[Article] www.whxb.pku.edu.cn

一种新多肽表征方法及支持向量机用于肽 HPLC 定量结构-保留建模预测

梁桂兆[1,2] 李志良[1,2*] 周 原[1,2] 何 留[1] 周 鹏[1]
([1] 重庆大学化学化工学院; [2] 重庆大学生物力学与组织工程教育部重点实验室, 重庆 400030)

摘要 从 20 种天然氨基酸的 1369 种性质参数经主成分分析得出一种新多肽序列表征方法——SZOTT. 将其用于 71 个不同长度肽序列表征, 以偏最小二乘(PLS)和支持向量机(SVM)建立定量结构-保留模型(QSRM). 研究表明, SZOTT 能够较好表征 71 个肽序列特征, 其含信息量大且易操作, 与 PLS 相比, SVM 对 lg*k* 建模预测表现出较强的拟合能力和良好外部预测能力, SZOTT 表征方法和 SVM 建模可进一步用于肽 HPLC 保留行为研究.

关键词: 肽, SVM, QSRM, SZOTT
中图分类号: O641

A New Peptide Sequences Representation Technique and Support Vector Machine for Quantitative Structure-Retention Modeling of Peptides in HPLC

LIANG, Gui-Zhao[1,2] LI, Zhi-Liang[1,2*] ZHOU, Yuan[1,2] He, Liu[1] ZHOU, Peng[1]
([1]*College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400030, P. R. China*;
[2]*Key Laboratory of Biomechanics and Tissue Engineering, MOE, Chongqing University, Chongqing 400030, P, R. China*)

Abstract A new representation technique for peptide sequences, namely SZOTT (scores vector of zero dimension, one dimension, two dimension, and three dimension), was derived from 1369 parameters of 20 coded amino acids using principle components analysis (PCA). It was then employed to express 71 peptide sequences with different lengths. Quantitative structure-retention modelings (QSRMs) were constructed by support vector machine (SVM) and partial least square (PLS). The results indicated that 71 peptide sequences could be preferably represented by SZOTT with many advantages, such as plentiful structural information and easy manipulation. Also simulative power for interior samples and predictive power for exterior samples by SVM were superior to those from PLS. SZOTT and SVM can be applied to develop QSRMs.

Keywords: Peptides, SVM, QSRM, SZOTT

肽在生物体中承担着多种功能, 故其分离、纯化和分析对药物开发等具重要意义. 在众多分离与分析方法中, 高压液相色谱(HPLC)技术具快捷、高效和高选择性等优点. 尽管其在肽和蛋白质分离与分析技术中得到广泛应用, 但对某一特定肽色谱条件挑选仍然费时和费力, 在此背景下, 定量结构-特性关系 (QSPR) 为研究分子色谱行为提供了思路, 用 QSPR 已成功预测许多分子特性, 但较少将其用于肽色谱保留行为定量研究中, 可能是因为肽结构太复杂而难于进行实验条件优化. 肽序列表征对

Received: January 10, 2006; Revised: March 21, 2006. *Correspondent, E-mail: zlli2662@163.com; Tel: 023-65106677.
重庆市应用基础基金[01-3-6]及湖南大学化学生物传感与计量学国家重点实验室基金(2005-12)资助项目

QSPR 至关重要. 目前对肽序列表征多集中于 2D 方面. 在相关研究[1-2]基础上, 在 20 种天然氨基酸的0D, 1D, 2D 及 3D 结构信息中共有 1369 种性质, 经主成分分析 (PCA) 得出一种新多肽序列表征方法——SZOTT, 其具有操作简便及所含信息量大等优点. 建模方法是 QSPR 研究的重要一环, 现已发展了较多建模方法. 自 20 世纪 90 年代, 一种新机器学习方法——SVM[3]引起研究者广泛关注和浓厚兴趣, 其能较好解决小样本、非线性、高维数和局部最小等问题. 本文将 SZOTT 用于 71 个不同长度肽序列表征, 以 SVM 建立定量结构-保留模型(QSRM), 获得较好建模和预测效果.

