

后叶催产素的定量构效关系研究

梅虎^{1,2}, 杨力^{1,2}, 舒茂^{1,2}, 刘丽², 李志良^{1,3}

(1. 重庆大学生物医学工程重庆市重点实验室, 2. 生物工程学院, 3. 化学化工学院, 重庆 400044)

摘要 基于氨基酸物化性质的描述子矢量 VHSE, 对 21 个后叶催产素类似物进行结构表征. 经逐步回归与偏最小二乘相结合的变量筛选技术, 根据模型的外部预测结果, 筛选得到一个最优的 9 变量组合. 应用该变量组合对 21 个后叶催产素类似物的促宫缩活性进行偏最小二乘建模, 模型复相关系数 R^2 为 92.6%, 留一法和留组法交叉验证 Q^2 分别为 78.3% 和 79.4%. 结果表明, 后叶催产素的促宫缩活性主要与第 3 号氨基酸残基的疏水性、立体结构和电性性质以及第 8 号氨基酸残基的电性特征密切相关.

关键词 VHSE; 后叶催产素; 肽; 定量构效关系; 偏最小二乘

中图分类号 O641 **文献标识码** A **文章编号** 0251-0790(2007)05-0964-04

后叶催产素 (Oxytocin, OT) 是脑下垂体后叶分泌的一种多肽荷尔蒙, 其主要作用是在分娩过程中促进子宫平滑肌收缩, 同时在哺乳期具有加速乳汁分泌等功能^[1~7]. 此外, 催产素还可在中枢神经系统中以神经递质的形式参与人的多种行为以及生理反应的调节^[8]. 研究表明: 催产素与早产的发生有着十分重要的关系^[9~11]. Fuchs^[12] 研究发现: 早产患者比同龄妇女具有较高的催产素敏感性和受体密度. 因此, 用催产素受体拮抗剂阻断催产素与其受体的亲和作用, 是预防早产的一种有效途径.

本文从催产素结构与促宫缩活性的定量构效关系 (Quantitative structure-activity relationships, QSAR) 研究入手, 通过偏最小二乘 (Partial least squares, PLS) 法建立结构-活性的定量模型, 从模型出发探讨和解释催产素与其受体结合的关键理化特征. 对于催产素作用机理的解释以及催产素拮抗剂的药物设计均具有重要的理论意义和实际应用价值.

1 原理与方法

1.1 结构表征

选用的 21 个后叶催产素类似物取自文献[13], 其促宫缩活性列于表 1. 采用 VHSE (Principle component score vector of hydrophilicity, steric, and electronic variables) 进行结构表征. VHSE 是一组基于氨

Table 1 Oxytocin sequences with observed and calculated activities

No.	Peptide	Obsd.	Calcd.	Res.	No.	Peptide	Obsd.	Calcd.	Res.
1	YYL	2.00	2.08	-0.08	12	YFH	3.18	4.03	-0.85
2	FFL	3.52	3.11	0.41	13	YIL	5.65	5.19	0.46
3	YWL	1.60	1.65	-0.05	14	FIL	4.50	3.96	0.54
4	YLL	4.65	4.34	0.31	15	YFL	4.30	4.34	-0.04
5	YVL	4.77	5.69	-0.92	16	YII	5.46	5.22	0.24
6	YIV	5.30	5.16	0.14	17	YFK	3.70	3.76	-0.06
7	FYK	1.00	0.26	0.74	18	YYK	1.00	1.50	-0.50
8	SIK	1.00	1.57	-0.57	19	FFK	2.48	2.52	-0.04
9	YWK	1.00	1.06	-0.06	20	YIK	4.89	4.61	0.28
10	FIK	3.00	3.37	-0.37	21	YFR	4.30	3.95	0.35
11	YIR	4.88	4.80	0.08					

收稿日期: 2006-01-12.

基金项目: 国家“八六三”计划(批准号: 2006AA02Z312)、教育部和国家外国专家局“高等学校学科创新引智计划”(111计划)资助.

联系人简介: 李志良(1962年出生), 男, 教授, 博士生导师, 主要从事药物化学及生物信息学研究. E-mail: zlli2662@163.com

基酸物理化学性质描述子的矢量, 来源于 20 种天然氨基酸的 50 种物理化学性质, 并按照疏水性、立体结构和电性特征分类后分别进行主成分特征提取而产生. 其中 $VHSE_1$ 和 $VHSE_2$ 代表氨基酸的疏水性特征; $VHSE_3$ 和 $VHSE_4$ 代表氨基酸的立体空间特征; $VHSE_5 \sim VHSE_8$ 则代表氨基酸的电性特征^[14]. 20 种天然氨基酸的 $VHSE$ 值列于表 2. 对于 21 个催产素类似物, 其仅在 2, 3 和 8 号氨基酸残基位发生了变化, 而在其它位置则保持不变. 因此采用 $VHSE$ 描述子对这 3 个氨基酸位进行结构表征, 共产生 24 个描述子变量, 依次用 $v_1 \sim v_{24}$ 表示.

