Hu Mei[1,2]
Zhi H. Liao[3]
Yuan Zhou[1]
Shengshi Z. Li[1,2]
[1] *College of Chemistry and Chemical Engineering, Chongqing University, Chongqing, 400044 People's Republic of China*

[2] *Key Laboratory of Biomedical Engineering of both Educational Ministry and Chongqing City, Chongqing, 400044 People's Republic of China*

[3] *School of Life Sciences, Southwest China Normal University, Chongqing, 400715 People's Republic of China*

A New Set of Amino Acid Descriptors and Its Application in Peptide QSARs

Received 14 September 2004;
revised 11 November 2004;
accepted 18 November 2004
Published online 13 May 2005 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/bip.20296

Abstract: *In this work, a new set of amino acid descriptors, i.e., VHSE (principal components score Vectors of Hydrophobic, Steric, and Electronic properties), is derived from principal components analysis (PCA) on independent families of 18 hydrophobic properties, 17 steric properties, and 15 electronic properties, respectively, which are included in total 50 physicochemical variables of 20 coded amino acids. Using the stepwise multiple regression (SMR) method combined with partial least squares (PLS), the VHSE scales are then applied to QSAR studies of bitter-tasting dipeptides (BTD), angiotensin-converting enzyme (ACE) inhibitors, and bradykinin-potentiating pentapeptides (BPP). To validate the predictive power of resulting models, external validation are also performed. A comparison of the results to those obtained with z scores and other two-dimensional (2D) or three-dimensional(3D) descriptors shows that the VHSE scales are comparable for parameterizing the structural variability of the peptide series.* © 2005 Wiley Periodicals, Inc. Biopoly 80: 775–786, 2005
This article was originally published online as an accepted preprint. The Published Online date corresponds to the preprint version. You can request a copy of the preprint by emailing the Biopolymers editorial office at biopolymers@wiley.com

Correspondence to: Shengshi Zhiliang Li; e-mail: lisz2662@163.com

Contract grant sponsor: State Chuihui Project Fund (SCPF), Fok Ying-Tung Educational Foundation (FYTF), and Chongqing University Innovation Fund (CUIF)
Contract grant number: 98-3-8 (SCPF), 98-7-6 (FYTF), and 03-5-6 (CUIF)

Biopolymers (Peptide Science), Vol. 80, 775–786 (2005)

Keywords: *amino acids; peptide; quantitative structure–activity relationship; principal components analysis; partial least squares; stepwise multiple regression*

INTRODUCTION

Peptides are very important in all living systems. They act as hormones, enzyme inhibitors, antibodies, olfaction and taste receptors, antimicrobial compounds or agents, and other biological functions. Hence, they have attracted considerable pharmacological interest in recent years.[1,2]

With the development of a peptide library, thousands of different peptides are designed, synthesized, and then subjected to a range of screening procedures and biological assays. To be effectively used, the biological data can be analyzed with multivariate quantitative structure–activity relationships (QSARs). For the properties of peptides, a precise amino acid sequence is required for a particular function or biological activity. A QSAR model will then indicate how the change in peptide sequence is correlated with the variation in biological activity and how to modify the sequence to achieve the improved activity. The basic assumption in QSAR is that the biological activity within a set of compounds is related to the structural variation of the compounds, i.e., the biological activity can be modeled as a function of molecular structure. In this context, quantitative amino acid descriptor variables have shown to be valuable.

Since the pioneering work of Sneath,[3] who derived amino acid descriptor variables from physicochemical semiqualitative data for the 20 coded amino acids and used them in a quantitative sequence–activity model (QSAM) analysis of oxytocin–vasopressine analogues, a number of quantitative amino acid descriptor variables have been proposed for the 20 coded amino acids.[4–10] A recent development in the QSAR field is the use of amino acid "*z* scores" obtained by principal components analysis (PCA) based on 29 physicochemical variables of 20 coded amino acids.[5] The three resulting principal components (PCs), so-called principal properties, are linear combinations of the primary parameters and primarily represent hydrophobicity, side-chain bulk, and electronic properties of amino acids. The *z* scores have proven to be useful for modeling a number of biological activities of small peptides as a function of the *z* scores. By using only 12 physicochemical variables, Jonsson et al.[7] took a first step toward expanding these scales to encompass 35 noncoded amino acids. More recently, the same approach was expanded to a larger set of amino acids (20 coded + 67 noncoded) and more parameters. Application of PCA resulted in a set of five orthogonal variables termed *zz* scores, of which the first three corresponded to the original *z* scores. The *zz* scores were applied to two peptide data sets, elastase substrates and neurotensin analogues, with good results obtained.[9]

However, all the amino acid descriptors mentioned above are derived from the PCA of the data matrix, which comprises hydrophobic, steric, and electronic properties of the amino acids. Thus each principal component is a linear combination of the properties of different categories and still limited to definite physiochemical meanings. One way to solve the problem is to apply PCA separately according to different categories of properties. In this article, a new set of amino acid descriptors, i.e., VHSE (principal components score V̲ectors of H̲ydrophobic, S̲teric, and E̲lectronic properties), is derived from the PCA individually on 18 hydrophobic properties, 17 steric properties, and 15 electronic properties, as three independent families included a total 50 physicochemical variables of 20 coded amino acids. As a new set of amino acid descriptors, VHSE is of relatively definite physicochemical meaning, easy interpretation and more information contained in comparison with z scales. VHSE scales are then applied to QSARs of three sets of peptides, and better results are obtained in comparison with those obtained with *z* scores and other two-dimensional (2D) or three-dimensional (3D) descriptors.

