

文章编号: 2095—0411 (2012) 04—0008—05

# 取代芳烃对发光菌急性毒性的 QSAR 研究<sup>\*</sup>

冯琳琳<sup>1</sup>, 张兆志<sup>2</sup>, 王新颖<sup>1</sup>, 陈海群<sup>1</sup>

(1. 常州大学 环境与安全工程学院, 江苏 常州 213164; 2. 衡水学院 应用化学系, 河北 衡水 053000)

**摘要:** 根据定量结构—活性相关性 (QSAR) 原理, 研究了 38 种取代芳烃对发光菌的急性毒性 ( $C_{e,50}$ ) 与其分子结构之间的构效关系。应用遗传算法筛选出 5 个与  $C_{e,50}$  最为相关的描述符, 并应用多元线性回归方法和支持向量机方法建立 QSAR 模型。两种模型的复相关系数、留一法交互验证系数分别为 0.988、0.979 和 0.991、0.981, 对外部预测样本的复相关系数和外部测试集交互验证系数分别为 0.913、0.904 和 0.924、0.906, 与相关文献比较, 所建 QSAR 模型均具有更好的预测能力和稳健性。

**关键词:** 定量结构—活性相关; 取代芳烃; 遗传算法; 支持向量机; 急性毒性

中图分类号: X 131

文献标识码: A

## QSAR Study on Acute Toxicity of Substituted Aromatic Compounds to *Photobacterium Phosphoreum*

FENG Lin—lin<sup>1</sup>, ZHANG Zhao—zhi<sup>2</sup>, WANG Xin—ying<sup>1</sup>, CHEN Hai—qun<sup>1</sup>

(1. School of Environmental and Safety Engineering, Changzhou University, Changzhou 213164, China;  
2. Department of Applied Chemistry, Hengshui University, Hengshui 053000, China)

**Abstract:** This QSAR study relates to the structure of 38 sorts of substituted aromatic compounds, in which a set of five descriptors were chosen by using the variable selection method of genetic algorithm. The five descriptors were used to establish the QSAR of the acute toxicity of substituted aromatic compounds to *photobacterium phosphoreum* by multiple linear regression and support vector machine. The statistical results indicate that the multiple correlation coefficient and cross validation using leave—one—out were 0.988, 0.979 and 0.991, 0.981, respectively. To validate the predictive power of the resulting models, external validation multiple correlation coefficient and cross validation were 0.913, 0.904 and 0.924, 0.906, respectively. Compared with pertinent literature, the QSAR models have more favorable estimation stability and better prediction power.

**Key words:** QSAR; substituted aromatic compounds; genetic algorithm; support vector machine; acute toxicity

取代芳烃类化合物作为重要的化工产品或中间体, 应用广泛, 但它们大多数具有毒性, 不仅对生物存在致癌、致畸和致突变的作用, 也会污染空气、土壤和水源, 给环境带来严重危害<sup>[1-2]</sup>。因

此, 对取代芳烃类化合物的毒性进行分析研究具有一定的应用价值。

定量结构—活性相关 (QSAR) 可仅根据化合物的分子结构参数, 利用适当的数学模型, 预测其

<sup>\*</sup> 收稿日期: 2012—09—30

作者简介: 冯琳琳 (1984—), 女, 江苏邳州人, 硕士生; 通讯联系人: 陈海群。

毒性、致突变性、致癌性大小等生物活性，克服了实验方法中的一些不足之处，比如费用昂贵、费时费力、滞后性等。选择合适的分子描述符和高效的建模方法是 QSAR 研究中的两大关键问题。目前，选取分子结构描述符的方法主要有：辛醇/水分配系数法、量子化学方法和遗传算法等优化算法等<sup>[3-7]</sup>；用于 QSAR 建模的方法主要有：多元线性回归（MLR）、偏最小二乘法（PLS）、人工神经网络法（ANN）和支持向量机（SVM）等<sup>[8-12]</sup>。其中，遗传算法（GA）是一种简洁高效的全局优化算法<sup>[13]</sup>，具有相当强的搜索能力，能够在有限的时间内搜索出与目标性质最为相关的分子描述符集合，因此在 QSAR 研究中得到了较好的应用；而 SVM 算法因其能解决小样本、非线性、高维数、局部极小值等实际问题，也在 QSAR 研究中得到了广泛的应用<sup>[14-16]</sup>。

