基于支持向量机的高维特征非线性快速筛选与肽 QSAR 建模

代志军^{1,2} 周 玮¹ 袁哲明^{1,2,*}

(¹湖南农业大学生物安全科学技术学院, 长沙 410128;

²湖南农业大学, 湖南省作物种质创新与资源利用重点实验室, 长沙 410128)

摘要: 以氨基酸的 531 个物理化学性质参数直接表征肽的结构, 基于支持向量回归发展了一种新的高维特征非线性快速筛选方法, 将其应用于苦味二肽和血管紧张素转化酶抑制剂 2 个肽体系的定量序效关系(QSAR)建模, 各筛选获得 10 个意义明确的保留描述子. 以保留描述子建立支持向量回归模型, 其拟合精度、留一法交叉测试精度和外部预测精度较文献报道结果均有较大幅度提升, 优势明显; 对所建模型进行了非线性回归显著性测验、单因子相对重要性显著性测验和单因子效应分析, 增强了模型的可解释性. 新方法在肽、蛋白质 QSAR 建模等高维数据回归预测领域有广泛应用前景.

关键词: 高维特征; 特征选择; 肽; 定量序效关系; 支持向量机

中图分类号: O641

A Novel Method of Nonlinear Rapid Feature Selection for High Dimensional Data and Its Application in Peptide QSAR Modeling Based on Support Vector Machine

DAI Zhi-Jun^{1,2} ZHOU Wei¹ YUAN Zhe-Ming^{1,2,*}

(¹College of Bio-safety Science and Technology, Hunan Agricultural University, Changsha 410128, P. R. China;

²Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Changsha 410128, P. R. China)

Abstract: Each amino acid residue of one peptide was characterized directly by 531 physicochemical property parameters. Based on support vector regression (SVR) we developed a new nonlinear rapid feature selection method for high dimensional data, which was applied to a quantitative sequence–activity relationship (QSAR) study of two peptide systems (bitter tasting thresholds and angiotensin converting enzyme inhibitors). In both systems, 10 descriptors with clear meaning were reserved. We established a SVR model for both peptide systems using the reserved descriptors of the peptides. For both models the accuracies of fitting, the leave-one-out cross validation, and the external prediction improved significantly compared with the results reported in literature. To enhance the interpretability of the models, significance tests of the nonlinear regression model, single-factor relative importance, and a single-factor effect analysis were carried out. The new method has broad application prospects for regression forecasting of high dimensional data such as QSAR modeling of peptide or proteins.

Key Words: High dimensional feature; Feature selection; Peptide; Quantitative sequence–activity relationship; Support vector machine

Received: March 24, 2011; Revised: May 10, 2011; Published on Web: June 3, 2011.

*Corresponding author. Email: zhmyuan@sina.com; Tel: +86-731-84613956.

The project was supported by the Science Foundation for Distinguished Young Scholars of Hunan Province, China (10JJ1005), Research Fund for the Doctoral Program of Higher Education of China (200805370002), Team Project for the Technology Innovation of Higher Education of Hunan Province, China, 2008.

湖南省杰出青年科学基金(10JJ1005), 高等学校博士点基金(200805370002)和湖南省 2008 年高校科技创新团队项目资助

© Editorial office of *Acta Physico-Chimica Sinica*

1 引言

肽、蛋白质功能本质上由其一级结构——氨基酸残基序列决定。蛋白质三级结构测定耗时费力, 而肽、蛋白质一级结构简便易得, 因此蛋白质、肽的定量序效关系(QSAR)研究极为重要, 在研发肽类新药、揭示蛋白质结构与功能关系方面有广泛应用前景。¹ 蛋白质、肽 QSAR 建模的一个重要内容是其一级结构表征。Sneath² 首先基于 20 种天然氨基酸的一些半定量描述子表征氨基酸残基, 用于后叶催产素类似物活性预测。Kidera 等³ 收集天然氨基酸的 188 种物理化学性质参数, 经因子分析得到 10 个正交因子, 用于肽或蛋白质的高级结构预测。Hellberg 等⁴ 采用正交变换处理氨基酸 29 种物理化学性质参数得到 Z 标度。Sandberg 等⁵ 进一步将 Z 标度扩展到包括天然氨基酸在内的 87 种氨基酸。李志良与其合作者⁶⁻¹⁰ 基于分子电性、氨基酸侧链残基表面势能、立体性质、拓扑指数等提出了一系列氨基酸描述子, 并应用于多个肽体系的 QSAR 研究。上述描述子多采用主成分分析降维、去噪、去冗余, 每一个主成分是氨基酸所选多个物理化学性质参数的线性加权, 所得模型解释性差; 对不同的肽体系, 往往需有针对性地精选氨基酸若干物理化学性质若干主成分, 所得描述子可推广性差。

