

• 研究论文 •

一种新型手性分子电性矩边矢量(Vmedc)的设计及其应用

陈国华^{*,a,b} 夏之宁^{*,a} 陆瑶^b 廖立敏^a
舒茂^a 孙家英^a 李志良^a

^a重庆大学生物工程学院/化学化工学院 重庆 400044)

^b四川理工学院材料与化学工程系 四川自贡 643000)

摘要 根据分子中不同类型原子间电相互作用的不同,文中提出了一种手性分子电矩边矢量(Vmedc),进一步拓展分子电矩边性矢量(Vmed)使用范围.为检测该手性描述矢量的结构表达特性和模型预测能力,分别对 32 个培哚普利拉类血管紧张素转化酶(ACE)抑制剂的对映结构体和 7 对苯基哌啶类 σ -受体抑制剂进行考察.32 个 ACE 抑制剂多元逐步回归系数 $R=0.913$ ($R^2=0.834$, $SD=0.768$, $F=33.875$),留一法交互检验为 $R_{cv}=0.877$ ($R_{cv}^2=0.769$, $SD_{cv}=0.906$, $F_{cv}=22.473$),具有较强预测能力;继而用 BP 神经网络,对 60 组随机样本(23:9)进行留分法分析取得较好结果,训练集平均为: $R_{Training}=0.931$ ($R_{Training}^2=0.967$),预测集为: $R_{cv}=0.918$ ($R_{cv}^2=0.842$);而对 14 个 σ -受体抑制剂多元回归($R=0.955$, $R_{cv}^2=0.849$)获得与文献一致结果.再用 Fisher 线性判别方法和 BP 神经网络对 ACE 抑制剂进行判别分析,其活性分类 88.89% 正确(仅 9 号错误),非活性分类 100.0% 正确,总分类正确率为 96.87%.两个数据集测试证明该方法与其它文献方法相当,这为定量构效关系(QSAR)研究提供一种新选择,扩充了 Vmed 描述矢量应用范围.

关键词 ACE 抑制剂;手性;Vmedc;苯基哌啶类;反传神经网络;Fisher 判别分析

A Vector of Molecular Electronegative Distance for Chiral Compounds (Vmedc) and Its Applications to Codification of Central Chirality

CHEN, Guo-Hua^{*,a,b} XIA, Zhi-Ning^{*,a} LU, Yao^b LIAO, Li-Min^a
SHU, Mao^a SUN, Jia-Ying^a LI, Zhi-Liang^a

^aCollege of Bioengineering/College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400044)

^bDepartment of Materials and Chemical Engineering, Sichuan University of Science and Engineering, Zigong 643000)

Abstract Based on the interaction between different atomic types, Vmedc, a novel vector of molecular electronegative distance (Vmed) has been defined and generalized in order to further codify chemical structural information for chiral drugs. Some quantitative structure-activity relationships (QSAR) have been modeled by Vmedc for both 32 stereoisomers of perindoprilate as angiotensin-converting enzyme ACE inhibitors and 7 pairs of chiral *N*-alkylated 3-(3-hydroxyphenyl)-piperidines that bind σ -receptors. Stepwise linear regression analysis was made forward to the 32 stereoisomers with good modeling results: $R=0.913$ ($R^2=0.834$, $SD=0.768$, $F=33.875$); $R_{cv}=0.877$ ($R_{cv}^2=0.769$, $SD_{cv}=0.906$, $F_{cv}=22.473$). Furthermore, average correlation coefficients (R) for random 60 groups with 23 training compounds for all the 32 ACE stereoisomers by backpropagation neural network (BPNN) were $R_{tr}=0.931$ ($R_{tr}^2=0.967$) and $R_{cv}=0.918$ ($R_{cv}^2=0.842$), except for four groups sampled unreasonably. Compared with literatures, Vmedc has also

* E-mail: zlli-cqu@163.com, chgh29@163.com, zlli2662@163.com

Received December 26, 2007; revised April 16, 2008; accepted May 20, 2008.

国家高技术研究发展计划(863 计划)专题(No. 2006AA02Z312), 重庆大学研究生创新团队项目科技创新基金(No. 200711C1A0010260)资助项目.

been applied to obtain good results for 14 samples with correlation coefficient being $R_{cv}=0.955$ ($R_{cv}^2=0.849$). Through both Fisher' linear discriminant analysis and BPNN, the 32 ACE stereoisomers were classified correctly into 88.89% active with one (#9) wrongly classified, 100.00% nonactive with no wrongly classified, and average classification of 96.87% globally. Good results obtained here were compared to those obtained with other chiral descriptors, when it was applied to the same 2 datasets, which shows that the Vmedc approach provides a powerful alternative QSAR technique for chiral compounds.

