，徐州 221111）

摘 要：为了研究三唑并嘧啶磺酰胺类除草剂对乙酰乳酸合成酶（aceto lactate synthase，ALS）抑制活性（pI_{50}）的定量构效关系，以电性距离矢量（M_k）表征了31种三唑并嘧啶磺酰胺类化合物的分子结构；利用最佳变量子集回归的方法建立了含有6个参数（M_2、M_{10}、M_{14}、M_{15}、M_{67}、M_{85}）的QSAR模型. 该模型的相关系数R及交叉验证相关系数R_{CV}分别为0.911、0.887，具有良好的稳健性和预测能力；以此6个参数为人工神经网络输入层，设定6∶3∶1的网络结构，构建人工神经网络的BP算法模型，相关系数R提升为0.988. 结果表明：影响三唑并吡啶磺胺类除草剂抑制活性pI_{50}的主要因素是-CH_3、-CH_2-、>C-、-O-、>N-及-X（-F，-Cl）等分子结构单元，且pI_{50}与M_2、M_{10}、M_{14}、M_{15}、M_{67}、M_{85}呈现良好的非线性关系，为设计高活性的ALS抑制剂提供理论依据.

关键词：分子电性距离矢量；乙酰乳酸合成酶抑制剂；抑制活性；定量结构-活性相关
中图分类号：O 626.41 **文献标志码**：A **文章编号**：0254-0037(2014)05-0771-05

QSAR Research of Triazolopyrimidine Herbcides by Electronegativity-distance Vector and Artificial Neural Network

CHEN Yan, FENG Chang-jun, DU Xi-hua
(School of Chemistry & Chemical Engineering, Xuzhou Institute of Technology, Xuzhou 221111, Jiangsu, China)

Abstract: To study the quantitative structure-activity relationship (QSAR) of the inhibited activity (pI_{50}) for aceto lactate synthase (ALS) inhibitor, molecular electronegativity-distance vector (M_k) was used to describe the molecular structure of 31 triazolopyrimidine herbcides in this paper. The six-parameter (M_2、M_{10}、M_{14}、M_{15}、M_{67}、M_{85}) QSAR model of pI_{50} for 31 the triazolopyrimidine compounds was constructed by leaps-and-bounds regression (LBR). The traditional correlation coefficient (R) and the cross-validation correlation coefficient (R_{CV}) were 0.911 and 0.887, respectively. The six structural parameters were used as the input neurons of artificial neural network, and a 6∶3∶1 network architecture was employed. A satisfied model could be constructed with the back-propagation algorithm, the correlation coefficient R was 0.988. The result demonstrates that the dominant influencing factors of inhibited activity are the molecular structure fragments: -CH_3、-CH_2-、>C-、-O-、>N-及-X(-F,-Cl), and there is a good non-linear relationship between the parameter M_2、M_{10}、M_{14}、M_{15}、M_{67}、M_{85} and pI_{50} of ALS. The model can provide some theoretical insights into the design of this series of ALS inhibitor with higher inhibited activity.

Key words: electronegativity distance vector; aceto lactate synthase inhibitor; inhibited activity; quantitative structure-activity relationship

收稿日期：2013-04-11
基金项目：国家自然科学基金资助项目(21272095)；徐州工程学院培育项目(XKY2011102)
作者简介：陈 艳(1968—)，女，教授，主要从事有机合成及物质构效学方面的研究，E-mail:chenyan681110@126.com

乙酰乳酸合成酶抑制剂通过抑制 ALS 酶而破坏植物体内氨基酸的合成，由于这一靶标不涉及人和动物，所以以 ALS 靶标设计开发的超高效除草剂对人和动物十分安全. 现在已开发出的 ALS 抑制剂除草剂共有 13 类 50 余种[1]，其中，以磺酰脲类除草剂为先导经过进一步的结构修饰开发的三唑并嘧啶磺酰胺类除草剂成为当前引人注目的化学除草剂类型之一，因其高活性、高安全性和高选择性而得到迅速发展[2]. 研究该除草剂对 ALS 酶的抑制活性的构效关系（quantitative structure activity relationship，QSAR）[3-4]有助于揭示其抑制机理，为设计和开发新的 ALS 抑制剂提供理论依据. 任天瑞等[3]利用比较分子场分析方法（CoMFA），在三维立体空间研究了这类化合物的构效关系，建立了有 6 个变量的 CoMFA 模型，取得了较为满意的结果. 本文基于刘树深等[5-7]的分子电性距离矢量（$\boldsymbol{M}_k$），运用最佳变量子集回归的方法（leaps-and-bounds regression，LBR）得到六元 QSAR 模型，并进一步把进入模型的 6 个变量作为神经网络的输入节点，建立了 6∶3∶1 型的 BP-ANN 的 QSAR 模型，大幅度提高了模型的相关程度、稳定性和预测能力，结果优于文献[3].