氨基酸指数(AAindex)数据库¹¹中的 531 个特征是氨基酸物理化学性质的较全面总结, 可直接作为描述子用于肽、蛋白质结构表征, 但导致特征维数巨增。无关、冗余的特征对模型不利, 从 m 个特征中选取 p ($p \leq m$) 个最优特征子集理论上共有 2^m 种可能, 已知这是一个完全多项式非确定性 (non-deterministic polynomial, NP) 问题, 在 m 较大时无法穷举。常用的特征选择方法按照特征子集的形成方式, 分为穷举法、启发法(包括向前选择、向后选择、决策树法、Relief 法等)和随机法(包括拉斯维加斯算法、遗传算法、模拟退火算法等)三类。¹² 按照评价函数与预测模型的关系, 分为筛选器(包括距离测度、信息测度、相关性测度和一致性测度等)与封装器两类。其中, 筛选器的评价函数与预测模型无关, 选取的特征冗余性高、精度低; 而封装器采用错误概率(分类)或均方误差(MSE)(回归)作为评价函数, 计算复杂度高。¹³ Relief 法不能去除冗余特征。业已证实, 独立最好的两个特征不一定是组合最好的两个特征,¹⁴ 独立最好的 K 个特征不一定是组合最好的 K 个特征,¹⁵ 因此特征筛选必需在一个统一的模型中进行(采用封装

器)。多数现有非穷举特征筛选方法存在易陷入局部最优的弊端。

Cortes 和 Vapnik¹⁶ 基于统计学习理论提出的支持向量机(SVM)是目前发展最快的机器学习方法, 其基于结构风险最小, 较好地解决了小样本、非线性、过拟合、维数灾和局部极小等问题, 泛化推广能力优异。SVM 包括分类(SVC)和回归(SVR)。Hou 及其合作者^{17,18} 基于分子相互作用组件与 SVM 用于蛋白质结构域 SH3 (Src homology 3) 的结合特异性预测。Wang 等¹⁹ 以 SVR 与 4 类氨基酸描述子预测主要组织相容性复合体 I 类分子与抗原肽的结合特异性。Zhang 等²⁰ 分别基于序列和结构表征肽序列, 以非线性 SVM 对肽的离子迁移谱漂移时间进行建模预测。然而, SVM 的主要缺点是可解释性差。袁哲明及其合作者²¹⁻²⁴ 针对特征个数 m 较小的情形, 基于 SVR 发展了多轮末尾淘汰的非线性特征筛选方法; 对保留特征进行非线性强制汰选, 可给出各保留特征影响预测精度的相对重要性排序, 从而使 SVR 具有了部分解释能力; 进一步基于 F 测验建立了一套较完整的 SVR 解释性体系。但当 m 较大时, 多轮末尾淘汰失效。

为了消除大量冗余、无关描述子对模型稳健性和泛化能力的影响, 本文基于 SVR 发展了一种新的高维特征非线性快速筛选方法。新方法具有非线性、全局最优与筛选快速的优点。将其应用于 2 个二肽体系的特征筛选与 QSAR 建模, 并以前期构建的 SVR 解释性体系对模型进行了解释。

2 原理和方法

2.1 肽结构表征

每条肽序列上的每个氨基酸残基由氨基酸指数数据库¹¹中的 531 个描述子表征(命名为 AA531), 依次串联排列。每个二肽含 1062 个描述子。

2.2 高维特征非线性快速筛选

首先, 设单因变量原始训练集为 (y_i, x_{ij}) , 有 n 个样本, m 个特征, $i=1, 2, \dots, n; j=1, 2, \dots, m$ 。每个特征有两种情形, 1(选取)和 0(不选取)。有条件产生一个 $K \times m$ 随机矩阵: K 为给定的组合数(本文取 $K=500$), 矩阵各元素为 1 或 0, 每列(每个特征)中 1 与 0 的个数相同。从随机矩阵的一行选取值为 1 的矩阵元素, 以选取元素的下标从原始训练集中选择特征。以 SVR 经交叉测试获得一个 MSE 值。重复此步骤, 获得 K 个 MSE 值。

其次,将 K 个 MSE 作为新的因变量 y' , $K \times m$ 随机矩阵作为新的自变量,构建新训练集.对 $K \times m$ 随机矩阵中的某一列元素替换产生全为 0 与全为 1 (其它列不变) 的两组测试集.对新训练集以 SVR 训练,并对两组测试集预测得到 $\hat{y}_0'$ 与 $\hat{y}_1'$, 比较 $\hat{y}_0'$ 与 $\hat{y}_1'$ 的均值,若 $\hat{y}_0'$ 的均值小于 $\hat{y}_1'$ 的均值,说明删除该列对应特征可以提高原始训练集交叉测试精度.重复此步骤,获得一轮保留特征.

