2006 年第 64 卷 第 5 期, 393～396
化 学 学 报 ACTA CHIMICA SINICA
Vol. 64, 2006 No. 5, 393～396

• 研究论文 •

一组新氨基酸描述子用于肽定量构效关系研究

梁桂兆[a,b,c] 周 鹏[a] 周 原[a,b,c] 张巧霞[a] 李志良*[,a,b]

([a]重庆大学化学化工学院 重庆 400044)

([b]湖南大学化学生物传感与计量学国家重点实验室 长沙 410082)

([c]重庆大学生物工程学院 重庆 400044)

摘要 用主成分分析从 20 种天然氨基酸 0D～3D 结构信息中收集到的共 1369 个描述子变量得到了一组新氨基酸描述子(SZOTT), 将其用于血管紧张素转化酶抑制剂和苦味二肽结构表征并以偏最小二乘法建立定量构效关系模型, 得复相关系数 R_{CU}^2 分别为 0.894 和 0.908, 留一法交互检验的复相关系数 R_{CV}^2 分别为 0.828 和 0.736, 估计均方根误差 RMS 分别为 0.331 和 0.195. 研究结果表明, SZOTT 描述子含信息量大, 操作简便, 结构表达能力强, 有望在多肽定量构效关系研究中得到进一步推广.

关键词 肽; 定量构效关系; 主成分分析; 遗传算法; 偏最小二乘法

New Descriptors of Aminoacids and Their Applications to Peptide Quantitative Structure-Activity Relationship

LIANG, Gui-Zhao[a,b,c] ZHOU, Peng[a] ZHOU, Yuan[a,b,c] ZHANG, Qiao-Xia[a] LI, Zhi-Liang*[,a,b]

([a] *College of Chemistry and Chemical Engineering*, *Chongqing University*, *Chongqing 400044*)

([b] *State Key Laboratory of Chemo/Biosensing and Chemometrics at Hunan University*, *Changsha 410082*)

([c] *College of Bioengineering*, *Chongqing University*, *Chongqing 400044*)

Abstract A new set of descriptors, namely vector of scores for zero dimension, one dimension, two dimension and three dimension (SZOTT), was derived from principle component analysis of a matrix of 1369 structural variables of 0D～3D information for 20 coded aminoacids. SZOTT scales were then employed to express structures of angiotensin-converting enzyme inhibitors and bitter tasting thresholds, and to construct QSAR models based on partial least square (PLS). The results obtained are as follows: the multiple correlation coefficient (R_{CU}^2) of 0.894 and 0.908, the leave one out cross validated R_{CV}^2 of 0.828 and 0.736, and root-mean-square error for estimated error (RMS) of 0.331 and 0.195, respectively. Satisfactory results showed that, as new aminoacid scales, data of SZOTT may be a useful structural expression methodology for study on peptide QSAR (quantitative structure-activity relationship) due to its many advantages such as plentiful structural information, easy manipulation, and high characterization competence.

Keywords peptide; quantitative structure-activity relationship; principal component analysis; genetic algorithm; partial least square

近年来, 数以万计的多肽和蛋白质的结构与功能引起了研究者的浓厚兴趣. 研究表明, 蛋白质的诸多生物学功能与特定肽链的氨基酸排列顺序密切相关, 另外在生物体内亦有相当一部分属于肽直接发挥生物学功能,

* E-mail: zlli2662@163.com

Received June 15, 2005; revised and accepted November 3, 2005.

国家“春晖计划”教育部启动基金(No. 99-4-4＋37)、霍英东基金(No. 98-7-6)、重庆直辖市应用基础基金(No. 01-3-6)、重庆大学自主创新科技攻关项目(No. 03-5-6＋04-10-10)及湖南大学化学生物传感与计量学国家重点实验室项目(No. 2005-12)资助项目.