1 数据与研究方法

1.1 31 种三唑并嘧啶磺酰胺类除草剂的分子结构及 pI_{50}

31 种三唑并嘧啶磺酰胺类除草剂的母体结构见图 1，其分子结构及相应的对 ALS 酶的抑制活性（pI_{50}）见表 1.

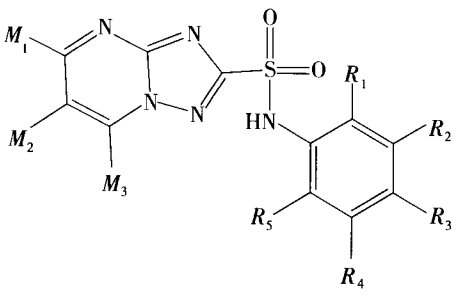

图 1 三唑并嘧啶磺酰胺类化合物的分子结构

Fig. 1 Structurals of triazolopyrimidine herbicides

1.2 分子电性距离矢量的建构方法

分子电性距离矢量（MEDV-13）是一种描述分子二维结构的拓扑描述子，由 13 种类型原子间的两两相互作用对构成 91 个元素（用 $\boldsymbol{M}_k$ 表示，k 代表 1～91）组成，通过引入原子类型和原子属性的概念，较为全面地反映了分子的拓扑、几何及电性特征，适用于含多个杂原子、饱和键与不饱和键、环和非环等结构. 用 ChemDraw Ultra 9.0 软件分别构建 31 个三唑并嘧啶磺酰胺类化合物的分子结构，存为 .mol 文件，然后在 Matlab 环境下，调用上述分子结构，应用文献[8-9]的方法编制程序，计算得到 91 种分子电性距离矢量，根据统计学原理，变量值的个数少于 5% 的自变量，其对因变量的贡献可以忽略不计. 因此，首先对自变量集合中自变量值少于 2（31×5%）的自变量予以剔除，用剩余 50 个描述子来表征分子的结构.

1.3 定量构效-活性相关分析方法

将每种化合物的 50 种描述子作为自变量 X，相应的 pI_{50} 作为因变量 Y，构建数据集，应用最佳子集回归选择最佳变量组合，建立相应 QSAR 模型. 为进一步提高模型的相关程度、稳定性和预测能力，本研究采用最佳变量子集回归选择的最佳组合的变量作为 BP 网络隐蔽层的结点，建立 pI_{50} 的神经网络模型.

2 结果与讨论

2.1 多元线性模型的建立

运用 SPSS13.0 软件中逐步回归（stepwise multiple regression，SMR）的方法对数据集进行多元线性回归，建立最佳六元数学模型为

$$pI_{50} = 0.211(\pm 0.056)\boldsymbol{M}_2 - 0.146(\pm 0.057)\boldsymbol{M}_{10} - 0.134(\pm 0.029)\boldsymbol{M}_{14} + 0.094(\pm 0.029)\boldsymbol{M}_{15} - 24.985(\pm 4.225)\boldsymbol{M}_{67} + 0.206(\pm 0.206)\boldsymbol{M}_{85} + 7.425(\pm 0.482) \quad (1)$$

$n = 31$，$R = 0.911$，$R^2 = 0.830$，$S = 0.370$，$F = 19.542$

式中：n、R、R^2、S、F 分别为样本容量、相关系数、判定系数（亦称削减误差比例）、估计标准误差和 Fischer 检验值. 按式（1）给出 31 个化合物的计算值，与其实验值较为接近.

2.2 模型的验证

1）模型所包含的化合物数与其变量数之比应大于 5，本研究的样本数为 31，变量数为 6，符合统计学规律.

2）对所建模型的外部样本预测能力和真实有效性进行验证，采用逐一剔除法（leave one out，LOO）交叉检验. 逐一剔除法的交叉验证系数（R^2_{CV}）是目前较为广泛使用的模型验证方法之一. 本文利用 Minitab 软件中的 LOO 法计算式（1）的交叉验证系数 Q^2，其值为 0.787，一般认为 $R^2_{CV} \geqslant 0.5$ 时，所建模型具有良好的稳健性与预测能力[10]，所以该模型具有可信的预报能力.

