2006 年第 64 卷 化 学 学 报 Vol. 64, 2006
第 18 期, 1889～1896 ACTA CHIMICA SINICA No. 18, 1889～1896

• 研究论文 •

基于估计均方差的统计量 Sp 的修正变量选择法用于持久性有机污染物毒性研究

易忠胜[a] 刘树深*,[a,b]
([a]桂林工学院材料与化学工程系 桂林 541004)
([b]同济大学环境学院 长江水环境教育部重点实验室 上海 200092)

摘要 提出并详细说明基于估计均方差的统计量 Sp 的修正变量选择(MVSSp)方法. 采用分子电性距离矢量(MEDV)及其二次项表征多氯二苯并呋喃(PCDFs), 多氯二苯并二噁英(PCDDs)和多氯联苯(PCBs)三种持久性有机污染物异构体分子结构, 结合 MVSSp 方法选择适当的描述子, 建立了三种持久性有机污染物分别对芳烃受体(AhR)亲合性, 芳香羟化酶(AHH), 7-乙氧基异吩唑酮−脱乙基酶(EROD)诱导作用的定量结构活性相关(QSAR)模型, 模型的质量优于文献或相当, 讨论了这些实验毒性与分子结构关系. 这对研究和预测这些持久性污染物的毒性具有一定的指导意义.

关键词 估计均方差的统计量 Sp; 变量选择; 分子电性距离矢量; 多氯二苯并二噁英; 多氯二苯并呋喃; 多氯联苯; 毒性

Modified Variable Selection Based on Estimated Root Mean Square Statistic Sp for Studying Toxicity of Persistent Organic Pollutants

YI, Zhong-Sheng[a] LIU, Shu-Shen*,[a,b]
([a] *Department of Material and Chemical Engineering, Guilin University of Technology, Guilin 541004*)
([b] *Key Laboratory of Yangtze Aquatic Environment, Ministry of Education, College of Environmental Science and Engineering, Tongji University, Shanghai 200092*)

Abstract A modified variable selection based on the estimated root mean square statistic Sp (MVSSp), has been developed. The molecular electronegativity distance vector (MEDV) based on 13 atom type (MEDV) and their quadratic terms were used to describe the molecular structures of polychlorinated dibenzofurans (PCDFs), polychlorinated dibenzo-*p*-dioxins (PCDDs) and polychlorinated biphenyls (PCBs) isomers. With the choices of some appropriate descriptors by MVSSp, several QSAR models, between the MEDV and the abilities of three types of persistent organic pollutants (POPs) bound to the cytosolic aryl hydrocarbon receptor (AhR), the stimulating induction to aryl hydrocarbon hydroxylase (AHH) and 7-ethoxyresorufin O-deethylase (EROD), were built. The qualities of those models are not worse than those of literature. And then the relations between the experiment toxicities and the molecular structures were discussed. This study will help to provide a useful guideline for modeling and predicting the toxicity of POPs.

Keywords estimated root mean square statistic Sp; variable selection; MEDV; PCDDs; PCDFs; PCBs; toxicity

多氯二苯并呋喃(polychlorinated dibenzofurans, PCDFs), 多氯二苯并二噁英(polychlorinated dibenzo-*p*-dioxins, PCDDs)和多氯联苯(polychlorinated biphenyls, PCBs)是近年来研究比较多的持久性有机污染物

* E-mail: ssliuhl@263.net or ssliu@nju.edu.cn
Received August 6, 2005; revised March 21, 2006; accepted June 19, 2006.
全国优秀博士论文作者基金(No. 200355)、国家自然科学基金(No. 20577023)和广西新世纪十百千人才计划(No. 2003208)资助项目.

(persistent Organic Pollutants, POPs), 这些有机污染物对人类健康造成的伤害和环境污染等问题已经引起了环境科学以及相关领域研究者的极大关注[1~7]. PCDFs, PCDDs和PCBs作为一类在自然环境中广泛分布的有机污染物, 目前的研究已经获得了主要包括肝中毒, 皮肤损伤, 胃损伤, 内分泌干扰, 免疫毒性, 致畸致癌等对人类的危害, 对各种酶的诱导作用如芳香羟化酶(aryl hydrocarbon hydroxylase, AHH)和 7-乙氧基异吩唑酮-脱乙基酶(7-ethoxyresorufin O-deethylase, EROD) 等[1]. 然而这些有机物对人类健康危害的毒性机理非常复杂, 到目前为止还不是十分清楚, 普遍接受的观点是这些有机物毒性作用并不是直接对生物体作用而产生, 而是通过一种称为芳烃受体(aryl hydrocarbon receptor, AhR)的特殊蛋白质介导产生毒性[8], 也就是说这些有机物进入生物体后, 与 AhR 亲合并产生一系列的生物作用从而导致生物和毒理学效应; 因此, 有机物与AhR的亲合是产生毒性作用的关键所在, 但这些蛋白质的结构目前还不十分清楚; 再者, 这些有机物的高毒性和大量的异构体, 甚至很多异构体目前还没有分离出来, 使得通过实验研究变得非常困难. 定量结构-活性相关(quantitative structure-activity relationships, QSAR)研究近年来在药物设计的研究中取得了很大的成功[9], 引入环境科学研究领域后也取得了大量成果[10]. 因此, 用 QSAR 建立有机物毒性模型并预测其毒性的研究目前相当活跃.