Table 2 VHSE scales for 20 coded amino acids

Amino acid	$VHSE_1$	$VHSE_2$	$VHSE_3$	$VHSE_4$	$VHSE_5$	$VHSE_6$	$VHSE_7$	$VHSE_8$
Ala A	0.15	-1.11	-1.35	-0.92	0.02	-0.91	0.36	-0.48
Arg R	-1.47	1.45	1.24	1.27	1.55	1.47	1.30	0.83
Asn N	-0.99	0.00	-0.37	0.69	-0.55	0.85	0.73	-0.80
Asp D	-1.15	0.67	-0.41	-0.01	-2.68	1.31	0.03	0.56
Cys C	0.18	-1.67	-0.46	-0.21	0.00	1.20	-1.61	-0.19
Gln Q	-0.96	0.12	0.18	0.16	0.09	0.42	-0.20	-0.41
Glu E	-1.18	0.40	0.10	0.36	-2.16	-0.17	0.91	0.02
Gly G	-0.20	-1.53	-2.63	2.28	-0.53	-1.18	2.01	-1.34
His H	-0.43	-0.25	0.37	0.19	0.51	1.28	0.93	0.65
Ile I	1.27	-0.14	0.30	-1.80	0.30	-1.61	-0.16	-0.13
Leu L	1.36	0.07	0.26	-0.80	0.22	-1.37	0.08	-0.62
Lys K	-1.17	0.70	0.70	0.80	1.64	0.67	1.63	0.13
Met M	1.01	-0.53	0.43	0.00	0.23	0.10	-0.86	-0.68
Phe F	1.52	0.61	0.96	-0.16	0.25	0.28	-1.33	-0.20
Pro P	0.22	-0.17	-0.50	0.05	-0.01	-1.34	-0.19	3.56
Ser S	-0.67	-0.86	-1.07	-0.41	-0.32	0.27	-0.64	0.11
Thr T	-0.34	-0.51	-0.55	-1.06	-0.06	-0.01	-0.79	0.39
Trp W	1.50	2.06	1.79	0.75	0.75	-0.13	-1.01	-0.85
Tyr Y	0.61	1.60	1.17	0.73	0.53	0.25	-0.96	-0.52
Val V	0.76	-0.92	-0.17	-1.91	0.22	-1.40	-0.24	-0.03

1.2 SMR-PLS 变量筛选

为了确定最优的变量组合, 确保在其基础上建立的 QSAR 模型具有较强的外部预测能力, 本文采用逐步回归 (Stepwise multiple regression, SMR) 与 PLS 相结合的方法, 即 SMR-PLS 进行变量筛选. 具体方法如下: (1) 采用 SMR 对 24 个描述子变量进行筛选, 得到候选变量组合; (2) 训练集与预测集的划分: 针对每个候选变量组合和活性数据产生的 $21 \times n$ (n 为筛选变量数) 矩阵, 采用 D 最优设计 (D -optimal) 的方法从 21 个样本中选择 12 个样本作为训练集, 余下的 9 个样本为预测集. D -optimal 是一种行列式最优算法, 其目的是寻求数据矩阵 $X^T X$ 行列式值 $D = \det(X^T X)$ 最大化, 确保训练集样本最大程度地覆盖整个样本空间; (3) 应用上述候选变量组合分别对训练集和预测集样本进行 PLS 建模和外部验证, 通过确定最优的 PLS 模型来选择最优的变量组合. 最优模型的选择按下式标准来进行^[15,16]:

$$Q^2 > 0.5, \quad R^2 > 0.6, \quad \frac{(R^2 - R_0^2)}{R^2} < 0.1 \quad \text{or} \quad \frac{(R^2 - R_0'^2)}{R^2} < 0.1,$$

$$0.85 \leq k \leq 1.15 \quad \text{or} \quad 0.85 \leq k' \leq 1.15$$

式中, Q^2 为训练集样本的交互检验值; R^2 为预测集样本实验值对预测值回归方程的决定系数, 为了明确起见用 R_{pred}^2 表示; R_0^2 和 $R_0'^2$ 分别为预测集样本实验值对预测值通过原点回归方程的决定系数 ($R_0^2 = \text{预测值} - \text{实验值}$; $R_0'^2 = \text{实验值} - \text{预测值}$); k 和 k' 分别为相应回归方程的斜率. 在文中非特殊说明, 均采用留一法 (Leave-one-out, LOO) 交互验证.

