一种新的氨基酸描述符SVHEHS在生物活性肽QSAR中的应用研究

彭剑秋[1]，刘 静[2]，管 骁[1,*]
(1.上海理工大学医疗器械与食品学院，上海 200093；2.上海海事大学信息工程学院，上海 200135)

摘 要：对20种氨基酸的457种性质参数按疏水性质、电性特征、氢键贡献和立体特征进行分类后，并各自进行主成分分析(PCA)，得到一种新的氨基酸结构描述符SVHEHS(score vector of hydrophilicity, electronic, hydrogen bond contribution and steric properties)。用该描述符分别对一系列血管紧张素转化酶抑制二肽以及苦味二肽进行序列表征，并用来与生物活性建立多元线性回归模型。血管紧张素转化酶抑制二肽、苦味二肽模型的相关系数、交叉验证相关系数、均方根误差分别为0.936、0.854、0.259和0.949、0.886、0.136，同时还对所得模型进行了外部验证。结果表明，该描述符建立的模型具有较好的拟合与预测能力，用于生物活性肽的定量构效关系研究是理想的。

关键词：氨基酸描述符；定量构效关系；主成分分析；多元线性回归

A New Amino Acid Descriptor SVHEHS and Its Application in QSAR of Bioactive Peptides

PENG Jian-qiu[1]，LIU Jing[2]，GUAN Xiao[1,*]
(1.School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China；2. College of Information Engineering, Shanghai Maritime University, Shanghai 200135, China)

Abstract: Totally 457 physicochemical variables of 20 kinds of natural amino acids were classified according to hydrophobic properties, electronic characteristics, hydrogen bonds contributions and steric properties, and analyzed by principal component analysis (PCA). A new amino acid structure descriptor, SVHEHS, was achieved. The descriptor was used to characterize the structures of a series of ACE inhibitory dipeptides and bitter dipeptides. A multiple linear regression (MLR) model was established on the basis of bioactivity. The correlation coefficients (R^2), cross-validation correlation coefficients (Q^2_{LOO}) and root mean square errors (RMSE) for the models established for ACE inhibitory dipeptides and bitter dipeptides were 0.936, 0.854, 0.259, and 0.949, 0.886, 0.136, respectively. External validation was also conducted to validate the prediction capability of the models. The results showed that the models obtained with the descriptor had good fitting and prediction capability and consequently could be used for QSAR studies.

Key words: amino acid descriptor; quantitative structure activity relationship; principal component analysis; multiple linear regression

中图分类号：Q517 文献标识码：A 文章编号：1002-6630(2012)07-0026-06

现代临床研究已证明，生物活性肽可能具有多种人体代谢和生理调节功能，如促进免疫、激素调节、抗菌、抗病毒、降血压、降血脂等，且易消化吸收，这在新药设计和开发上能产生可观的效益。随着肽库技术的飞速发展，人们已经能有目的设计、合成各种生理活性肽，并进行一系列的生理活性实验。为了能有效利用肽，定量构效关系(quantitative structure activity relationship，QSAR)已被引入到该研究热点，它不仅应用在药物化学和药理学上，同时也是药物设计的依据。近年来，QSAR已被不同研究领域科学家广泛用于研究

收稿日期：2011-04-14
基金项目：国家自然科学基金项目(31101348)；上海市晨光计划项目(2008CG055；2009CG50)；
江南大学食品科学与技术国家重点实验室开放课题(SKLF-KF-201106)
作者简介：彭剑秋(1987—)，女，硕士研究生，研究方向为食品营养与功能。E-mail：jjqqiu@sina.com
* 通信作者：管骁(1979—)，男，副教授，博士，研究方向为食品功能与营养。E-mail：gnxo790521@yahoo.com.cn