重复上述过程,经过多轮筛选,至没有特征可删除为止.最后以非线性多轮末尾淘汰法²⁵精细筛选,获得 m' 个保留特征,用于后续 QSAR 建模 ($m' \ll m$).

2.3 模型验证

为便于比较,采用 D-optimal 优化方法²⁶将样本分为训练集和预测集两个部分,训练集拟合、训练集留一法交叉测试、测试集预测精度分别用决定系数 R^2 、 Q^2 、 R_{ext}^2 评价. Tropsha 等²⁶的研究结果显示, Q^2 的大小与模型的独立预测能力并没有明显的相关关系,对模型预测能力的评价只能通过外部样本集即测试集来进行,并进一步提出用 Q_{ext}^2 来衡量模型的外部预测能力:

$$Q_{\text{ext}}^2 = 1 - \frac{\sum_{i=1}^{\text{test}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{\text{test}} (y_i - \bar{y}_{\text{tr}})^2}$$

其中, y_i 为测试集中样本活性的实验值, $\hat{y}_i$ 为测试集样本活性的预测值, $\bar{y}_{\text{tr}}$ 为训练集样本活性实验值的平均值, test 指测试集中的样本个数.

2.4 模型解释

以全部样本、 m' 个保留特征基于 SVR 建立肽 QSAR 模型,参照 SVR 解释性体系^{23,24}进行 SVR 回归显著性测验、单因子重要性显著性测验(非线性偏回归)和单因子效应分析.

3 结果与讨论

3.1 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.²⁷ 48 个苦味二肽 (BTT) 为一个经典的 QSAR 样本集,²⁸ 其活性用其浓度 T (单位: $\text{mol} \cdot \text{L}^{-1}$) 的负对数 $\text{p}T$ 表示,氨基酸序列及其生物活性值与拟合值(对训练集)、预测值(对测试集)见表 1.

基于 SVR 对 48 个样本, 1062 个特征进行 5 次交叉测试,其 MSE 为 0.0375. 应用高维特征非线性快速筛选,第 1 轮即筛去 663 个特征,5 次交叉测试 MSE 降至 0.0103; 7 轮筛选后保留 18 个特征,5 次交叉测试 MSE 降为 0.0028. 继续以非线性多轮末尾淘汰法²⁵精细筛选,获得一个仅包含 10 个特征的最优特征子集. 基于 SVR 与最优特征子集,对训练集 (Nos.1–24) 建立 QSAR 模型,预测测试集 (Nos.25–48),其 R^2 、 Q^2 、 R_{ext}^2 和 Q_{ext}^2 分别为 0.975、0.927、0.934 和 0.904,与文献^{6,7,9,29,30}报道结果比较(表 2),可见本文结果无论拟合精度、留一法交叉测试精度还是外部预测精度均有较大幅度提升,优势明显.

以全部 48 个样本、10 个保留特征基于 SVR 建立肽 QSAR 模型,参照 SVR 解释性体系^{23,24}进行 SVR 回归显著性测验,其 $R^2=0.97$, 显著性 F 测验中 $F=142.88$, 表明 SVR 模型非线性回归极显著 ($F_{0.01}$

表 1 BTT 的氨基酸序列及其生物活性值²⁸与预测值

Table 1 Amino acid sequences of bitter tasting thresholds (BTT) with measured²⁸ and predicted activity