PCDFs, PCDDs 和 PCBs 分别与老鼠体外肝细胞的AhR亲合性, AHH和EROD的诱导作用已经建立了很多QSAR 模型[3,6,11~15]. 这些模型主要采用了比较分子力场分析(comparative molecular field analyses, CoMFA) 3D-描述子[3,15], 分子表面性质自相关描述子[11], 电性特征值[12], 量子化学描述子或再加上有机物的理化性质[6]等描述子表征分子结构. 从建模的效果来看, 这些描述子都各有自己的优点, 并且所建立的模型质量大致相当, 也得出了很多有益结论, 如 Safe 等[16]通过线性自由能相关性研究得出在没有较大的取代官能团时, 空间因素对 PCB 与 AhR 的亲合性并没有多大影响; Waller 等[3]得出了 PCBs 对 EROD 的诱导作用很难采用 CoMFA 方法建立简单的线性模型.

基于 13 种原子类型的分子电性距离矢量(molecular electronegativity distance vector based on 13 atoms type, MEDV)在有机物 QSAR 研究中已经取得了部分研究成果[5,17~25]. 因此, 本文尝试采用 MEDV 描述子表征 PCDFs, PCDDs 和 PCBs 异构体分子结构, 研究描述子与这三种持久性有机污染物对 AhR 的亲合性, AHH 和 EROD 的诱导作用的相关关系; 因为 MEDV 描述子的数量比较多, 建模过程中, 考虑到模型的稳定性与预测能力, 需要对这些描述子进行选择. 本文对基于估计均方差的统计量 Sp[26]的变量选择方法的计算过程进行修改, 提出一种基于估计均方差的统计量 Sp 的修正变量选择方法(modified variable selection based on estimated root mean square Sp statistic, MVSSp), 其基本思路是从变量数为 1 时开始, 计算不同变量数下的 Sp 值, 然后再比较 Sp 值的大小, 根据建立稳健多元线性模型的经验要求, 计算到稍大于满足要求的变量数即可停止计算, 而不是原方法中的计算全部变量组合, 这样可以大大减少计算量. 并对 MVSSp 方法的实现过程及变量选择的原则进行详细的描述; 然后应用该方法从 MEDV 中选择最优子集, 构建最优的 QSAR 模型; 结果表明, 这种方法选择的描述子组合建立的模型具有较高的留一法(Leave-One-Out, LOO)交互检验相关系数 q^2 和模型相关系数 R^2.

1 计算方法

1.1 数据集

三种持久性有机污染物分别对 AhR, AHH 和 EROD 的实验毒性参数(pEC_{50})列在表 1 中, 这些数据均来自文献, 包括 PCDFs[16,27], PCDDs[1]和 PCBs[3]对 AhR 的亲合性, 与 AHH 和 EROD 的诱导作用. 其中 PCDFs 的实验数据个数分别为 34 (AhR), 31 (AHH), 31 (EROD); PCDDs 分别为 14 (AhR), 13 (AHH), 13 (EROD); PCBs 分别为 14 (AhR), 9 (AHH), 9 (EROD).

1.2 MEDV 描述子的计算

根据文献[19～21,28], 首先将有机化合物中出现的非氢原子 C, N, O, S, P, F, Cl, Br, I, 根据原子所处不同的环境划分为 13 种类型, 每一种类型规定一个识别号(ID), 识别号定义为: ID=4×(ν−4)+该原子连接的非氢原子数, 其中ν为价电子层电子数, 并根据分子中的非氢原子所处的环境不同, 参照由 Hall 和 Kier 等[30]提出并发展的 *E*-状态指数划分为 43 种属性, 对不同原子属性原子引入修改过的 *E*-状态指数也就是原子固有状态 I 表征不同原子电负性的变化, 其值定义为 $I=\sqrt{\frac{\nu}{4}\frac{(2/n)^2\delta^{\nu}+1}{\delta}}$, 43 种原子属性的划分方案和相应的 I 值见相关的文献, 这里不再一一列出. MEDV 的计算公式为:

$$x_r=m_{kl}=\sum_{i\in k,\ j\in l}\frac{q_iq_j}{d_{ij}^2}$$

(k, l=1, 2, …, 13; $l\geqslant k$; r=1, 2, …, 91) (1)

式中 k, l 为各非氢原子类型, d_{ij} 表示非氢原子 i, j 之间的

表 1 PCDFs, PCDDs 和 PCBs 异构体及其对 AhR 的亲合性, AHH, EROD 诱导作用

Table 1 Structures and binding affinity data for AhRs, induction for AHH and EROD data set of PCDFs, PCDDs and PCBs