化合物与其生物活性的关系[1-4]。然而建立准确合理的QSAR模型目前仍存在一些问题，因为生物的复杂多样性使得对分子结构进行简单有效的表征是QSAR研究的瓶颈。在多肽的定量构效关系研究中，普遍采用组成多肽的氨基酸残基的结构参数定量描述多肽的化学结构。目前已见报道的氨基酸残基描述符有*Z*-scales[5-6]、ISA-ECI[7]、MSW-scores[2]、*T*-values[8]、MHDV[3]、ST-scale[9]、V[10]、E[11]和G-scale[12]等。*Z*-scales是从29个实验测定的物化参数中通过主成分分析提取出的信息，并涉及到分子的亲水性、立体大小和电荷性质；ISA-ECI是由分子建模并经过量子化学计算得到的表征分子的疏水性和侧链极性的信息；MS-WHIM则是由36个量化计算的参数并经主成分分析得到的主要表征分子静电势及其影响的信息；MHDV描述符只涉及非氢原子之间的距离并且与三维结构无关。这些描述符都已成功地描述了部分多肽类似物的结构，但是由于多肽分子的复杂性及较大的柔性，如何进一步完善这些参数使它们更适应多肽的结构信息仍是个任重道远的过程。

本实验在上述研究工作的基础上，通过收集20种天然氨基酸的457种物化性质参数，并按疏水性质、电性特征、氢键贡献和立体特征进行分类，对它们分别进行主成分分析(PCA)，得到一种新的氨基酸残基结构描述符SVHEHS(scores vector of hydrophobic, electronic, hydrogen bonds and steric properties)。用该描述符对血管紧张素转化酶(angiotensin I converting enzyme，ACE)抑制二肽、苦味二肽序列进行表征，经多元线性回归(multiple linear regression，MLR)建模，获得较好的建模结果。

1 原理与方法

1.1 SVHEHS描述符建立及肽序列表征

从AA index数据库[13]共收集20种氨基酸的457个物化性质参数变量，根据氨基酸结构信息分成疏水、电性、氢键贡献及立体特征共4类参数：其中疏水性参数119个、电性参数25个、氢键贡献参数5个、立体参数308个。分别对4类参数变量进行主成分分析[14]以剔除原始数据矩阵中的噪音信息。由统计分析可知，每类物化性质变量的前2、4、2、5个主成分可分别解释原始数据矩阵75.93%、76.15%、77.15%和74.56%的方差。从而认为这13个分类主成分已经能够表征各自原始数据矩阵中绝大多数有用信息，所以可用这13个主成分得分矢量替代原始变量矩阵，并称该得分矢量为SVHEHS。其中SVHEHS 1～2代表氨基酸的疏水性特征，SVHEHS 3～6代表氨基酸的电性特征，SVHEHS 7～8代表氨基酸的氢键贡献参数特征，SVHEHS 9～13代表氨基酸的立体特征。表1所示为20种天然氨基酸的SVHEHS值。每个肽的结构可根据其氨基酸残基序列用SVHEHS得分矢量表征。如一个二肽分子可得到13 × 2=26个变量，具有*n*个氨基酸残基的肽，其一级序列结构可用13 × *n*个SVHEHS变量表征。主成分分析用SPSS 18.0软件完成。

