A New Descriptor for Amino Acids and Its Applications in Peptide QSAR

JIANBO TONG[1,2]*, JIA CHANG[1,2], XIAMENG XU[1,2], SHULING LIU[1,2], MIM BAI[1,2]

[1]College of Chemistry and Chemical Engineering, Shaanxi University of Science & Technology, Xi'an 710021, PR China

[2] Key Laboratory of Auxiliary Chemistry & Technology for Chemical Industry, Ministry of Education. Shaanxi University of Science & Technology, Xi'an 710021, PR China

To establish a new amino acid structure descriptor that can be applied in peptide quantitative structure activity relationship (QSAR) studies, a new descriptor, named SVMW, was derived by principal components analysis of the matrix of 160 MoRSE descriptors and 99 WHIM descriptors of amino acids. The scale was then applied in two panels of peptide QSAR that were molded by partial least square regression. The correlation coefficients (R_{cum}^2), cross validation correlation coefficients (Q_{LOO}^2) were 0.821 and 0.783 for angiotensin-converting enzyme inhibitors (dipeptide), 0.991 and 0.969 for angiotensin-converting enzyme inhibitors (tri-peptides), 0.868 and 0.807 for bitter tasting thresholds, respectively. In addition, the estimation capability and generalization ability of the models were analyzed by external validation. The correlation coefficients of predicted values versus experimental ones of external samples (Q_{ext}^2) were 0.821, 0.774 and 0.964. Satisfactory results showed that information related to biological activity could be systemically expressed by SVWG scales, which may be a useful structural expression methodology for study on peptides QSAR.

Keywords: *amino acids, peptides, quantitative structure-activity relationship (QSAR), SVMW descriptor*

Peptides are essential substance to sustain life [1,2]. They have high activity, high selectivity and fewer side effects. The peptide is one of the hot spots of drug research [3,4]. However, most of experimental methods are inefficient and expensive. Therefore, the computational methods such as quantitative structure-activity relationship (QSAR) have been brought into the spotlight, involving in not only the key idea of pharmaceutical chemistry and pharmacology but also the foundation of drug design. Molecular structural characterization (MSC) is critical to the success of QSAR. However, in recent years, a good descriptor should contain as much chemical information relating to biological activities as possible.

In this paper, SVMW, which derived by principal components analysis of the matrix of 160 MoRSE descriptors and 99 WHIM descriptors of amino acids, were examined through principal component analysis (PCA). Applying SVMW to 58 angiotensin-converting enzyme inhibitors (dipeptide), 55 angiotensin-converting enzyme inhibitors (tri-peptides) and 48 bitter tasting thresholds, satisfying results were obtained from the constructed QSAR models.

Experimental part

Principle and Methodology

Principal component analysis (PCA)

Based on quantum chemistry calculation level of density function theory (DFT) [5, 6], Berny energetic gradient and generalized gradient approximation (GGA) were employed to optimize the spatial conformation of 20 coded amino acids. The descriptor calculation software of Dragon 5.2 was utilized to generate the 160 MoRSE descriptors and 99 WHIM descriptors for each single amino acid [7]. Among these, the calculation of DFT was achieved by Gaussian 98, and the input file of structures for natural amino acids was automatically generated by Chemoffice 8.0. Thus, the matrix of 20×160 dimensions consisting of 160 MoRSE descriptors, and 20×99 dimensions consisting of 99 WHIM descriptors was obtained. Primarily, the original variable matrix $X_{20\times160}$ and $X_{20\times99}$ were subject to autoscaling, then PCA was applied to generate corresponding scores of principal components and characterized vector for every principal component. For original 160 MoRSE descriptors there generated 7 prominent PCs with eigenvalues > 1, cumulatively explaining 83.05% variances. For original 99 WHIM descriptors, the former 4 PCs with eigenvalues > 1 cumulatively explained 87.55% variances were selected. A total of 11 PCs, which MoRSE descriptors and WHIM descriptors, for 20 coded amino acids were termed as vector of principal component score (SVMW) (table 1). Statistics software SPSS 13.0 implemented PCA program.

Partial least square

Partial least square (PLS) [8] regression is a widely employed modeling method at the present time, which has advantage of effectively overcoming multicollinearity issues and especially suits for condition of sample size smaller than variable number. Even more, PLS has the desirable property that the precision of the model parameters is improved with the increasing number of relevant variables and observations [9, 10].

Stepwise multiple regression (SMR) was carried out for variable selection because it was less time-consuming and easy to implement. PLS was implemented by software of Simca-P 10.0. Matlab 7.0 was used for PCA, and SPSS 10.0 was used for stepwise multiple variable selection.