METHODS

Principal Components Analysis

First, a total 50 physicochemical variables of 20 coded amino acids were divided into three groups according to different categories of properties; 18 hydrophobic properties, 17 steric properties, and 15 electronic properties were obtained respectively (Table I). Then, these three categories of property matrices (available as supporting information) were analyzed by PCA separately. For the matrices of hydrophobic, steric, and electronic properties, the first 2, 2, and 4 principal components accounted for 74.33, 78.68, and 77.97% variance of original data matrices, respectively. That is to say, the most information in the three original matrices can be replaced by the first 2, 2, and 4 principal component scores, respectively. Hence, the hydrophobic, steric, and electronic properties of 20 coded amino acids can

Table I Variables Used to Characterize the 20 Coded Amino Acids

Variable No.	Description	Accession ID[a]
	Hydrophobic Property	
1	Retention coefficient in TFA	BROC820101
2	Free energy of solution in water	CHAM820102
3	Solvation free energy	EISD860101
4	Melting point	FASG760102
5	Number of hydrogen-bond donors	FAUJ880109
6	Number of full nonbonding orbitals	FAUJ880110
7	Partition energy	GUYH850101
8	Hydration number	HOPA770101
9	Retention coefficient in high performance liquid chromatography (HPLC), pH 7.4	MEEJ800101
10	Retention coefficient in HPLC, pH 2.1	MEEJ800102
11	Partition coefficient in thin-layer chromatography	PLIV810101
12	Retention coefficient at pH 2	GUOD860101
13	R_f for 1-N-(4-nitrobenzofurazono)-amino acids in ethyl acetate/pyridine/water	—[b]
14	dG of tranfer from organic solvent to water	—
15	Hydration potential or free energy of tranfer from vapor phase to water	—
16	R_f salt chromatoghaphy	—
17	Log D, partition coefficient at pH 7.1 for acetylamide derivatives of amino acids in octanol/water	—
18	dG $= RT \ln f, f =$ fraction buried/accessible amino acids in 22 proteins	—
	Steric Property	
19	Average volume of buried residue	CHOC750101
20	Residue accessible surface area in tripeptide	CHOC760101
21	Graph shape index	FAUJ880101
22	Normalized van der Waals volume	FAUJ880103
23	STERIMOL length of the side chain	FAUJ880104
24	STERIMOL minimum width of the side chain	FAUJ880105
25	STERIMOL maximum width of the side chain	FAUJ880106
26	Average accessible surface area	JANJ780101
27	Distance between Cα and centroid of side chain	LEVM760102
28	Side-chain angle θ	LEVM760103
29	side chain torsion angle ϕ	LEVM760104
30	Radius of gyration of side chain	LEVM760105
31	van der Waals parameter R_0	LEVM760106
32	van der Waals parameter epsilon	LEVM760107
33	Refractivity	MCMT640101
34	Value of θ (i)	RACS820113
35	Substituent van der Waals volume	—
	Electronic Property	
36	αCH chemical shifts	ANDN920101
37	αNH chemical shifts	BUNA790101
38	A parameter of charge transfer capability	CHAM830107
39	A parameter of charge transfer donor capability	CHAM830108
40	Nuclear magnetic resonance (NMR) chemical shift of α carbon	FAUJ880107
41	Localized electrical effect	FAUJ880108
42	Positive charge	FAUJ880111
43	Negative charge	FAUJ880112
44	Polarity	GRAR740102
45	Net charge	KLEP840101
46	Amphiphilicity index	MITS020101
47	Isoelectric point	ZIMJ680104
48	Electron-ion interaction potential values	COSI940101
49	pKNH$_2$(NH$_2$ on C_alpha)	FASG760104
50	pKCOOH(COOH on C_alpha)	FASG760105

[a] The accession ID of Amino Acid Index Database[1–3].

[b] The supplementary material of S. Hellberg et al., 1987[5].

Table II VHSE Scales for 20 Coded Amino Acids

Amino Acids	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

be expressed by the 8 principal component scores with less information lost. Here, we tentatively call these 8 score vectors VHSE (principal components score Vectors of Hydrophobic, Steric, and Electronic properties). For amino acids, VHSE_1 and VHSE_2 are related to hydrophobic properties, VHSE_3 and VHSE_4 to steric properties, and VHSE_5–VHSE_8 to electronic properties (Table II). Coefficient matrices of principal component score are also available as supporting information.

Structural Description and Variable Selection

For a set of peptide analogues, the chemical structure can now be quantified by describing each varied amino acid position with 8 VHSE variables. So a chemical structure of a dipeptide, for example, can be described by 2×8 variables. Thus, a set of peptide analogues varied in m positions can be described by $m \times 8$ variables.

For a QSAR data set, not all the structural descriptors are relevant to biological activity. So those redundant descriptors should be deleted from model in order to promote model robustness and its predictive capability especially when the number of variables is very large. Several variable selection methods including stepwise multiple regression (SMR), genetic algorithms (GAs),[14–16] generalized simulated annealing,[17,18] and evolutionary algorithms[19,20] have been widely used to eliminate irrelevant variables. If the number of variables is not too large, the classic SMR method should be the best because it is easy to implement and it is less time-consuming.

Partial Least Square

Partial least square (PLS) is the most used latent regression method for relating two data matrices, X and Y, by a linear multivariate model. Compared to those traditional methods, PLS can analyze the data with many, noisy, collinear, and even incomplete variables in both X and Y. The desirable property of PLS is that the precision of the model parameters improves with the increasing number of relevant variables and observations. The PLS regression algorithm consists of outer relations (X and Y block individually) and an inner relation linking both blocks:

$$x_{ik} = \sum_{a=1}^{A} t_{ia} p_{ak} + e_{ik} \tag{1}$$

$$y_{im} = \sum_{a=1}^{A} u_{ia} c_{am} + g_{im} \tag{2}$$

The t and u latent variables are correlated through the inner relation given below, which leads to the estimation of the y from the x:

$$\hat{u} = bt \tag{3}$$

Computational and other details are given in Refs. 21 and 22.