表1 20种天然氨基酸的SVHEHS描述符

Table 1 SVHEHS descriptors for 20 natural amino acids

氨基酸	得分矢量SVHEHS												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Ala	0.65	－5.37	－1.61	－1.82	－0.46	0.44	－1.89	0	2.53	9.12	9.91	－0.8	－4
Arg	－11.92	6.82	2.82	4.48	－4.08	－0.22	4.29	0.05	－0.35	3.88	－5.45	1.8	6.41
Asn	－8.26	－0.55	2.36	－0.28	0.15	0.03	1.69	－0.19	－11.37	－0.99	－2.85	－5.03	1.89
Asp	－9.56	－0.88	6.78	－2.49	2.45	0.26	1.21	－0.15	－12.37	5.21	－2.71	－1.19	－2.07
Cys	7.08	－4.93	0.47	2.48	3.63	－0.59	－1.65	1.54	2.03	－9.24	－3.27	－5.5	－10.81
Gln	－8.41	1.03	1.05	0.32	－0.44	1	1.88	0.82	－0.95	5.3	－4.69	0.99	－0.35
Glu	－9.61	0.43	3.77	－4.56	－0.69	0.64	1.38	1.24	－2.57	16.03	－0.98	1.27	－3.87
Gly	－2.46	－7.21	－1.4	－4	－2.17	0.37	－2.1	－0.74	－18.02	－6.73	10	－7.69	2.82
His	－3.21	1.47	1.67	2.42	－0.99	－1.41	0.42	0.14	2.14	0.29	－7.66	－3.62	－0.64
Ile	12.46	－0.57	－3.79	－1.44	－0.52	0.62	－1.08	－0.73	13.53	－5.41	5.26	4.05	1.36
Leu	11.27	－0.36	－3.3	－1.27	－0.26	0.9	－1.01	－1.75	12.1	4.57	10.13	1.3	1.18
Lys	－10.97	3.46	0.56	2.01	－4.51	0.36	2.13	－1.62	－2.73	10.16	－2.99	0.33	6.87
Met	7.49	－0.88	－1.24	1.05	1.46	1.08	－0.97	－1.06	13.35	3.09	－3.33	－1.81	－4.09
Phe	12.35	2.81	－0.94	2.4	2.35	0.43	－1.36	－0.27	11.09	－4.08	－0.9	0.38	0.7
Pro	－1.97	－1.3	－1.76	－1.08	0.39	－6.4	－1.62	1.75	－18.4	－5.85	－2.14	14.8	－4.45
Ser	－5.42	－3.62	0.73	－0.49	0.98	－0.21	－0.12	0.68	－10.47	－2.03	4.7	－1.8	0.95
Thr	－3.12	－2.47	－0.05	0.02	0.26	－0.07	0.17	0.07	－3.37	－4.17	3.58	0.19	0.68
Trp	10.99	8.5	－1.93	2.28	1.83	1.51	－0.22	－1.38	9.84	－4.52	－10.5	－1.44	－0.09
Tyr	3.73	6.32	－0.92	1.56	1.25	0.72	0.24	1.3	2.65	－9.55	－5.39	0.93	6.59
Val	8.88	－2.71	－3.26	－1.6	－0.62	0.54	－1.4	0.3	11.34	－5.1	9.27	2.86	0.93

1.2 MLR 建模和模型验证

采用多元线性回归方法进行 QSAR 建模，同时计算评价回归方程拟合能力的标准回归系数(R^2)和均方根误差(*RMSE*)，模型的验证采用留一交叉验证法(leave-one-out crossvalidation，LOO-CV)，外部验证系数 Q^2_{ext} 用于评估模型的预测能力。多元线性回归用 SPSS 18.0 软件完成。各系数的计算公式如下：

$$R^2=1-\frac{\sum_{i=1}^{\text{训练集}}(y_{obs}-y_{est})^2}{\sum_{i=1}^{\text{测试集}}(y_{obs}-y_{trave})^2} \quad (1)$$

$$Q^2_{LOO}=1-\frac{\sum_{i=1}^{\text{训练集}}(y_{obs}-y_{pre})^2}{\sum_{i=1}^{\text{测试集}}(y_{obs}-y_{ave})^2} \quad (2)$$

$$RMSE=\sqrt{\frac{\sum_{i=1}^{\text{训练集}}(y_{obs}-y_{est})^2}{n}} \quad (3)$$

$$Q^2_{ext}=1-\frac{\sum_{i=1}^{\text{训练集}}(y_{obs}-y_{est})^2}{\sum_{i=1}^{\text{测试集}}(y_{obs}-y_{trave})^2} \quad (4)$$

式中：y_{obs} 表示肽的活性测量值；y_{est} 表示肽活性估计值；y_{trave} 表示训练集的活性平均值；y_{pre} 表示留一交叉验证法得到的活性预测值；y_{ave} 表示整个数据集中肽活性的平均值。

交叉验证用于训练集，所得统计量 Q^2_{LOO} 体现模型的预测能力，它是对模型进行内部验证；Q^2_{ext} 为模型的外部验证统计量，也是评估模型预测能力。