本文依据 QSAR 基本原理，从分子结构角度出发，研究了 38 种取代芳烃对发光菌的急性毒性与其分子结构之间的构效关系，分别采用 MLR 和 SVM 方法建立了 QSAR 模型，并比较 2 种方法的模型。

## 1 研究方法

### 1.1 数据来源

取代芳烃对发光菌急性毒性 $-\lg C_{e,50}$ 引自文献[17]， $C_{e,50}$ 指取代芳烃对发光菌 15min 半数发光抑制浓度， $-\lg C_{e,50}$ 越大，化合物的毒性越大。为了与原文献进行比较，测试集和训练集的划分也同文献一致，见表 1。

### 1.2 试验步骤

首先将化合物的二维结构导入化学软件 Hyperchem7.5 进行优化，先采用软件中的分子力学方法 MM+ 进行初步优化，在此基础上，用半经验的量子化学方法 AM1 进行进一步的几何优化，之后将优化好的分子结构文件输入到 Dragon5.4 软件中计算化合物的各种结构参数。通过计算，每个化合物均得到 1 664 种结构参数。对于如此多的结构参数，为避免“机会相关”现象，要对其进行初步删减，以删除一些不能为模型提供有用信息的参数，将常数或者近似常数的描述符以及相关系数大于 0.95 的其中 1 个描述符都删除。经初步删除，共剩余 459 种结构参数。之后，对剩余的结构参数

采用遗传—偏最小二乘法（GA—PLS）进行特征变量选择，筛选出与目标性质最为相关的结构参数集合，以此作为 MLR 和 SVM 建模的输入变量，建立 QSAR 模型，并对模型进行内外部验证。

表 1 取代芳烃对发光菌急性毒性的实验值（ $-\lg C_{e,50}$ ）和 MLR、SVM 的预测值（Pre）

Table 1 Experimental and predicted toxicity values of substituted aromatic compounds to *photobacterium phosphoreum* ( $-\lg C_{e,50}$ ) by MLR and SVM

| 序号               | 化合物         | $-\lg C_{e,50}$ | MLR—Pre | SVM—Pre |
|------------------|-------------|-----------------|---------|---------|
| 1                | 六氯苯         | 6.31            | 6.292   | 6.248   |
| 2                | 1,2,4,5-四氯苯 | 5.51            | 5.543   | 5.516   |
| 3                | 1,2,3-三氯苯   | 4.53            | 4.521   | 4.541   |
| 4                | 对二氯苯        | 4.39            | 4.395   | 4.375   |
| 5                | 间二氯苯        | 4.24            | 4.327   | 4.324   |
| 6                | 邻二氯苯        | 4.38            | 4.260   | 4.263   |
| 7                | 对二溴苯        | 4.54            | 4.561   | 4.539   |
| 8                | 间二溴苯        | 4.99            | 4.947   | 4.927   |
| 9                | 溴苯          | 3.78            | 3.862   | 3.843   |
| 10               | 1-氯-4-溴苯    | 4.50            | 4.461   | 4.438   |
| 11               | 2,5-二氯甲苯    | 4.38            | 4.239   | 4.318   |
| 12               | 4-氯甲苯       | 3.88            | 3.885   | 3.880   |
| 13               | 对二甲苯        | 3.68            | 3.543   | 3.617   |
| 14               | 间二甲苯        | 3.65            | 3.655   | 3.588   |
| 15               | 甲苯          | 3.08            | 3.226   | 3.200   |
| 16               | 2,4,6-三氯苯胺  | 4.51            | 4.617   | 4.573   |
| 17               | 2,6-二氯苯胺    | 4.16            | 4.083   | 4.097   |
| 18               | 2,4-二氯苯胺    | 4.09            | 4.114   | 4.126   |
| 19               | 3,4-二氯苯胺    | 4.20            | 4.205   | 4.218   |
| 20               | 对溴苯胺        | 3.92            | 4.019   | 4.018   |
| 21               | 2-氯-4-硝基苯胺  | 3.99            | 4.046   | 4.027   |
| 22               | 3-硝基苯胺      | 3.77            | 3.819   | 3.833   |
| 23               | 苯胺          | 3.28            | 3.224   | 3.217   |
| 24               | 2,4-二氯苯酚    | 4.45            | 4.455   | 4.487   |
| 25               | 对氯苯酚        | 4.48            | 4.449   | 4.417   |
| 26               | 邻氯苯酚        | 4.14            | 4.139   | 4.123   |
| 27               | 邻甲基苯酚       | 3.75            | 3.734   | 3.724   |
| 28               | 苯酚          | 3.64            | 3.561   | 3.558   |
| 29 <sup>1)</sup> | 1,2,4-三氯苯   | 4.50            | 4.383   | 4.542   |
| 30 <sup>1)</sup> | 氯苯          | 3.86            | 3.651   | 3.658   |
| 31 <sup>1)</sup> | 苯           | 3.34            | 3.656   | 3.668   |
| 32 <sup>1)</sup> | 2,4,5-三氯甲苯  | 4.86            | 4.980   | 4.950   |
| 33 <sup>1)</sup> | 对氯苯胺        | 3.57            | 3.951   | 3.938   |
| 34 <sup>1)</sup> | 2,4-二硝基苯胺   | 4.16            | 4.183   | 4.203   |
| 35 <sup>1)</sup> | 2-硝基苯胺      | 3.70            | 3.794   | 3.845   |
| 36 <sup>1)</sup> | 五氯苯酚        | 5.69            | 5.762   | 5.712   |
| 37 <sup>1)</sup> | 对硝基苯酚       | 4.05            | 3.802   | 3.811   |
| 38 <sup>1)</sup> | 间苯二酚        | 3.00            | 3.404   | 3.413   |