因此研究肽定量构效关系(QSAR)具有重要意义. 在肽的QSAR研究中, 结构表征是重要环节. 自Neath[1]基于20 种天然氨基酸的物化性质提出了一些半定量氨基酸描述子, 并将其用于建立一些脑下垂体-抗利尿激素类似物定量序效模型(QSAM)以来, 已产生了许多氨基酸定量描述子[2~8]并取得了成功的应用. 在本课题组研究[9]基础上, 从20种天然氨基酸的0D～3D信息共1369个描述子变量经过主成分分析, 得到了一组新氨基酸描述子SZOTT. 该描述子信息含量大, 操作简便且不需实验数据. 将其用于血管紧张素转化酶抑制剂及苦味二肽的QSAR研究, 应用偏最小二乘法建模, 获得较好结果.

1 原理与方法

1.1 主成分分析与 SZOTT 描述子

共收集了 20 种天然氨基酸的 1369 个描述子变量, 包括 31 个 0D 描述子[10], 69 个 1D 描述子[10], 640 个 2D 描述子[11~15], 629 个 3D 描述子[16~21]. 因描述子变量之间可能高度相关, 故采用主成分分析(PCA)[22]压缩描述子数量. 经 PCA 变换后, 其前 13 个主成分得分矩阵累计解释了原始变量数据矩阵(20×1369) 96.19%的方差. 因此可用此 13 个主成分得分矩阵替代原始变量矩阵. 为方便, 称13个得分矢量为SZOTT. PCA由变量矩阵经过自定标法(autoscaling)进行标准化处理后用软件Matlab 7.0 完成.

1.2 肽结构表征和变量挑选

每个肽可根据氨基酸顺序用13个SZOTT描述子表达, 不同长度肽链具有不同描述子个数, 即具有 n 个氨基酸残基的肽, 其一级序列结构可被 $13n$ 个变量表征. 为提高模型稳健性和预测能力, 应将与生物活性无关的描述子剔除, 使模型更易解释. 我们采用比较流行的遗传算法(GA)-偏最小二乘(PLS)法[23]挑选变量[24]. 模型内部预测能力以适应度函数 R_{CV}^2 评价: $R_{CV}^2 = 1 - \mathrm{PRESS/SSY}$, 式中, R_{CV}^2 为留一法交互验证(LOO-CV)的复相关系数 R^2; PRESS 为 CV 预测残差平方和; SSY 指 Y 值与其均值之差平方和. GA-PLS 程序用 Matlab 7.0 编写.

1.3 建立 QSAR 模型

应用 PLS 建立 QSAR 模型, PLS 回归主要适用于建立多自变量对多因变量线性回归模型, 可避免因变量多重相关性所造成的危害, 特别适用于样本数目小于变量数目情况下回归建模, 并且集中了回归建模、PCA 和典型相关分析的优点. 将数据进行标准化处理后, PLS 回归由 Simca-p10.0 软件完成.

2 结果与讨论

2.1 血管紧张素转化酶抑制剂 QSAR 研究

58个二肽血管紧张素转化酶(ACE)抑制剂是一个经典抗高血压肽类物质样本集[8], 其活性用 $\log(1/IC_{50})$表征. 首先用 SZOTT 描述子表征 58 个二肽结构, 每个氨基酸用 13 个 SZOTT 描述子表征, 二个氨基酸共有 26 个 SZOTT 描述子. 用 GA-PLS 挑选变量, 参数设置如下: 初始群体大小 200, 最大遗传代数 200, 收敛标准 80%, 交叉频率 50%, 变异概率 0.5%. 从 10 个训练模型中, 确定一个相对最优包含 14 个变量的模型, 14 个变量为: v_1, v_2, v_3, v_4, v_6, v_8, v_{10}, v_{14}, v_{15}, v_{16}, v_{19}, v_{20}, v_{22} 和 v_{25}, 用 PLS 建模得两个显著主成分, R_{CU}^2=0.894, R_{CV}^2=0.828, 估计值均方根误差(RMS)为 0.331. 将 ACE 抑制剂实验值与计算值进行回归(图 1), 表明模型预测值与实验值较接近. 为寻找模型特异点, 绘制ACE抑制剂PLS得分图(图 2), 可看出编号 1, 2, 4 和 11 的四个化合物位于 Hotelling T^2 椭圆置信区间外, 是模型异常点. 对样本 X 空间标准化模型距离(图3)进行分析发现仅有49号样本到模型 X 空间中心标准化距离大于 5%显著性检验临界值 1.474, 说明 13 个 SZOTT 描述子对此化合物重构质量较差. Hellberg[7]用 z scale, Cocchi[5]用 GRID 法, Collantes[8]用 ISA-ECI 指数, Zaliani[25]用 MS-WHIM scores, Liu[26]用 MHDV 描述子等对 ACE 抑制剂 QSAR 进行研究, 其QSAR模型结果见表1, 经比较得出, 我们所建模型结果不同程度优于已有报道, 且预测能力较强.