No.	Peptide	$\text{p}T$		No.	Peptide	$\text{p}T$		No.	Peptide	$\text{p}T$		No.	Peptide	$\text{p}T$	
		Obsd.	Calc.			Obsd.	Calc.			Obsd.	Calc.			Obsd.	Calc.
1	GL	1.68	1.68	13	IK	1.65	1.65	25	GV	1.13	1.29	37	IA	1.68	1.78
2	GP	1.35	1.35	14	PA	1.32	1.32	26	GI	1.70	1.75	38	IV	2.05	2.20
3	GF	1.80	1.80	15	PL	2.22	2.22	27	AV	1.16	1.43	39	IL	2.26	2.43
4	GW	1.89	1.89	16	PI	2.33	2.33	28	AL	1.70	1.78	40	II	2.26	2.50
5	GY	1.77	1.77	17	PY	1.80	2.17	29	AF	1.72	1.77	41	IN	1.49	1.34
6	LG	1.72	1.72	18	FG	1.77	1.77	30	VG	1.19	1.36	42	IQ	1.49	1.50
7	LL	2.35	2.64	19	FL	2.87	2.87	31	VA	1.16	1.51	43	IE	1.37	1.31
8	LW	3.40	3.40	20	FP	2.70	2.70	32	VV	1.71	2.03	44	IS	1.49	1.73
9	IG	1.68	1.68	21	FY	3.13	2.92	33	VL	2.00	2.33	45	IT	1.49	1.65
10	IP	2.40	2.40	22	WE	1.56	1.56	34	LA	1.72	1.90	46	PF	2.80	2.64
11	IW	3.05	3.05	23	WW	3.60	3.60	35	LF	2.75	3.13	47	FF	3.10	3.24
12	ID	1.37	1.37	24	YL	2.40	2.40	36	LY	2.46	2.64	48	SL	1.49	1.47

training set: Nos.1–24, test set: Nos.25–48; $\text{p}T$: negative logarithm of the molarity (in $\text{mol} \cdot \text{L}^{-1}$) of BTT

表2 BTT的QSAR模型比较

Table 2 Comparison between QSAR models of BTT

Descriptor	Model	A	R ²	Q ²	R ² _{ext}	Q ² _{ext}
3D-HoVAIF ⁶	PLSR	3	0.936	0.849	—	—
VHSE ²⁹	PLSR	3	0.881	0.843	—	—
SZOTT ⁹	PLSR	2	0.908	0.736	—	—
VHSE ³⁰	PLSR	1	0.744	0.643	0.758	0.707
VHSE ³⁰	SVR	16	0.948	0.729	0.898	0.875
ASSPF ⁷	MLR	9	0.846	0.767	—	—
AA531	SVR	10	0.975	0.927	0.934	0.904

3D-HoVAIF: 3D holographic vector of atomic interaction field;
VHSE: vector of hydrophobic, steric, electronic properties; SZOTT:
vector of scores for zero dimension, one dimension, two dimension
and three dimension; ASSPF: amino acid side chain surface potential
field; PLSR: partial least squares regression; MLR: multiple linear
regression; A: number of descriptors or principal components; R²:
determination coefficient of fitting; Q²: determination coefficient of
leave-one-out cross validation for training set; R²_{ext}: determination
coefficient of external prediction; Q²_{ext}: determination coefficient of
external prediction proposed by Tropsha²⁶

(10,37)=2.85). 10个保留特征(描述子)的单因子相对重要性的显著性测验 F 值及其排序见表 3, 除 1-PLIV810101(ID 号前的序号表示二肽序列中残基位置,下同)描述子的相对重要性为显著外($F_{0.05}(1,$

37)=4.11), 其余均达极显著($F_{0.01}(1,37)=7.37$).
各保留描述子的单因子效应见图 1, 对各因子 x_i 分别通过 $x'_i=(x_i-x_{\max})/(x_{\max}-x_{\min})$ 得到归一化的 x 轴坐标值, 其中 $x_{\max}$ 、 $x_{\min}$ 分别为各因子的上、下限. 结果表明, 苦味二肽的活性随 2-GOLD730101、1-MEEJ810102、2-KHAG800101 与 1-PLIV810101 的值增加而增加; 随 2-BULH740101、2-PONP800104 的值增加而减小; 随 1-WERD780102、1-YUTK870101、2-RICJ880111 与 1-CORJ870106 的值先减少后增加呈开口向上抛物线变化.