Table III Results of Partial Least-Square Regression on the Variables Selected by the Stepwise Multiple Regression

Model	b0	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10	b11	b12	b13	b14	b15	b16	*A*	R^2	Q^2
1	1.952										0.358							1[a]	0.337	0.302
2	2.043		0.342								0.317							1	0.543	0.502
3	1.853		0.319							0.318	0.296							1	0.756	0.731
4	1.443	0.383	0.246							0.464	0.193							3	0.881	0.843
5	1.297	0.616	0.224			−0.457				0.464	0.205							4	0.899	0.855
6	1.387	0.422	0.213			−0.005			0.019	0.494	0.177							2	0.844	0.796
7	1.466	0.358	0.223			0.088			0.024	0.469	0.215						0.121	3	0.894	0.810
8	1.457	0.370	0.219			0.071			0.025	0.496	0.199		0.054				0.142	3	0.910	0.816
9	1.634	0.214	0.215			0.244			0.020	0.311	0.233		0.024	0.234			0.036	2	0.862	0.797
10	1.706	0.159	0.176	0.088		0.166			0.016	0.294	0.229		0.019	0.239			0.022	2	0.855	0.802
11	1.724	0.168	0.181	0.094	0.040	0.195			0.016	0.283	0.224		0.021	0.232			0.019	2	0.859	0.805
12	1.620	0.296	0.180	0.112	0.057	0.059			0.024	0.400	0.083	0.055	0.074	0.187			0.118	3	0.909	0.816
13	1.938	0.149	0.154	0.090	0.006	0.248	0.135		0.014	0.153	0.142	0.121	0.024	0.136			−0.003	1	0.784	0.732
14	1.500	0.353	0.118	0.086	0.026	0.044	−0.039	−0.33	0.019	0.391	0.107	0.069	0.081	0.208			0.121	3	0.907	0.760
15	1.940	0.124	0.127	0.074	0.005	0.206	0.112	−0.086	0.011	0.127	0.118	0.101	0.020	0.113		−0.107	−0.002	1	0.775	0.727
16	1.456	0.359	0.128	0.086	0.040	0.018	−0.043	−0.024	0.019	0.331	0.129	0.061	0.152	0.154	−0.124	−0.104	0.077	4	0.919	0.804

[a] Not significant according to cross-validation.

Table IV The Observed and Calculated Biological Activities of Bitter-Tasting Dipeptides

No.	Peptide	Obsd	Calcd	Resd	No.	Peptide	Obsd	Calcd	Resd	No.	Peptid.	Obsd	Calcd	Resd
1	GV	1.13	1.06	0.07	17	LL	2.35	2.53	−0.18	33	IS	1.49	1.41	0.08
2	GL	1.68	1.53	0.15	18	LF	2.75	2.81	−0.06	34	IT	1.49	1.64	−0.15
3	GI	1.70	1.46	0.24	19	LW	3.40	3.05	0.35	35	PA	1.32	1.32	0.00
4	GP	1.35	1.56	−0.21	20	LY	2.46	2.56	−0.10	36	PL	2.22	2.15	0.07
5	GF	1.80	1.81	−0.01	21	IG	1.68	1.44	0.24	37	PI	2.33	2.07	0.26
6	GW	1.89	2.05	−0.16	22	IA	1.68	1.65	0.03	38	PY	1.80	2.17	−0.37
7	GY	1.77	1.56	0.21	23	IV	2.05	2.00	0.05	39	PF	2.80	2.43	0.37
8	AV	1.16	1.35	−0.19	24	IL	2.26	2.47	−0.21	40	FG	1.77	1.69	0.08
9	AL	1.70	1.82	−0.12	25	II	2.26	2.40	−0.14	41	FL	2.87	2.72	0.15
10	AF	1.72	2.10	−0.38	26	IP	2.40	2.50	−0.10	42	FP	2.70	2.75	−0.05
11	VG	1.19	1.08	0.11	27	IW	3.05	2.99	0.06	43	FF	3.10	3.00	0.10
12	VA	1.16	1.29	−0.13	28	IN	1.49	1.35	0.14	44	FY	3.13	2.75	0.38
13	VV	1.71	1.64	0.07	29	ID	1.37	1.56	−0.19	45	WE	1.56	2.01	−0.45
14	VL	2.00	2.11	−0.11	30	IQ	1.49	1.41	0.08	46	WW	3.60	3.57	0.03
15	LG	1.72	1.50	0.22	31	IE	1.37	1.43	−0.06	47	YL	2.40	2.61	−0.21
16	LA	1.72	1.71	0.01	32	IK	1.65	1.53	0.12	48	SL	1.49	1.56	−0.07

Software Used

SIMCA-P 10.0 was used for the PLS analysis. The variables of the *X* matrix were mean centered and scaled to unit variants prior to PLS analysis. Cross-validation was done by the leave one out procedure. SPSS 10.0 was used for stepwise multiple variable selection.

RESULTS AND DISCUSSION

Bitter-Tasting Dipeptides

A set of 48 dipeptides with reported bitter-tasting thresholds was obtained from the report of Collantes et al.[8] The biological activity was modeled as negative logarithm of the concentration. At first, amino acids in the two positions of dipeptide were quantified by 16 VHSE descriptors, which were denoted by V_1–V_{16} in sequence. Then, the 16 variables were screened by SMR. At last, the variables selected in each step together with the response were modeled by PLS. Variables selected in each step and correspondent results of PLS are show in Table III.

From Table III, we can see that all models established, except for 1 and 2, have satisfied values of R^2 and Q^2. However, It has been pointed out in recent literature that Q^2 is an insufficient parameter to deter-