2 结果与讨论

2.1 变量筛选

通过 SMR-PLS 变量筛选及候选模型的多重验证, 得到一个最优的 9 变量组合: $v_5, v_8, v_9, v_{10}, v_{11}$,

v_{16} , v_{18} , v_{21} 和 v_{23} . 模型的内部和外部的验证结果(表3)显示:该模型不仅具有较好的交互验证结果,同时也具备较高的外部预测能力.

Table 3 Results of internal and external validations of the optimal PLS model

A	R^2	Q^2	R_{ext}^2	Q_{ext}^2	R_0^2	$R_0'^2$	k	k'
2	0.909	0.667	0.876	0.968	0.945	0.923	1.139	0.873

2.2 偏最小二乘模型分析

采用上述最优的9变量组合,对21个后叶催产素类似物进行PLS建模和定量构效关系研究. PLS建模产生了两个显著的主成分,复相关系数 R^2 为92.6%,其中第一和第二个主成分分别解释原始数据83.4%和9.2%的变异,可以看出第一个主成分已经解释了原始数据绝大多数的信息. 模型LOO交互验证 Q^2 为78.3%;当采用更为严格的留组法交互验证(n -fold cross-validation, n 设定为5)时,相应的 Q^2 为79.4%. 以上结果说明,模型不仅具有较好的拟合能力,同时也具有较好的稳定性和预测能力. 此外,对样本标准化残差累积概率研究发现:残差基本呈正态分布,各样本的标准化残差值均小于2,无特异点出现. 样本活性的观测值和计算值参见表1.

图1为PLS主成分得分图,可以看出,所有样本均匀地分布于Hotelling T^2 椭圆置信区间内,无特异点出现. 除了第8号样本外,其它样本到模型 X 空间的规格化距离都位于95%置信区间内.

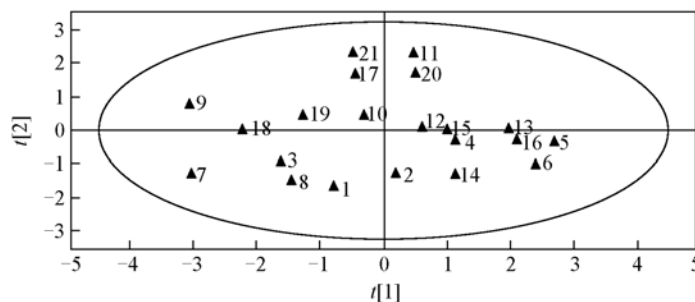


Fig. 1 PLS scores of the oxytocin analogues

从PLS载荷(图2)中可以看出,各描述子变量与活性 Y 之间的相关性. 在第一个主成分中, v_5 , v_9 和 v_{16} 与 Y 呈正相关性,分别对应着2号氨基酸残基位的 $VHSE_5$ (电性)和3号氨基酸残基位的 $VHSE_1$ (疏水性)和 $VHSE_8$ (电性),其中以3号氨基酸残基位的电性性质(v_{16})与 Y 的相关性最强;而 v_8 , v_{10} , v_{11} , v_{18} , v_{21} 和 v_{23} 与 Y 呈负的相关性,其中又以 v_{10} 和 v_{11} 与 Y 的相关性最强,分别对应着3号氨基酸残基

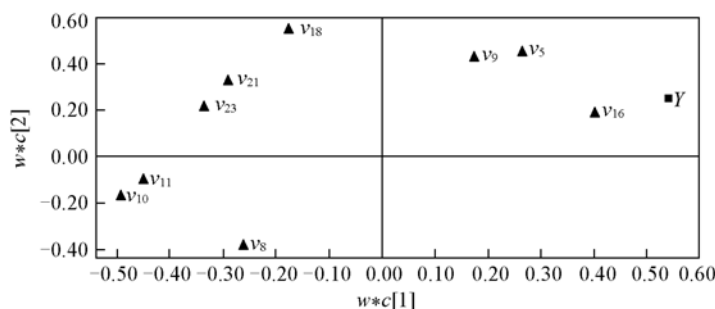


Fig. 2 PLS loading plot of the oxytocin analogues

的 $VHSE_2$ (疏水性)和 $VHSE_3$ (立体). 对于第二个主成分, v_{18} 与 Y 的正相关性最大,其对应第8号氨基酸残基的疏水性质,而 v_8 与 Y 的负相关性最大,其对应第2号氨基酸残基的疏水性质.