1 原理与方法

1.1 SZOTT 的提出

氨基酸构成及其物理化学性质与拓扑结构对肽功能特性至关重要, 故收集 20 种天然氨基酸的 1369 种性质参数: 0D 参数包括分子构成[4]; 1D 参数包括官能团数目、原子中心碎片和分子特性等[4-5]; 2D参数包括由本室提出的分子电性作用矢量、分子电距矢量和全息分子电距矢量、拓扑、拓扑电荷指数、运转和路径数目、边缘邻接指数、Burden特征值、自相关、连接性指数、信息指数、特征值指数等[6-10]; 3D 参数包括 Randic 分子剖面, 几何、RDF、MoRSE、WHIMs 和 GETAWAY 等变量[11-15].

因各变量之间可能高度相关, 故用 PCA[16]压缩参数数量, 得到原始变量矩阵(20×1369)得分矩阵的前 13 个主成分累计解释其 96.19%方差, 故可用此 13 个主成分得分矩阵(20×13)替代原始变量矩阵. 为方便, 称此 13 个得分矢量为 SZOTT(scores vector of zero dimension, one dimension, two dimension and three dimension)(表1). 每个肽可据其氨基酸顺序用 13×n个(n为氨基酸残基数目)SZOTT变量串联表征.

1.2 数据集选取与结构表征

从文献[17]选取 71 个肽和相应保留指数(lg$k$)(附表 1, 可从 www.whxb.pku.edu.cn 免费下载). 用 SZOTT 表征含 n 个氨基酸残基的肽可产生 13×n 个变量, 因所选 71 个肽含 2~20 个氨基酸残基, 经表征将得不同数目变量, 故用自交叉协方差(ACCs)[18]处理, 使各肽表征变量数目一致, ACCs 考虑肽链不同位点氨基酸参数之间交互效应, 可较大程度降低数据信息损失. 不同长度肽经 ACCs 处理后变量数目为 13^2×l 个(l 为步长, 其小于最短肽链长度).

1.3 变量挑选与模型验证

表 1 经 PLA 处理的 20 种天然氨基酸的 13 个 SZOTT(t_i)得分矢量

Table 1 13 SZOTT(t_i)score vectors by PCA for 20 coded amino acids

AA	abbr.	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}
Ala	A	−34.27	8.26	−5.60	5.73	−3.88	−3.64	2.37	−0.15	0.86	−0.11	4.20	−7.54	1.57
Glu	E	1.85	−4.28	13.13	−6.33	1.05	0.71	3.46	7.98	−2.27	−6.01	−0.93	−3.34	−10.24
Leu	L	0.71	−17.60	−11.02	2.62	1.87	−2.50	9.24	3.60	−5.12	1.19	−5.92	5.90	5.35
Ser	S	−26.53	6.57	0.83	−2.83	−6.39	−1.81	−2.80	−3.60	−5.04	−0.55	8.87	0.23	2.64
Arg	R	24.44	−15.80	16.78	12.97	−7.78	−4.93	−4.85	−11.69	1.50	9.76	−0.85	−2.91	−3.22
Gln	Q	3.87	−5.66	12.25	−3.74	0.28	0.60	1.80	6.48	−1.78	−3.44	0.36	−4.60	−2.27
Lys	K	10.56	−21.76	7.85	16.16	0.78	3.90	−5.22	2.97	−6.11	−10.33	−0.29	4.26	6.61
Thr	T	−15.66	−1.88	−3.96	−7.52	−6.86	0.22	−1.35	−2.51	−6.64	−1.14	6.36	1.96	1.58
Asn	N	−7.95	1.24	11.00	−11.27	−3.26	−0.47	3.76	1.71	0.63	5.10	−4.57	2.41	6.78
Gly	G	−47.81	22.58	4.10	21.66	1.12	−1.03	6.50	4.17	3.22	1.40	−2.53	5.15	−2.76
Met	M	3.84	−4.77	0.77	−1.61	24.62	−9.02	−6.52	6.07	1.41	6.48	6.71	1.14	0.90
Trp	W	62.17	22.91	−10.09	2.64	−10.91	−9.38	−7.46	9.89	1.43	−0.74	−1.38	0.23	1.29
Asp	D	−10.09	4.08	12.01	−14.08	−3.09	0.37	3.96	2.38	−1.82	4.15	−4.79	3.75	2.81
His	H	15.31	7.02	8.35	−4.85	−0.62	10.05	−1.66	−4.14	18.64	−4.70	5.93	0.86	0.91
Phe	F	28.81	8.70	−8.05	2.32	10.52	7.74	9.17	−8.69	−1.27	−2.45	−1.62	−0.54	1.71
Tyr	Y	38.59	13.52	−1.55	0.08	4.21	2.50	9.35	−6.37	−10.12	1.66	1.83	0.81	−2.46
Cys	C	−23.81	8.84	−3.15	−7.02	6.56	−9.29	−11.98	−11.58	0.24	−7.63	−9.35	−0.61	−1.44
Ile	I	−0.20	−20.36	−16.33	−3.08	−4.25	−2.22	2.57	−0.31	3.81	1.15	−0.04	11.06	−6.06
Pro	P	−12.00	1.96	−10.48	0.81	0.68	20.48	−13.56	4.42	−3.21	6.95	−3.78	−2.33	−1.82
Val	V	−11.85	−13.58	−16.85	−2.68	−4.65	−2.29	3.23	−0.64	1.40	−0.74	1.77	−4.08	−1.86