Results and discussions

QSAR model for angiotensin-converting enzyme inhibitors (dipeptide)

Angiotensin converting enzyme inhibitor (dipeptide) is an inhibitor of angiotensin-converting enzyme (ACE)

* *email: jianbotong@aliyun.com; jianbotong@yahoo.com.cn*

Amio acids	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}
Ala A	-4.8673	4.8567	0.2935	-1.1984	-1.7882	0.1084	-0.9456	-11.6344	-1.89677	1.97792	-2.60584
Arg R	2.8074	-5.7328	-10.0304	0.1795	-1.8715	-2.2739	1.9078	11.87107	-2.87032	2.74823	1.25733
Asn N	-4.2352	4.4868	-1.581	-0.5684	2.9533	-1.0546	0.4217	-5.34951	7.68317	4.11656	4.1737
Asp D	-3.675	5.6066	-1.3744	2.375	2.2126	1.3277	2.1233	-4.02676	2.99296	-3.35851	-3.77013
Cys C	-4.8303	6.3381	0.7404	-3.6652	-1.4237	-1.5163	-2.8869	-5.65032	-2.87927	-2.99048	2.3435
Gln Q	-4.7011	-0.8588	-3.5331	2.8551	0.7795	-1.1897	0.1375	2.17649	-2.40014	0.84542	3.57161
Glu E	-2.5167	-0.6925	-1.4272	-0.5017	3.8605	1.0923	-0.405	2.36689	0.15236	-4.04847	0.80392
Gly G	-4.6947	9.543	-0.1147	-0.1988	-0.4047	1.5034	0.8204	-11.7823	-13.6975	3.47023	0.20083
His H	4.3481	3.6123	1.8679	3.8021	0.8011	5.1879	-5.1697	2.33867	0.36099	-1.56522	-1.07602
Ile I	-2.213	-5.2	3.6825	5.6186	-5.5525	2.8169	1.2975	0.41167	6.40422	-1.24429	-1.62207
Leu L	-3.1883	-9.1722	5.8668	4.8209	-0.3621	-0.8706	3.2602	0.26884	8.11636	2.89694	0.98231
Lys K	-0.5855	-11.934	-4.8895	-1.5481	3.8189	2.5185	-2.3203	9.00644	-2.09657	-3.35485	2.39215
Met M	-3.9573	-4.0717	2.3255	-4.3088	-5.9397	-3.4378	-3.8467	4.36305	-1.66543	-3.97675	-1.02332
Phe F	13.8796	-2.8069	2.0969	-6.5485	-0.7813	0.7353	0.795	7.26373	-4.36604	-1.09058	1.62148
Pro P	-2.7762	-3.1282	7.5571	-2.5701	7.0349	-3.0942	0.809	-5.3069	3.18399	0.59462	4.27735
Ser S	-3.3416	6.8606	-0.9756	0.8169	-1.6716	-2.4715	3.4412	-9.15545	2.32007	-0.49944	-2.26929
Thr T	-2.7966	0.8798	-1.07	-1.2773	-0.1541	2.2859	0.1757	-4.22009	-0.27234	-1.39095	-2.53752
Trp W	19.2063	3.7382	-0.0853	6.6477	1.0446	-4.6764	-2.6452	11.70248	0.16164	5.62047	-4.91875
Tyr Y	13.2974	2.2367	0.8869	-4.4618	-1.3706	3.1242	4.2243	8.54042	-1.52583	1.74131	-1.28473
Val V	-5.1598	-4.5618	-0.2362	-0.2687	-1.1856	-0.1157	-1.1943	-3.18401	2.29446	-0.49218	-0.51649