表 1　三唑并嘧啶磺酰胺类除草剂的分子结构及对 ALS 酶的抑制活性（R_3，R_4 = H）

Table 1　Structurals of triazolopyrimidine herbicides and the inhibited activity（pI_{50}）for ALS

编号	M_1	M_2	M_3	R_1	R_2	R_5	PI_{50}		
							实验值	预测值	误差
1	CH_3	H	CH_3	Cl	H	Cl	7.22	7.22	0.00
2	H	H	CH_3	Cl	H	Cl	6.96	6.76	-0.02
3	CH_3	H	H	Cl	H	Cl	5.74	5.70	-0.04
4	H	Cl	H	Cl	H	Cl	6.32	6.39	0.07
5	CF_3	H	CH_3	Cl	H	Cl	6.64	7.00	0.36
6	H	H	H	Cl	H	Cl	5.18	5.14	-0.04
7	H	CH_3	H	Cl	H	Cl	6.72	6.63	-0.09
8	CH_3	CH_3	CH_3	Cl	H	Cl	6.39	6.27	-0.12
9	CH_2CH_3	CH_3	CH_3	Cl	H	Cl	7.35	7.03	-0.32
10	H	CH_3	CH_3	Cl	H	Cl	6.62	6.76	0.14
11	CH_3	Cl	CH_3	Cl	H	Cl	6.00	6.08	0.08
12	H	H	C_2H_5	Cl	H	Cl	6.57	6.69	0.12
13	H	H	H	Cl	H	Cl	6.62	6.69	0.07
14	OCH_3	H	H	Cl	H	Cl	8.55	8.55	0.00
15	OCH_3	H	H	Cl	H	Cl	9.00	9.00	0.00
16	H	H	H	CF_3	H	OCH_2CF_3	7.48	7.48	0.00
17	H	H	CH_3	CF_3	H	OCH_2CH_3	7.08	7.07	-0.01
18	H	H	CH_3	CF_3	H	CH_3	6.87	6.96	0.09
19	H	H	CH_3	F	CF_3	F	7.39	7.25	-0.14
20	H	H	CH_3	F	H	$COOCF_3$	6.80	6.72	-0.08
21	H	H	CH_3	F	H	F	6.71	6.44	-0.27
22	H	H	CH_3	F	H	SCH_3	6.65	6.73	0.08
23	H	H	CH_3	F	H	COO-i-Pr	6.54	6.63	0.09
24	H	H	CH_3	F	H	$COOC_2H_5$	6.52	6.52	0.00
25	H	H	CH_3	F	H	Cl	6.47	6.34	-0.13
26	H	H	CH_3	F	H	CF_3	5.67	5.85	0.18
27	H	H	CH_3	F	H	CH_2CH_3	6.92	6.91	-0.01
28	H	H	CH_3	F	H	OCH_3	5.63	5.66	0.03
29	H	H	CH_3	F	H	CN	5.38	5.53	0.15
30	H	H	CH_3	F	H	CH_3	6.16	6.01	-0.15
31	H	H	CH_3	F	OCH_3	F	7.10	7.25	0.15

3）用变异膨胀因子（variance inflation factors，VIF）评价模型中各自变量间的多重相关性，VIF 的定义式为[11]

$$VIF = 1/(1 - R^2) \quad (2)$$

式中 R^2 为模型中一个自变量与余下自变量的判定系数. 当 VIF = 1，表明模型中各自变量间不存在相

关性;VIF<5时,各自变量间相关性很弱,该模型是稳定可以接受的;VIF>5则存在明显的相关性,所建模型不能用于估算和预测. 式(1)中M_2、M_{10}、M_{14}、M_{15}、M_{67}、M_{85}的VIF值分别为1.098、3.397、1.129、1.149、3、436和1.037,均小于5,证明式(1)中各自变量间不存在自相关性.

2.3 人工神经网络模型

为了进一步提高相关性和预测能力,本文采用误差反向传输人工神经网络即BP(back-propagation)算法构建预测三唑并嘧啶磺胺类化合物对ALS酶的抑制活性(pI_{50})模型,其前馈多层神经网络隐蔽层的激活函数为Sigmoid函数,输出层的转移函数为线性转移函数. 所有网络都是由一个输入层、一个隐蔽层和一个输出层构成,利用Levenberg-Marquardt函数进行训练. 输入层单元选取为对抑制活性起主要作用的6个结构参数M_2、M_{10}、M_{14}、M_{15}、M_{67}、M_{85}. 根据许禄等[12]的建议规则寻找最佳隐蔽层的单元数(H),即

