Vol. 27 高等学校化学学报 No. 10
2006年10月 CHEMICAL JOURNAL OF CHINESE UNIVERSITIES 1900~1902

[研究简报]

氨基酸描述子 SZOTT 用于多肽定量序效建模研究

梁桂兆[1,2,3], 梅 虎[1,3], 周 原[1,2,3], 杨善彬[1,3], 吴世容[1,2], 李志良[1,2]
(1. 重庆大学化学化工学院药学系, 重庆 400044;
2. 湖南大学化学生物传感与计量学国家重点实验室, 湖南 410082;
3. 重庆大学生物力学与组织工程教育部重点实验室, 重庆 400030)

关键词 氨基酸描述子(SZOTT); 肽; 定量序效建模(QSAM); 偏最小二乘(PLS)
中图分类号 O629; R914 **文献标识码** A **文章编号** 0251-0790(2006)10-1900-03

肽在生命体系中至关重要[1], 故研究肽的定量序效建模(QSAM), 对了解其作用机理, 开发肽类新药, 理解蛋白结构与功能皆有具重要意义. 然而, 目前对肽类物质的 QSAM 研究报道不多, 一是因肽结构的相对复杂性, 二是则受到合成、分离和纯化等技术的制约. 序列表征是 QSAM 研究的关键内容之一.

本文在相关研究的基础上[2,3], 提出一新的氨基酸描述子 SZOTT, 该描述子所含信息量大, 且操作简便. 将其用于两类肽体系序列表征, 用偏最小二乘和正交信号纠正-偏最小二乘建模, 获得较好的建模结果.

1 原理与方法

1.1 *SZOTT 的提出及肽序列表征* 收集 20 种天然氨基酸的 1 369 个性质变量: 0D 描述子共 31 种[4]; 1D 描述子共 69 种[4,5]; 2D 描述子共 640 种[6~8]; 3D 描述子共 629 种[9~11].

为剔除原始数据矩阵中的噪声信息, 经主成分分析(PCA)[12]变换, 由其前 13 个主成分得分矩阵累计可解释原始变量数据矩阵 96.19% 的方差, 故可用此 13 个主成分得分替代原始变量矩阵. 为方便起见, 称 13 个得分矢量为 SZOTT. 这样, 每个肽的结构可根据其氨基酸残基顺序用 13 个 SZOTT 得分矢量表达. 不同长度的肽具有不同的变量个数, 故一个具有 n 个氨基酸残基的肽, 其一级序列结构可用 $13\times n$ 个 SZOTT 变量表征. 但对含有不同氨基酸残基数目的肽表征将得到不同数目的变量, 为使各肽的变量个数一致, 用自交叉协方差(ACCs)[13]处理, 使各肽的变量数目皆为 $13^2\times l$(l 为步长)个.

PCA 由 Matlab 7.0 软件完成. ACCs 由 C 语言编写实现.

1.2 *PLS 建模* PLS[13]主要适用于多自变量对多因变量的线性回归建模, 具有很多普通多元线性回归方法所没有的优点, 如可避免因变量多重相关性造成的危害, 且特别适合于样本数目小于变量数目的建模, 另外, 并具有集成回归建模、PCA 和典型相关分析的优点.

PLS 由 Simca-p10.0 软件完成.

1.3 *模型验证* 对模型的预测能力的评价通常采用留一法交互验证(LOOCV), 用统计量 Q^2 来表示:

$$Q^2 = 1\text{-PRESS/SSY}$$

式中, Q^2 为 LOOCV 的复相关系数(R^2); PRESS 指 LOOCV 预测残差平方和; SSY 指 Y 值与其均值之差平方和. 通常, Q^2 越大, 模型的预测能力越强.

收稿日期: 2005-08-15.
基金项目: 国家"春晖计划"教育部启动基金(批准号: 99-8-7)、霍英东基金(批准号: 98-7-6)、湖南大学化学生物传感与计量学国家重点实验室项目(批准号: 2005-12)及重庆市应用基础基金(批准号: 01-3-6)资助.
联系人简介: 李志良(1962年出生), 男, 博士, 教授, 博士生导师, 从事药物设计和药物生物信息学研究.
E-mail: zlli2662@163.com

2　结果与讨论

2.1　*促凝血酶原激酶抑制剂的* QSAM *研究*　促凝血酶原激酶抑制剂样本[14]为 20 个具有 6 ~ 12 个氨基酸残基数目的多肽，其活性大小由 50% 抑制浓度即 $IC_{50}/(\mu mol \cdot L^{-1})$ 表示. 用 SZOTT 表征，然后经 ACCs 处理后($l=5$)，每个样本由 845 个变量表征.