No.	Isomer	pEC_{50}			No.	Isomer	pEC_{50}		
		AhR	AHH	EROD			AhR	AHH	EROD
	PCDFs				32	1,2,3,6,7,8-HCDF	6.569	8.833	8.907
1	PCDF	3.000			33	1,2,4,6,7,8-HCDF	5.081	7.373	7.533
2	2-MCDF	3.553			34	2,3,4,6,7,8-HCDF	7.328	9.163	9.240
3	3-MCDF	4.377				PCDDs			
4	4-MCDF	3.000	5.000	4.767	1	1-MCDD	4.0000	4.0000	4.0000
5	2,3-DCDF	5.326	5.600	5.315	2	2,8-DCDD	5.4949	4.0000	4.0000
6	2,6-DCDF	3.609	4.210	4.200	3	1,2,4-TrCDD	4.8861	4.3188	5.6576
7	2,8-DCDF	3.590	4.403	4.398	4	1,7,8-TrCDD	6.6576		
8	1,2,7-TrCDF	6.347	5.553	5.504	5	2,3,7-TrCDD	7.1487	6.4437	6.8539
9	1,3,6-TrCDF	5.357	5.597	5.472	6	1,2,3,4-TCDD	5.8861	5.4318	5.6198
10	1,3,8-TrCDF	4.071	4.712	4.520	7	1,2,7,8-TCDD	6.1024	7.2147	7.9586
11	2,3,4-TrCDF	4.721	6.821	6.606	8	1,3,7,8-TCDD	6.7959	6.2291	6.4949
12	2,3,8-TrCDF	6.000	5.604	5.807	9	2,3,7,8-TCDD	8.0000	10.1427	9.7212
13	1,2,3,6-TCDF	6.451	4.000	4.000	10	1,2,3,4,7-TCDD	5.1938	6.1805	6.0862
14	1,2,3,7-TCDF	6.951	4.569	4.201	11	1,2,3,7,8-TCDD	7.1024	7.9586	7.7696
15	1,2,4,8-TCDF	5.000	4.921	4.033	12	1,2,4,7,8-TCDD	5.9586	7.6778	7.9586
16	2,3,4,6-TCDF	6.456	5.879	5.947	13	1,2,3,4,7,8-TCDD	6.5528	8.6778	8.3872
17	2,3,4,7-TCDF	7.600	7.747	7.830	14	1,2,3,4,6,7,8,9-TCDD	5.0000	6.5086	6.1549
18	2,3,4,8-TCDF	6.699	7.383	7.425		PCBs			
19	2,3,6,8-TCDF	6.658	5.983	6.108	1	2,2',4,4'- TCB	3.886		
20	2,3,7,8-TCDF	7.387	8.408	8.695	2	2,3,4,5- TCB	3.854		
21	1,2,3,4,8-PeCDF	6.921	6.680	6.788	3	3,3',4,4'- TCB	6.149	7.55	7.05
22	1,2,3,7,8-PeCDF	7.128	8.595	8.514	4	3,4,4',5- TCB	4.553		
23	1,2,3,7,9-PeCDF	6.400	7.066	7.066	5	2,3,3',4,4'- PeCB	5.367	7.06	6.92
24	1,2,4,6,7-PeCDF	7.169	6.488	6.458	6	2,3,4,4',5- PeCB	5.387	6.02	6.25
25	1,2,4,6,8-PeCDF	5.510	5.000	4.921	7	2,3',4,4',5'-PeCB	4.854	5.41	5.95
26	1,2,4,7,8-PeCDF	5.886	6.975	6.830	8	2,3',4,4',5- PeCB	5.041	4.94	5.05
27	1,2,4,7,9-PeCDF	4.699	7.424	7.416	9	3,3',4,4',5- PeCB	6.886	9.62	9.61
28	1,3,4,7,8-PeCDF	6.699	8.796	8.854	10	2,2',4,4',5,5'- HeCB	4.102		
29	2,3,4,7,8-PeCDF	7.824	9.592	9.873	11	2,3,3',4,4',5- HeCB	5.149	5.68	6.05
30	2,3,4,7,9-PeCDF	6.699	8.102	8.237	12	2,3,3',4,4',5'-HeCB	5.301	6.15	5.9
31	1,2,3,4,7,8-HCDF	6.638	9.449	9.421	13	2,3',4,4',5,5'- HeCB	4.796	4.88	5.05
					14	2,3',4,4',5',6-HeCB	4.004		

最短拓扑距离即从原子 i 到原子 j 的各个路径中化学键数的加和的最小值. q_i 与 q_j 是原子在实际分子环境中的 E-状态指数:

$$q_i = I_i + \sum_{j \neq i}^{all\ j} \frac{(I_i - I_j)}{d_{ij}^2} \tag{2}$$

这样 13 种原子类型的两两相互作用对便构成 MEDV 描述子, 共有 91 个. 根据 MEDV 的定义, PCDDs 和 PCDFs 分子中只有第 2(～C～), 3(≈C～), 10(—O—), 13(Cl—)四种类型, 其中“～”表示共轭单键, “≈”表示共轭双键, “—”表示普通单键, 这四种类型两两相互作用对构成的描述子序号为: 14, 15, 22, 25, 26, 33, 36, 82, 85, 91, 它们分别对应于 13 种原子类型中 2-2, 2-3, 2-10, 2-13, 3-3, 3-10, 3-13, 10-10, 10-13, 13-13 的两两相互作用对, 因为 PCDFs 异构体分子中第 10 种类型(—O—)只出现一次, 所以该类型相互作用对(也就是描述子序号 82)全部为零, 这样 MEDV 描述子共有 9 个; PCDDs 的 MEDV 描述子共有 10 个; 而 PCBs 只有第 2(～C～), 3(≈C～), 13(Cl—)三种类型, 其 MEDV 描述子序号为:

14, 15, 25, 26, 36, 91, 它们分别对应于 13 种原子类型中 2-2, 2-3, 2-13, 3-3, 3-13, 13-13 的两两相互作用对, 共 6 个.

1.3 基于 Sp 统计量的修正变量选择方法

目前, QSAR 建模研究中, 从大量的描述子中选择几个适当的变量(描述子)显得越来越重要了. QSAR 建模常用的变量选择技术(方法)已经报道了很多, 如逐步回归[31], 偏最小二乘/主成分分析[32], 遗传算法[33]以及我们实验室提出的 VSMP 方法[23,28,34,35]等. 本文选择基于 n 个有机物(样本)活性估计均方差应最小而提出的 Sp 统计量[26]标准, 其定义如下:

$$S_p=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{(n-p-2)(n-p-1)} \quad (3)$$