说明：1) 为测试集。

## 2 结果与讨论

### 2.1 GA—PLS 筛选结果

运用 GA—PLS 筛选方法，获取了 5 个与取代

芳烃对发光菌急性毒性最为密切的分子描述符，见 表 2。

表 2 GA—PLS 筛选出的分子描述符  
Table 2 The molecular descriptors selected by GA—PLS

| 序号 | 描述符   | 定义                             | 类型           |
|----|-------|--------------------------------|--------------|
| 1  | $x_1$ | 3D—MoRSE—signal 04 / 未加权       | 3D—MoRSE 描述符 |
| 2  | $x_2$ | H 总体指数/根据原子范德瓦尔斯体积加权           | GETAWAY 描述符  |
| 3  | $x_3$ | K 总体形状指数/根据原子范德瓦尔斯体积加权         | WHIM 描述符     |
| 4  | $x_4$ | 第 1 成分可达性定向 WHIM 指数/根据原子电拓扑性加权 | WHIM 描述符     |
| 5  | $x_5$ | 等距离度上的平均信息量                    | 信息指数         |

上述描述符中， $x_1$  为 3D—MoRSE 描述符，它以电子衍射为基础表征分子的 3D 结构特征，与其他由分子图论计算而得到的描述符不同，它明确考虑了原子的 3D 排列，因此能够较好的表征分子的空间结构特征。 $x_2$  属于 3 维结构特征 GETAWAY 描述符，可用氢原子自相关性总体指数来表征分子形状、大小、原子分布，主要反映原子在 3 维立体环境中的分布。 $x_3$  与  $x_4$  均为 WHIM 描述符，它们与分子中的原子分布情况有关。 $x_5$  为信息指数，主要描述分子的链接信息，表征有机物的空间效应。

## 2.2 MLR 模型

以筛选出的 5 个描述符作为输入变量，针对训练集样本构建相应的 MLR 模型：

$$-\lg C_{50} = 4.596 + 0.365x_1 + 0.56x_2 - 4.209x_3 + 0.385x_4 - 0.286x_5 \quad (1)$$

$$n=28, R^2=0.988, F=353.130, S=0.079, P<0.001$$

式中， $n$  为训练集样本数， $R^2$  为复相关系数， $F$  为 Fisher 检验值， $S$  为标准误差， $P$  为方程显著性概率。由方程（1）可知，模型具有较高的相关系数和较低的标准偏差，说明模型是可靠的；显著性概率远小于 0.05，表明回归方程具有统计学意义。同时，从方程中可以看出，取代芳烃化合物的急性毒性与  $x_1$ 、 $x_2$ 、 $x_4$  呈正相关，与  $x_3$ 、 $x_5$  呈负相关。其各个参数对目标值的影响程度要根据它们在方程中的标准回归系数确定，方程中 5 个描述符在方程中的标准回归系数分别为 0.155、0.825、-0.354、0.133、-0.331。正负号表示影响方向，其绝对值越大影响越大，所以这 5 个描述符对  $-\lg C_{e,50}$  的影响程度依次为  $x_2 > x_3 > x_5 > x_1 > x_4$ 。