Table 1
SVMW DESCRIPTORS FOR AMINO ACIDS

No.	Peptide	Obsd	Calcd[a]	Pred[b]	No.	Peptide	Obsd	Calcd[a]	Pred[b]
1	VW*	5.80	5.11	5.22	30	KG	2.49	2.45	2.33
2	IW	5.70	5.62	5.75	31	FG	2.43	2.51	2.34
3	IY	5.43	4.48	4.45	32	GS	2.42	2.79	2.91
4	AW	5.00	4.81	4.91	33	GV*	2.34	2.37	2.57
5	RW*	4.80	5.16	5.44	34	MG	2.32	2.20	3.14
6	VY	4.66	3.96	3.92	35	GK	2.27	2.27	2.29
7	GW	4.52	4.34	4.49	36	GE	2.27	2.41	2.41
8	VF	4.28	4.12	3.88	37	GT*	2.24	2.27	2.55
9	AY*	4.06	3.66	3.60	38	WG	2.23	2.27	2.10
10	IP	3.89	1.11	3.83	39	HG	2.2	2.16	2.03
11	RP	3.74	3.66	3.51	40	GQ	2.15	1.40	2.17
12	AF	3.72	3.81	3.55	41	GG*	2.14	1.95	1.83
13	GY*	3.68	3.20	3.18	42	QG	2.13	2.18	2.15
14	AP	3.64	3.30	2.98	43	SG	2.07	2.29	2.16
15	RF	3.64	4.17	4.08	44	LG	2.06	2.79	2.59
16	VP	3.38	5.59	3.30	45	GD*	2.04	2.41	2.73
17	GP*	3.35	2.83	2.55	46	TG	2.00	2.43	2.28
18	GF	3.20	3.35	3.14	47	EG	2.00	1.98	1.87
19	IF	3.03	4.63	4.40	48	DG	1.85	2.02	1.88
20	VG	2.96	2.72	2.57	49	PG*	1.77	1.64	1.38
21	IG*	2.92	3.23	3.10	50	LA	3.51	3.18	3.31
22	GI	2.92	2.39	2.70	51	KA	3.42	2.83	3.04
23	GM	2.85	2.67	2.85	52	RA	3.34	3.16	3.49
24	GA	2.70	2.33	2.55	53	YA*	3.34	2.83	3.03
25	YG*	2.70	2.45	2.32	54	AA	3.21	2.80	2.97
26	GL	2.60	2.68	2.81	55	FR	3.04	3.50	3.39
27	AG	2.60	2.42	2.25	56	HL	2.49	2.88	3.00
28	GH	2.51	2.49	2.55	57	DA*	2.42	2.41	2.60
29	GR*	2.49	2.94	2.89	58	EA	2.00	2.35	2.57

Table 2
SEQUENCES OF ACE INHIBITORS (DIPEPTIDE) WITH THEIR OBSERVED AND CALCULATED ACTIVITIES

activity of compounds. Angiotensinogen [11,12], which is produced by liver, is catalyzed by rennin to disrupt into inactive angiotensin I which is further catalyzed by angiotensin converting enzyme to rupture into angiotensin II. ACE inhibitors combined with the structural characteristics of angiotensin I to compete with the ACE and inhibited its effective biological activity, so that to achieve the purpose of control and lower blood pressure. Thus, ACE inhibitors have prepared the premedicant [13] to treat blood pressure, heart disease and diabetes kidney disease. Data sets of ACE inhibitory with experimentally determined values were originally from Cushman *et al.*[14].

Firstly, as a classical sample set in QSAR studies [15-17], 58 ACE inhibitors (dipeptide) are often utilized to validate the efficiency of amino acid descriptors. For a set of peptide analogues, the chemical structure can now be quantified by describing each varied amino acid position. So the chemical structure of each dipeptide ACE inhibitor can be described by 2×11 variables. Not all structural descriptors are relevant to bioactivities for a QSAR data set. SMR was employed to delete the irrelevant and redundant descriptors. Q_{LOO}^2 was shown to achieve the maximum at the seventh step (ν_{12}, ν_{17}, ν_2, ν_{20}, ν_{11}, ν_{13}, ν_5). Ultimately, PLS model in which two PLS components were enough to account for 90.6% variances of *Y* variables with cross-validation achieving 88.5% and RMSE achieving 0.432. There are sequences of ACE inhibitors (dipeptide) which sorted from low activity to high activity with their observed and calculated activities in table 2.

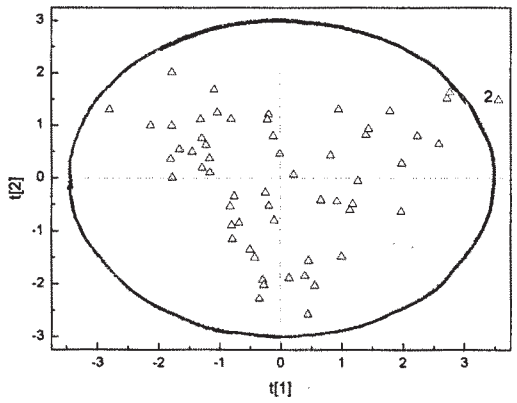

Fig. 1. PLS scores of ACE inhibitors (dipeptide)

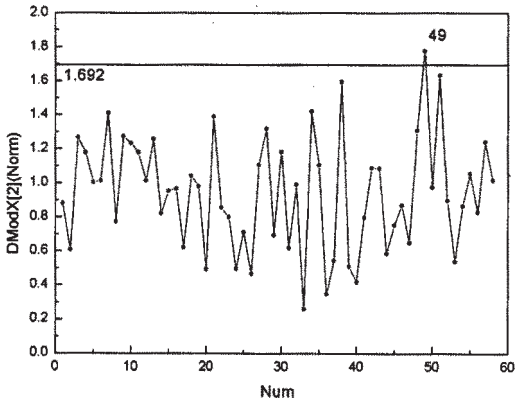

Fig. 2. Distance to PLS model in X space of ACE inhibitors (dipeptide)