Keywords ACE inhibitor; chiral; Vmedc; *N*-alkylated-3-(3-hydroxyphenyl)piperidine; backpropagation neural network; linear discriminant analysis

1 前言

物质的分子结构决定其性质, 在三维空间中, 分子立体结构通常被特征化为各原子之间的键长、键角和二面角等特征表示. 而在二维空间中, 分子结构特征被简化为组成分子中的各原子拓扑连接关系, 即分子图. 由相同原子组成的分子, 由于具有不同的拓扑连接关系, 常被称为同分异构体; 然而, 即使两个分子具有完全相同的分子连接拓扑关系, 由于其中一个原子的手性不同, 对应结构也不同, 即具有不同的3D空间结构, 若对应分子与其镜像不能重叠则称它们为对映异构体^[1,2]. 在本研究中, 组成不对称的原子被称作手性原子, 而具有手性原子的分子被称为手性分子, 物质的性质不仅与分子的二维拓扑连接关系有关, 而且与它的三维空间结构也具有重要的关系. 在药物分子中, 具有相同分子拓扑结构而立体结构不同的分子, 其药物活性可能完全不同, 有的甚至还可能具有毒副作用, 例如: 震惊中外的“反应停”事件^[3], 因此分子性质不仅决定于它的分子拓扑关系而且受制于它的空间立体结构.

1.1 定量构效关系(QSAR)

定量构效关系(quantitative structure-activity relationship, QSAR)是应用最为广泛的药物设计手段与技术之一, 它是采用数理统计方法, 建立起化合物性质、生物活性或药物药效与其结构参数之间的定量关系^[4]. 通过这些定量关系, 可预测化合物的生理活性或物理性质, 指导我们设计出具有更高活性的化合物^[5]. 常用QSAR方法有二维定量构效关系(2D-QSAR)^[6,7], 三维定量构效关系(3D-QSAR)^[8~10]以及更高维的定量结构关系^[11,12]. 定量构效关系最早可追溯到 Crum-Brow 和 Fraser^[13]于 1867 年提出, 到 20 世纪前后, Overton 和 Meyer 等^[14,15]提出了麻醉作用的类脂学说, 即化学结构各异的麻醉剂其活性会随着脂-水分配系数增加而增加, 这可能是最早提出化合物生理活性和物理化学性质之间的定量关系模型, 因而得到大量的应用. 特别是在 1962 年, Hansch^[16]提出同系列化合物某些生物活性的变

化是和它们某些可测量物理化学性质的变化相联系的, 这些可测量的特性包括疏水性质、电性质和空间立体性质等, 都有可能影响化合物的生物活性, 这一重要发现促进了 QSAR 方法的进一步蓬勃发展. 但 2D-QSAR 是基于分子结构的拓扑关系而建立起来的, 物理意义不十分明确. 随着我们对三维空间研究能力的发展, 其三维定量构效关系(3D-QSAR)应运而生, 譬如分子形状分析(molecular shape analysis, MSA)^[17]、距离几何方法(distance geometry, DG)^[18]和比较分子场分析(comparative molecular field analysis, CoMFA)方法^[19]等. 这些 3D-QSAR 方法在物理化学上的意义较为明确, 能间接反映药物分子和靶点之间的非键相互作用特征, 因此近十多年来 3D-QSAR 方法得到了迅速的发展和广泛的应用, 其中 CoMFA 方法是目前最为成熟且应用最为广泛的方法^[20]. 虽然这些 QSAR 方法在药物设计方面已取得巨大成功, 但是对于手性化合物, 传统 2D-QSAR 方法无法进行区分^[21,22], 而基于空间结构 3D-QSAR 方法, 如 CoMFA, 对于手性化合物, 由于其自身空间关系不同而被考虑, 但是对于那些高度柔韧的化合物, 其处理起来就非常困难, 即使这一缺陷被 Grid 方法所克服, 但是其依然存在处理数据量太大这一难题^[23,24]. 因此开展手性化合物的 QSAR 方法研究显得极为重要, 经过多年的努力, 现已取得比较令人满意的结果^[21,22]. 本研究室提出 Vmed 方法在 QSAR 研究方面已经取得了比较满意的结果^[25,26], 本研究就是在原 Vmed 方法的基础上, 进行手性分子的修订, 从而扩展其使用范围.

1.2 Vmedc 描述矢量

本研究室根据组成分子拓扑结构关系, 不同类型原子划分方案和分子中各类原子的相对电性和原子之间相对键距的不同, 提出了 Vmed 描述矢量. 在此基础上, 经过不断的努力, 发展出系列二维新型电性矩边描述矢量(MED/MDE/MEE/MHD), 广泛应用于生物与药物及毒物分子结构性质关系的表达中^[25~27]. 但是该方法属于 2D-QSAR 方法, 因而无法区分对映异构体, 为了进一步拓展 Vmed 的分析范围, 在本研究中对其计算方法