自变量中可能含与活性相关性较小的信息, 在建模前用遗传算法(GA)-PLS 剔除. GA 模拟自然界中"适者生存, 不适者被淘汰"原理, 在 GA-PLS 中, 染色体对应一组变量, 其物种适应性由 PLS 模型控制, 包括五个步骤[19]: (1)基因编码与群体初始化; (2)基因评价; (3)基因选择; (4)遗传操作, 包括杂交、变异和复制; (5)反复执行(2)~(4)步骤, 直到达到终止条件, 选择最佳个体作为结果. 模型预测能力由如下适应度函数评价: Q^2_{cv}=1−PRESS/SSY, 式中, Q^2_{cv} 为留一法(LOO)-交互验证(CV)复相关系数R^2; PRESS为预测残差平方和; SSY为Y值(活性)离差平方和. 对内部样本预测能力验证采取 LOO-CV, 用 Q^2_{cv} 表示. 一般讲, Q^2_{cv}越大, 模型预测能力越强. 但研究[20]表明: 仅以内部验证得出模型具较高预测能力可能对外部样本验证失败, 模型外部预测能力只能经外部验证得知, 可用 Q^2_{ext} 评价, 用 D-最优算法划分样本, 外部验证结果常常较优. 该法是将 Fisher 信息矩阵$|X'X|$行列式最大化的一种优化实验设计方法. X 指自变量, 或自变量与因变量矩阵, 使$|X'X|$占据整个数据点空间, 这些点构成训练集, 其它视为测试集, 此算法最大保证数据集空间与结果多变性平衡. 用 D-最优算法将 71 个样本划分为 50 个训练集样本与 21 个测试集样本.

1.4 PLS 建模

PLS[21]主要适于建立多自变量对多因变量线性回归, 具较多优点, 如可避免变量相关性危害等, 特别适于样本数目小于变量数目情况下回归, 集回归建模、PCA 和典型相关等.

1.5 SVM 建模

SVM 解决线性回归问题就是求一超平面, 使所有样本点到超平面距离最小. 对非线性问题, 首先经一非线性映射 Φ, 将样本映射到一高维特征空间, 然后用线性方法解决. 高维映射经核函数: $K(x, x_i)=\Phi(x)\cdot\Phi(x_i)$实现. 因SVM引入核函数, 故可有效避免维数灾难、计算复杂性等问题. 目前常用核函数主要有线性核: $K(x, x_i)=x\cdot x_i$; 多项式核: $K(x, x_i)=(\alpha_1 x\cdot x_i+\alpha_2)^p$; 径向基函数(RBF)核: $K(x, x_i)=\exp(-\gamma \| x-x_i \|^2)$; sigmoid核: $K(x, x_i)=\tanh(\alpha_1 x\cdot x_i+\alpha_2)$. 为与线性 PLS 建模方法比较, 选择非线性 RBF 核 SVM 建立 QSRM.

1.6 软件实现

ACCs 由 C 语言程序编写, PCA、GA-PLS、D-最优、PLS 和 SVM 回归均由 Matlab 7.0 实现.