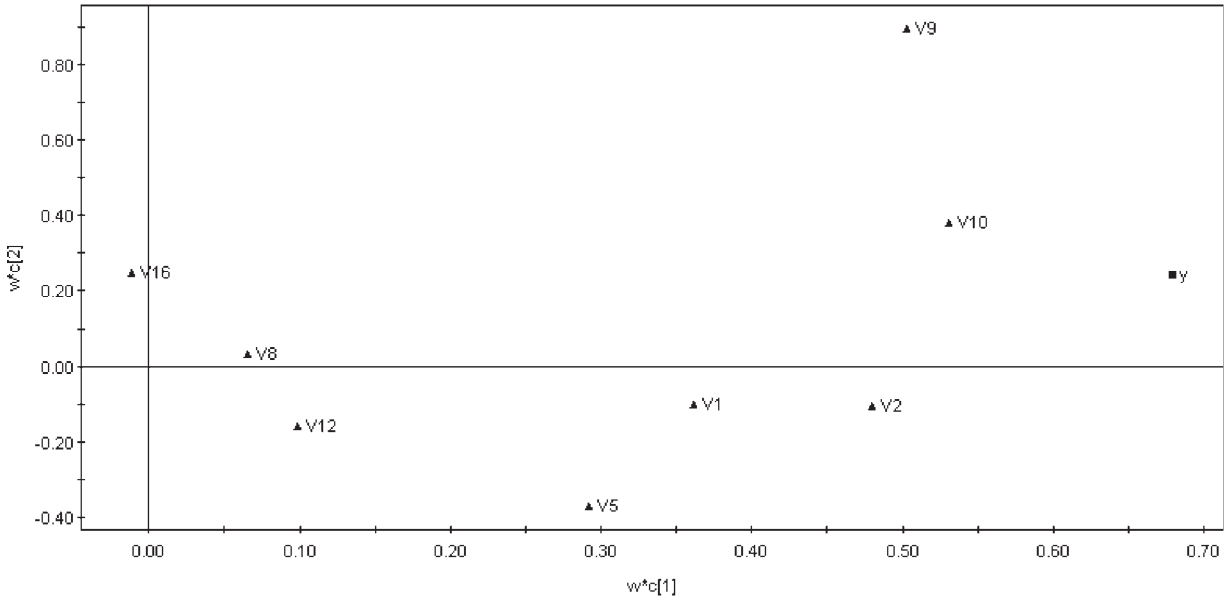

FIGURE 1 The PLS loading plot for bitter-tasting dipeptides.

Table V Statistical Parameters of PLS Model in QSAR of Bitter-Tasting Dipeptides[a]

Model	Descriptor	Sum of Descriptors	*A*	R^2_{cum}	Q^2_{cum}	*RSD*	*RMSEE*
1	Z scale[26]	6	2	0.824	nd	0.26	nd
2	ISA-ECI[8]	4	2	0.848	nd	0.24	nd
3	MS-WHIM[10]	6	3	0.754	0.710	nd	nd
4	VHSE	8	3	0.910	0.816	0.19	0.20

[a] R^2_{cum}: cumulative multiple correlation coefficient; Q^2_{cum}: cumulative cross-validated R^2; *A*: number of components in a PLS model; *RSD*: residual standard deviation; *RMSEE*: root mean square error of estimation for the training set; nd: not determined.

mine actual predictive power of a QSAR model and that external validation is required.[23–25] Corresponding criteria for a QSAR model to have high predictive power are also proposed as follows[23]:

$$Q^2 > 0.5$$

$$R^2 > 0.6$$

$$\frac{(R^2 - R_0^2)}{R^2} < 0.1 \text{ or } \frac{(R^2 - R_0'^2)}{R^2} < 0.1 \qquad (6)$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15 \qquad (7)$$

where Q^2 is the cross-validation R^2 for the training set, R^2 is the coefficients of determination for the

Table VI The Observed and Calculated Activities of ACE Inhibitors

No.	Peptide	Obsd	Calcd	Resd	No.	Peptide	Obsd	Calcd	Resd
1	VW	5.80	4.95	0.85	30	KG	2.49	2.65	−0.16
2	IW	5.70	4.95	0.75	31	FG	2.43	2.45	−0.02
3	IY	5.43	4.45	0.98	32	GS	2.42	2.16	0.26
4	AW	5.00	4.74	0.26	33	GV	2.34	2.49	−0.15
5	RW	4.80	4.78	0.02	34	MG	2.32	2.42	−0.10
6	VY	4.66	4.44	0.22	35	GK	2.27	2.56	−0.29
7	GW	4.52	4.11	0.41	36	GE	2.27	2.30	−0.03
8	VF	4.28	4.51	−0.23	37	GT	2.24	2.38	−0.14
9	AY	4.06	4.24	−0.18	38	WG	2.23	2.43	−0.20
10	IP	3.89	3.36	0.53	39	HG	2.20	2.46	−0.26
11	RP	3.74	3.20	0.54	40	GQ	2.15	2.40	−0.25
12	AF	3.72	4.31	−0.59	41	GG	2.14	1.88	0.26
13	GY	3.68	3.61	0.07	42	QG	2.13	2.36	−0.23
14	AP	3.64	3.16	0.48	43	SG	2.07	2.35	−0.28
15	RF	3.64	4.35	−0.71	44	LG	2.06	2.54	−0.48
16	VP	3.38	3.36	0.02	45	GD	2.04	2.67	−0.63
17	GP	3.35	2.53	0.82	46	TG	2.00	2.51	−0.51
18	GF	3.20	3.68	−0.48	47	EG	2.00	1.77	0.23
19	IF	3.03	4.51	−1.48	48	DG	1.85	1.70	0.15
20	VG	2.96	2.71	0.25	49	PG	1.77	2.35	−0.58
21	IG	2.92	2.72	0.20	50	LA	3.51	2.90	0.61
22	GI	2.92	2.97	−0.05	51	KA	3.42	3.00	0.42
23	GM	2.85	3.01	−0.16	52	RA	3.34	2.91	0.43
24	GA	2.70	2.24	0.46	53	YA	3.34	2.74	0.60
25	YG	2.70	2.38	0.32	54	AA	3.21	2.87	0.34
26	GL	2.60	3.12	−0.52	55	FR	3.04	3.40	−0.36
27	AG	2.60	2.51	0.09	56	HL	2.49	3.71	−1.22
28	GH	2.51	2.66	−0.15	57	DA	2.42	2.06	0.36
29	GR	2.49	2.83	−0.34	58	EA	2.00	2.13	−0.13

regression of observed vs. predicted activities of the test set, R_0^2 and $R_0'^2$ are the coefficients of determination for the regression through the origin (predicted vs. observed activities R_0^2, and observed vs. predicted activities $R_0'^2$), and k together with k' are the corresponding slope of theregression line through the origin.