2.2 苦味二肽 QSAR 研究

苦味是人类重要味觉之一, 可保护人类免受潜在有毒植物和其它环境毒素影响[5], 我们取48个苦味活性二肽(BTT)[8]验证 SZOTT 描述子对其构效表征的有效性, 同时建立 BTT 的 QSAR 模型, 其活性用浓度负对数

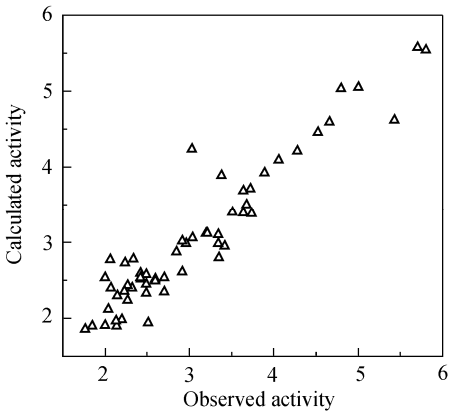

图 1 ACE 抑制剂 PLS 模型实验值与计算值回归

Figure 1 Regression between calculated and observed activities of ACE inhibitors

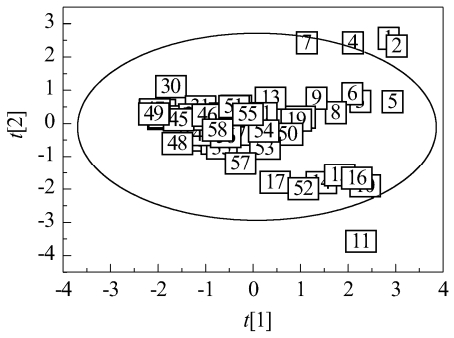

图 2　ACE 抑制剂的 PLS 模型得分图

Figure 2　PLS scores of ACE inhibitors

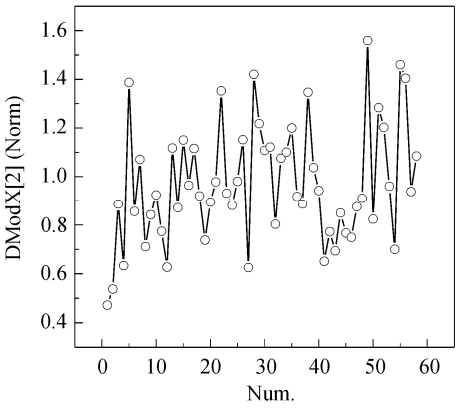

图 3　ACE 抑制剂的 PLS 模型标准化距离

Figure 3　Standardization distance to PLS model for ACE inhibitors in *X* space

p*T* 表示. 同样每个苦味二肽可用 26 个 SZOTT 描述子变量表征, 采用 GA-PLS 挑选变量, 参数设置为: 初始群体大小 200, 最大遗传代数 200, 收敛标准 80%, 交叉频率 50%, 变异概率 0.5%. 从 10 个被训练模型中, 选择一个相对最优模型, 所选 13 个变量为: v_1, v_4, v_8, v_{11}, v_{12}, v_{14}, v_{15}, v_{16}, v_{17}, v_{18}, v_{19}, v_{22}和v_{23}. 以PLS建模得两个显著的主成分, R_{CU}^2＝0.908, R_{CV}^2＝0.736. 两个主成分分别解释 *Y* 变量 82.7%和 8.1%方差, RMS 为 0.195. 图 4 所示BTT实验值与计算值回归图表明所建模型对其活性拟合效果较好. 关于BTT的PLS得分图(图5)表明所有样本都位于 Hotelling T^2 椭圆置信区间内, 无特异点. 对样本 *X* 空间标准化模型距离(图 6)分析发现所有样本到模型 *X* 空间中心标准化距离都小于 5%显著性检验临界值 1.509, 说明用 SZOTT 描述子能够较好表征 48 个 BTT 结构并拟合其活性. 将 SZOTT 描述子建模结果与相关文献报道 QSAR 建模结果列于表 1, 可看出我们所建模型结果显著优于其它同类建模方法, 而与我们用 MHDV 描述子基于主成分回归(PCR)所建模型结果[26]相当.