2 结果与讨论

2.1 PLS 建模分析

用 SZOTT 表征 71 个肽序列, 以 ACCs 对产生的变量作数目一致化处理, 经 ACCs(步长 l=1), 每个肽由 169 个变量表征. GA-PLS 参数设置如下, 初始群体大小: 500; 最大遗传代数: 200; 收敛标准: 80%; 交叉频率: 50%; 变异概率: 0.5%; 从训练的 10 个模型中, 确定一包含 45 个自变量的最优模型. 将 71 个肽用 D-最优划分为 50(附表 1 中编号 1~50)个训练集样本和21(附表1中编号51~71)个测试集样本. 以所选 45 个变量用 PLS 建模, 以累计 Q^2_{cv} 对模型边际贡献确定 PLS 模型主成分数目, 若 Q^2_{cv} 增加小于 0.0975, 则该主成分未显著解释序效关系而被拒绝. 得 2 个显著主成分, 其累计解释 Y 变量 85.0%方差, 均方根误差(RMS)为 0. 227. 对 21 个测试集预测得 Q^2_{ext}=0.277, RMS_{ext}=0.434, 预测值(附表 1)及建模结果表明, 模型拟合能力较高, 但其对外部样本预测能力较差. 从 21 个外部测试集样本中发现 52 号样本预测误差较大(图1), 故将此样本视为离群值, 剔除此样本后重新对 20 个测试集样本进行预测得 Q^2_{ext}=0.693, RMS_{ext}=0.252, 外部预测效果得到较大提高.

2.2 SVM 建模分析

参数选择对 SVM 建模成败至关重要, 比如, 惩罚系数 C、不敏感损失函数中的 ε, 核函数类型 K 及相应参数选择等. 因参数选择无严格标准, 故借鉴正交实验设计思想, 据所建模型对外部预测集验证的 Q^2_{ext} 确定 C、ε 和 γ. 研究发现, 以 169 个原始自变量作为 SVM 的输入, 无论建模还是外部预测效果都较差, 故以 GA-PLS 所选 45 个变量作为 SVM 输入, 且考察 GA-PLS 所选变量对 SVM 建模的有效性.