So, in this article, external validation of the models established above was also performed. The data set of each model was first divided into training and test sets using D-optimal design on the space of descriptors as well as response. D-optimal design provides an approach for selecting the most dissimilar molecular structural and response information in the data set. Therefore, it can guarantee that the training data sets have well balanced structural diversity and are also representative of the entire range of response variable.[25] Here, 24 samples were selected as the training set by D-optimal design, the remaining 24 as the test set. From the results of external validation, we selected model 8 using 8 VHSE descriptors as an optimal model, of which the corresponding Q^2, R^2, k, k', R_0^2, and $R_0'^2$ were 0.787, 0.883, 0.993, 0.996, 0.846, and 0.879, respectively.

Model 8 was then employed to calculate the activity of the bitter dipeptides; the differences between calculated and observed activities were very small, with an overall residual standard deviation of 0.19 (see Table IV). For model 8, the first 2 components explained 74.1 and 11.7% variance of Y, respectively. From the loading plot of the first two components (Figure 1), we can see that activity of bitter dipeptides is mainly related with hydrophobicity in both amino acid (AA) positions as well as electronic property in position 1. This conclusion is corroborant with earlier studies.[8]

Table V summarizes the most important statistical parameters of the model based on 8 VHSE variables together with those obtained with z scales, ISA-ECI scales, and MS-WHIM scales. From Table V, it can be seen clearly that the model derived from 8 VHSE descriptors is the best of all.

ACE Dipeptide Inhibitors

A series of 58 dipeptides of angiotensin-converting enzyme (ACE) inhibitors was taken from Hellberg et al.[26] For each dipeptide, the structure was first quantified by 16 VHSE descriptors denoted by V_1–V_{16} in sequence. Then SMR was used to screen variables and the variables selected in each step together with response were modeled by PLS. At last, in order to determine an optimal model, the resulting PLS models were further put into external validation. Here, 29 samples were selected as the training set by D-optimal design, and remains as the test set.

The results showed that the R^2, Q^2, and the number of component of the optimal PLS model were 0.770, 0.745, and 1, respectively, and that the variables selected were V_4, V_5, V_9, V_{10}, and V_{14}. The corresponding results of external validation, i.e., Q^2, R^2, k, k', R_0^2 and $R_0'^2$, were 0.761, 0.688, 0.939, 1.042, 0.614, and 0.683, respectively. This optimal model was then

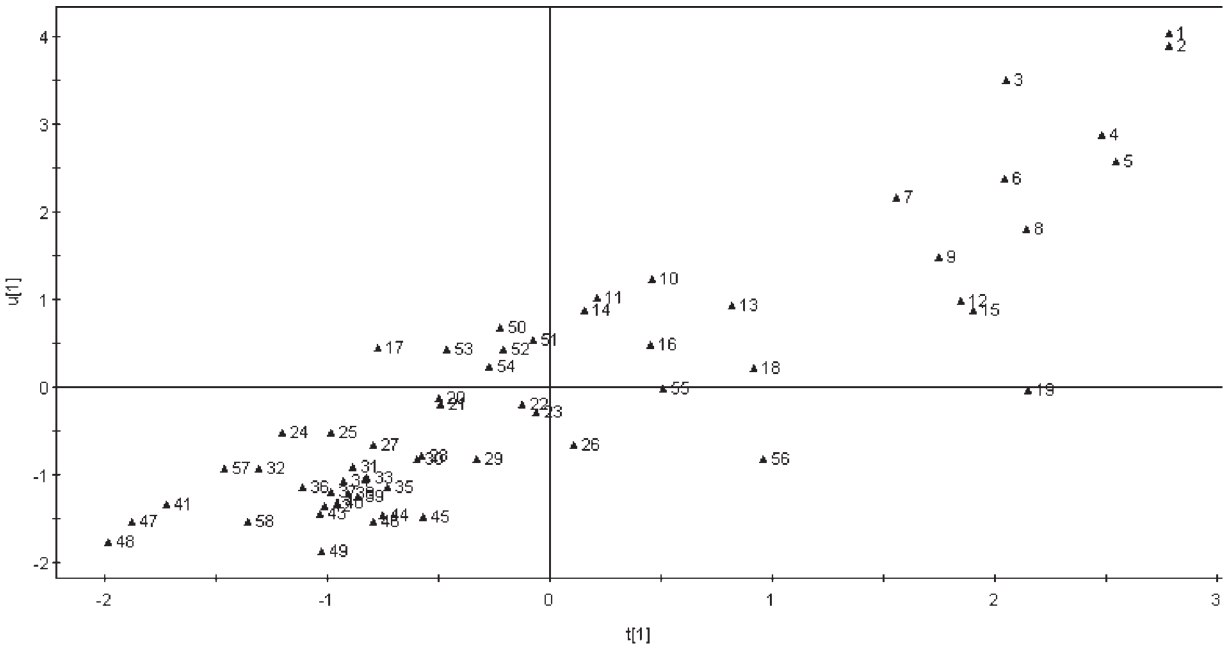

FIGURE 2 The PLS scores t_1 and u_1 of ACE inhibitor dipeptides.