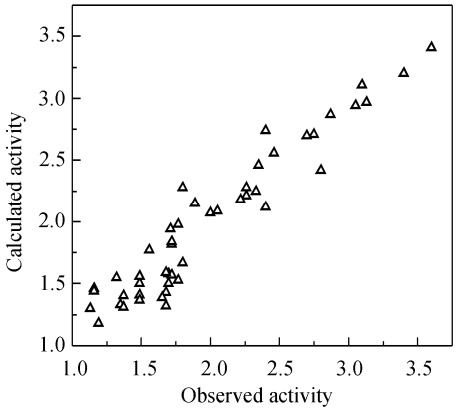

图 4　BTT 的 PLS 模型实验值与计算值回归

Figure 4　Regression between calculated and observed activities of BTT

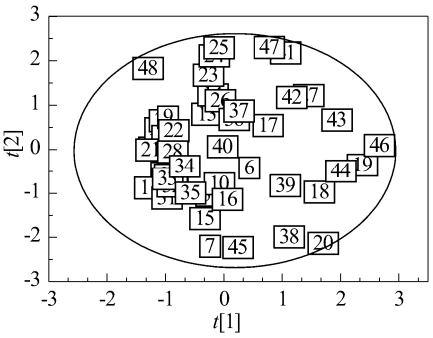

图 5　BTT 的 PLS 模型得分图

Figure 5　PLS scores for BTT

表 1　ACE 抑制剂及 BTT 的 QSAR 模型比较 [a]

Table 1　Comparison between QSAR models of ACE inhibitors and BTT[a]

No.	Descriptor	Model	A^a		R_{CU}^2		R_{CV}^2		RMS	
			ACE	BTT	ACE	BTT	ACE	BTT	ACE	BTT
1	*z* scale[7]	PLS	2	2	0.770	0.824	—	—	—	0.260
2	GRID (*t* scores)[5]	PLS	1	1	0.744	—	—	0.780	0.500	—
3	ISA-ECI[8]	PLS	2	2	0.700	0.847	—	—	—	—
4	MS-WHIM (rotameric)[25]	PLS	6	3	0.657	0.704	0.541	0.633	—	—
5	MS-WHIM (extended)[25]	PLS	2	3	0.708	0.754	0.637	0.710	0.540	0.320
6	MHDV[26]	PCR	19	10	0.878	0.919	0.753	0.864	0.350	0.180
7	SZOTT	PLS	2	2	0.894	0.908	0.828	0.736	0.331	0.195

[a] Number of principal component.

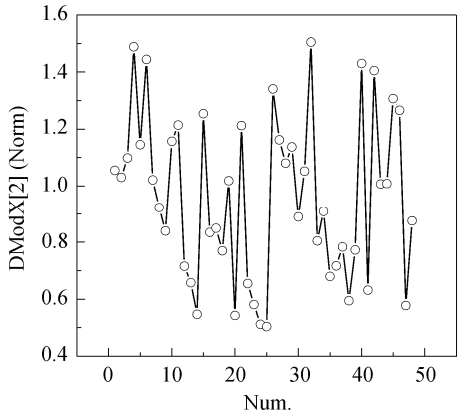

图 6 BTT 的 PLS 模型标准化距离

Figure 6 Standardization distance to PLS model for BTT in *X* space

此外, SZOTT 还可用于其它小分子肽的结构表征, 例如, 对 CTL 表位[27](T-cell epitopes)进行结构表征, 用 PLS 建模得 R_{CU}^2=0.778, R_{CV}^2=0.515, RMS=0.407, 效果良好. 对于上述三个肽体系的 QSAR 研究结果表明 SZOTT 对小分子肽结构表征具有普适性.