$$\beta(=N/M)\geqslant 1 \tag{3}$$

式中:N、M分别是样本数和网络总权重. M定义为

$$M=(I+1)H+(H+1)Q \tag{4}$$

式中:I、H、Q分别是输入层、隐蔽层和输出层的单元数. 本文的$I=6$,$Q=1$及$N=31$,可得$H\leqslant 3.75$. 至此,本文采用6:3:1的网络结构建立模型,在BP算法中,避免过拟合和过训练,将样本分为3个集:训练集、验证集和测试集,各集化合物数依次为21、5、5. 由此建立的模型其训练集的$R=0.985$、验证集的$R=0.992$、测试集的$R=0.992$,对于31个三唑并嘧啶磺胺类化合物的$R=0.988$,这3个集的相关系数值和总体的相关系数值均很接近,说明模型具有很高的稳健性. 该模型给出的预测值列于表1,与实验值非常接近,平均误差为0.104. 它们的相关性见图2,图2显示模型具有很高的预测能力. 该模型的权重和偏置列于表2.

表2 BP-ANN模型的权重和偏置

Table 2 Weights and biases of BP-ANN model

层间变化	权重						偏置
从输入层到隐蔽层	-0.390 65	1.587 10	-0.599 45	0.405 9	1.178 0	-0.590 8	-0.255 1
	-0.048 34	0.205 39	1.149 90	-1.956 9	-2.901 7	1.494 6	2.813 5
	-0.356 14	2.395 30	-1.211 60	0.850 4	1.362 7	-3.418 0	-0.420 0
从隐蔽层到输入层	-3.837 40	0.9796 4	3.486 20				-0.629 3

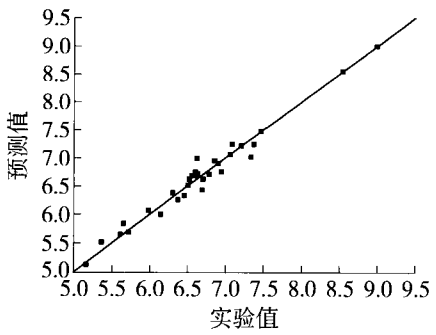

图2 31个三唑并嘧啶磺酰胺类化合物实验值和预测值的相关性

Fig. 2 Plot of calculated against experiment values of retention pI_{50} of 31 triazolopyrimidine herbcides

2.4 模型解释

ALS抑制剂除草剂是在作用机理上通过ALS酶与ALS抑制剂形成复合物,除草剂通过与酶通道内部残基连接,阻塞底物进入活性位点,从而抑制了酶的活性[13]. 进入模型的6个结构参数M_2、M_{10}、M_{14}、M_{15}、M_{67}、M_{85}分别是第1类碳原子(-CH_3)与第2类碳原子(-CH_2-)、第1类碳原子(-CH_3)与第10类氧原子(-O-)、第2类碳原子与第2类碳原子、第2类碳原子与第3类碳原子(>C-)、第7类氮原子(>N-)与第10类氧原子、第10类氧原子与第13类卤原子(-F,-Cl)相互作用的电性距离矢量,所以影响三唑并嘧啶磺酰胺类除草剂抑制活性的主要因素是-CH_3、-CH_2-、>C-、-O-、>N-及-X(-F,-Cl)分子结构单元,其中,-CH_3、-CH_2-、>C-等基团与受体以疏水作用相结合,-O-、>N-及-X与受体以氢键相结合,模型揭示了除草剂与靶标ALS酶之间的疏水作用和氢键作用. 这2个作用是影响三唑并嘧啶磺酰胺类除草剂抑制活性的本质因素.

3 结论

1)所建模型经过逐一剔除法交叉检验和自变量间的多重相关性检验证明具有良好的稳定性和预测能力.

2）分析进入最佳六元数学模型的结构参数可知，影响三唑并嘧啶磺酰胺类除草剂抑制活性的主要因素是疏水性和氢键作用，所对应的分子结构单元为-CH_3、-CH_2-、>C-、-O-、>N-及-X(-F，-Cl).

3）运用神经网络的 BP 算法，采用 6∶3∶1的网络结构建立的模型的相关性($R=0.988$)明显优于多元线性回归模型的相关性($R=0.911$)，证明了所选网络结构的合理性，也进一步验证了 M_2、M_{10}、M_{14}、M_{15}、M_{67}、M_{85}可揭示影响三唑并嘧啶磺酰胺类化合物 pI_{50}的本质因素，显示了两者之间具有良好的非线性关系，为设计高抑制活性的 ALS 抑制剂提供理论依据.