2 结果与分析

2.1 血管紧张素转化酶抑制肽的 QSAR 研究

ACE 是一种含锌二肽羧酶，在血压调节中扮演重要角色，通过肾素 - 血管紧张素系统和激肽释放酶 - 激肽系统发挥作用，体内活性过高可导致高血压，因此 ACE 抑制剂便构成抑制 ACE 活性的抗高血压药物。采用 58 个 ACE 抑制二肽[7-8]作为样本(表 2)，用多肽描述符 SVHEHS 对 ACE 抑制二肽进行 QSAR 研究，肽活性值用 $\lg(1/IC_{50})$表示。二肽中的每个氨基酸残基用 SVHEHS 描述符进行定量描述。选取其中 29 个肽样本作为训练集用于构建模型(表 2)，29 个二肽就形成一个 29 行 26 列的结构描述矩阵，二肽模型的建立采用多元线性回归统计方法，通过 LOO-CV 检验模型稳定性与预测能力。其余 29 个肽用于模型的外部验证。用建立的 ACE 抑制肽模型计算其系列活性，活性的观测值和计算值见表 2。

表 2 58 个 ACE 抑制二肽序列及其活性的观测与计算值

Table 2 Sequences of 58 ACE inhibitory dipeptides and observed and calculated activities

序号	肽	测量值	估计值	序号	肽	测量值	估计值
1	VW	5.80	5.32	30	KG$^{\Delta}$	2.49	2.62
2	IW$^{\Delta}$	5.70	5.40	31	FG	2.43	2.70
3	IY	5.43	4.74	32	GS$^{\Delta}$	2.42	2.47
4	AW$^{\Delta}$	5.00	4.87	33	GV	2.34	2.38
5	RW	4.80	5.26	34	MG$^{\Delta}$	2.32	2.54
6	VY$^{\Delta}$	4.66	4.66	35	GK	2.27	3.45
7	GW	4.52	4.64	36	GE$^{\Delta}$	2.27	3.45
8	VF$^{\Delta}$	4.28	3.54	37	GT	2.24	2.32
9	AY	4.06	4.21	38	WG$^{\Delta}$	2.23	2.63
10	IP$^{\Delta}$	3.89	3.97	39	HG	2.20	2.41
11	RP	3.74	3.84	40	GQ$^{\Delta}$	2.15	3.56
12	AF$^{\Delta}$	3.72	3.09	41	GG	2.14	2.04
13	GY	3.68	3.98	42	QG$^{\Delta}$	2.13	2.34
14	AP$^{\Delta}$	3.64	3.44	43	SG	2.07	2.15
15	RF	3.64	3.48	44	LG$^{\Delta}$	2.06	2.75
16	VP$^{\Delta}$	3.38	3.90	45	GD	2.04	2.03
17	GP	3.35	3.22	46	TG$^{\Delta}$	2.00	2.32
18	GF$^{\Delta}$	3.20	2.86	47	EG	2.00	2.14
19	IF	3.03	3.62	48	DG$^{\Delta}$	1.85	1.97
20	VG$^{\Delta}$	2.96	2.72	49	PG	1.77	1.70
21	IG	2.92	2.79	50	LA$^{\Delta}$	3.51	3.31
22	GI$^{\Delta}$	2.92	2.42	51	KA	3.42	3.18
23	GM	2.85	2.66	52	RA$^{\Delta}$	3.34	3.22
24	GA$^{\Delta}$	2.70	2.60	53	YA	3.34	3.31
25	YG	2.70	2.60	54	AA$^{\Delta}$	3.21	2.83
26	GL$^{\Delta}$	2.70	2.75	55	FR	3.04	3.18
27	AG	2.60	2.27	56	HL$^{\Delta}$	2.49	2.55
28	GH$^{\Delta}$	2.51	2.17	57	DA	2.42	2.53
29	GR	2.49	2.52	58	EA$^{\Delta}$	2.00	2.70

注：Δ. 用于测试集。

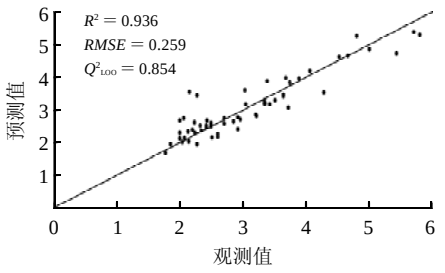

图 1 ACE 抑制肽的活性观测值和预测值关系图

Fig.1 Correlation plot for the calculated and predicted activities of ACE inhibitors