天然蛋白质的疏水性基团一般都包含在分子结构内部, 从而不会呈现苦味. 当蛋白质水解成小分子多肽时, 就会曝露出其疏水性氨基酸残基, 此类氨基酸残基刺激味蕾, 即呈现苦味. 保留的 10 个描述子均与氨基酸的疏水性相关, 其中 2-GOLD730101 与 2-PONP800104 均为疏水性相关物化性质. 转移自由能 (2-BULH740101、1-WERD780102、1-YUTK870101) 决定疏水性强度. 疏水性残基具有更强的截留能力(1-MEEJ810102). Richardson 等³¹发现在蛋白质二级结构 α 螺旋的

表3 BTT的保留描述子

Table 3 Reserved descriptors in BTT QSAR model

Residue position-AAindex ID	Description	F-value
2-GOLD730101	hydrophobicity factor	99.51**
2-BULH740101	transfer free energy to surface	38.41**
1-MEEJ810102	retention coefficient in NaH ₂ PO ₄	29.54**
2-KHAG800101	the Kerr-constant increments	22.49**
1-WERD780102	free energy change of epsilon(i) to epsilon(ex)	16.02**
1-YUTK870101	unfolding Gibbs energy in water, pH7. 0	14.28**
2-PONP800104	surrounding hydrophobicity in alpha-helix	11.52**
2-RICJ880111	relative preference value at C4	11.15**
1-CORJ870106	ALTLS index	8.98**
1-PLIV810101	partition coefficient	4.58*

*: $p<0.05$; **: $p<0.01$

图1 苦味二肽单因子效应折线图
Fig.1 Single-factor effects of BTT
predicted activity of BTT: negative logarithm of the molarity (in mol·L⁻¹) of BTT

表4 ACE抑制剂的氨基酸序列及其生物活性值⁴与预测值Table 4 Amino acid sequences of angiotensin converting enzyme (ACE) inhibitors with measured⁴ and predicted activity

No.	Peptide	pIC ₅₀		No.	Peptide	pIC ₅₀		No.	Peptide	pIC ₅₀		No.	Peptide	pIC ₅₀	
		Obsd.	Calc.			Obsd.	Calc.			Obsd.	Calc.			Obsd.	Calc.
1	VW	5.80	5.68	16	WG	2.23	2.35	30	IW	5.70	5.39	45	GL	2.60	2.78
2	RW	4.80	4.91	17	HG	2.20	2.08	31	IY	5.43	4.56	46	AG	2.60	2.41
3	GW	4.52	4.57	18	GQ	2.15	2.30	32	AW	5.00	4.82	47	KG	2.49	2.53
4	VF	4.28	4.16	19	SG	2.07	2.19	33	VY	4.66	4.85	48	FG	2.43	2.80
5	IP	3.89	3.99	20	GD	2.04	1.98	34	AY	4.06	4.01	49	GV	2.34	2.65
6	RP	3.74	3.65	21	EG	2.00	1.88	35	AF	3.72	3.34	50	GK	2.27	2.24
7	GP	3.35	3.40	22	DG	1.85	1.94	36	GY	3.68	3.79	51	GT	2.24	2.40
8	IG	2.92	2.79	23	PG	1.77	1.71	37	AP	3.64	3.56	52	GG	2.14	2.29
9	GI	2.92	2.87	24	KA	3.42	3.04	38	RF	3.64	3.44	53	QG	2.13	2.20
10	GM	2.85	2.72	25	YA	3.34	3.42	39	VP	3.38	4.29	54	LG	2.06	2.52
11	GH	2.51	2.63	26	FR	3.04	3.17	40	GF	3.20	3.13	55	TG	2.00	2.25
12	GR	2.49	2.42	27	HL	2.49	2.61	41	IF	3.03	3.87	56	LA	3.51	3.22
13	GS	2.42	2.33	28	DA	2.42	2.29	42	VG	2.96	3.10	57	RA	3.34	3.06
14	MG	2.32	2.52	29	EA	2.00	2.13	43	GA	2.70	2.75	58	AA	3.21	2.94
15	GE	2.27	2.33					44	YG	2.70	2.77				

training set: Nos.1–29, test set: Nos.30–58; pIC₅₀: negative logarithm of the IC₅₀ (mol·L⁻¹) of ACE inhibitors

C-cap-4 端(2-RICJ880111)疏水性残基具有偏好峰。Cornette 等³²指出可变螺旋最小二乘估计值(1-CORJ870106)为一种探测跨膜两性蛋白分子指数。Pliška 等³³利用薄层色谱法测定的正辛醇-水分配系数(1-PLIV810101)作为疏水性的量度参数。

以48个苦味二肽样本训练建模,以所建模型预测剩余的352个苦味二肽活性值,预测活性值大于3.00的二肽序列为FW与WF,其活性值分别为3.51与3.31,均未超过真实样本集中二肽WW的活性值(3.60)。结果表明,若要获得更高的活性值,可能需要设计更长序列的多肽。