进行部分的修改. 在本研究中, 依然应用隐氢分子图进行计算^[25], 但是在原子分类方法时, 先将分子按原来 Vmed 方法划分为 C 族, O 族, N 族, X 族(卤素)四类, 在此基础上进一步根据分子中原子轨道杂化状态(sp³, sp², sp¹)的不同进一步划分为 9 类 C, N (sp³, sp², sp¹), O (sp³, sp²)和 X (sp³), 电负性分别为 2.55, 3.44, 3.04 和 3.98 (3.16, 2.96, 2.66); 半径为 1.70, 1.55, 1.60 和 1.50 (1.80, 1.95, 2.10). Vmed 定义各类原子之间相互作用为^[25]:

$$M_{kl} = \sum_{i \in k, j \in l} (q_i q_j / d_{ij}^2) \quad (1)$$

$$d_{i,j} = \sum_{i,j} (d'_i) \quad (2)$$

其中 K 和 l 分别代表不同分子类型, i 和 j 分子图中的连接编码, d'_i 为第 i 原子到 j 原子之间的相对键长, 而 q_i 和 q_j 分别代表第 i 号原子和 j 号原子的相对电负性. 为了拓展 Vmed 对手性化合物的处理能力, 在相对路径计算时, 先按照原方法搜索出两原子之间连接的最短路径, 然后根据最短路径中, 其经过的手性碳原子构型不同进行不同的计算. 根据各键包含手性碳原子的不同, 将式(2)按照下面规则进行手性变形处理:

$$d_{i,j} = \sum_{i,j} [(d'_i) * (1+w)] \quad (3)$$

对于含有手性碳原子 i 有:

$$d'_{(i-1)} * (1+w_1) + d'_{(i+1)} * (1+w_2)$$

(1)如果 i 的手性为 R 型, 则 $w_1=0, w_2=0$. (2)如果中心原子 i 为手性为 S 型, 则: (a)如果从 i 到 $(i-1)$ 的方向其最近手性为 S 型, 则 $w_1=0$; 如果无手性或者手性为 R 型, 则 $w_1=1$. (b)如果从 i 到 $(i+1)$ 的方向其最近手性为 S 型, 则 $w_1=0$; 如果无手性或者手性为 R 型, 则 $w_1=1$. 如图 1, 不考虑原子的手性, 其 $RSSRS$ 构型分子与 $RSSRR$ 构型分子中 4 号原子到 15 号原子 Vmed 计算用最短路径为: 4, 3a, 7a, 1, 8, 9, 10, 11, 15, 则(2)式相对最短距离计算如下:

$$d_{4,15} = L_{4-3a}/1.54 + L_{3a-7a}/1.54 + L_{7a-1}/1.54 + L_{1-8}/1.54 + L_{8-9}/1.54 + L_{9-10}/1.54 + L_{10-11}/1.54 + L_{11-15}/1.54 = 1.54/1.54 + 1.54/1.54 + 1.47/1.54 + 1.367/1.54 + 1.54/1.54 + 1.47/1.54 + 1.47/1.54 + 1.54/1.54 = 7.751$$

而对于按照手性相对路径计算方法, 由式(3)其 $RSSRS$ 构型分子与 $RSSRR$ 构型分子计算分别如下:

$$d_{4,15}(RSSRS) = (1.54/1.54) * (1+0) + (1.54/1.54) * (1+1) + (1.47/1.54) * (1+1) + (1.367/1.54) * (1+0) + (1.54/1.54) * (1+0) + (1.47/1.54) * (1+0) + (1.47/1.54) * (1+1) + (1.54/1.54) * (1+1) = 11.660$$

$$d_{4,15}(RSSRS) = (1.54/1.54) * (1+0) + (1.54/1.54) * (1+1) + (1.47/1.54) * (1+1) + (1.367/1.54) * (1+0) + (1.54/1.54) * (1+0) + (1.47/1.54) * (1+0) + (1.47/1.54) * (1+0) + (1.54/1.54) * (1+0) = 9.70584$$

则 sp³ 杂化碳原子之间的 Vmedc 为: $M_{(RSSRS)} = 15.442$; $M_{(RSSRR)} = 14.584$. 由上面的计算可知, 对于手性构型分别为 $RSSRS$ 和 $RSSRR$ 的分子, 不管分子中四个手性原子构型如何, Vmed 计算的最小相对距离 $d_{4,15}$ 都相同, 无法对手性化合物进行分辨识别; 而按照修订的手性相对最短路径计算式(3)计算, 结果完全不同, 从而达到了对手性原子的区分目的.

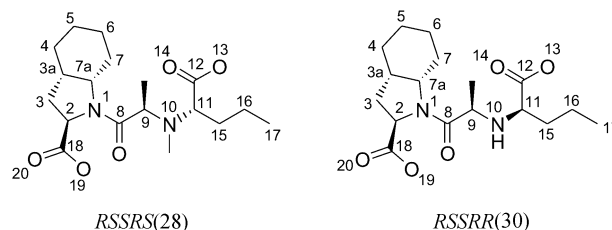


图 1 ACE 抑制剂分子结构图
Figure 1 The sketches of ACE inhibitor

2 试验方法

2.1 Vmedc 描述矢量的数据准备

为检验该手性描述矢量的预测和分辨分子结构与活性的能力, 分别选取如图 1 所示的 32 个血管紧张素转换酶(ACE)抑制剂对映异构体^[28]和 14 个苯基哌啶类(HPP)手性化合物进行研究^[29]. 首先在 Chem3D 10.0 下生成分子的立体结构, 然后采用 Gaussian98, 使用密度泛函(DFT)的 B3PW91 方法, 在 6-31G 基组上进行分子结构优化^[30], 然后利用优化后的分子实际键长代替标准键长, 按照式(1)和(3)计算 Vmedc 描述子. 在分类研究中, 设定 32 个 ACE 抑制剂的 IC₅₀ 值不高于 110 nmol/L 为活性化合物类别, 用 ACE_{inactv} = 1 表示; 而对于抑制浓度高于 110 nmol/L 的设为非活性化合物类别, 用 ACE_{inactv} = -1 表示; 所有的计算都在 Pentium IV 3.0 上进行.