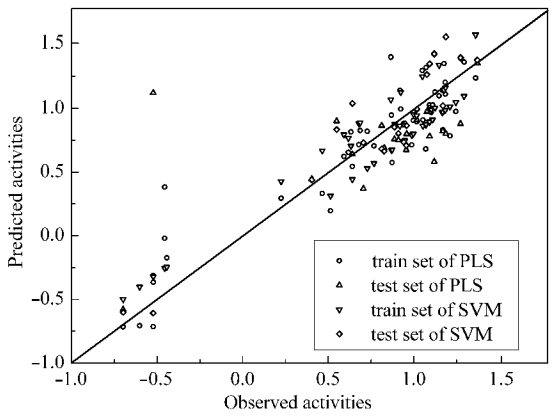

图 1 观测值与预测值相关

Fig.1 Regression between observed activities and predicted activities

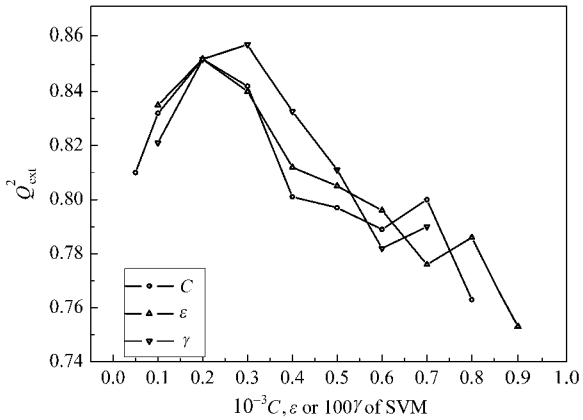

图 2 C, ε, γ 与 Q^2_{ext} 相关

Fig.2 C, ε, and γ *vs* Q^2_{ext} by SVM

以单因素轮换法确定各参数合适的范围(图 2), 首先固定 ε=0.2 和 γ=0.002, 然后分别取 C=50, 100, 200, 300, 400, 500, 600, 700 和 800, 可看出当 C=200 时, Q^2_{ext}值最大(0.852), 故选择C=100~400. 固定C=200 和 γ=0.002, 然后取 ε=0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8和0.9, 当ε=0.2时对应的Q^2_{ext}最大(0.852), 故选择 ε=0.1~0.4. 固定 C=200, ε=0.2, 然后取 γ=0.001, 0.002, 0.003, 0.004, 0.005, 0.006 和 0.007, 当 γ=0.003 时, 最大 Q^2_{ext}=0.855, 故选择 γ=0.001~0.004. 在确定 C=100~400, ε 在 0.1~0.4, γ 在 0.001~0.004 条件下, 设计三因素四水平正交实验优化 C、ε 和 γ(附表 2, 可从 www.whxb.pku.edu.cn 免费下载), 从对外部预测集 16 次验证结果看, 当 C=300, ε=0. 2 及 γ=0.004 时具相对最大 Q^2_{ext}=0.857. 故在此条件下建模得 R^2=0.904, RMS=0.182, Q^2_{ext}=0.857, RMS_{ext}=0.194. 图 1 关于观测值与计算值回归表明, 用非线性径向基核函数 SVM 可获得优于 PLS 的 QSRM, 所建模型表现出更强的拟合能力和外部样本预测能力.

3 结束语

序列表征与建模方法是肽 QSRM 研究的重要内容. 提出 SZOTT 描述子表征 71 个不同长度肽序列结构, 分别用 PLS 回归和 RBF 核 SVM 获得较好的 QSRM. 研究表明, SZOTT 可较好地表征 71 个肽结构, SZOTT 描述子含信息量大且易解释等优点, 可进一步用于肽序列表征. 相对于 PLS, SVM 表现出较强模型拟合能力和稳健外部样本预测能力, SVM 在 QSRM 研究中具有广阔应用前景. 用 SVM 可获得较好 QSRM, 可为解决目标问题提供良好思路, 但是因建模时所选参数较多且复杂, 要进一步在回归方面获得更为令人满意和成功应用, 需要对 SVM 原理和应用技巧做进一步探讨.

References

1 Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. *J. Med. Chem.,* **1987, 30**: 1126

2 Mei, H.; Zhou, Y.; Sun, L. L.; Li, Z. L. *Acta Phys. -Chim. Sin.,* **2004, 20**(8): 821 [梅 虎, 周 原, 孙立力, 李志良. 物理化学学报(*Wuli Huaxue Xuebao*), **2004, 20**(8): 821]

3 Cortes, C.; Vapnik, V. *Machine Learning,* **1995, 20**: 273

4 Tdeschini, R.; Consonni, V. Handbook of molecular descriptors. Weinheim: Wiley-VCH, 2000: 268

5 Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.,* **2000, 43**: 3714

6 Liu, S. S.; Liu, H. L.; Xia, Z. N.; Cao, C. Z.; Li, Z. L. *J. Chem. Inf. Comput. Sci.,* **1999, 39**(6): 951

7 Liu, S. S.; Cai, S. X.; Cao, C. Z.; Li, Z. L. *J. Chem. Inf. Comput. Sci.,* **2001, 40**(6): 1337

8 Gilvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. *J. Chem. Inf. Comput. Sci.,* **1994, 34**: 520

9 Rucker, G.; Rucker, C. *J. Chem. Inf. Comput. Sci.,* **1993, 33**: 683

10 Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. *J. Chem. Inf. Comput. Sci.,* **1991, 31**: 517

11 Diudea, M. V.; Horvath, D.; Graovac, A. *J. Chem. Inf. Comput. Sci.,* **1995, 35**: 129

12 Randic, M.; Kleiner, A. F.; DeAlba, L. M. *J. Chem. Inf. Comput. Sci.,* **1994, 34**: 277

13 Schuur, J. H.; Selzer, P.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.,* **1996, 36**: 334

14 Todeschini, R.; Gramatica, P.; Provenzani, R.; Marengo, E. *Chemon. Intell. Lab. Syst.,* **1995, 27**: 221

15 Consonni, V.; Todeschini, R.; Pavan, M. J. *Chem. Inf. Comput. Sci.,* **2002, 42**: 682

16 Kim, D.; Lee, I. B. *Chemon. Intell. Lab. Syst*., **2003, 67**: 109

17 Yamaki, S.; Isobe, T.; Okuyama, T.; Shinoda, T. *J. Chromatogr. A,* **1996, 729**: 143

18 Nyström, A.; Andersson, P. M.; Lundstedt, T. *Quant. Struct. -Act. Relat.,* **2000, 19**: 264

19 Hasegawa, K.; Miyashita, Y.; Funatsu, K. *J. Chem. Inf. Comput. Sci.,* **1997, 37**: 306

20 Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb. Sci.,* **2003, 22**: 69

21 Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.,* **2001, 58**: 109