其中 p 为模型中描述子(自变量)个数, y 为有机物活性(因变量), $\hat{y}$ 为某一变量组合线性模型对活性的估计值, n 为样本数. 其基本思想为: 如果 Sp 在所有的描述子组合中取得最小值, 那么相应的描述子组合具有最好的建模性能. 但是在实际操作过程中, 选择全局具有最小 Sp 值的变量组合(共有 2^{p-1} 个组合)时, 如果变量数较多, 其计算量相当大. 因此, 本文利用统计量 Sp 变量选择标准建立一种新的变量选择方法, 称之为基于估计均方差的统计量 Sp 的修正变量选择方法(modified variable selection based on the estimated root mean square statistic Sp, MVSSp), 其基本思路是从变量数为 1 时开始, 计算不同变量数下的 Sp 值, 然后再比较 Sp 值的大小, 根据建立稳健多元线性模型的一般要求, 样本数最低要大于变量数的 3 倍以上, 实际过程中不需要计算全部的变量组合, 这样可以大大减少计算量. 具体操作方法如下:

(1) 确定指定变量数 vn, 然后通过枚举法从自变量矩阵 $x(n, p)$ 中选出一个子集 $x(n, vn)$ 计算 vn 个变量之间的自相关系数矩阵 **R**, 如果 **R** 中的元素存在大于预先设置的标准 r_{cri}, 那么不计算该变量组合的 Sp 值而是直接用一个很大的值 Sp_{cri} 替代, 这样避免出现自变量的自相关系数大于自变量与因变量之间相关系数的情况; 如果 **R** 中元素都小于预设的 r_{cri}, 则计算并保留其 Sp 统计量, 然后与下一个相同变量数子集的 Sp 统计量进行比较, 保留最小的 Sp 值, 直到变量数 vn 下所有组合计算完毕, 此时具有最小 Sp 值的变量组合即为当前变量数的最优子集.

(2) 计算得到的最优子集的相关统计参数: 模型的相关系数 R, 均方根误差 RMSE, F 统计量以及 LOO 交互检验的相关系数 q, 均方根误差 RMSV.

(3) 如果还存在没有被选择的子集, 返回步骤 1, 继续计算; 反之进入下一步.

(4) 从不同的变量数 vn (1, 2, 3, 4, …)的最优子集中确定全局最优子集. (3)式中的分子将随着变量数的增大而逐渐变小, 而分母则随着变量数的增加而减小, 因此, 总存在适当的 p 值, 使 Sp 达到最小; 同时, 在模型进行校验时, 相关系数总是随着变量数 vn 的增加而增加; 而在 LOO 交互检验时, 相关系数 q 随 vn 的增加到一定值后会有所下降; RMSV 的变化规律与 Sp 类似. 因此, 选择最优子集时, 在满足样本数最低大于变量数的 3 倍以上的条件下, 首先考虑 Sp 统计量是否达到最小, 如果是则选择对应的子集作为最优子集; 如果 Sp 一直下降, 也就是还没有找到全局最小的 Sp 值, 此时, 需要综合考虑 Sp, q 和 RMSV 三个统计量值, 也就是权衡三者的大小, 选择具有较小的 Sp, 较大的 q 和较小的 RMSV 的变量组合, 并且变量数尽可能少的子集作为全局最优子集, 并以此子集采用多元线性回归(Multi-Linear Regression, MLR)方法[32]建立最终的 QSAR 模型. 所有的计算都是通过自编程序实现.

2 结果与讨论

2.1 基于 MEDV 描述子的最优子集选择

根据 MVSSp 方法原理, 以 PCDFs 为例说明该方法选择最优子集的过程. 从 PCDFs 异构体的 MEDV 描述子中挑选出指定数量描述子(1, 2, 3, …)下分别与 AhR, AHH 和 EROD 的实验毒性 pEC_{50} 值建立多元线性模型, 然后选择 Sp 统计量最小的子集, 并以此子集建立最终的模型. 表 2 和图 1, 2 分别为 PCDFs 的各异构体分别与 AhR, AHH 和 EROD 的实验毒性 pEC_{50} 值和 MEDV 描述子数目的 Sp 统计量和 RMSV 的计算结果及变化趋势图, 可以看到其 Sp 在变量数达到 6 时有最小值, 因此, 该模型就选择 6 个描述子(描述子序号为 **25**, **26**, **33**, **36**, **85**, **91**)用于建模. 同样我们可以得到对 31 个 PCDFs 异构体对 AHH 的实验毒性建模时, 变量数为 4 (描述子序号为 **14**, **22**, **25**, **33**)时 Sp 值最小, 因此确定了其最优子集的变量数为 4; 同理可以得到 31 个 PCDFs 异构体对 EROD 的实验毒性的最优模型为 4 个变量(描述子序号为 **14**, **22**, **25**, **33**). 另外从图 2 和表 2 中我们可以看到模型校正时的均方根误差(RMSE)一直在下降.

按照同样的方法, 我们可以得到 PCDDs, PCBs 和全部三类化合物分别对 AhR, AHH 和 EROD 的实验毒性与 MEDV 描述子之间的最优子集, 限于篇幅只将最优子集的结果列于表 3. 挑选出最优描述子组合后, 对不同实验毒性数据用多元线性回归方法, 就可以分别与各自挑选的描述子建立模型. 在表 2 和表 3 中我们可以看

表 2　PCDFs 三种实验毒性与不同数量 MEDV 描述子组合的模型统计量

Table 2　The statistic of various combinations of the descriptors for experiments of PCDFs