3.2 血管紧张素转化酶抑制剂QSAR研究

肾素血管紧张素系统(RAS)在调节人体血压中起关键作用。由肝脏产生血管紧张素原经肾素催化断裂为非活性十肽血管紧张素I,再经血管紧张素转化酶(ACE)催化裂解为具有极强血管收缩功能的八肽血管紧张素II,因此ACE是降血压药物作用靶点。³⁴ACE抑制剂通过模拟结合底物血管紧张素I的活性部位结构特征,从而达到竞争抑制ACE有效生物活性的目的。ACE抑制剂为一组58个二肽QSAR经典样本集,⁴其活性值用IC₅₀(mol·L⁻¹)的负对数表示,氨基酸序列及其生物活性值与拟合值(对训练集)、预测值(对测试集)见表4。

基于SVR对58个样本、1062个特征进行5次交叉测试,其MSE为0.7517。应用高维特征非线性快速筛选,第1轮即筛去693个特征,5次交叉测试MSE降至0.3277;9轮筛选后保留21个特征,5次交

叉测试MSE降为0.0827。继续以非线性多轮末尾淘汰法²⁵精细筛选,获得一个仅包含10个特征的最优特征子集。基于SVR与最优特征子集,对训练集(Nos.1–29)建立QSAR模型,预测测试集(Nos.30–58),其 R^2 、 Q^2 、 R^2_{ext} 和 Q^2_{ext} 分别为0.983、0.949、0.880和0.885,与文献报道^{8–10,30}结果比较(表5),本文结果无论拟合精度、留一法交叉测试精度还是外部预测精度均有较大幅度提升,优势明显。

以全部58个样本、10个保留特征基于SVR建立肽QSAR模型,参照SVR解释性体系^{23,24}进行SVR回归显著性测验,其 $R^2=0.9596$, $F=110.21$,SVR模型非线性回归极显著($F_{0.01}(10,47)=2.72$)。10个保留特征(描述子)的单因子相对重要性显著性测验 F 值及其排序见表6,除2-LEVM780106描述子相对重要性为显著外($F_{0.05}(1,47)=4.05$),其余均达极显著($F_{0.01}(1,47)=7.21$)。

各保留描述子的单因子效应如图2,结果表

表5 ACE抑制剂QSAR模型比较

Table 5 Comparison between QSAR models of ACE inhibitors

Descriptor	Model	A	R^2	Q^2	R^2_{ext}	Q^2_{ext}
VSTV ¹⁰	SVR	12	0.850	0.802	—	—
VHSEH ⁸	PLSR	3	0.877	0.838	—	—
VHSE ³⁰	SVM	16	0.876	0.767	0.761	0.773
SZOTT ⁹	PLSR	2	0.894	0.828	—	—
AA531	SVR	10	0.983	0.949	0.880	0.885

VSTV: principal component score vector of structural and topological variables; VHSEH: vector of hydrophobic, steric, electronic properties and hydrogen bonds contributions

表6 ACE抑制剂保留描述子的QSAR模型
Table 6 QSAR model of reserved descriptors in ACE inhibitors

Residue position-AAindex ID	Description	F-value
2-RICJ880104	relative preference value at N1	95.18**
2-WIMW960101	free energies of transfer of AcWI-X-LL peptides from bilayer interface to water	64.32**
1-BROC820102	retention coefficient in HFBA	47.34**
2-NAKH920103	AA composition of EXT of single-spanning proteins	25.04**
1-GEIM800106	beta-strand indices for beta-proteins	22.72**
1-ROBB760103	information measure for middle helix	20.44**
1-WILM950103	hydrophobicity coefficient in RP-HPLC	18.74**
1-FINA910104	helix termination parameter at position $j+1$	15.76**
1-CIDH920104	normalized hydrophobicity scales for alpha/beta-proteins	7.86**
2-LEVM780106	normalized frequency of reverse turn, unweighted	5.64*

*: $p < 0.05$; **: $p < 0.01$

图2 ACE抑制剂单因子效应折线图
Fig.2 Single-factor effects of ACE inhibitors

明, ACE 抑制剂活性随 1-BROC820102、2-NAKH920103、1-WILM950103 与 2-LEVM780106 的值增加而减小, 随其余 6 个保留描述子的值增加而增加. 可见, ACE 抑制剂的生物活性与第一残基的疏水性质(1-BROC820102、1-WILM950103)负相关, 立体性质(1-GEIM800106、1-ROBB760103、1-FINA910104、1-CIDH920104)正相关; 与第二残基疏水相关性质(2-RICJ880104、2-WIMW960101)正相关, 氨基酸组成(2-NAKH920103)和立体性质(2-LEVM780106)负相关.