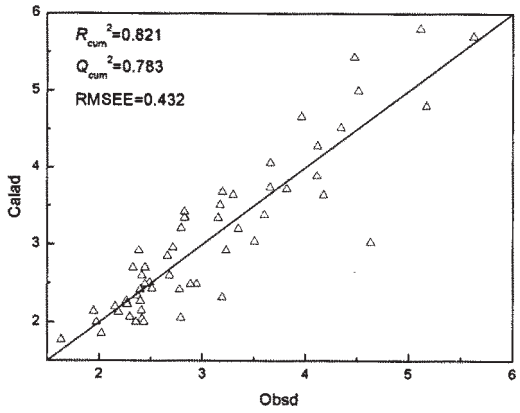

Fig. 3. Plots of calculated versus observed activities for ACE inhibitors

Figure 1 presented that most of the samples were falling into Hotelling's T^2 [18,19] confidence ellipse with 95% confidence except sample #2, which indicated that SVMW could provide a satisfied result to the model of ACE inhibitors (dipeptide). The distance to the PLS model in the *X* space was described by the solid line in figure 2 in order to investigate efficiency on recombination. It can be seen that the normalized distance to *X* for most samples were smaller than the critical value of 1.692 (significance level = 5%) except sample #49. A 7-variable PLS model was constructed for the training set with its fitting correlative coefficient R_{cum}^2=0.821, and the cross-validation Q_{LOO}^2=0.783. In order to prove the validity and stability of the model, the whole data set is systematically divided into two subsets, from table 2 samples were chosen regularly every four as test set, thus 43 samples were treated as training set which were utilized to construct QSAR model and the remaining 15 samples were regarded as test set (samples in test set are highlighted with "*" in table 2). The rest 15 samples were utilized to validate the external prediction power of the model developed. Consequently, the correlation coefficients of predicted values versus experimental ones of external samples Q_{ext}^2=0.821. There are plots of calculated versus observed activities for ACE inhibitors (dipeptide) in figure 3. We, also compared between QSAR models of ACE inhibitors (dipeptide) in table 3. The results are similar to or better than those in the literatures.

QSAR model for angiotensin-converting enzyme inhibitors (tri-peptides)

Research on angiotensin-converting enzyme inhibitors (tri-peptides) is an active field in medicine exploitation in recent years. A series of 55 peptides of ACE inhibitors (tri-peptides) peptides were taken from the data by Z.H. Lin et al.[20].

The processes of establish model of ACE inhibitors (tripeptide) are similar to ACE inhibitors (dipeptide). There are sequences of ACE inhibitors (tripeptide) which sorted from low activity to high activity with their observed and calculated activities in table 4. PLS model in which two PLS components were enough to account for 99.9% variances of Y variables with cross-validation achieving 96.5% and RMSE achieving 0.07. Figure 4 presented that all of the samples were falling into Hotelling's T^2 confidence ellipse with 95% confidence. Except sample #29, the distance to the PLS model in the *X* space was described by the solid line in figure 5. The normalized distance to *X* for

Table 3
COMPARISON BETWEEN QSAR MODELS OF ACE INHIBITORS (DIPEPTIDE)

No	descriptors	model	A[a]	R_{cum}^{2b}	Q_{cum}^{2c}	RMSEE[d]
1	zscale	PLS	2	0.770	nd[e]	nd[e]
2	GRID(tscores)	PLS	1	0.744	nd[e]	0.500
3	ISA-ECI	PLS	2	0.700	nd[e]	nd[e]
4	MS2WHIM(rotameric)	PLS	3	0.657	0.541	nd[e]
5	MS2WHIM(extended)	PLS	3	0.708	0.637	0.54
6	MHDV	PCR	19	0.878	0.753	0.350
7	MEEV(M1)	MLR	10	0.711	0.475	0.340
8	MEEV(M2)	MLR	3	0.649	0.570	0.370
9	MEEV(M3)	MLR	10	0.773	0.588	0.330
10	MEEV(M4)	MLR	3	0.735	0.677	0.320
11	VHSE	PLS	1	0.770	0.745	0.480
12	SVTV	PLS	1	0 789	0.767	0.460
13	c*scales	G/PLS	3	0.806	0.761	nd[e]
14	SVMW	PLS	3	0.906	0.885	0.432

[a] principal components; [b] cumulative multiple correlation coefficient; [c] cumulative cross-validated R_{cum}^2; [d] root mean square error; [e] not determined.