应用方程（1）分别对训练集和测试集进行预测，所得急性毒性预测值与实验值的比较见表 1，模型的主要性能参数见表 3。

表 3 MLR 和 SVM 模型的主要性能参数比较

Table 3 Performance comparison between models obtained by MLR and SVM

| 模型  | 训练集   |             |       | 测试集   |             |       |
|-----|-------|-------------|-------|-------|-------------|-------|
|     | $R^2$ | $Q^2_{loo}$ | RMS   | $R^2$ | $Q^2_{ext}$ | RMS   |
| MLR | 0.988 | 0.979       | 0.070 | 0.913 | 0.904       | 0.236 |
| SVM | 0.991 | 0.981       | 0.061 | 0.924 | 0.906       | 0.232 |

## 2.3 SVM 分析

SVM 为获得最佳的泛化能力，在建模过程中需要调节相应的参数组合，即选择合适的核函数、确定核函数参数、惩罚系数  $C$  及  $\epsilon$ —不敏感损失函数中  $\epsilon$  的大小<sup>[18]</sup>。本文选择目前应用最多的径向基形式的核函数，对核函数的宽度  $\gamma$ 、惩罚系数  $C$  以及  $\epsilon$  3 个最佳参数的确定采用格点搜索方法获得。

将 GA—PLS 筛选出的 5 个描述符作为 SVM 模型的输入变量，通过格点搜索方法得到的 SVM 最优参数值如下： $C=256$ ， $\gamma=0.007\ 812\ 5$ ， $\epsilon=0.062\ 5$ 。应用所建模型分别对训练集和测试集样本进行预测，所得急性毒性预测值与实验值的比较见表 1，模型的主要性能参数见表 3。

## 2.4 模型验证

对模型的稳定性、预测能力以及泛化能力的验证是 QSAR 研究中非常重要的环节。常用的验证方法有内部验证和外部验证两种。内部验证常采用“留一法”交互验证，以交互验证系数  $Q^2_{loo}$  表示。内部验证主要是为了检验模型的稳定性及其内部预测能力，而相关研究表明<sup>[19]</sup>，内部验证结果的好坏并不能说明其外部预测能力的大小，对模型预测能力的评价必须通过对未参与训练的测试集进行预测。模型的外部预测能力通常用外部交互验证系数  $Q^2_{ext}$  衡量。MLR 和 SVM 模型验证的主要性能参数见表 3。

此外，本文还给出了两种模型的预测残差分布图，如图 1 和图 2。从图中可以看出，散点呈现随机分布状态且均匀分布于 0 基准线的两侧，可见所

建预测模型是合适的，在建立过程中未产生系统误差。

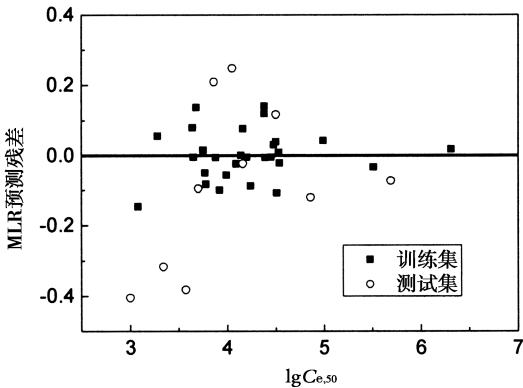


图 1 MLR 模型 -lgC<sub>e,50</sub> 预测残差图

Fig. 1 Plot of the residuals versus the experiment -lgC<sub>e,50</sub> values for the MLR model