2.2 预测分析

2.2.1 多元线性回归

多元线性回归常被应用于一个因变量和一个或者几个自变量之间的线性关系研究. 当样本数 n 相对于变量数 m 不是很大时, 会很容易地获得较大的 R 值, 即容易产生所谓偶然相关, 使变量发生重叠; 为了使所得的定量数学模型具有良好的相关性及较好的均方根偏差, 但含有较少的变量, 需要对变量进行选择 and 压缩, 即建

立“最优”回归方程. 所谓“最优”回归方程, 主要是指, 在回归方程中包含所有对因变量 y 影响显著的自变量, 而不包含对 y 影响不显著的自变量的回归方程. 逐步回归分析正是根据这种原则提出来的一种回归分析方法, 逐步回归法的基本思想是将变量逐一引入, 即以偏回归平方和 F 检验显著的变量作为被引入变量, 新变量引入后还要重新对所有已引入的变量进行检验, 不显著者从方程中剔除, 直到没有变量可剔除也没有变量可引入时为止, 最后对所选定变量建立线性回归方程^[31,32], 本文使用 SPSS15.0 进行多元线性回归分析.

2.2.2 反向传播人工神经网络(BPNN)

人工神经网络(Artificial neural network, ANN)是一种非线性预测工具^[33], 它的数学模型是通过神经网络对人脑或自然神经网络若干基本特性的抽象和模拟, 其目的在于模拟大脑的某些机理与机制, 实现某个方面的功能^[34]. 一般的神经网络采用 I-H-O 的三层网络结构, 其中 I 为输入层神经元数目, H 代表隐藏层神经元数目, O 代表输出层神经元数目. 目前已经有了很多的神经网络方法, 如: 多层网络 BP 算法^[35], Hopfield 网络模型^[36], 自适应共振理论^[37], 自组织特征映射理论^[38]等. 反向传播(Backpropagation, BP)模型是一种前向多层神经网络的误差反向传播学习算法, 它亦采用 I-H-O 的三层网络结构, 输入数据经作用函数后, 再将隐节点的输出信号传递到输出节点, 最后给出输出结果. 该算法的学习过程由信息的前向传播和误差的反向传播组成. 在前向传播的过程中, 输入信息从输入层经隐含层逐层处理, 并传向输出层; 第一层神经元的状态只影响下一层神经元的状态. 如果在输出层尚得不到期望的输出结果, 则转入反向传播, 将误差信号(目标值与网络输出之差)沿原来的连接通道返回, 通过修改各层神经元权值, 使得误差均方最小. 由于神经网络映射关系不限于线性关系, 常被应用在非线性和曲线方面的分析. 通常研究的生物现象, 常常不具有线性关系, 本研究采用自编 BP 神经网络程序, 先将输入神经网络的数据按照公式: $X = (X - X_{\min}) / (X_{\max} - X_{\min})$ 归一化处理后,

输入神经网络进行分析.

2.2.3 分类判别分析

LDA 线性判别分析的基本原理就是在 n 维样本空间, 按照(4)式的线性函数通过一判决边界直线或超平面型判别边界将任意的两个类别区分开来^[21,22].

$$Y = a_1X_1 + a_2X_2 + a_3X_3 + \dots + a_mX_m \quad (4)$$

其中: Y 为判别分数(判别值); $X_1, X_2, X_3, \dots, X_m$ 为反映研究对象特征的变量, a_1, a_2, a_3, a_m 为各变量的贡献大小, 亦称判别系数. 判别分析就是通过判别函数进行判别分析. 根据样本空间类别数量的不同, 依据这些描述变量, 样本空间被一个判别函数将样本空间划分两类别, 通过计算各判别函数的得分来决定预测样本的类别. 判别模型的质量通过检测 wilk 的 λ 统计量, F 值, p 水平和各类分类正确率进行 F 值的评价来入选变量. 本文通过检测典则回归系数, χ 和 p 水平进行分析. 为了判定哪些变量具有最优的组间判别效力, 采用前向逐步判别方法对研究中 Vmedc 描述矢量进行选择. 所有分析均采用 SPSS15.0 程序进行, 在此基础上, 利用上述判别分析得到的主要影响变量, 采用一个 4-9-1 的自编 BP 神经网络程序进一步的对培哌普利类对映异构体进行分类分析.