描述子数	描述子	q^2	R^2	RMSE	RMSV	Sp	F
				PCDFs, AhR, n=34			
1	14	0.5165	0.5592	0.9100	0.9541	0.02838	41.8695
2	26, 36	0.6204	0.6897	0.7636	0.8484	0.02131	36.6669
3	26, 36, 91	0.7033	0.7684	0.6597	0.7543	0.01700	36.4915
4	14, 15, 22, 85	0.7162	0.8068	0.6025	0.7450	0.01519	34.4578
5	15, 22, 33, 36, 91	0.7371	0.8288	0.5671	0.7173	0.01446	31.9549
6	25, 26, 33, 36, 85, 91	0.7669	0.8604	0.5122	0.6856	0.01270	33.8915
7	22, 25, 26, 33, 36, 85, 91	0.7432	0.8619	0.5094	0.7238	0.01357	29.4163
8	14, 22, 25, 26, 33, 36, 85, 91	0.6890	0.8621	0.5090	0.8101	0.01467	25.7937
9	14, 15, 22, 25, 26, 33, 36, 85, 91	0.6617	0.8621	0.5090	0.8550	0.01595	22.9289
				PCDFs, AHH, n=31			
1	91	0.4794	0.5215	1.1408	1.1919	0.04968	32.6975
2	26, 36	0.5646	0.6327	0.9996	1.0927	0.04096	25.8370
3	26, 85, 91	0.5960	0.6902	0.9180	1.0590	0.03721	22.2745
4	14, 22, 25, 33	0.6296	0.7431	0.8360	1.0218	0.03332	21.6929
5	14, 22, 33, 36, 85	0.5899	0.7461	0.8310	1.1120	0.03567	17.6344
6	14, 22, 33, 36, 85, 91	0.5687	0.7465	0.8304	1.1533	0.03872	14.7252
7	14, 22, 25, 33, 36, 85, 91	0.5355	0.7476	0.8286	1.2156	0.04206	12.6943
8	14, 15, 22, 25, 33, 36, 85, 91	0.5180	0.7482	0.8276	1.2490	0.04595	11.1423
9	14, 15, 22, 25, 26, 33, 36, 85, 91	0.4721	0.7496	0.8253	1.3428	0.05027	9.9779
				PCDFs, EROD, n=31			
1	91	0.4644	0.5062	1.2396	1.2928	0.05866	30.7506
2	26, 36	0.5776	0.6434	1.0534	1.1509	0.04549	27.0642
3	26, 85, 91	0.6150	0.7065	0.9557	1.1058	0.04033	24.0663
4	14, 22, 25, 33	0.6454	0.7540	0.8748	1.0685	0.03649	22.9933
5	14, 15, 22, 25, 33	0.6281	0.7542	0.8746	1.0975	0.03951	18.4085
6	14, 15, 22, 25, 26, 33	0.5689	0.7567	0.8701	1.2069	0.04251	15.5510
7	14, 15, 22, 25, 26, 33, 36	0.5477	0.7586	0.8667	1.2580	0.04602	13.4670
8	14, 15, 22, 25, 26, 33, 36, 91	0.5283	0.7594	0.8653	1.3001	0.05023	11.8351
9	14, 15, 22, 25, 26, 33, 36, 85, 91	0.4917	0.7596	0.8649	1.4071	0.05520	10.5334

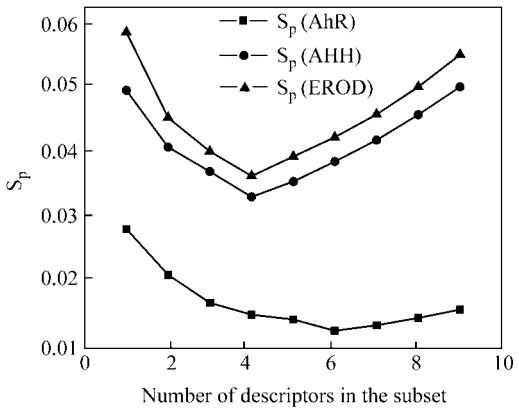

图 1　PCDFs 子集优化过程 Sp 随变量数的变化

Figure 1　Variation of the Sp with the number of descriptors in the optimal subset of PCDFs

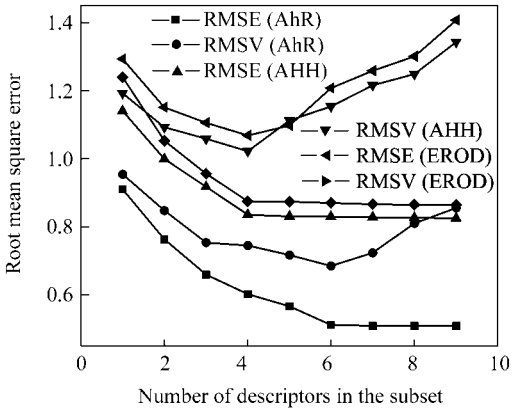

图 2　PCDFs 子集优化过程均方根误差随变量数的变化

Figure 2　Variation of the root mean square error with the number of descriptors in the optimal subset of PCDFs

表 3 PCDDs, PCBs 及全部三类化合物的三种实验毒性与最优 MEDV 描述子组合的模型统计量

Table 3 The statistic of optimized combinations of the MEDV descriptors for experimental toxicity of PCDDs, PCBs and PCDFs＋PCDDs＋PCBs

体系		样本数	描述子数	描述子	q^2	R^2	RMSE	RMSV	Sp	F
	AhR	14	3	14,25,91	0.7361	0.9003	0.3273	0.5859	0.01665	39.1168
PCDDs	AHH	13	3	15,25,26	0.7883	0.8901	0.5860	0.8214	0.06200	32.3858
	EROD	13	3	14,15,36	0.7674	0.9035	0.5011	0.7975	0.04533	37.4710
	AhR	14	4	15,26,36,91	0.5564	0.8068	0.3688	0.6036	0.02645	13.5750
PCBs	AHH	9	2	15,25	0.4783	0.7998	0.6374	1.0648	0.12189	15.9783
	EROD	9	2	15,25	0.1933	0.6858	0.7293	1.2841	0.15956	8.7296
PCDFs	AhR	62	7	15,22,26,36,82,85,91	0.6267	0.7078	0.6814	0.7749	0.01005	21.1087
PCDDs	AHH	54	6	14,25,26,33,36,91	0.5238	0.6424	0.9849	1.1504	0.02483	15.5697
PCBs	EROD	54	6	14,25,26,33,36,91	0.4996	0.6257	1.0149	1.1904	0.02637	14.4894

到, 所建立的模型中最少的样本数为 9, 变量数为 2, 模型中样本数都是变量数的 3 倍以上, 满足通常进行多元线性回归建模要求样本数是变量数 3 倍以上的要求.