以 58 个 ACE 抑制剂样本训练建模, 预测剩余 342 个 ACE 抑制剂生物活性值, 预测活性值大于 5.00 的二肽序列为 FW、LW、MW、WW 与 YW, 其活性值分别为 5.41、5.36、5.28、5.29 与 5.28, 均未超过真实样本集中二肽 VW 的活性值(5.80). 显示若要获得更高的 ACE 抑制剂活性值, 可能需要设计更长序列的多肽.

4 结论

以氨基酸指数数据库中的 531 个特征参数直接

作为描述子用于肽结构表征, 充分利用 SVR 非线性、适于小样本、有效解决维数灾难、有效克服局部极小的优点, 基于 SVR 发展了新的高维特征非线性快速筛选方法, 应用于苦味二肽和 ACE 抑制剂 2 个肽体系的 QSAR 建模, 各筛选获得 10 个意义明确的保留描述子. 所建 QSAR 模型拟合精度、留一法交叉测试精度、外部预测精度较文献报道结果均有较大幅度提升, 优势明显; 进一步参照 SVR 解释性体系进行了 SVR 回归显著性测验、单因子相对重要性显著性测验、单因子效应分析, 并给出了特征参数所对应的真正生物化学意义, 增强了所建模型的可解释性; 最后根据所建模型预测活性未知的二肽, 可为实验提供参考. 新方法在肽、蛋白质 QSAR 建模等高维数据回归预测领域有广泛应用前景.

References

- (1) Ding, J. L.; Ho, B. *Drug Dev. Res.* **2004**, 62 (4), 317.
- (2) Sneath, P. H. A. *J. Theor. Biol.* **1966**, 12 (2), 157.
- (3) Kidera, A.; Konishi, Y.; Oka, M.; Ooi, T.; Scheraga, H. A. *J. Protein Chem.* **1985**, 4 (1), 23.
- (4) Hellberg, S.; Eriksson, L.; Jonsson, J.; Lindgren, F.; Sjöström,