3 结果与讨论

3.1 预测分析结果

多元线性逐步回归常被应用在生物系统的建模方面, 本文分别对 32 个 ACE 抑制剂的 19 个 Vmedc 描述矢量和 7 对 HPP 手性衍生物的 8 个 Vmedc 描述矢量进行逐步回归分析. ACE 抑制剂的 19 个 Vmedc 描述矢量的前向分析结果如表 1, 随着变量的增加($R = 0.777 \sim 0.920$), 其相关系数逐步增加, 最终由 4 个 Vmedc 描述变量 $v_2(\text{C}_{\text{SP}2} - \text{C}_{\text{SP}3}, b_1)$, $v_{16}(\text{N}_{\text{SP}3} - \text{O}_{\text{SP}3}, b_2)$, $v_{12}(\text{N}_{\text{SP}2} - \text{N}_{\text{SP}3}, b_3)$, $v_4(\text{C}_{\text{SP}2} - \text{N}_{\text{SP}3}, b_4)$, $v_{19}(\text{O}_{\text{SP}3} - \text{O}_{\text{SP}3}, b_5)$ 达到最佳预测能力(b_0 为拟合常数), 经 LOO 方法检验 $R_{\text{cv}} = 0.877$, $SD_{\text{cv}} = 0.906$, 模型具有比较好的预测分析能力.

表 1 采用多元线性逐步回归分析获得的 ACE 抑制剂的 QSAR 结果
Table 1 QSAR results of variables selected for ACE inhibitors by SMR

Model	b_0	b_1	b_2	b_3	b_4	b_5	R	R^2	R_{adj}^2	SD	F	U	Q	$R_{\text{cv}}^2(01)$	$SD(01)$	$F(01)$	$U(01)$	$Q(01)$
1	9.452	-1.384					0.777	0.604	0.591	1.125	45.736	57.927	37.996	0.554	1.194	37.275	53.148	42.775
2	8.500	-1.773			13.447		0.821	0.675	0.652	1.038	30.049	64.702	31.221	0.609	1.138	22.562	58.395	37.528
3	10.922	-1.548			15.522	-194.270	0.872	0.760	0.734	0.907	29.528	72.885	23.038	0.701	1.012	21.884	67.244	28.679
4	12.043	-1.177		-11.251	11.613	-215.233	0.893	0.798	0.768	0.847	26.658	76.542	19.381	0.717	1.003	17.101	68.777	27.146
5	37.348	-0.653	149.284	-336.633	-291.159	-96.356	0.920	0.846	0.817	0.753	28.629	81.178	14.745	0.751	0.959	15.653	72.004	23.919
6	44.592	-0.579	196.688	-439.157	-387.635		0.913	0.834	0.809	0.768	33.875	79.985	15.938	0.769	0.906	22.473	73.766	22.157

逐步回归方程为($N=32$, $R=0.913$ ($R^2=0.834$), $SD=0.768$, $F=33.875$, $U=79.985$, $Q=15.938$; LOO 交互检验预测: $R_{cv}=0.877$ ($R_{cv}^2=0.769$), $SD_{cv}=0.906$, $F_{cv}=22.473$):

$$\log(\text{IC}_{50}) = 44.592 - 0.579v_2 + 196.687v_4 - 439.153v_{12} - 387.632v_{16}$$

在多元线性逐步回归变量选择的基础上, 利用被选中的 4 个 Vmedc 描述变量 v_2 , v_{16} , v_{12} 和 v_4 作为输入, 选取一个 4-9-1 的三层 BP 神经网络进行预测分析, 设定神经学习效率因子 0.01, 学习动量为 0.01, 最大学习次数为 5 百万次, 为了评估模型的预测能力, 采用留分法对神经网络进行训练测试, 随机抽取 60 组样本, 其中 23 个样本作为训练集, 9 个作为预测集, 采用 BP 方法进行训练预测, 结果如表 2. 其中除 2 组[49 ($R=0.962$, $S=0.481$, $F=169.343$, $R_{cv}^2=-0.033$), 57 ($R=0.942$, $S=0.569$, $F=110.079$, $R_{cv}^2=0.282$)]由于抽样分布不太合理外, 其它各组的相关系数都不低于 0.500; 进一步说明该方法预测能力比较稳定, 其训练和测试集的平均相关系数为: $R=0.931$ ($R^2=0.867$), $R_{cv}=0.918$ ($R_{cv}^2=0.842$), 其总体结果优于线性回归结果. 本文中进一步研究只含有一个手性碳原子的苯基哌啶类手性化合物, σ -受体抑制剂的逐步回归结果如表 3(Vmedc 描述矢量: $v_1(\text{C}_{\text{SP}2}-\text{C}_{\text{SP}2}, b_1)$, $v_3(\text{C}_{\text{SP}2}-\text{N}_{\text{SP}3}, b_2)$, $v_5(\text{C}_{\text{SP}3}-\text{C}_{\text{SP}3}, b_3)$, $v_6(\text{C}_{\text{SP}3}-\text{N}_{\text{SP}3}, b_4)$, b 为拟合常数), 其回归方程为($n=14$, $R=0.955$ ($R^2=0.912$), $SD=0.343$, $F=34.536$, $U=12.169$, $Q=1.175$, LOO 预测: $n=14$, $R_{cv}=0.921$ ($R_{cv}^2=0.849$), $SD_{cv}=0.449$, $F_{cv}=18.701$, $U_{cv}=11.325$, $Q_{cv}=2.019$):

$$\log(1/\text{IC}_{50}) = 4.603(\pm 0.539) - 0.539(\pm 0.065)*v_1 + 8.515(\pm 1.304)*v_3 - 0.442(\pm 0.054)*v_5$$

与文献^[22]比较, Vmedc 方法回归 R 值与文献报道相当($R=0.969$, $F=85.826$, $S=0.270$, $S_{cv}=0.289$, $p<0.000$), 说明该手性 Vmedc 方法是完全具备对手性化合物的预测分析能力, 拓展了其方法的预测范围, 其 14 个化合物的估计值与观测值的回归和交互校验结果如图 2 所示.