由图 1 可知，ACE 抑制肽的活性所建模型的 R^2 和 Q^2_{LOO} 分别为 0.936 和 0.854，*RMSE* 为 0.259，外部验证的相关系数为 0.737，说明实验值与模型的预测值之间有较好的相关性。

模型的 R^2 和 *RMSE* 体现模型对样本集数据的拟合能

力，R^2越大、*RMSE*越小表明模型的拟合能力越好；交叉验证相关系数和外部验证相关系数评价模型的预测能力。由图1可以看出，建立的ACE抑制肽模型对样本数据的拟合和预测能力都较好。ACE抑制二肽库是一个经典的QSAR研究用肽库，已有研究者采用不同的描述符(*Z*-scale，GRID，ISA-ECI等)及统计方法(PLS、SMR、MLR等)进行过相关研究，部分结果如表3所示，对比各个模型的相关系数、交叉验证相关系数及均方根误差发现，本实验的描述符在拟合及预测能力上都优于其他模型(较高的R^2、Q^2_{Loo}与较低的*RMSE*值，见表3)，说明用该描述符建立的模型是可靠的，可以用来对生物活性肽进行定量构效关系研究。

表3 ACE抑制二肽的QSAR模型对比

Table 3 Comparison among QSAR model for ACE inhibitory dipeptides

序号	描述符	建模方法	主成分数	R^2	Q^2_{Loo}	*RMSE*
1	*Z*-scale[15]	PLS	2	0.770	nd	nd
2	GRID[8]	PLS	1	0.744	nd	0.50
3	ISA-ECI[16]	PLS	2	0.700	nd	nd
4	MS-WHIM[2]	PLS	3	0.657	0.541	nd
5	MS-WHIM[2]	PLS	3	0.708	0.637	0.54
6	MHDV[3]	PCR	19	0.878	0.753	0.35
7	MEEV[f][17]	MLR	10	0.711	0.475	0.34
8	MEEV[b][17]	MLR	3	0.649	0.570	0.37
9	MEEV[b][17]	MLR	10	0.773	0.588	0.33
10	MEEV[b,c][17]	MLR	3	0.735	0.677	0.32
11	VSTV[b,c][18]	PLS	1	0.789	0.767	0.46
12	SSIA-AM1[19]	PLS	1	0.769	0.699	0.49
13	SSIA-PM3[19]	PLS	4	0.789	0.773	0.47
14	SSIA-HF[19]	PLS	2	0.797	0.760	0.46
15	SSIA-DFT[19]	PLS	2	0.734	0.678	0.52
16	SZOTT[20]	PLS	2	0.878	0.753	0.33
17	T-scale[21]	PLS	2	0.845	0.786	0.39
18	VSW[22]	PLS	2	0.868	0.784	0.37
19	VSW[22]	PLSd	1	0.861	0.835	0.38
20	SVHEHS	MLR	8	0.936	0.854	0.259

注：nd.没有测定；b.用SMR(stepwise multivariate regression)-MLR方法建立的模型；c.该模型只有47个样本；d.采用了SMR方法。下同。

剔除对活性的不显著($P > 0.1$)影响变量，ACE抑制二肽的活性(y)拟合模型如式(1)所示。

$$y = 3.114 + 0.026S_9 + 0.045S_{13} + 0.235S_{15} - 0.261S_{16} + 0.375S_{19} + 0.393S_{21} - 0.068S_{22} - 0.169S_{26} \quad (1)$$

其中，S_1～S_{13}表示N端氨基酸残基的性质，S_{14}～S_{26}表示C端残基的性质(一一对应于S_1～S_{13})。比较各标准化偏回归系数值的大小可知，ACE抑制二肽的活性主要与S_{15}、S_{19}、S_{21}呈正相关，与S_{16}呈负相关，亦即肽活性受C末端氨基酸残基的性质影响，主要为氢键贡献、疏水性质以及立体结构的影响，其中疏水性质对活性的影响为正效应，表明C端疏水性氨基酸残基有利于肽的活性；而立体结构对活性的影响为负效应，这与前人的研究结果是相符的。