所建立的模型质量与文献[3,6,11～15]相比(见表 4, 其中 n 后面括弧中的数值为变量数或主成分数), 在相同样本数条件下, LOO 交互验证的 q^2, 模型的相关系数 R^2 有所改善或相当. 其中, Tuppurainen[12]的研究中, PCDFs 对 AhR 受体毒性模型的 q^2 和 R^2 都比本文的要大, 但 PLS 模型中采用了 11 个主成分, 而 PCBs 的模型比本文稍差, 其主成分数为 3, 本文为 4 个变量, 从这一点来看, 二者的模型质量应该相当; PCDDs 的模型中比本文多了 11 个含溴化合物, 因而样本数不一致, 不好比较. 文献[3,11,13,14]等所采用的数据中, PCDFs 的异构体有重复, PCDDs 中也包含了 11 个含溴化合物, Bradley[14] 另外增加了 21 个其它的化合物, 这些结果中, 有些比本文的要好, 但因其中包含了相同的样本, 有可能使得统计参数产生虚高, 加之样本数不一致, 不能简单地进行比较.

2.2 分子结构与持久性有机污染物毒性的关系

通过 MVSSp 方法, 我们得到了基于 MEDV 描述子的 PCDF, PCDDs, PCBs 以及全部三类化合物分别与 AhR, AHH 和 EROD 实验毒性的最优 QSAR 模型. 虽然

表 4 不同方法对 PCDFs, PCDDs, PCBs 和全部三种类型化合物的三种实验毒性建模结果对照

Table 4 The result of different models for PCDFs, PCDDs, PCBs and PCDFs＋PCDDs＋PCBs

体系		Tuppurainen[12]			Waller[3]			Arulmozhiraja[6]			本文		
		n	q^2	R^2	n	q^2	R^2	n	q^2	R^2	n	q^2	R^2
	AhR	34 (11)	0.795	0.966	39 (5)	0.742	0.858	33 (3)		0.776	34 (6)	0.7669	0.8604
PCDFs	AHH				30 (3)	0.695	0.848	31 (2)		0.670	31 (4)	0.6296	0.7431
	EROD				30 (3)	0.663	0.838	31 (2)		0.691	31 (4)	0.6454	0.7540
	AhR	25 (4)	0.862	0.946	25 (4)	0.715	0.919				14 (3)	0.7361	0.9003
PCDDs	AHH				13 (1)	0.424	0.698				13 (3)	0.7883	0.8901
	EROD				13 (1)	0.232	0.633				13 (3)	0.7674	0.9035
	AhR	14 (3)	0.549	0.823	14 (2)	0.534	0.823				14 (4)	0.5564	0.8068
PCBs	AHH				9 (3)	0.273	0.926				9 (2)	0.4783	0.7998
	EROD				9 (1)	0.021	0.490				9 (2)	0.1933	0.6858
	AhR	73 (8)	0.818	0.912	78 (6)	0.724	0.878				62 (7)	0.6267	0.7078
PCDFs, PCDDs, PCBs	AHH				52 (4)	0.572	0.788				54 (6)	0.5238	0.6424
	EROD				52 (4)	0.489	0.745				54 (6)	0.4996	0.6257
体系		Wagener[11]			Bravi[13]			Bradley[14]			Waller[15]		
		n	q^2	R^2	n	q^2	R^2	n	q^2	R^2	n	q^2	R^2
	AhR	78 (12)	0.83	0.792	78 (2)	0.732	0.794	99 (6)	0.72	0.91	78 (10)	0.711	0.903
PCDFs, PCDDs, PCBs	AHH												
	EROD												

MEDV 描述子不能直接反映化合物与其受体蛋白之间的相互作用，但是从各自挑选出的 MEDV 描述子中，我们还是可以间接地了解分子结构与毒性之间存在的关系.

(1) PCDDs 异构体对 AhR 的亲合性最优模型中，选择了序号为 14, 25, 91 的三个描述子，根据 MEDV 的计算原理，我们知道它们分别对应原子类型～C～与～C～之间的相互作用，～C～与 Cl—之间的相互作用，Cl—与 Cl—相互作用，这从一个侧面反映了 PCDDs 对 AhR 的亲合性与苯环共轭体系中的 C 原子之间，共轭体系中 C 原子与取代 Cl 之间，取代 Cl 相互之间的作用有着密切的联系，而与分子中的氧原子的作用关系不大；异构体对 AHH 的诱导作用与第 15, 25, 26 三个描述子存在线性关系，反映了诱导作用与苯环共轭体系中的 C 之间的相互作用有关，而与取代 Cl 和氧原子的关系相对来说不大；异构体对 EROD 的诱导作用中第 14, 15, 36 三个描述子比较重要，也就说明诱导作用与苯环共轭体系中的 C 原子之间的相互作用以及苯环共轭体系中的 C 与取代 Cl 之间的相互作用有着密切的关系.