用 PLS 建 QSAM，得 2 个显著主成分，其累计解释了 Y 变量 98.9% 的方差，LOOCV 累计解释了 Y 变量 74.8% 的方差. 结果表明，模型的拟合能力较高，但其预测能力相对较低.

为提高模型的预测能力，对原始 X 变量矩阵进行正交信号纠正(OSC)[13]处理，将与 Y 变量正交的信息滤除，再用 PLS 模建，结果得 3 个显著主成分，R^2 为 0.994，Q^2 为 0.936. 可见，经 OSC 处理，模型的拟合能力(图 1)与预测能力皆得到显著提高. Anderson 等[14]对该样本集经 PLS 建模得 $R^2=0.886$，$Q^2=0.490$，经 OSC-PLS 建模得 $R^2=0.881$，$Q^2=0.706$，经比较，无论应用 PLS，还是 OCS-PLS，模型的拟合能力和预测能力明显高于 Anderson 等[14]所得结果.

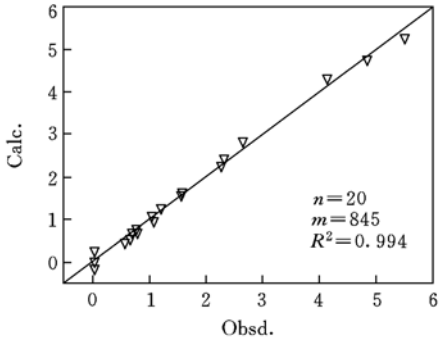

Fig. 1　Regression between observed and calculated abilities of thromboplastin inhibitors by OSC-PLS

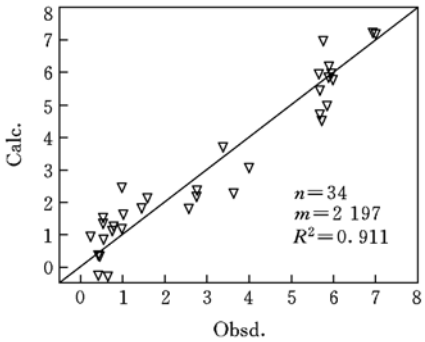

Fig. 2　Regresion between observed and calculated of bactericidal peptides by OSC-PLS

2.2　*抗菌肽的* QSAM *研究*　研究抗菌肽类似物的序效关系为了解其作用机理，进一步设计更有效的抗菌药物提供了可能. 34 个抗菌肽的活性值[15]用 2 h 内杀死金黄色葡萄球菌的数量的对数值表示. 经 SZOTT 描述子表征，然后用 ACCs($l=13$)处理，得到 2 197 个变量表征每个肽. 用 PLS 建立 QSAM，得到 1 个显著主成分，解释了 Y 变量 61.9% 的方差，LOOCV 累计解释了 Y 变量 40.6% 的方差.

从建模结果可看出，模型的拟合能力及预测能力均较低，可能是因 2 197 个变量中存在较多的信息噪声所致. 为提高模型的拟合及预测能力，对原始 X 变量矩阵进行 OSC 处理，再经 PLS 模建，得到 2 个显著主成分，R^2 上升到 0.911，Q^2 增至 0.503. 由此可见，经 OSC 后，模型的拟合能力(图 2)和预测能力都得到较显著提高. 对肽的 QSAM 研究来说，制约其发展的仍然是序列结构表征问题. 其各种方法和技术还处于不断尝试和发展之中，因为肽的结构相对较复杂，故目前有关研究主要集中在 2D 层次.

收集 20 种天然氨基酸的 0D ~ 3D 信息的 1 369 个变量经 PCA 得到一新氨基酸描述子-SZOTT，将其用于两类多肽序列表征，分别经 PLS 和 OSC-PLS 建模，结果表明，SZOTT 描述子具有相对简单而方便，不需实验数据及包含信息量大等优点，有望进一步得到推广和应用.