参考文献：

[1] 薛思佳，邹金山. 新型乙酰乳酸合成酶(ALS)抑制剂的研究进展[J]. 化学世界，2000(8)：399-403.
XUE Si-jia, ZOU Jin-shan. Progress on new acetolactate synthase (ALS) inhibitor[J]. Chemical World, 2000(8): 399-403. (in Chinese)

[2] 赵青山，付颖，叶非，等. 三唑并嘧啶磺酰胺类除草剂的研究概况[J]. 植物保护，2011，37(2)：14-19.
ZHAO Qing-shan, FU Ying, YE Fei, et al. Study summary of triazolo[1, 5-a] pyrimidine-2-sulfonanilide herbicides[J]. Plant Protection, 2011, 37(2): 14-19. (in Chinese)

[3] 任天瑞，谢桂荣，周家驹，等. 三唑并嘧啶磺酰胺类除草剂的 QSAR 研究(II)[J]. 计算机与应用化学，1998，15(5)：281-284.
REN Tian-rui, XIE Gui-rong, ZHOU Jia-ju, et al. QSAR research of triazolopyrimidine herbcides (II)[J]. Computers and Applied Chemistry, 1998, 15(5): 281-284. (in Chinese)

[4] 任天瑞，陈红明，谢桂荣. 三唑并嘧啶磺酰胺类除草剂的比较分子场法分析[J]. 高等学校化学学报，1998，19(12)：1950-1953.
REN Tian-rui, CHEN Hong-ming, XIE Gui-rong. Comparative molecular field analysis of triazolopyrimidine sulfonaniline herbicides[J]. Chemical Journal of Chinese Universities, 1998, 19(12): 1950-1953. (in Chinese)

[5] 刘树深，刘堰，李志良，等. 一个新的分子电性距离矢量(MEDV)[J]. 化学学报，2000，58(11)：1353-1357.
LIU Shu-shen, LIU Yan, LI Zhi-liang, et al. A novel molecular eletrogetivity-distance vector (MEDV)[J]. Acta Chemica Sinica, 2000, 58(11): 1353-1357. (in Chinese)

[6] LIU S S, YIN C S, LI Z L, et al. QSAR study of steroid benchmark and dipeptides based on MEDV-13[J]. Journal of Chemical Information and Computer Sciences, 2001, 41(2): 321-329.

[7] LIU S S, LIU H L, YIN C S, et al. VSMP: a novel variable selection and modeling method based on the prediction[J]. Journal of Chemical Informationand Computer Sciences, 2003, 43(3): 964-969.

[8] 胡黔楠，梁逸曾，王亚丽，等. 直观队列命名法的基本原理及其在矩阵与拓扑指数计算中的应用[J]. 计算机与应用化学，2003，20(4)：386-390.
HU Qian-nan, LIANG Yi-zeng, WANG Ya-li, et al. The basic principles of heuristic queue notation and its applications in calculation of matrix and topological index for topological graphs[J]. Computers and Applied Chemistry, 2003, 20(4): 386-390. (in Chinese)

[9] 张婷，梁逸曾，赵晨曦，等. 基于分子结构预测气相色谱程序升温保留指数[J]. 分析化学，2006，34(11)：1607-1610.
ZHANG Ting, LIANG Yi-zeng, ZHAO Chen-xi, et al. Prediction of temperature-programmed retention indices from molecule structures[J]. Analytical Chemistry, 2006, 34(11): 1607-1610. (in Chinese)

[10] SWAMINATHAN S, FRUEY W, PLETCHER J, et al. Crystalstructure of staphylococcal enterotoxin B, a superan- tigen[J]. Nature, 1992, 359(6389): 801-806.

[11] 冯长君. 手性有机酸保留指数的手性指数及原子类型电拓扑指数模型[J]. 物理化学学报，2010，26(1)：193-198.
FENG Chang-jun. Mathematical model for retention indices, chiral index and electrotopological state indices for atom types of chiral organic acids[J]. Acta Physico-Chimica Sinica, 2010, 26(1): 193-198. (in Chinese)

[12] 许禄，邵学广. 化学计量学方法[M]. 2 版. 北京：科学出版社，2004：251-261.

[13] 郑培忠，沈健英. 新型乙酰乳酸合成酶(ALS)抑制剂作用机理的研究进展[J]. 杂草科学，2009(3)：1-5.
ZHENG Pei-zhong, SHEN Jian-ying. Progress on the mechanism of new acetolactate synthase (ALS) inhibitor[J]. Weed Science, 2009(3): 1-5. (in Chinese)

（责任编辑　刘　潇）