(2) PCDFs 异构体对 AhR 的亲合性与第 25, 26, 33, 36, 85, 91 共六个描述子有着良好的线性关系，说明苯环共轭体系中的 C 之间，共轭体系中的 C 与取代 Cl 之间，共轭体系中 C 与呋喃杂环上的 O 之间，取代 Cl 与呋喃杂环上的 O 之间，取代 Cl 之间的相互作用密切相关；而异构体对 AHH 和 EROD 的诱导作用都与苯环共轭体系中的 C 之间，共轭体系中 C 与呋喃杂环上的 O 之间，共轭体系中的 C 与取代 Cl 之间的相互作用关系密切.

(3) PCBs 异构体对 AhR 的亲合性与 MEDV 的关系与 PCDDs 和 PCDFs 相比效果较差. 从选择的变量看主要与苯环共轭体系中的 C 之间，共轭体系中的 C 与取代 Cl 之间，取代 Cl 之间的相互作用关系较大；从 AHH 的模型两变量模型来看，其诱导作用主要与苯环上的 C 之间以及苯环共轭体系中的 C 与取代 Cl 之间的关系比较紧密；而 EROD 的 2 变量模型 q^2 仅为 0.1933，所得模型的预测能力非常低，就该模型来说其诱导作用与共轭体系中的 C 之间以及共轭体系中的 C 与取代 Cl 之间的关系比较有关. 文献[3]用 CoMFA 得到的 PCBs 对 EROD 的诱导作用模型，其预测能力也比较差，该文指出该体系可能很难用二维和三维描述子建立简单线性模型，必须寻求其它更为合适的描述子建立线性关系，或者非线性模型.

2.3　MEDV 及其二次项的最优子集选择

前面讨论了 PCDDs, PCDFs, PCBs 和全部三类化合物对 AhR 的亲合性，AHH 和 EROD 诱导作用与 MEDV 的相关模型的建立. 考虑到生物活性有可能与描述子之间存在非线性关系，因此，我们将 MEDV 描述子的平方项以及所有描述子相互乘积(描述子的交互效应)等二次项引入到自变量矩阵中，简单测试非线性关系. 这样 PCDFs, PCDDs, PCBs 以及全部三类化合物对应的描述子个数分别为 54, 65, 27 和 66 个，同时需要把这些描述子中全部为零或者方差为零的删除(这里并没有出现，因此，描述子个数并没有减少)，然后再通过 MVSSp 方法进行最佳子集的选择并建立多元线性模型，结果如表 5 所示. 与表 2, 3 相比所建立的模型质量有了很大的改善. 从表 5 中我们可以看到，除了 PCDDs 对 EROD 的诱导作用还是与 MEDV 的一次项相关性较强外，一次项仅在 PCBs 对 AhR 的亲合性，全部化合物对 AhR 的亲合性以及对 EROD 的诱导作用模型中分别出现一次，其它

表 5　PCDDs, PCDFs, PCBs 和全部化合物的三种实验毒性与最优 MEDV 及其二次项组合的模型统计量

Table 5　The statistics of optimized combinations of the MEDV descriptors including quadratic terms for experimental toxicity of PCDFs, PCDDs, PCBs and PCDFs＋PCDDs＋PCBs

		描述子	q^2	R^2	RMSE	RMSV	Sp	F
	AhR	22×22, 14×22	0.7586	0.8195	0.2742	0.6400	0.01169	29.5124
PCDDs	AHH	15×25, 15×26, 15×36	0.8088	0.9300	0.5234	0.7770	0.04946	41.6118
	EROD	14, 15, 36	0.7674	0.9035	0.5011	0.7975	0.04533	37.4710
	AhR	85×85, 15×26, 22×36, 26×85	0.8669	0.9269	0.3706	0.5114	0.00665	66.7478
PCDFs	AHH	14×36, 22×26, 22×91	0.6405	0.7268	0.8621	1.0017	0.03281	26.6031
	EROD	15×26, 15×36, 25×85	0.6655	0.7452	0.8904	1.0302	0.03500	29.2513
	AhR	26, 15×15, 14×15	0.7403	0.8720	0.3002	0.4529	0.01752	14.8190
PCBs	AHH	25×25, 15×25	0.5172	0.8195	0.6053	1.0134	0.10991	18.1568
	EROD	14×26, 14×36	0.3075	0.7367	0.6676	1.3093	0.13372	11.1889
PCDFs	AhR	15, 14×91, 22×33, 33×91, 82×85	0.7288	0.7639	0.6126	0.6571	0.00755	39.4628
PCDDs	AHH	14×14, 14×22, 14×26, 15×91, 22×33, 25×91	0.6854	0.7464	0.8295	0.9281	0.01761	25.5036
PCBs	EROD	36, 14×26, 15×85, 25×26, 36×36, 82×82	0.6647	0.7371	0.8506	0.9656	0.01852	24.3043

的几种生物活性都是与 MEDV 的二次项密切相关, 也就是说大部分生物活性与 MEDV 描述子之间存在比较强的非线性关系.

3 结论

本文提出了一种新的变量选择方法——基于估计均方差的统计量 Sp 的修正变量选择(MVSSp)方法, 对该方法的具体实现过程以及选择最优变量组合(最优子集)的原则作了详细描述. 并采用 MEDV 描述子表征 PCDDs, PCDFs 和 PCBs 异构体分子, 结合 MVSSp 方法, 建立了 PCDDs, PCDFs, PCBs 以及全部三类化合物分别对 AhR 的亲合性, AHH 和 EROD 的诱导作用与 MEDV 以及描述子二次项之间比较稳定的 QSAR 模型, LOO 交互验证的 q^2, 模型的相关系数 R^2 与文献相比, 在样本数相同的情况下都有了改善或相当, 并且 MEDV 描述子的计算简单易行, 不需要昂贵的商业计算软件或实验测定; 讨论了这些有机污染物毒性与分子结构关系; 研究表明 MVSSp 方法确实能够相对快速地从大尺寸自变量矩阵中找到最优子集, 对研究和建立 QSAR 模型有较强的指导意义.