No.	Peptides	Obsd	Calcd[a]	Pred[b]	No.	Peptides	Obsd	Calcd[a]	Pred[b]
1	VVV*	1.63	1.72	1.88	29	GVV	1.82	1.97	1.89
2	RPG	3.09	3.08	3.05	30	PPG	3.18	3.18	3.10
3	GRP	0.48	0.55	0.51	31	PGG*	3.14	3.09	3.06
4	LLL	1.35	1.33	1.42	32	PGP	1.82	1.80	1.80
5	GLG	2.45	2.43	2.44	33	GPG	2.65	2.65	2.66
6	LGL*	1.52	1.46	1.59	34	GGP	1.28	1.27	1.35
7	FGG	2.79	2.78	2.97	35	PGI	2.23	2.35	2.25
8	GFG	2.53	2.46	2.43	36	KPK*	2.63	2.68	2.59
9	GGF	1.11	1.08	1.09	37	ADA	2.17	1.93	2.04
10	FFG	2.71	2.67	2.78	38	GEG	2.28	2.32	2.20
11	FGF*	1.29	2.29	1.45	39	LEL	1.19	1.22	1.18
12	GFF	1.02	0.99	0.89	40	RGP	1.73	1.70	1.74
13	GGG	2.61	2.56	2.61	41	PIP*	1.69	1.69	1.58
14	GYG	2.33	2.34	2.34	42	FPF	1.32	1.39	1.48
15	GGY	1.35	1.32	1.30	43	KPF	1.51	1.64	1.67
16	YGY*	1.82	1.75	1.74	44	VYP	0.82	0.82	1.07
17	GYY	1.07	1.11	1.03	45	YPF	1.60	1.61	1.58
18	YYY	1.54	1.54	1.47	46	LGG*	2.49	2.54	2.68
19	FIV	2.04	2.22	2.30	47	GGL	1.63	1.47	1.53
20	FPP	1.50	1.58	1.75	48	LLG	2.33	2.41	2.42
21	FPK*	2.45	2.53	2.40	49	GLL	1.47	1.34	1.36
22	PFP	1.74	1.70	1.61	50	YGG	3.07	1.99	3.06
23	RRR	1.77	1.79	1.65	51	YYG*	2.79	2.77	2.79
24	PPP	1.86	1.89	1.84	52	LDL	1.42	1.36	1.36
25	FFF	1.20	1.20	1.25	53	VIF	0.78	0.75	0.85
26	RGP*	1.73	1.70	1.74	54	RPF	1.59	1.60	1.52
27	PGR	2.67	2.62	2.55	55	PPF	1.68	1.71	1.58
28	GGV	1.99	2.11	2.16					

[a]calculated values; [b]predicted values; *test set.

Table 4
SEQUENCES OF ACE INHIBITORS (TRI-PEPTIDES) WITH THEIR OBSERVED AND CALCULATED ACTIVITIES

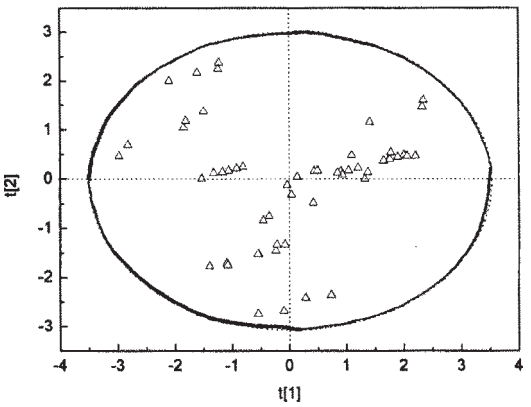

Fig. 4. PLS scores of ACE inhibitors (tri-peptides)

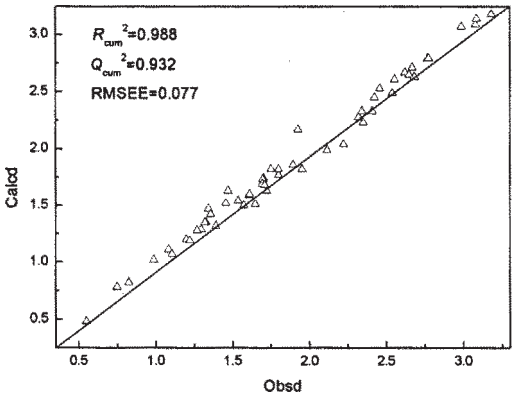

Fig. 6. Plots of calculated versus observed activities for 55 ACE inhibitors (tri-peptides)

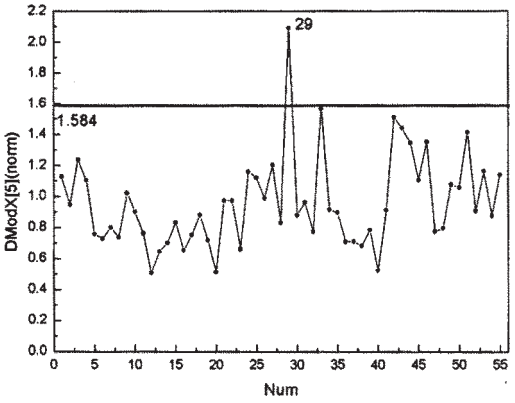

Fig. 5. Distance to PLS model in X space of ACE inhibitors (tri-peptides)