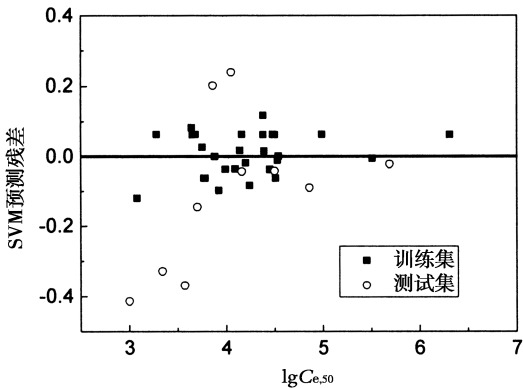


图 2 SVM 模型 -lgC<sub>e,50</sub> 预测残差图

Fig. 2 Plot of the residuals versus the experiment -lgC<sub>e,50</sub> values for the SVM model

## 2.5 模型的比较

### 2.5.1 MLR 方法与 SVM 方法的比较

从表 3 中可以看出，MLR 和 SVM 模型无论对于训练集还是测试集都具有较大的复相关系数和较小的预测误差，这说明模型具有良好的稳定性和较好的外部预测能力。此外，两种预测模型，其训练集和测试集的各性能参数均较为接近，说明模型还具有良好的泛化能力。总体来说，SVM 模型各项性能略优于 MLR 模型，但若是“黑箱”模型，无法得知各参数对目标值的影响大小，而 MLR 则可以直观的反映出各参数的重要度，两者各有优缺点且具有一定的互补性。

### 2.5.2 与其他模型的对比

为了进一步验证本文预测模型的优越性，将其与顾云兰等<sup>[17]</sup>运用同样样本集所建立的对发光菌急性毒性预测模型进行比较，各性能参数的比较见

表 4。

表 4 取代芳烃化合物 QSAR 模型比较

Table 4 Comparison of the QSAR models of substituted aromatic compounds

| 文献     | 变量数 | R <sup>2</sup> | Q <sup>2</sup> <sub>loo</sub> | R <sup>2</sup> <sub>ext</sub> | Q <sup>2</sup> <sub>ext</sub> |
|--------|-----|----------------|-------------------------------|-------------------------------|-------------------------------|
| 方程 (1) | 5   | 0.988          | 0.979                         | 0.913                         | 0.903                         |
| 方程 (2) | 3   | 0.947          | 0.925                         | 0.899                         | 0.879                         |
| [17]   | 3   | 0.920          | 0.896                         | 0.870                         | 0.849                         |

从表 4 中可以看出，相对于文献，本文所建模型的各项性能参数优于原文献。为了进一步对比，本文选用与原文献相同数量的描述符，建立如下的回归方程：

$$-\lg C_{e,50} = 3.740 + 0.606x_2 - 2.714x_3 - 0.21x_5 \quad (2)$$

$n = 28$ ,  $R^2 = 0.947$ ,  $F = 143.224$ ,  $S = 0.157$ ,  $P < 0.001$ ,  $Q^2_{loo} = 0.925$ ,  $Q^2_{ext} = 0.879$

方程 (2) 的各项性能参数均优于原文献，可见无论是从相关系数还是模型的稳定性、预测能力以及泛化能力，本文所建模型均优于文献。

## 3 结 论

本文从分子结构角度出发，研究了取代芳烃对发光菌的急性毒性与其分子结构之间的构效关系。应用遗传算法从众多分子描述符中筛选出了 5 个与取代芳烃急性毒性最为相关的描述符，结合 MLR 和 SVM 方法建立了 QSAR 模型，并对所建模型进行了内外部验证。内外部验证结果表明本文所建模型均具有较好的预测能力和稳定性，且与 MLR 模型相比，SVM 模型性能略胜一筹。由此可见，SVM 可用于预测取代芳烃急性毒性。但应用 SVM 方法建立的是一种“黑箱”模型，不能给出直观的数学模型，而 MLR 方法所得出的是一个直观、准确的数学关系模型，它能准确给出各结构参数对模型的贡献值。两种方法各有优缺点，实际应用中可根据需要选择合适的建模方法。

## 参考文献：

[1] Nimrod A C, Benson W H. Environmental estrogenic effects of alkylphenol ethoxylates [J]. Critical Reviews in Toxicology, 1996, 2 (3): 335-364.