2.2 苦味二肽的定量构效关系研究

苦味作为一种味觉感受，能防止人和有机体受到有毒物质的伤害。研究表明，味觉信号在味觉受体细胞中的传导包括一系列复杂的过程，该过程由耦合G蛋白的受体介导完成[5]。48个苦味二肽是QSAR研究中验证描述符有效性的经典样本。用文献报道的48个苦味二肽[5](表4)和自建的多肽描述符，对苦味二肽进行QSAR研究。二肽中的每个氨基酸用SVHEHS描述符进行定量描述，其中24个肽用于构建模型(表4)，这样24个二肽就形成一个24行26列的结构描述矩阵，二肽模型的建立采用多元线性回归统计方法，通过LOO-CV检验模型稳定性与预测能力。其余24个肽用于模型的外部验证。用建立的苦味二肽模型计算其系列活性，活性的观测值和计算值如表4所示。活性值用$\lg(1/IC_{50})$表示。

表4 48个苦味二肽序列及其活性的观测与计算值

Table 4 Sequences of 48 bitter dipeptides and observed and calculated activities

序号	肽	测量值	估计值	序号	肽	测量值	估计值
1	GV	1.13	1.50	25	II	2.26	2.50
2	GL$^{\Delta}$	1.68	1.66	26	IP$^{\Delta}$	2.40	2.18
3	GI	1.70	1.79	27	IW	3.05	2.66
4	GP$^{\Delta}$	1.35	1.47	28	IN$^{\Delta}$	1.49	1.33
5	GF	1.80	1.82	29	ID	1.37	1.38
6	GW$^{\Delta}$	1.89	1.95	30	IQ$^{\Delta}$	1.49	1.58
7	GY	1.77	1.62	31	IE	1.37	1.51
8	AV$^{\Delta}$	1.16	1.29	32	IK$^{\Delta}$	1.65	1.55
9	AL	1.70	1.45	33	IS	1.49	1.41
10	AF$^{\Delta}$	1.72	1.61	34	IT$^{\Delta}$	1.49	1.63
11	VG	1.19	0.91	35	PA	1.32	1.28
12	VA$^{\Delta}$	1.16	1.34	36	PL$^{\Delta}$	2.22	2.02
13	VV	1.71	1.92	37	PI	2.33	2.15
14	VL$^{\Delta}$	2.00	2.08	38	PY$^{\Delta}$	1.80	1.98
15	LG	1.72	1.2	39	PF	2.80	2.18
16	LA$^{\Delta}$	1.72	1.63	40	FG$^{\Delta}$	1.77	1.81
17	LL	2.35	2.37	41	FL	2.87	2.98
18	LF$^{\Delta}$	2.75	2.53	42	FP$^{\Delta}$	2.70	2.79
19	LW	3.40	2.67	43	FF	3.10	3.14
20	LY$^{\Delta}$	2.46	2.37	44	FY$^{\Delta}$	3.13	2.94
21	IG	1.68	1.20	45	WE	1.56	2.68
22	IA$^{\Delta}$	1.68	1.63	46	WW$^{\Delta}$	3.60	3.83
23	IV	2.05	2.21	47	YL	2.40	3.48
24	IL$^{\Delta}$	2.26	2.37	48	SL$^{\Delta}$	1.49	1.54

注：$^{\Delta}$.用于训练集。

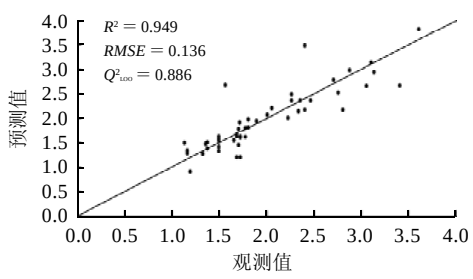

图2 苦味二肽的活性观测值和预测值关系图

Fig.2 Correlation plot for the calculated and predicted activities of bitter dipeptides

由图2可知，苦味二肽所建模型的R^2和Q^2_{Loo}分别为0.949和0.886，*RMSE*为0.136，外部验证的相关系数为0.543。可以看出建立的苦味二肽模型对样本数据的拟合和内部预测能力较好，但外部预测能力稍有不足。