感谢杨　力、蔡绍皙、李声时、黄　莺、周　鹏及李根容等提供的帮助.

参　考　文　献

[1] Ding J. L., Ho B.. Drug Development Research[J], 2004, **62**: 317—335
[2] Hellberg S., Sjöström M., Skagerberg B. *et al*.. J. Med. Chem.[J], 1987, **30**: 1126—1135
[3] Mei H., Zhou Y. Li S. Z. *et al*.. Pept. Sci.[J], 2005, **80**(6): 775—786
[4] Tdeschini R., Consonni V.. Handbook of Molecular Descriptors[M], Weinheim(Germany): Wiley-VCH, 2000
[5] Ertl P., Rohde B., Selzer P.. J. Med. Chem.[J], 2000, **43**: 3714—3717
[6] Liu S. S., Cai S. X., Li Z. *et al*.. J. Chem. Inf. Comput. Sci.[J], 2001, **40**(6): 1337—1348

[7] Rucker G., Rucker C. J.. Chem. Inf. Comput. Sci. [J], 1993, **33**: 683—695
[8] Balaban A. T., Ciubotariu D., Medeleanu M. J.. Chem. Inf. Comput. Sci. [J], 1991, **31**: 517—523
[9] Randic M., Kleiner A. F., DeAlba L. M.. J. Chem. Inf. Comput. Sci. [J], 1994, **34**: 277—286
[10] Schuur J. H., Selzer P., Gasteiger J. J.. Chem. Inf. Comput. Sci. [J], 1996, **36**: 334—344
[11] Consonni V., Todeschini R., Pavan M. J.. Chem. Inf. Comput. Sci. [J], 2002, **42**: 682—692
[12] Kim D., Lee I. B.. Chemon. Intell. Lab. Syst. [J], 2003, **67**: 109—123
[13] Wold S., Trygg J., Berglund A. *et al*.. Chemom. Intell. Lab. Syst. [J], 2001, **58**: 131—150
[14] Andersson P. M., Sjöstrom M., Lundstedt T.. Chemom. Intell. Lab. Syst. [J], 1998, **42**: 41—50
[15] Patel S., Stott I. P., Bhakko M. *et al*.. J. Comput. Aid. Mol. Des. [J], 1998, **12**: 543—556

Using SZOTT Descriptors for the Development of QSAMs of Peptides

LIANG Gui-Zhao[1,2,3], MEI Hu[1,3], ZHOU Yuan[1,2,3], YANG Shan-Bin[1,3], WU Shi-Rong[1,2], LI Zhi-Liang[1,2*]

(1. *Department of pharmacy, College of Chemistry and Chemical Engineering, Chongqing University, Chongqing* 400044, *China*;
2. *State Key Laboratory for Chemobiosensors and Chemobiometrics under MOST, Hunan University, Changsha* 410012, *China*;
3. *MOE Key Laboratory of Biomechanics and Tissue Engineering, Chongqing University, Chongqing* 400030, *China*)

Abstract A new descriptor, namely scores vector of zero dimension, one dimension, two dimension and three dimension (SZOTT), was derived from principle components analysis of a matrix of 1 369 structural variables including 0D, 1D, 2D and 3D information for 20 coded amino acids. SZOTT scales were then employed to express structures of 20 thromboplastin inhibitors and 34 bactericidal peptides. The correlation coefficients of both whole calibration ($R^2 = R^2_{cu}$) and of cross validation ($Q^2 = R^2_{cv}$) for the multiple-variable models by classical partial least squares (PLS) and orthogonal signal correction-partial least squares (OSC-PLS) of 20 thromboplastin inhibitors were 0.989 and 0.748, 0.994 and 0.936, respectively. R^2 and Q^2 for the models by PLS and OSC-PLS of 34 bactericidal peptides were 0.619 and 0.406, 0.910 and 0.503, respectively. Satisfactory results obtained showed that structural information related to biological activity in both data sets could be described by SZOTT which included plentiful information related to biological activity, and which was conveniently operated and easy interpreted., also predictive capability of models were relative robust. There is a high prospect for SZOTT wide applications on quantitative sequence-activity modeling (QSAM) of peptides.

Keywords SZOTT descriptor of amino acid; Peptide; Quantitative sequence-activity modeling (QSAM); Partial least squares (PLS)

(Ed.: H, J, Z)