[2] Cronin M T D, Schultz T W. Development of quantitative structure - activity relationships for the toxicity of aromatic compounds to *tetrahymena pyriformis*: comparative assessment of the methodologies [J]. Chemical Research in Toxicology, 2001, 14 (9): 1284-1295.

- [3] Wang X D, Dong Y Y, Wang L S, et al. Acute toxicity of substituted phenols to *rana japonica* tadpoles and mechanism — based quantitative structure — activity relationship (QSAR) study [J]. Chemosphere, 2001, 44 (3): 447—455.
- [4] Malakhata A Turabekova, Bakhtiyor F Rasulev, Mikhail G Levkovich, et al. *Aconitum* and *delphinium* sp. Alkaloids as antagonist modulators of voltage-gated  $\text{Na}^+$  channels AM1/DFT electronic structure investigations and QSAR studies [J]. Computational Biology and Chemistry, 2008, 32 (2): 88—101.
- [5] Zhu Menjun, Ge Fei, Zhu Runliang, et al. A DFT — based QSAR study of the toxicity of quaternary ammonium compounds on *chlorella vulgaris* [J]. Chemosphere, 2010, 80 (1): 46—52.
- [6] Aziz Habibi — Yangjeh • Mohammad Danandeh — Jenagharad. Application of a genetic algorithm and an artificial neural network for global prediction of the toxicity of phenols to *tetrahymena pyriformis* [J]. Monatsh Chem, 2009, 140 (11): 1279—1288.
- [7] 刘天宝, 彭艳芬, 汪新. DFT 法研究取代酚对黑头呆鱼的急性毒性 [J]. 计算机与应用化学, 2012, 29 (2): 175—177.
- [8] Zhu M J, Ge F, Zhu R L, et al. A DFT — based QSAR study of the toxicity of quaternary ammonium compounds on *chlorella vulgaris* [J]. Chemosphere, 2010, 80 (1): 46—52.
- [9] Qin Y, Deng H F, Yan H, et al. An accurate nonlinear QSAR model for the antitumor activities of chloroethylnitrosoureas using neural networks [J]. Journal of Molecular Graphics and Modelling, 2011, 29 (6): 826—833.
- [10] 崔秀君, 王志欣, 袁星, 等. 支持向量机用于酚类化合物毒性的 QSAR 研究 [J]. 计算机与应用化学, 2008, 25 (3): 298—302.
- [11] 崔毅, 蒋军成, 潘勇, 等. 基于支持向量机的脂肪族化合物急性毒性的 QSAR 研究 [J]. 安全与环境学报, 2009, 9 (5): 19—24.
- [12] Hemmateenejad B, Mehdipour A R, Miri R, et al. Comparative QSAR studies on toxicity of phenol derivatives using quantum topological molecular similarity indices [J]. Chemical Biology & Drug Design, 2010, 75 (5): 521—531.
- [13] 刘付喜, 钱苏翔, 曹坚. 基于遗传算法的 BP 神经网络在声音智能监控中的应用 [J]. 常州大学学报: 自然科学版, 2012, 24 (3): 70—74.
- [14] 崔毅, 蒋军成, 潘勇, 等. 羧酸以其衍生物急性毒性的 QSAR 研究 [J]. 环境科学与技术, 2010, 33 (4): 29—34.
- [15] 张克俊, 孙守迁, 柴春雷, 等. 基于启发式算法和支持向量机算法预测醛类化合物急性毒性 [J]. 分析化学研究报告, 2007, 35 (9): 1263—1268.
- [16] Zhao C Y, Zhang H X, Zhang X Y, et al. Application of support vector machine (SVM) for prediction toxic activity of different data sets [J]. Toxicology, 2006, 217 (2—3): 105—119.
- [17] 顾云兰, 陶建清, 费正皓, 等. DFT 法研究取代芳烃结构与毒性的定量关系 [J]. 计算机与应用化学, 2009, 26 (10): 1303—1306.
- [18] 蒋军成, 潘勇. 有机化合物的分子结构与危险特性 [M]. 北京: 科学出版社, 2011: 159—160.
- [19] Tropsha A, Gramatica P, Gombar V K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSAR models [J]. QSAR & Combinatorial Science, 2003, 22 (1): 69—77.