致谢 感谢桂林工学院青年基金对研究工作的资助.

References

1 Safe, S. H. *Annu. Rev. Pharmacol. Toxicol*. **1986**, *26*, 371.
2 Safe, S. H. *Crit. Rev. Toxicol*. **1990**, *21*(1), 51.
3 Waller, C. L.; McKinney, J. D. *J. Med. Chem.* **1992**, *35*, 3660.
4 Andersson, P. L.; Burght, A. S. A. M. V. D.; Berg, M. V. D.; Tysklind, M. *Environ. Toxicol. Chem*. **2000**, *19*(5), 1454.
5 Liu, S. S.; Cui, S. H.; Wang, L. S. *Chin. Chem. Lett*. **2004**, *15*(4), 467.
6 Arulmozhiraja, S.; Morita, M. *Chem. Res. Toxicol.* **2004**, *17*, 348.
7 Buzatu, D. A.; Beger, R. D.; Wilkes, J. G.; Jackson, O. L. Jr. *Environ. Toxicol. Chem*. **2004**, *23*(1), 24.
8 Mekenyan, O. G.; Veith, G. D.; Call, D. J.; Ankley, G. T. *Environ. Health Perspect.* **1996**, *104*, 1302.
9 Selassie, C. D.; Mekapati, S. B.; Verma, R. P. *Curr. Top. Med. Chem*. **2002**, *2*(12), 1357.
10 Schultz, T. W.; Cronin, M. T. D.; Netzeva, T. I. *J. Mol. Struct.* (*Theochem*) **2003**, *622*, 23.
11 Wagener, M.; Sadowski, J.; Gasteiger, J. *J. Am. Chem. Soc.* **1995**, *117*, 7769.
12 Tuppurainen, K.; Ruuskanen, J. *Chemosphere* **2000**, *41*, 843.
13 Bravi, G.; Wikel, J. H. *Quant. Struct.-Act. Relat*. **2000**, *19*, 29.
14 Bradley, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1301.
15 Waller, C. L.; McKinney, J. D. *Chem. Res. Toxicol*. **1995**, *8*, 847.
16 Mason, G.; Sawyer, T.; Keys, B.; Bandiera, S.; Romkes, M.; Piskorskapliszcynska, J.; Zmudzka, B.; Safe, S. *Toxicology* **1985**, *37*, 1.
17 Liu, S. S.; Liu, Y.; Li, Z. L.; Cai, S. X. *Acta Chim. Sinica* **2000**, *58*(11), 1353 (in Chinese).
(刘树深, 刘堰, 李志良, 蔡绍晢, 化学学报, **2000**, *58*, (11), 1353.)
18 Liu, S. S.; Yin, C. S.; Shi, Y. Y.; Cai, S. X.; Li, Z. L. *Chin. J. Chem*. **2001**, *19*, 751.
19 Liu, S. S.; Yin, C. S.; Li, Z. L.; Cai, S. X. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 321.
20 Liu, S. S.; Liu, H. L.; Shi, Y. Y.; Wang, L. S. *Internet Electron. J. Mol. Des*. **2002**, *1*(6), 310.
21 Liu, S. S.; Cui, S. H.; Yin, D. Q.; Shi, Y. Y.; Wang, L. S. *Chin. J. Chem*. **2003**, *21*, 1510.
22 Liu, S. S.; Cai, S. X.; Cao, C. Z.; Li, Z. L. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1337.
23 Liu, S. S.; Cui, S. H.; Shi, Y. Y.; Wang, L. S. *Internet Electron. J. Mol. Des*. **2002**, *1*(11), 610.
24 Liu, S. S.; Cao, C. Z.; Li, Z. L. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 387.
25 Liu, S. S.; Liu, H. L.; Xia, Z. N.; Cao, C. Z.; Li, Z. L. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 951.
26 Fang, K. T.; Quan, H.; Chen, Q. Y. *Applied Regression Analysis*, China Science Press, Beijing, **1988** (in Chinese).
(方开泰, 全辉, 陈庆云, 实用回归分析, 科学出版社, 北京, **1988**.)
27 Bandiera, S.; Sawyer, T.; Romkes, M.; Zmudzka, B.; Safe, L.; Mason, G.; Keys, B.; Safe, S. *Toxicology* **1984**, *32*, 131.
28 Liu, S. S. *Structural Characterization of Organic Compounds by the Molecular Electronegativity Distance Vector* (*MEDV*), Higher Education Press, Beijing, **2005** (in Chinese).
(刘树深, 有机物分子电性距离矢量表征及其应用, 高等教育出版社, 北京, **2005**.)
29 Hall, L. H. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 1039.
30 Kier, L. B.; Hall, L. H. *Molecular Structure Description—The Electrotopological State*, Academic Press, San Diego, **1999**.
31 Westfall, P. H.; Young, S. S.; Lin, D. K. *J. Stat. Sinica* **1998**, *8*, 101.
32 Liu, S. S.; Yi, Z. S. *Base Chemometrics*, Science Press, Beijing, **1999** (in Chinese).
(刘树深, 易忠胜, 基础化学计量学, 科学出版社, 北京, **1999**.)
33 Cho, S. J.; Hermsmeier, M. A. *J. Chem. Inf. Comput. Sci*. **2002**, *42*(4), 927.
34 Liu, S. S.; Liu, H. L.; Yin, C. S.; Wang, L. S. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 964.
35 Liu, S.-S.; Yin, D.-Q.; Cui, S.-H.; Wang, L.-S. *Chin. J. Chem*. **2005**, *25*(5), 622.

(A0508067 QIN, X. Q.; FAN, Y. Y.)