表 3 对 HPP 的 Vmedc 作用项前向逐步回归变量筛选结果

Table 3 QSAR results of forward stepwise MLR for the Vmedc indices of HPP

Model	b_0	b_1	b_2	b_3	b_4	R	R^2	R_{adj}^2	SD	F	U	Q	$R_{cv}^2(01)$	$SD(01)$	$F(01)$	$U(01)$	$Q(01)$
1	4.619				-1.139	0.706	0.499	0.457	0.747	11.945	6.657	6.687					
2	5.519	-0.120			-1.054	0.823	0.677	0.618	0.626	11.527	9.034	4.310	0.527	0.758	6.124	7.030	6.314
3	6.203	-0.319	4.550		-1.143	0.903	0.815	0.760	0.496	14.718	10.880	2.464	0.709	0.623	8.111	9.457	3.887
4	5.378	-0.490	7.691	-0.340	-0.355	0.963	0.927	0.895	0.329	28.654	12.372	0.972	0.864	0.448	14.345	11.535	1.809
5	4.603	-0.539	8.515	-0.442		0.955	0.912	0.886	0.343	34.536	12.169	1.175	0.849	0.449	18.693	11.325	2.019

表 2 60 组 ACE 抑制剂随机样本训练统计结果

Table 2 QSAR Statistical results of random 60 sets for the 32 ACE Inhibitors

	R_{training}	S_{training}	F_{training}	PRESS	R_{cv}^2
Mean	0.931	0.554	99.956	3.153	0.842
Max	0.964	0.635	181.467	6.124	0.985
Min	0.866	0.446	43.748	0.607	-0.033

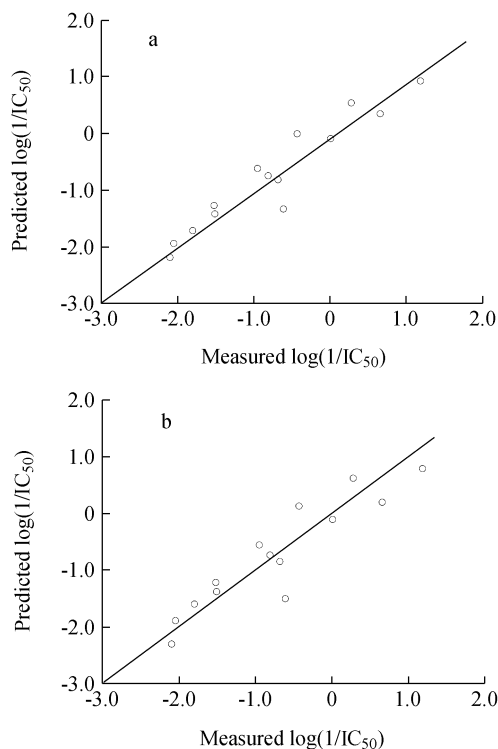


图 2 苯基哌啶类及其留一法计算值与实验值

Figure 2 Correlation between experimental and calculated (a) or validated (b) $\log(1/\text{IC}_{50})$ of 14 derivatives of HPP

3.2 分类结果

3.2.1 Fisher 线性判别分析结果

在本研究中, LDA 和 BPN 被分别采用对 ACE 的活性进行分类分析, 检验该 Vmedc 的分类能力. LDA 采用一个简单的线性 QSAR 模型去拟合分类函数, 为了获得高的统计特性和较少的参数, 采用 Wilks's lambda 方法进行前向逐步判别分析, 得到下面方程 $n=32$, $\lambda=0.175$,

$\chi^2=48.875, p<0.000$:

$$\text{ACE}_{\text{inactive}} = 123.62 \cdot v_3 + 235.877 \cdot v_{12} - 527.89 \cdot v_{17} + 2956.005 \cdot v_{19} - 238.693$$

其中 p 为检验显著性水平, λ 取值范围 0 到 1, 其中 0 代表完全分类, 1 为最差分类. 其特征值为 4.729, 典型相关系数 R 为 0.909, 显著性概率 $\text{Sig}=0.000, \text{Sig}<0.01$ 从而认为判别函数有效. 方差百分比和方差累计百分比均为 100%. 典则判别结果 88.89% 的活性化合物(8/9)分类正确, 而非活性化合物 100.0% 的分类正确, 总的分类正确率为 96.87% (31/32), 仅有一个化合物 9 号分类错误. 为了和现有方法进行比较^[21,22], 也为了检验其方法的预测能力, 按照文献方式分别采用 LOO 检验法和留分法得到相同的分析判别能力, 该手性描述子得到一致的结果(表 4).