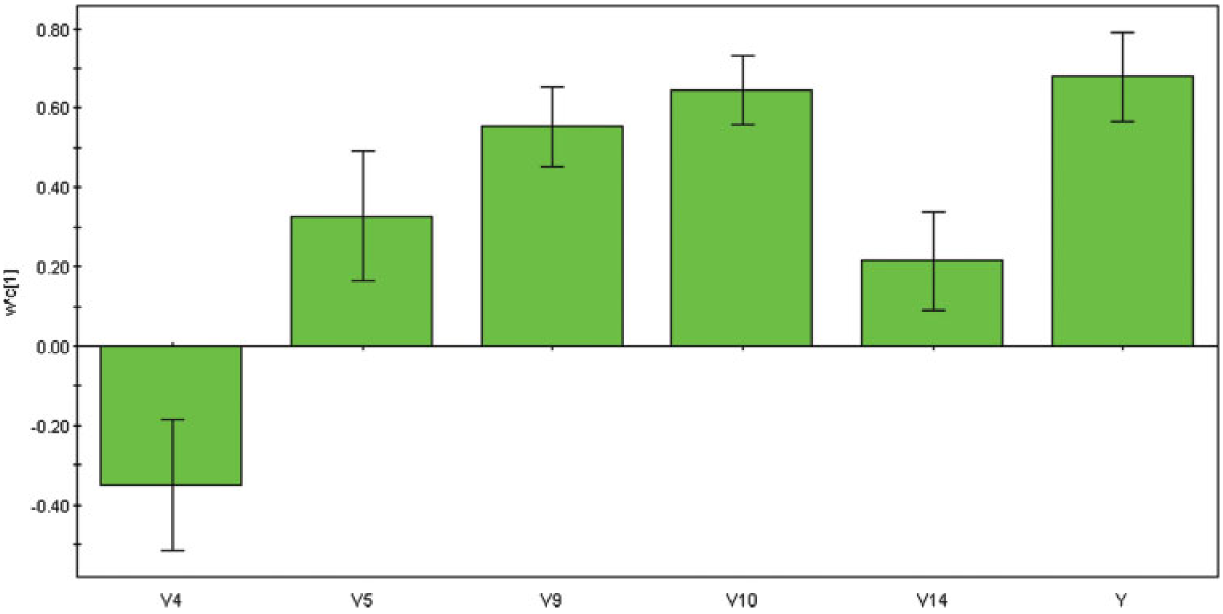

FIGURE 3 Plot of PLS loading for ACE inhibitor dipeptides.

employed to calculate activities of the dipeptides, given in Table VI.

From the variables selected, it can be deduced that the activity of ACE inhibitors is mostly related with steric and electronic properties in position 1 and hydrophobic and electronic properties in position 2. From the plot of the first component, t_1 vs. u_1 given in Figure 2, it can be seen that dipeptides 1–9, 12, 15, and 19 are separate from the others clearly. This cluster of dipeptides with high activities has Tyr(Y), Trp(W), and Phe(F) in position 2. Figure 3 is a loading plot of the first component, from which we can see clearly that the hydrophobic property in position 2 is the most important to the activity of ACE inhibitors. These results are in agreement with earlier findings that the presence of an aromatic amino acid residue in position 2 is essential for high activity.[8,26]

TableVII summarizes the most important parameters of PLS compared to those obtained with other descriptors. From Table VII, it can be seen that the best of all is the VHSE descriptors based model with only one principal component.

Bradykinin-Potentiating Pentapeptides

The brakykinin-potentiating activity for 31 pentapeptides were reported by Ufkes et al.[28,29] The activities of first 15 pentapeptides were determined in 1978 and those of last the 16, including one inactive, were measured in 1982. The biological activities were expressed as the logarithm of the relative activity index compared to the first peptide VESSK. The sequences and activities of the 31 pentapeptides are shown in Table VIII.

First, all pentapeptides were described by the VHSE scales for the five positions, giving an X matrix with $8 \times 5 = 40$ descriptor variables. Then, the X variables selected by SMR together with the Y response were modeled by PLS. The predictive capabilities of the resulting models were further tested by external validation. By means of D-optimal design, 16 samples were selected as the training set, the remaining 15 as the testing set.

An optimal PLS was derived from the results of external validation. The variables included were V_1, V_5, V_{12}, V_{15}, V_{18}, V_{19}, V_{21}, V_{25}, and V_{36}, R^2,

Table VII Statistical Parameters of PLS Model in QSAR of ACE Inhibitors

Model	Descriptor	Sum of Descriptors	A	R^2_{cum}	Q^2_{cum}	*RMSEE*
1	z scale[26]	6	2	0.770	nd	nd
2	ISA-EC[8]	4	2	0.700	nd	nd
3	MS-WHIM[10]	6	2	0.708	0.637	nd
4	MEEV[27]	10	2	0.742	0.711	nd
5	VHSE	5	1	0.770	0.745	0.48

Table VIII The observed and Calculated Activities (Log RAI) of Bradykinin-Potentiating Pentapeptides

No.	Peptide	Obsd	Calcd	Resd	No.	Peptide	Obsd	Calcd	Resd
1	VESSK	0	−0.04	0.04	16	AAAAA	−0.1	−0.56	0.46
2	VESAK	0.28	0.11	0.17	17	AAYAA	0.46	0.81	−0.35
3	VEASK	0.2	0.10	0.10	18	AAWAA	0.75	1.17	−0.42
4	VEAAK	0.51	0.25	0.26	19	VAWAA	1.43	1.31	0.12
5	VKAAK	0.11	0.48	−0.37	20	VAWAK	1.45	1.63	−0.18
6	VEWAK	2.73	1.98	0.75	21	VKWAA	1.71	1.89	−0.18
7	VEAAP	0.18	0.10	0.08	22	VWAAK	0.04	−0.11	0.15
8	VEHAK	1.53	1.09	0.44	23	VAAWK	0.23	0.14	0.09
9	VAAAK	−0.1	−0.10	0.00	24	EKWAP	1.3	1.38	−0.08
10	GEAAK	−0.52	−0.03	−0.49	25	VKWAP	2.35	2.08	0.27
11	LEAAK	0.4	0.34	0.06	26	RKWAP	1.98	1.92	0.06
12	FEAAK	0.3	0.38	−0.08	27	VEWVK	1.71	2.09	−0.38
13	VEGGK	−1	−0.54	−0.46	28	PGFSP	0.9	1.36	−0.46
14	VEFAK	1.57	1.19	0.38	29	FSPFR	0.64	0.56	0.08
15	VELAK	0.59	0.93	−0.34	30	RYLPT	0.4	0.09	0.31
					31	GGGGG	0	0.05	−0.05

Q^2, and the number of component of this optimal PLS model were 0.869, 0.824, and 2 respectively. Corresponding results of external validation, i.e., Q^2, R^2, k, k', R_0^2 and $R_0'^2$, were 0.732, 0.866, 0.944, 0.980, 0.866, and 0.860, respectively. The optimal model was then employed to calculate the activities of 31 peptapeptides (Table VIII). Figure 4 is a plot of observed against calculated activities of peptapeptides. We can see that the activities for 31 pentapeptides are modeled well, although there is some systematic difference between them. From the loading plot of the first 2 components given in Figure 5, the highest contributions to the model are hydrophobic, steric, and electronic properties in position 3.