- M.; Skagerberg, B.; Wold, S.; Andrews, P. *Int. J. Pept. Protein Res.* **1991**, 37 (5), 414.
- (5) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. *J. Med. Chem.* **1998**, 41 (14), 2481.
- (6) Liang, G. Z.; Mei, H.; Zhou, P.; Zhou, Y.; Li, Z. L. *Acta Phys. -Chim. Sin.* **2006**, 22 (3), 388. [梁桂兆, 梅 虎, 周 鹏, 周 原, 李志良. 物理化学学报, **2006**, 22 (3), 388.]
- (7) Zhou, Y.; Mei, H.; Yang, L.; Zhou, P.; Yang, S. B.; Li, Z. L. *Chem. J. Chin. Univ.* **2007**, 28 (7), 1263. [周 原, 梅 虎, 杨 力, 周 鹏, 杨善斌, 李志良. 高等学校化学学报, **2007**, 28 (7), 1263.]
- (8) Yang, S. B.; Xia, Z. N.; Shu, M.; Mei, H.; Lü, F. L.; Zhang, M.; Wu, Y. Q.; Li, Z. L. *Chem. J. Chin. Univ.* **2008**, 29 (11), 2213. [杨善彬, 夏之宁, 舒 茂, 梅 虎, 吕凤林, 张 梅, 吴玉乾, 李志良. 高等学校化学学报, **2008**, 29 (11), 2213.]
- (9) Liang, G. Z.; Zhou, P.; Zhou, Y.; Zhang, Q. X.; Li, Z. L. *Acta Chim. Sin.* **2006**, 64 (5), 393. [梁桂兆, 周 鹏, 周 原, 张巧霞, 李志良. 化学学报, **2006**, 64 (5), 393.]
- (10) Li, Z. L.; Li, G. R.; Shu, M.; Sun, J. Y.; Yang, S. B.; Mei, H.; Zhang, M. J.; Zhou, P.; Wu, S. R.; Chen, G. H.; Lü, F. L.; Lü, T. T. *Sci. China Ser. B: Chem.* **2008**, 38 (8), 745. [李志良, 李根容, 舒 茂, 孙家英, 杨善斌, 梅 虎, 张梦军, 周 萍, 吴世荣, 陈国华, 吕凤林, 吕廷亭. 中国科学B辑: 化学, **2008**, 38 (8), 745.]
- (11) Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. *Nucleic Acids Res.* **2008**, 36, D202.
- (12) Dash, M.; Liu, H. *Intell. Data Anal.* **1997**, 1 (3), 131.
- (13) Kohavi, R.; John, G. H. *Artif. Intel.* **1997**, 97 (1-2), 273.
- (14) Cover, T. M. *IEEE Trans. Syst. Man Cybern.* **1974**, 4 (2), 116.
- (15) Colannino, J.; Juban, J. The Best K Measurements are Not the K Best. <http://cgm.cs.mcgill.ca/~athens/cs644/Projects/2004/JustinColannino-JeremieJuban/best.html>, (accessed Mar 22, 2011).
- (16) Cortes, C.; Vapnik, V. *Mach. Learn.* **1995**, 20 (3), 273.
- (17) Hou, T. J.; Xu, Z.; Zhang, W.; McLaughlin, A. W.; Case, D. A.; Xu, Y.; Wang, W. *Mol. Cell. Proteomics* **2009**, 8, 639.
- (18) Hou, T. J.; Zhang, W.; Case, D. A.; Wang, W. *J. Mol. Biol.* **2008**, 376, 1201.
- (19) Wang, H.; Zhang, B. J.; Liu, X. Z.; Luo, D. Z.; Zhong, S. B. *Adv. Mater. Res.* **2011**, 143, 1254.
- (20) Zhang, Y. M.; Jin, Q.; Wang, S. T.; Ren, R. *Comput. Biol. Med.* **2011**, 41 (5), 271.
- (21) Yuan, Z. M.; Zhang, Y. S.; Xiong, J. Y. *Scientia Agricultura Sinica* **2008**, 41 (8), 2485. [袁哲明, 张永生, 熊洁仪. 中国农业科学, **2008**, 41 (8), 2485.]
- (22) Zhang, Y. S.; Yuan, Z. M.; Xiong, J. Y.; Zhou, T. J. *Acta Ecologica Sinica* **2007**, 27 (6), 2419. [张永生, 袁哲明, 熊洁仪, 周铁军. 生态学报, **2007**, 27 (6), 2419.]
- (23) Tan, X. S.; Wang, Z. M.; Tan, S. Q.; Yuan, Z. M.; Xiong, X. Y.; *Journal of System Simulation* **2009**, 21 (24), 7795. [谭显胜, 王志明, 谭泗桥, 袁哲明, 熊兴耀. 系统仿真学报, **2009**, 21 (24), 7795.]
- (24) Yuan, Z. M.; Tan, X. S. *Acta Agronomica Sinica* **2010**, 36 (7), 1176. [袁哲明, 谭显胜. 作物学报, **2010**, 36 (7), 1176.]
- (25) Tan, X. S.; Yuan, Z. M.; Zhou, T. J.; Wang, C. J.; Xiong, J. Y. *Chem. J. Chin. Univ.* **2008**, 29 (1), 95. [谭显胜, 袁哲明, 周铁军, 王春娟, 熊洁仪. 高等学校化学学报, **2008**, 29 (1), 95.]
- (26) Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb. Sci.* **2003**, 22 (1), 69.
- (27) Cocchi, M.; Johansson, E. *Quant. Struct. -Act. Relat.* **1993**, 12 (1), 1.
- (28) Collantes, E. R.; Dunn, W. J. *J. Med. Chem.* **1995**, 38 (14), 2705.
- (29) Mei, H.; Zhou, Y.; Sun, L. L.; Li, Z. L. *Chemistry* **2005**, (7), 534. [梅 虎, 周 原, 孙立力, 李志良. 化学通报, **2005**, (7), 534.]
- (30) Mei, H.; Liang, G. Z.; Zhou, Y.; Li, Z. L. *Chin. Sci. Bull.* **2005**, 50 (16), 1703. [梅 虎, 梁桂兆, 周 原, 李志良. 科学通报, **2005**, 50 (16), 1703.]
- (31) Richardson, J. S.; Richardson, D. C. *Science* **1988**, 240 (4859), 1648.
- (32) Cornette, J. L.; Cease, K. B.; Margalit, H.; Spouge, J. L.; Berzofsky, J. A.; DeLisi, C. *J. Mol. Biol.* **1987**, 195 (3), 659.
- (33) Pliška, V.; Schmidt, M.; Fauchère, J. L. *J. Chromatogr. A* **1981**, 216, 79.
- (34) Hassell, C. H.; Krohn, A.; Moody, C. J.; Thomas, W. A. *J. Chem. Soc. Perkin. Trans. 1* **1984**, 23, 155.