3.2.2 BP 分类结果

在此基础上, 依据 Fisher 线性逐步判别结果, 以 v_3, v_{12}, v_{17} 和 v_{19} 作为输入, 选取一个 4-9-1 的三层 BP 神经网络进行分类分析. 设定神经网络的学习效率因子

0.01, 学习动量为 0.01, 最大学习次数为 5 百万次, 随机产生 60 组训练预测样本(23:9), 分别对神经进行训练测试, 结果与 Fisher 判别分类结果一致. 同时在该 60 组样本中仅发现 9 号发生分类错误, 因此将 9 号样本作为一个离群值, 再随机选择 60 组样本进行分类训练测试, 则无任何分类错误.

4 结论

传统的 2D-QSAR 方法无法对手性化合物进行分辨处理, 发展高效 2D-和 3D-QSAR 方法显得越来越迫切. 本研究提出的 Vmedc 描述子不仅能够准确地对手性化合物进行分类分析, 而且能够比较准确地预测手性化合物的生物活性值, 但是更进一步的对手性化合物其立体结构与生物活性的关系研究还需努力.

致谢 感谢新加坡国立技术学院生命科学技术中心(新加坡 139651)的陈刚博士提供的 Gaussian98 计算结果.

表 4 对 32 个 ACE 抑制剂的 LDA 分类结果^[21,22]

Table 4 QSAR results of LDA classification to 32 ACE inhibitors with canonical score (s_{cano})

ID	Compound ^a	IC ₅₀ ^b /(nmol·L ⁻¹)	P ^c /%	P _{CV} ^d /%	s_{cano} ^e	P ^{f,g} /%	s_{cano} ^h	ID	Compound ^a	IC ₅₀ ^b /(nmol·L ⁻¹)	P ^c /%	P _{CV} ^d /%	s_{cano} ^e	P ^{f,g} /%	s_{cano} ^h
1	SSRSS ^p	1.1	1.000	0.999	2.890	0.999	2.551	17	SRRRR	33×10 ³	0.990	0.978	0.041	0.956	0.341
2	SRSSS ^p	1.2	1.000	1.000	4.529	1.000	4.276	18	RSSSR	36×10 ³	1.000	1.000	-1.547	1.000	-1.270
3	SSSSS	1.5	1.000	1.000	4.858	1.000	4.661	19	RSRSR	47×10 ³	1.000	1.000	-1.340	1.000	-1.052
4	SRRSS ^p	3.3	1.000	1.000	3.486	1.000	3.183	20	RSRSS ^p	60×10 ³	0.991	0.968	0.015	0.991	-0.033
5	SSSSR	12.2	1.000	1.000	3.811	1.000	3.910	21	RRRRR	10 ⁵	1.000	1.000	-1.820	1.000	-1.391
6	SSRSR	29.4	1.000	1.000	3.352	1.000	3.389	22	SRRRS	10 ⁵	1.000	1.000	-2.590	1.000	-2.336
7	SRRSR	39.8	1.000	1.000	3.356	1.000	3.390	23	RRRSS	10 ⁵	0.967	0.890	0.304	0.970	0.252
8	SRSSR	54.0	1.000	1.000	3.473	1.000	3.508	24	SRSRR ^p	10 ⁵	0.999	0.997	-0.378	0.994	-0.118
9	RRSSS ^e	108.0	0.907	0.999	0.538	0.908	0.522	25	RRRRS	10 ⁵	1.000	1.000	-2.087	1.000	-1.684
10	SSSSS	1.1×10 ³	1.000	1.000	-1.772	1.000	-1.444	26	RRSRR	10 ⁵	1.000	1.000	-2.094	1.000	-1.676
11	RSSSS	1.9×10 ³	0.950	0.876	0.397	0.948	0.379	27	SSSRR	10 ⁵	1.000	0.999	-0.720	0.999	-0.488
12	SSRRR ^p	2.6×10 ³	1.000	0.999	-0.681	0.998	-0.441	28	RSSRS ^p	10 ⁵	1.000	1.000	-1.781	1.000	-1.364
13	RRSSR	5.5×10 ³	1.000	1.000	-1.244	1.000	-0.945	29	RRRSR	10 ⁵	1.000	1.000	-1.375	1.000	-1.095
14	SSRRS	7.1×10 ³	1.000	1.000	-2.393	1.000	-2.089	30	RSSRR	10 ⁵	1.000	1.000	-2.257	1.000	-1.864
15	RRRS	7.8×10 ³	1.000	1.000	-2.250	1.000	-1.853	31	SRRRS	10 ⁵	1.000	1.000	-1.627	1.000	-1.164

1~10 号为活性化合物, 11~32 号为非活性化合物; * 错误分类化合物; ^a ACE 抑制剂的手性碳原子的编号为 C2, C3a, C7a, C9, C11; ^b ACE 抑制剂化合物的 IC₅₀ 值, 其值取自文献[21]和[28]; ^c 32 个手性化合物的预测后验概率; ^d 留一法预测的预测后验概率; ^e 留一法典则分析的典则得分; ^f 留分法中训练集的预测后验概率; ^g 留分法中预测集的预测后验概率; ^h 留分法典则分析的典则得分; ^p 留分法中预测集化合物.