Table 5
COMPARISON BETWEEN QSAR MODELS OF ACE INHIBITORS (TRI-PEPTIDES)

No	descriptors	model	A[a]	R_{cum}^{2b}	Q_{cum}^{2c}	RMSEE[d]
1	zscale	PLS	2	0.770	nd[e]	nd[e]
2	GRID(tscores)	PLS	1	0.744	nd[e]	0.500
3	ISA-ECI	PLS	2	0.700	nd[e]	nd[e]
4	VSTV	PLS	3	0.789	0.767	0.460
5	SSIA-AMI	PLS	3	0.769	0.699	0.490
6	SSIA-PM3	PLS	3	0.789	0.773	0.470
7	SZOTT	PLS	3	0.878	0.753	0.330
8	T-scales	PLS	3	0.845	0.786	0.390
9	VSW	PLS	3	0.868	0.784	0.370
10	G-scales	PLS	3	0.870	0.831	0.370
11	SVMW	PLS	5	0.998	0.931	0.077

[a] principal components; [b] cumulative multiple correlation coefficient; [c] cumulative cross-validated R_{cum}^2; [d] root mean square error; [e] not determined.

QSAR model for bitter tasting thresholds

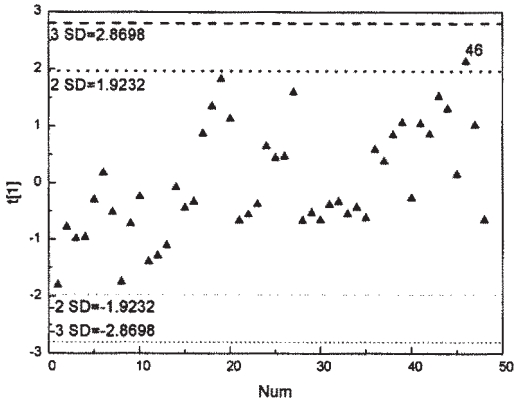

Fig. 7. PLS scores of BTT

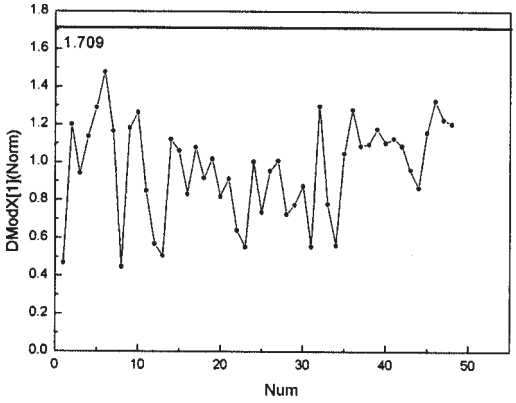

Fig. 8. Distance to PLS model in the X space of BTT

Table 6
SEQUENCES OF BTT INHIBITORS WITH THEIR OBSERVED AND CALCULATED ACTIVITIES

No.	Peptide	Obsd	Calcd[a]	Pred[b]	No	Peptide	Obsd	Calcd[a]	Pred[b]
1	GV*	1.13	0.88	0.93	25	II*	2.26	2.25	2.13
2	GL	1.68	1.50	1.62	26	IP	2.40	2.28	2.50
3	GI	1.70	1.37	1.44	27	IW	3.05	2.96	2.66
4	GP*	1.35	1.39	1.8	28	IN*	1.49	1.58	1.52
5	GF	1.80	1.79	1.96	29	ID	1.37	1.66	1.57
6	GW	1.89	2.09	1.96	30	IQ	1.49	1.58	1.31
7	GY*	1.77	1.67	1.81	31	IE*	1.37	1.75	1.60
8	AV	1.16	0.91	0.95	32	IK	1.65	1.78	1.34
9	AL	1.70	1.54	1.64	33	IS	1.49	1.65	1.62
10	AF*	1.72	1.83	1.98	34	IT*	1.49	1.72	1.63
11	VG	1.19	1.13	1.33	35	PA	1.32	1.60	1.59
12	VA	1.16	1.20	1.36	36	PL	2.22	2.34	2.22
13	VV*	1.71	1.31	1.30	37	PI*	2.33	2.22	1.05
14	VL	2.00	1.94	1.99	38	PY	1.80	2.50	2.41
15	LG	1.72	1.71	1.73	39	PF	2.80	2.64	2.56
16	LA*	1.72	1.78	1.76	40	FG*	1.77	1.82	1.88
17	LL	2.35	2.52	2.39	41	FL	2.87	2.62	2.55
18	LF	2.75	2.81	2.73	42	FP	2.70	2.51	2.73
19	LW*	3.40	3.10	2.75	43	FF*	3.10	2.91	2.89
20	LY	2.46	2.68	2.59	44	FY	3.13	2.78	2.74
21	IG	1.68	1.57	1.65	45	WE	1.56	2.07	1.94
22	IA*	1.68	1.64	1.68	46	WW*	3.60	3.29	3.00
23	IV	2.05	1.75	1.62	47	YL	2.40	2.61	2.55
24	IL	2.26	2.38	2.31	48	SL	1.49	1.58	1.69