By using *z* scales and Isotropic Surface Area-Electronic Charge Index (ISA-ECI) scales respectively, Hellberg[5] and Collantes[8] used the first 15 pentapeptides as a training set to develop QSAR models and the last 16 as a test set to validate predictive capabilities of the models developed. Here, for better comparison with the results of those, we rebuilt the PLS model with training and test data sets the same as those in the literature. From Statistical Parameters of PLS Models (Table IX),

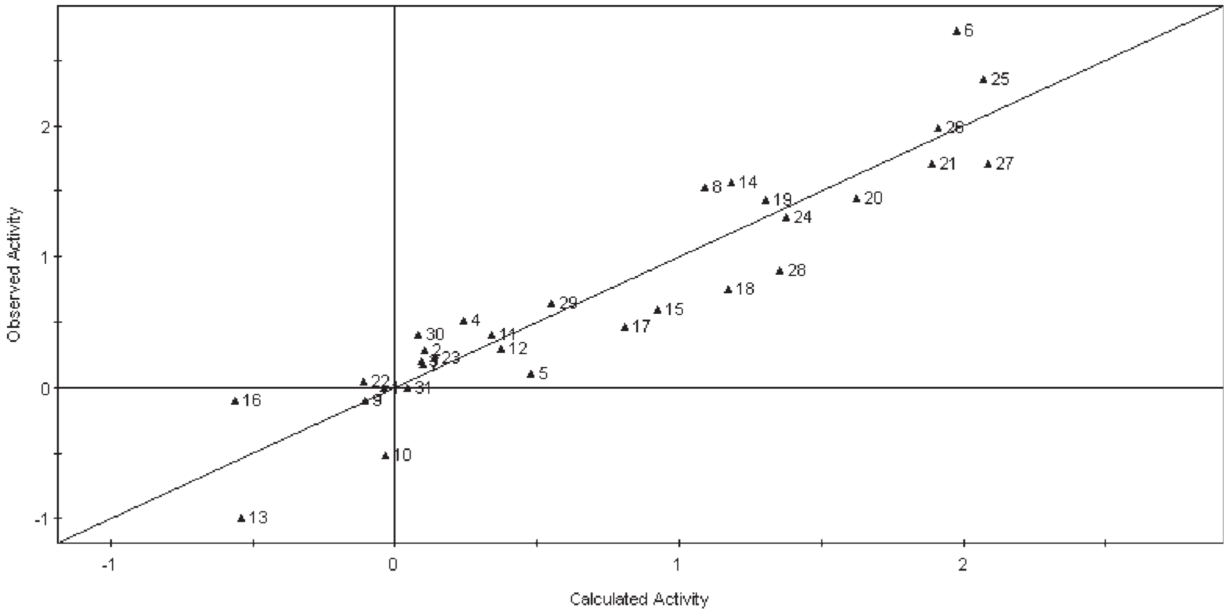

FIGURE 4 Plot of observed and calculated activities for bradykinin-potentiating pentapeptides.

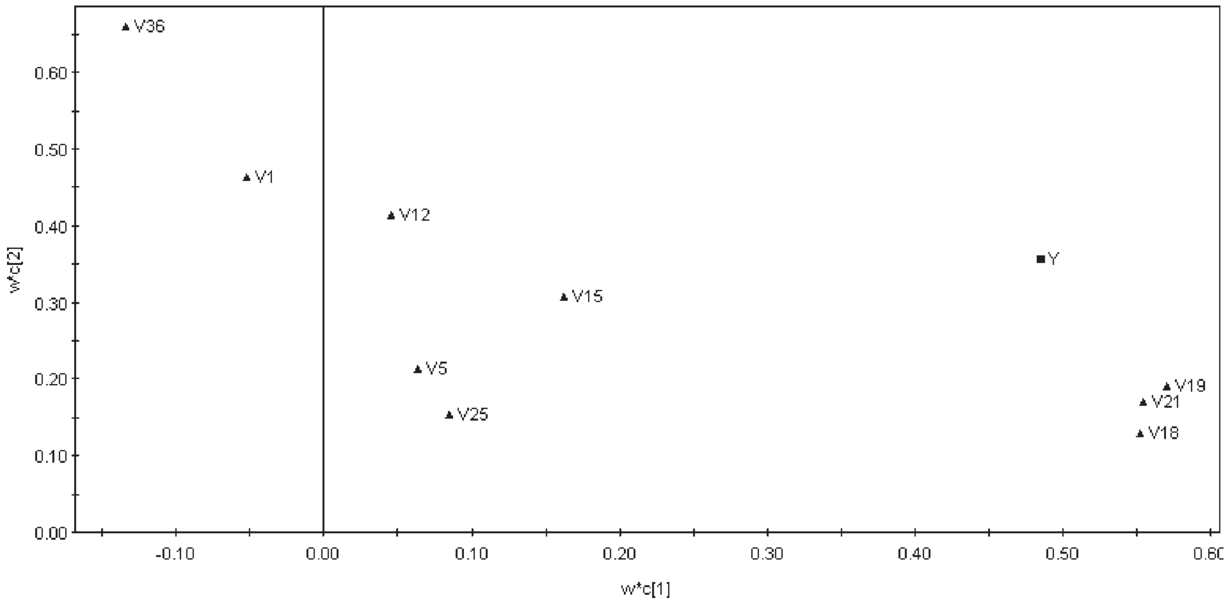

FIGURE 5 Plot of PLS loading for bradykinin-potentiating pentapeptides.

it can be seen that model 3 using 9 VHSE descriptors has the best predictive capability for the test set with the Relative Standard Deviation (RSD) of 0.80.