References

- Helgakert, T.; Jorgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*, Wiley, Chichester, 2000.
- Estrada, E.; Uriarte, E. *Curr. Med. Chem.* 2001, 8, 1573.
- Melchert, M.; List, A. *Int. J. Biochem. Cell Biol.* 2007, 39, 1489.
- Balaban, A. T. *SAR QSAR Environ. Res.* 1998, 8, 1.
- Randic, M. J. *Chem. Inf. Comput. Sci.* 1997, 37, 672.
- Zissimos, A. M.; Abraham, M. H.; Barker, M. C.; Box, K. J.; Tam, K. Y. *J. Chem. Soc., Perkin Trans. II* 2002, (3), 470.
- Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*, Academic Press, New York, 1999.
- Kubinyi, H.; Folkers, G.; Martin, Y. C. *3D QSAR in Drug*

- Design, Recent Advances*, Vol. 3, Kluwer/ESCOM, Dordrecht, **1998**.
- 9 Kubinyi, H. *3D QSAR in Drug Design, Theory, Methods and Applications*, ESCOM, Leiden, **1993**.
- 10 Kubinyi, H.; Folkers, G.; Martin, Y. C. *3D QSAR in Drug Design, Ligand Protein Interactions and Molecular Similarity*, Vol. 2, Kluwer/ESCOM, Dordrecht, **1998**.
- 11 Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B. Q.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. *J. Am. Chem. Soc.* **1997**, *119*, 10509.
- 12 Vedani, A. *J. Med. Chem.* **2002**, *45*, 2139.
- 13 Crum-Brown, A.; Fraser, T. R. *Trans. R. Soc. Edinburgh* **1868**, *25*, 151.
- 14 Overton, E. *Studien Uber die Narkose*, Fischer, Jena, Germany, **1901**.
- 15 Meyer, H. *Arch. Exp. Pathol. Pharmacol.* **1899**, *42*, 109.
- 16 Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. *Nature* **1962**, *194*, 178.
- 17 Hopfinger, A. J. *J. Am. Chem. Soc.* **1980**, *102*, 7196.
- 18 Crippen, G. M. *J. Med. Chem.* **1980**, *23*, 599.
- 19 Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959.
- 20 Baurin, N.; Vangrevelinghe, E. *J. Med. Chem.* **2000**, *43*, 1109.
- 21 Díaz, H. G.; Sánchez, I. H.; Uriarte, E.; Santana, L. *Comput. Biol. Chem.* **2003**, *27*, 217.
- 22 Ponce, Y. M.; Díaz, H. G.; Zaldivar, V. R.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2004**, *12*, 5331.
- 23 Golbraikh, A.; Bonchev, D.; Golbraik, A.; Bonchev, D.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 147.
- 24 Todeschini, R.; Moro, G.; Boggia, R.; Bonati, L.; Cosentino, U.; Lasagni, M.; Pitea, D. *Chemom. Intell. Lab. Syst.* **1997**, *36*, 65.
- 25 Liu, S.S.; Cao, C. Z.; Li, Z. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (2), 387.
- 26 Liu, S. S.; Cai, S. X.; Cao, C. Z.; Li, Z. *J. Chem. Inf. Comput. Sci.* **2000**, *40*(6), 1337.
- 27 Liu, S. S.; Yin, C. S.; Cai, S. X.; Li, Z. *J. Chin. Chem. Soc.* **2001**, *48*(2), 253
- 28 Vicent, M.; Marchand, B.; Rémond, G.; Jaquein-Guinamant, S.; Damien, G.; Portevin, B.; Baumal, J.; Volland, J.; Bouchet, J.; Lambert, P.; Serkiz, B.; Luitjen, W.; Laubie, M.; Schiavi, P. *Drug Des. Discov.* **1992**, *9*, 11.
- 29 de Julián-Ortiz, J. V.; de Alapont, C. G.; Ríos-Santamarina, I.; García-Doménech, R.; Gálvez, J. *J. Mol. Graphics Mod.* **1998**, *16*(1), 14.
- 30 Frisch, M. J. Trucks, G. W. H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, Replogle, E. S. and J. A. Pople, *Gaussian 98*, Gaussian, Inc., Pittsburgh PA, **1998**.
- 31 Draper, N. R.; Smith, H. *Applied Regression Analysis*, 2nd Ed., Wiley, New York, **1981**.
- 32 Miller, J. N.; Miller, J. C. *Statistics and Chemometrics for Analytical Chemistry*, Prentice Hall, London, **2000**.
- 33 Sadeghi, B. H. M. *J. Mater. Process. Technol.* **2000**, *103*, 411.
- 34 Zupan, J.; Gasteiger, J. *Anal. Chim. Acta* **1991**, *248*, 1.
- 35 Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. *Parallel Distributed Processing: Exploration in the Microstructures of Cognition*, MIT Press, Cambridge, **1986**.
- 36 Hopfield, J. J. *J. Proc. Natl. Acad. Sci. U. S. A.* **1982**, *79*, 2554.
- 37 Carpenter, G. A.; Grossberg, S. *Appl. Opt.* **1989**, *26*, 4919.
- 38 Fernández, M.; Caballero, J. *Bioorg. Med. Chem.* **2006**, *14*, 280.

(A0712262 CHEN, J. X.; LU, Z. S.)