表5 苦味二肽的QSAR模型对比

Table 5 Comparison among QSAR models for bitter dipeptides

序号	描述符	建模方法	主成分数	R^2	Q^2_{Loo}	*RMSE*
1	Z-scale[15]	PLS	2	0.824	nd	0.26
2	GRID[8]	PLS	1	nd	0.780	nd
3	ISA-ECI[16]	PLS	2	0.847	nd	nd
4	MS-WHIM[2]	PLS	3	0.704	0.633	nd
5	MS-WHIM[2]	PLS	3	0.754	0.710	0.32
6	MHDY[17]	PVR	10	0.919	0.864	0.18
7	MEEV[b][17]	MLR	10	0.711	0.475	0.34
8	MEEV[b][17]	MLR	3	0.649	0.570	0.37
9	MEEV[b,c][17]	MLR	10	0.773	0.588	0.33
10	MEEV[b,c][17]	MLR	3	0.735	0.677	0.32
11	SSIA-AM1[19]	PLS	1	0.850	0.837	0.25
12	SSIA-PM3[19]	PLS	4	0.888	0.829	0.22
13	SSIA-HF[19]	PLS	2	0.844	0.798	0.25
14	SSIA-DFT[19]	PLS	2	0.856	0.741	0.24
15	SZOTT[20]	PLS	2	0.908	0.736	0.20
16	VSW[22]	PLS	2	0.868	0.696	0.24
17	VSW[22]	PLSd	2	0.873	0.751	0.23
18	SVHEHS	MLR	6	0.949	0.886	0.136

将该模型与已有的文献报道值相比较，结果如表5所示。对比各个模型的相关系数、交叉验证相关系数及均方根误差发现，本研究建立模型的相关统计量均优于大部分文献报道(较高的R^2、Q^2_{Loo}与较低的*RMSE*值，见表5)，表明本研究建立的描述符以及模型统计方法是可靠的，可用来对苦味肽进行定量构效关系研究。

剔除对活性的不显著($P > 0.1$)影响变量，本实验建立的苦味二肽的活性(y)模型如式(2)所示。

$$y=1.542+0.168S_2-0.301S_7+0.067S_{13}+0.043S_{14}+0.044S_{15}+0.031S_{25} \quad (2)$$

比较各项标准化偏回归系数可知，苦味二肽的活性主要与S_2呈正相关，与S_7呈负相关，即对应于N端氨基酸残基疏水性以及氢键特性的影响，该位置氨基酸残基疏水性越大，氢键贡献越小对苦味二肽的活性越有利。

从以上论述不难看出，采用氨基酸描述符SVHEHS进行生物活性肽的QSAR研究，不仅能提供未知肽的较好的活性预测能力，而且模型中各自变量均具有较明确的物理化学含义，这无疑也有助于新的活性肽的分子定向设计。

3 结 论

实验通过收集AA index中各种氨基酸的多种理化参数，通过主成分分析方法建立了一种新的氨基酸结构描述符SVHEHS，此描述符包含了氨基酸的疏水、电性、氢键贡献、立体结构等信息，能对组成肽的氨基酸残基进行定量结构描述。相比其他现有的氨基酸描述符，SVHEHS包含的氨基酸物化性质信息更加丰富，尤其是立体结构方面的信息远远多于其他描述符的原始数据矩阵，无疑更有助于对肽的结构更加准确全面的描述，从而有利于提高模型的拟合度与预测能力。将此描述符应用于ACE抑制二肽、苦味二肽的QSAR研究中，均建立了预测能力较好的统计模型，而且模型的相关统计量较前期文献报道值更优。可见，此描述符应用于肽的QSAR研究，能进行肽分子活性的准确预测，并有望成为多肽QSAR研究中一个有用的结构表征方法；且该描述符物理化学意义明确，包含氨基酸信息较丰富，对正确指导新的活性肽的分子设计也能提供实际指导意义。

参考文献：

[1] KIDERA A, KONISHI Y, OKA M, et al. A statistical analysis of the physical properties of the 20 naturally occuring amino acids[J]. J Protein Chem, 1985, 4(1): 23-55.

[2] ZALIANI A, GANCIA E. MS-WHIM scores for amino acids: a new 3D description for peptide QSAR and QSPR studies[J]. J Chem Inf Comput Sci, 1999, 39(3): 525-533.