变量投影重要性(Variable importance in the projection, VIP)是反映各变量对 Y 解释能力大小的一个指标(图3). 通常VIP值大于1

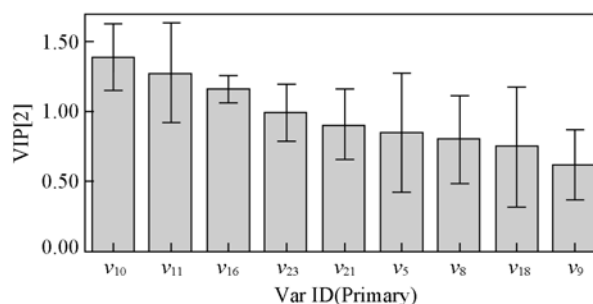


Fig. 3 VIP plot for the oxytocin analogues

的变量与 Y 具有较大的相关性, 相应的解释能力也越大. 对于该体系, v_{10} , v_{11} , v_{23} 和 v_{16} 的 VIP 值均大于 1, 因此, 这 4 个变量对 Y 的解释能力相对较大. 其中 v_{10} , v_{11} 和 v_{16} 分别表征第 3 号氨基酸残基的疏水性、立体结构和电性特征, 这说明后叶催产素的活性主要与 3 号氨基酸残基的性质密切相关; 而 v_{23} 则对应着第 8 号氨基酸残基的电性特征.

应用 Y 随机排序测试(Random permutation test, RPT)对模型的有效性作了进一步验证. 50 次随机排序结果如下: $R^2 = (0.0, 0.0595)$, $Q^2 = (0.0, -0.385)$, 可以判断该模型无明显过拟合现象发生.

参 考 文 献

- [1] Antonijevic I. A. , Leng G. , Luckman S. M. , *et al.* . Endocrinology[J], 1995, **136**: 154—163
- [2] Douglas A. J. , Dye S. , Leng G. , *et al.* . J. Neuroendocrinol[J], 1993, **5**: 307—314
- [3] Luckman S. M. , Antonijevic I. , Leng G. , *et al.* . J. Neuroendocrinol[J], 1993, **5**: 7—12
- [4] Russell J. A. , Leng G. , Douglas A. J. . Neuroendocrinol[J], 2003, **24**: 27—61
- [5] Crowley W. R. , Armstrong W. E. . Endocr. Rev. [J], 1992, **13**: 33—65
- [6] Insel T. R. , Gingrich B. S. , Young L. J. . Prog. Brain Res. [J], 2001, **133**: 59—66
- [7] Young W. S. I. , Shepard E. , Amico J. , *et al.* . J. Neuroendocrinol[J], 1996, **8**: 847—853
- [8] Argiolas A. , Gessa G. L. . Neuroscience & Biobehavioral Rev. [J], 1991, **15**(2): 217—231
- [9] Manning M. , Stoeva S. , Chenga L. L. , *et al.* . J. Peptide Sci. [J], 2001, **7**: 449—465
- [10] Fuchs A. R. , Fuchs F. , Husslein P. , *et al.* . Science[J], 1982, **215**: 1396—1398
- [11] Antonijevic I. A. , Douglas A. J. , Dye S. , *et al.* . J. Endocrinol. [J], 1995, **145**: 97—103
- [12] Fuchs A. R. . The Physiology and Biochemistry of the Uterus in Pregnancy and Labor[M], Florida: CRC Press, 1986: 125—131
- [13] Adenot M. , Sarrauste D. M. C. , Chavanieu A. , *et al.* . J. Mol. Graphics Mod. [J], 1999, **17**: 292—309
- [14] Mei H. , Liao Z. H. , Zhou Y. , *et al.* . Biopolymers(Pept. Sci.) [J], 2005, **80**(6): 775—786
- [15] Golbraikh A. , Tropsha A. . J. Mol. Graphics Mod. [J], 2002, **20**: 269—276
- [16] Tropsha A. , Gramatica P. , Gombar V. K. . QSAR Comb. Sci. [J], 2003, **22**: 69—77

Quantitative Structure-activity Relationships Studies on Oxytocin Analogues

MEI Hu^{1,2}, YANG Li^{1,2}, SHU Mao^{1,2}, LIU Li², LI Zhi-Liang^{1,3*}

(1. Key Laboratory of Biomedical Engineering of Chongqing, 2. College of Biological Engineering, 3. College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400044, China)

Abstract As a new set of amino acid descriptors, VHSE(principle component score vector of hydrophilicity, steric, and electronic variables) scales were applied to QSAR studies on 21 oxytocin analogues. Firstly, step-wise multiple regression combined with partial least squares was used to screen variables. According to the result of external validation, an optimal variable subset with 9 variables was obtained. Then, this optimal variable subset was employed to establish PLS model on all 21 samples with the results of the squared multiple correlation coefficient (R^2), leave-one-out cross-validation R^2 , and n -fold cross-validation ($n = 5$) R^2 were 92.6%, 78.3%, and 79.4%, respectively. The results show that oxytocic activity was closely related to the hydrophobic, steric, and electronic properties in positions 3 and electronic property in position 8 of oxytocin analogues.

Keywords VHSE; Oxytocin; Peptide; Quantitative structure-activity relationships; Partial least squares
(Ed. : H, J, Z)