CONCLUSIONS

Structural description is critical to the success of QSARs. As it is well known, a good descriptor should contain as much chemical information relating to biological activities as possible. In this work, 50 physiochemical properties of 20 coded amino acids were selected in total and grouped carefully into 18 hydrophobic properties, 17 steric properties, and 15 electronic properties. The resulting three families of independent variables were examined individually through principal components analysis. A total of 8 principal component scores were selected for each amino acid from the PCA of these three groups and applied as new vector of descriptors, namely VHSE scales, to QSAR studies of three peptide data sets. Good results were obtained in comparison with those obtained with *z* scales, ISA-ECI scales, MS-Weighted Holistic Invariant Molecular (WHIM) scales, and Molecular Electronegativity Edge Vector (MEEV) scales. As a new set of amino acid descriptors, VHSE was derived from a broad spectrum of physicochemical descriptors; thus, it contains more structural information, and it is easier to interpret. However, with the increasing of the length of the peptide chain, 3D structural information of the whole peptide chain becomes more important. Thus, other 3D descriptors based on the whole peptide chain are also required for the success of peptide QSARs. Last, it needs to be pointed out that no matter how robust, significant, and validated a QSAR may be, it cannot be expected to reliably predict the modeled property for the entire universe of compounds. So, before a QSAR is employed to screen new compounds, its domain of applications should be defined well and reliable predictions may be made only for compounds that fall in the well-defined domain. Several algorithms have been proposed to define the applicability domain of a QSAR model.[24,25]

Table IX Statistical Parameters of PLS Models in QSAR of Bradykinin-Potentiating Pentapeptides

Model	Descriptor	*A*	R^2_{cum}	Q^2_{cum}	RSD for Test Set
1	*z* scale[5]	3	0.970	nd	0.84
2	ISA-ECI[8]	2	0.920	nd	0.82
3	VHSE	1	0.934	0.861	0.80

The authors thank the State Chuihui Project Fund (SCPF, 98-3-8), Fok Ying-Tung Educational Foundation (FYTF, 98-7-6) and Chongqing University Innovation Fund (CUIF, 03-5-6) for financial supports.

REFERENCES

1. Jirácek, J.; Yiotakis, A.; Vincent, B.; Lecoq, A.; Checler, F.; Dive, V. J Biol Chem 1995, 270, 21701–21706.
2. Marraud, M.; Aubry, A. Biopolymers (Peptide Sci) 1996, 40, 45–83.

3. Sneath, P. H. A. J Theor Biol 1966, 12, 157–195.
4. Kidera, A.; Konishi, Y.; Oka, M.; Ooi, T.; Scheraga, J. A. J Protein Chem 1985, 4, 23–55.
5. Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. J Med Chem 1987, 30, 1126–1135.
6. Fauchère, J.-L.; Charton, M.; Kier, L. B.; Verlooop, A.; Pliska, V. Int J Pept Protein Res 1988, 32, 269–278.
7. Jonsson, J.; Eriksson, L.; Hellberg, S.; Sjöström, M.; Wold, S. Quant Struct-Act Relat 1989, 8, 204–209.
8. Collantes, E. R.; Dunn, W. J. J Med Chem 1995, 38, 2705–2713.
9. Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. J Med Chem 1998, 41, 2481–2491
10. Zaliani, A.; Gancia E. J Chem Inf Comput Sci 1999, 39, 525–533.
11. Kawashima, S.; Kanehisa, M. Nucleic Acids Res. 2000, 28, 374.
12. Tomii, K.; Kanehisa, M. Protein Eng 1996, 9, 27–36.
13. Nakai, K.; Kidera, A.; Kanehisa, M. Protein Eng 1988, 2, 93–100.
14. Rogers, D.; Hopfinger, A. J. J Chem Inf Comput Sci 1994, 34, 854–866.
15. Hou, T. J.; Wang, J. M.; Liao, N.; Xu, X. J. J Chem Inf Comput Sci 1999, 39, 775–781.
16. Hasegawa, K.; Kimura, T.; Funatsu, K. Quant Struct-Act Relat 1999, 18, 262–272.
17. Sutter, J. M.; Dixon, S. L.; Jurs, P. C. J Chem Inf Comput Sci 1995, 35, 77–84.
18. Sutter, J. M.; Kalivas, J. H. Microchem J 1993, 47, 60–66.
19. Kubinyi, H. Quant Struct Act Relat 1994, 13, 285–294.
20. Luke, B. T. J Chem Inf Comput Sci 1994, 34, 1279–1287.
21. Wold, S.; Sjöström, M.; Eriksson, L. Chemom Intell Lab Syst 2001, 58, 109–130.
22. Wold, S.; Trygg, J.; Berglund, A.; Antti, H. Chemom Intell Lab Syst 2001, 58, 131–150.
23. Golbraikh, A.; Tropsha, A. J Mol Graphics Mod 2002, 20, 269–276.
24. Tropsha, A.; Gramatica, P.; Gombar, V. K. QSAR Comb Sci 2003, 22, 69–77.
25. Gramatica, P.; Pilutti, P.; Papa, E. J Chem Inf Comput Sci 2004, 44, 1794–1802.
26. Hellberg, S.; Eriksson, L.; Jonsson, J.; Lindgren, F.; Sjöström, M.; Skagerberg, B.; Wold, S.; Andrews, P. Int J Pept Protein Res 1991, 37, 414–424.
27. Li, S. Z.; Fu, B.; Wang, Y.; Liu, S. J Chin Chem Soc 2001, 48, 937–944.
28. Ufkes, J. G. R.; Visser, R. J.; Heuver, G.; van der Meer, C. Eur J Pharmacol 1978, 50, 119–122.
29. Ufkes, J. G. R.; Visser, R. J.; Heuver, G.; Wynne, H. J.; van der Meer, C. Eur J Pharmacol 1982, 79, 155–158.