[3] LIU Shushen, YIN Chunsheng, CAI Shaoxi, et al. A novel MHDV descriptor for dipeptide QSAR studies[J]. J Chin Chem Soc, 2001, 48 (2): 253-260.

[4] RAYCHAUDHURY C, BANERJEE A, BAG P, et al. Topological shape and size of peptides: identification of potential allele specific helper T cell antigenic sites[J]. J Chem Inf Comput Sci, 1999, 39(2): 248-254.

[5] HELLBERG S, SJOESTROEM M, SKAGERBERG B, et a1. Peptide quantitative structure-activity relationships, a multivariate approach[J]. J Med Chem, 1987, 30(7): 1126-1135.

[6] JONSSON J, ERIKSSON L, HELLBERG S, et a1. Multivariate parametrization of 55 coded and non-coded amino acids[J]. Quant Struct-Act Relat, 1989, 8(3): 204-209.

[7] ELIZABETH R C, WILLIAM J D. Amino acid side chain descriptors for quantitative structure-activity ralationship of peptide analogues[J]. J Med Chem, 1995, 38(4): 2705-2713.

[8] COCCHI M, JOHANSSON E. Amino acids characterization by GRID and multivariate data analysis[J]. Quant Struct Act Relat, 1993, 12(1): 1-8.

[9] YANG Li, SHU Mao, MA Kaiwang, et al. ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues [J]. Amino Acids, 2010, 38: 805-816.

[10] LONG Haixia, WANG Yuanqiang, LIN Yong, et al. QSAR study on ACE inhibitors by using OSC-PLS algorithm[J]. Journal of the Chinese Chemical Society, 2010, 57(3A): 417-422.

[11] YIN Jiajian, DIAO Yuanbo, WEN Zhining, et al. Studying peptides biological activities based on multidimensional descriptors using sup-

port vector regression[J]. Int J Pept Res Ther, 2010, 16(2): 111-121.

[12] WANG Xiaoyu, DIAO Yuanbo, WEN Zhining, et al. QSAR study on angiotensin-converting enzyme inhibitor oligopeptides based on a novel set of sequence information descriptors[J]. J Mol Model, 2010, 17(7): 1599-1606.

[13] KAWASHIMA S, KANEHISA M. AAindex: amino acid index database [J]. Nucleic Acids Res, 2000, 28(1): 374.

[14] KIM D, LEE I B. Process monitoring based on probabilistic PCA[J]. Chemometrics and Intelligent Laboratory Systems, 2003, 67(2): 109-123.

[15] 梁桂兆, 周鹏, 周原, 等. 一组新氨基酸描述子用于肽定量构效关系研究[J]. 化学学报, 2006, 64(5): 393-396.

[16] LIANG Guizhao, YANG Li, KANG Lifang, et al. Using multidimensional patterns of amino acid attributes for QSAR analysis of peptides[J]. Amino Acids, 2009, 37(4): 583-591.

[17] 周鹏, 田菲菲, 李波, 等. 一种基于三维原子场相互作用矢量的新型氨基酸结构信息描述子[J]. 科学通报, 2006, 51(1): 34-39.

[18] 杨善斌, 夏之宁, 舒茂, 等. 氨基酸描述子 VHSEH 用于多肽定量序效建模研究[J]. 高等学校化学学报, 2008, 29(11): 2213-2217.

[19] ZHOU Peng, ZHOU Yuan, WU Shirong, et al. A new descriptor of amino acids based on the three-dimensional vector of atomic interaction field[J]. Chin Sci Bull, 2006, 51(5): 524-529.

[20] LIANG Guizhao, ZHOU Peng, ZHOU Yuan, et al. New descriptors of amino acids and their applications to peptide quantitative structure activity relationship[J]. Acta Chim Sin, 2006, 64(5): 393-396.

[21] TIAN Feifei, ZHOU Peng, LI Zhiliang. T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides[J]. Journal of Molecular Structure, 2007, 830: 106-115.

[22] TONG Jianbo, LIU Shuling, ZHOU Peng, et al. A novel descriptor of amino acids and its application in peptide QSAR[J]. Journal of Theoretical Biology, 2008, 253(1): 90-97.