[a]calculated values; [b]predicted values; *test set.

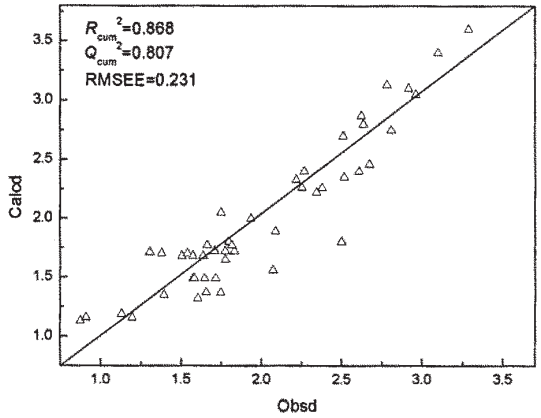

Fig. 9. Plots of calculated versus observed activities for 48 BBT

most samples was smaller than critical value of 1.584. We get the PLS model was constructed for the training set with its fitting correlative coefficient R_{cum}^2=0.991, cross-validation Q_{LOO}^2=0.969. After that, the whole data set is systematically divided into two subsets, from table 4 samples were chosen regularly every five as test set, thus 44 samples were treated as training set which were utilized to construct QSAR model and the remaining 11 samples were regarded as test set (samples in test set are highlighted with "*" in table 4). As a result, the correlation coefficients of predicted values versus experimental ones of external samples Q_{ext}^2=0.964. There are plots of calculated versus observed activities for ACE inhibitors (tripeptide) in figure 6. Table 5 is compared between QSAR models of ACE inhibitors (tripeptide). We also get that the results are similar to or better than those in the literatures.

QSAR model for bitter tasting thresholds

Bitter sensitivity, as one of gustatory sensitivities, protects humans and organisms from injury by toxic substances. Studies indicate that conduction of taste signal in taste receptor cells involves in a series of complicated processes mediated by G protein-coupled receptors [21]. As a classical sample set in QSAR studies, 48 bitter tasting thresholds (BTT) reported by Collantes [22], with its activity expressed by negative logarithm of concentration (pT), are often utilized to validate the efficiency of amino acid descriptors. Utilizing the 2×11 SVMW scales to describe each BBT, the resulting PLS model in which three PLS components were enough to account for 93.2% variances of *Y* variables with cross-validation achieving 89.8% and *RMSE* achieving 0.231. From figure 7, except sample #46, they are smaller than twice the standard deviation. However, they are all smaller than triple standard deviation.

No	descriptors	model	A[a]	R_{cum}^{2b}	Q_{cum}^{2c}	RMSEE[d]
1	zscale	PLS	2	0.824	nd[e]	0.260
2	GRID(tscores)	PLS	1	ndg	0.780	nd[e]
3	ISA-ECI	PLS	2	0.847	nd[e]	nd[e]
4	MS2WHIM(rotameric)	PLS	3	0.704	0.633	nd[e]
5	MS2WHIM(extended)	PLS	3	0.754	0.710	0.320
6	MHDV	PCR	10	0.919	0.864	0.180
7	MEEV(M1)	MLR	10	0.711	0.475	0.340
8	MEEV(M2)	MLR	3	0.649	0.570	0.370
9	MEEV(M3)	MLR	10	0.773	0.588	0.330
10	MEEV(M4)	MLR	3	0.735	0.677	0.320
11	VHSE	PLS	3	0.881	0.843	0.220
12	c*scales	G/ PLS	3	0.847	0.776	nd[e]
13	SVMW	PLS	1	0.869	0.806	0.432

Table 7
COMPARISON BETWEEN QSAR MODELS OF BTT

[a] principal components; [b] cumulative multiple correlation coefficient; [c] cumulative cross-validated R_{cum}^2; [d] root mean square error; [e] not determined.

The distance to the PLS model in the X space was described by the solid line in figure 8. The normalized distance to X for all samples was smaller than critical value of 1.709. We get that the PLS model was constructed for the training set with its fitting correlative coefficient R_{cum}^2=0.868, cross-validation Q_{LOO}^2=0.807.