3 结论

多年来, 肽的 QSAR 研究得到了一定进步和发展, 应该说, 肽结构表征是其 QSAR 研究中的瓶颈问题, 纵观以前研究, 因其结构的复杂性, 大多数肽描述子集中于 2D 方面. 我们收集了天然氨基酸的 1369 个描述子经 PCA 得到一组新氨基酸描述子, 将其用于 ACE 抑制剂和 BTT 的 QSAR 研究, 以 PLS 回归建模, 建立简单而有效 QSAR 模型并均取得优于或与文献相当的结果. 研究表明, SZOTT 能够避免一些传统描述子所含信息量少、难于解释、不易操作等缺点, 其可望为肽 QSAR 研究提供一个强有力工具.

References

1 Neath, P. H. A. *J. Theor. Biol.* **1966**, *12*, 157.
2 Asao, M.; Iwamura, H.; Akamatsu, M.; Fujita, T. *J. Med. Chem.* **1987**, *30*, 1873.
3 Depriest, S. A.; Mayer, D.; Naylor, C. D.; Marshall, G. R. *J. Am. Chem. Soc.* **1993**, *115*, 5372.
4 Waller, C. L.; Oprea, T. L.; Giolitti, A.; Marshall, G. R. *J. Med. Chem.* **1993**, *36*, 4152.
5 Cocchi, M.; Johansson, E. *Quant. Struct.-Act. Relat.* **1993**, *12*, 1.
6 Hellberg, S.; Sjostrom, M.; Skagerberg, B.; Wold, S. *J. Med. Chem.* **1987**, *30*, 1126.
7 Hellberg, S.; Eriksson, L.; Jonsson, J.; Lindgren, F.; Sjostrom, M.; Skagerberg, B.; Wold, S.; Andrews, P. *Int. J. Pept. Protein Res.* **1991**, *37*, 414.
8 Collantes, E. R.; Dunn, W. J. *J. Med. Chem.* **1995**, *38*, 2705.
9 Mei, H.; Liao, Z. H.; Zhou, Y.; Li, S. Z. *Pept. Sci.* **2005**, *80*(6), 775.
10 Tdeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, Germany, **2000**.
11 Liu, S. S.; Cao, C. Z.; Li, Z. *J. Chem. Inf. Comput. Sci.* **1998**, *38*(3), 387.
12 Liu, S. S.; Liu, H. L.; Xia, Z. N.; Cao, C. Z.; Li, Z. *J. Chem. Inf. Comput. Sci.* **1999**, *39*(6), 951.
13 Liu, S. S.; Cai, S. X.; Cao, C. Z.; Li, Z. *J. Chem. Inf. Comput. Sci.* **2001**, *40*(6), 1337.
14 Rucker, G.; Rucker, C. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683.
15 Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 517.
16 Diudea, M. V.; Horvath, D.; Graovac, A. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 129.
17 Randic, M.; Kleiner, A. F.; DeAlba, L. M. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277.
18 Schuur, J. H.; Selzer, P.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334.
19 Gasteiger, J.; Sadowski, J.; Schuur, J. Selzer, P.; Steinhauer, L.; Steinhauer, V. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1030.
20 Todeschini, R.; Gramatica, P.; Provenzani, R.; Marengo, E. *Chemom. Intell. Lab. Syst.* **1995**, *27*, 221.
21 Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682.
22 Kim, D.; Lee, I.-B. *Chemom. Intell. Lab. Syst.* **2003**, *67*, 109.
23 Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109.
24 Hasegawa, K.; Miyashita, Y.; Funatsu, K. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 306.
25 Zaliani, A.; Gancia, E. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 525.
26 Liu, S.; Yin, C.; Cai, S.; Li, Z. *J. Chin. Chem. Soc.* **2001**, *48*, 253.
27 Doytchinova, I. A.; Flower, D. R. *J. Med. Chem.* **2001**, *44*, 3572.

(A0506153 ZHAO, C. H.)