Moreover, the whole data set is systematically divided into two subsets, from table 6 samples were chosen regularly every three as test set, thus 32 samples were treated as training set which were utilized to construct QSAR model and the remaining 16 samples were regarded as test set (samples in test set are highlighted with "*" in table 6). Consequently, the constructed model was then utilized to predict test set with the result of Q_{ext}^2 =0.774. There are plots of calculated versus observed activities for BBT in figure 9. At the same time, we compared between QSAR models of BBT in table 7. The results are similar to or better than those in the literatures.

Conclusions

The amino acid descriptors, SVMW, were derived from principal component analysis of 160 MoRSE descriptors and 99 WHIM descriptors only by theoretical calculation. Applying SVMW scales into peptide QSAR studies for three kinds of classical peptide analogues, the results are similar to or better than those in the literatures. Thus it is suggested the SVMW scales have multiple advantages, such as plentiful structural information, easy to get and good structural characterization ability. This method can be used widely in forecasting QSAR studies.

Acknowledgements: The authors appreciate the financial support from the National Natural Science Foundation of China (21301113)(21275094), the Scientific Research Planning Program of the Education Department of Shaanxi Province (2013JK0684), the Scientific Research Planning Program of Key laboratory of Shaanxi Province of China (2011SZS007), and the Graduate Innovation Fund of Shaanxi University of Science and Technology.

References

1. JIRÁCEK, J., YIOTAKIS, A., VINCENT, B., LECOQ, A., J. Biol. Chem., **270**, 1995, p. 21701.
2. MARRAUD, M., AUBRY, A., Biopolymers, **40**, 1996, p. 45.
3. ZALIANI, A., GANCIA, E., J. Chem. Inf. Comput. Sci., **39**, 1999, p. 525.
4. ZHOU, X., LI, Z.C., DAI, Z., ZOU, X.Y., J. Mol. Graph. Model, **10**, 2010, p. 1016.
5. JEZIERSKA, A., MACZY SKI, M., KOLL, A., RYNG, S., Arch. Pharm., **337**, 2004, p. 81.
6. KAR, S., HARDING, A.P., ROY, K., POPELIER, P.L.A., SAR QSAR Environ. Res., **21** 2010, p. 149.
7. TODESCINI, R., CONSONNI, V., MAURI AND, A., PAVAN, M., Dragon-web version 5.2, 2004.
8. TONG, J., CHEN, Y., LIU, S., J Chemometr., **26**,2012, p. 549.
9. WOLD, S., SJÖSTRÖM, M., ERIKSSON, L., Chemom. Intell. Lab. Syst., **58**, 2001, p. 109.
10. S. WOLD, J. TRYGG, A. BERGLUND, H. ANTTI, Chemom. Intell. Lab. Syst., **58**, 2001, p. 131.
11. JEUNEMAITRE, X., SOUBRIER, F., KOTELEVTSEV, Y.V., LIFTON, R.P., Cell, **71**, 1992, p. 169.
12. SAWA, H., KAWAGUCHI, H., MOCHIZUKI, N., ENDO, Mol. Y., Cell. Biochem., **132**, 1994, p. 15.
13. HASSALL, C.H., KRÖHN, A., MOODY, C.J., THOMAS, W.A., J. Chem. Soc. Perkin Trans. I. **23**, 1984, p. 155.
14. CUSHMAN, D.W., Mechanisms of action and clinical implications, 1981, p. 3.
15. LIANG, G. Z., ZHOU, P., ZHOU, Y. Q., ZHANG, X., Acta Chimica Sinica., **64**, 2006, p. 393.
16. ZHOU, P., ZHOU,Y., WU, S.R., LI, B., Chin. Sci. Bull., **51**, 2006, p. 524.
17. LONG, H.X., WANG, Y.Q., LIN, Y., LIN, Z.H., J. Chin. Chem. Soc., **57**, 2010, p. 417.
18. IWASHITA, T, Stat. Plan. Infer., **61**, 1997, p. 85.
19. GEORGE, J.P., CHEN, Z., SHAW, P., WASET, **50**, 2009, p. 970.
20. LIN, Z.H., LONG, H.X., BO, Z., WANG, Y.Q., WU, Y.Z., J. peptides., **29**, 2008, p. 1798.
21. ARMAS, R.R., DÍAZ, H.G., MOLINA, R., GONZÁLEZ, M.P., URIARTE, E., Bioorg. Med. Chem., **12**, 2004, p. 4815.
22. Collantes, E.R., Dunn, W.J., J. Med. Chem., **38**, 1995, p. 2705

Manuscript received: 29.04.2013