Amino Acids (2010) 38:805–816
DOI 10.1007/s00726-009-0287-y

ORIGINAL ARTICLE

ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues

Li Yang · Mao Shu · Kaiwang Ma · Hu Mei · Yongjun Jiang · Zhiliang Li

Received: 5 July 2008 / Accepted: 25 March 2009 / Published online: 17 April 2009

Abstract In this study, structural topology scale (ST-scale) was recruited as a novel structural topological descriptor derived from principal component analysis on 827 structural variables of 167 amino acids. By using partial least squares (PLS), we applied ST-scale for the study of quantitative sequence-activity models (QSAMs) on three peptide datasets (58 angiotensin-converting enzyme (ACE) inhibitors, 34 antimicrobial peptides (AMPs) and 89 elastase substrates (ES)). The results of QSAMs were superior to that of the earlier studies, with determination coefficient (r^2) and cross-validated (q^2) equal to 0.855, 0.774; 0.79, 0.371 (OSC-PLS: 0.995, 0.848) and 0.846, 0.747, respectively. Therefore, ST-scale descriptors were considered to be competent to extract information from 827 structural variables and relate with their bioactivities.

Keywords Peptides · Structural topological scale (ST-scale) · Principal component analysis (PCA) · Partial least squares regression (PLS) · Quantitative sequence-activity models (QSAM)

Introduction

Peptides and their analogues as multifunctional bioactive substances such as hormones, enzyme inhibitors, growth promoters, neurotransmitters, taste receptors, antibacterial agents and so on, have been paid considerable attention by pharmacologists. With the development of peptide library, thousands of different peptides have been designed, synthesized, and then subjected to a series of screening procedures and biological assays. To be effectively used, the biological data should be analyzed with multivariate quantitative structure-activity relationships (QSARs). For the properties of peptides, the precise amino acid sequence must be the key to a particular function or bioactivity of the peptide. The change of amino acid sequence in a peptide will impact its bioactivities. QSAMs, as a crucial case of QSARs, can explain the relationship between the amino acid positions as related to its bioactivities. The essence of QSAR is to express the relation of structural features with bioactivities; hence, structural representation is the key for success of QSAR. Sneath (1966) were the first to report the QSAR study of peptides, in which the variables of amino acid descriptors were expressed with physicochemical properties of 20 coded amino acids. Their study used QSAM to analyze oxytocin vasopressin analogues (Sneath 1966). Since then a number of quantitative

Electronic supplementary material The online version of this article (doi:10.1007/s00726-009-0287-y) contains supplementary material, which is available to authorized users.

L. Yang · M. Shu (✉) · H. Mei · Z. Li
Key Laboratory of Biorheological Science and Technology
(Chongqing University), Ministry of Education,
400030 Chongqing, People's Republic of China
e-mail: sm7507@126.com

L. Yang · M. Shu · H. Mei · Z. Li
College of Bioengineering, Chongqing University,
400030 Chongqing, People's Republic of China

K. Ma
College of Medical Technology and Engineering,
Henan University of Science and Technology,
471003 Luoyang, People's Republic of China

Y. Jiang
Key Laboratory for Molecular Design and Nutrition Engineering
of Ningbo City, Ningbo Technology Institute of Zhejiang
University, 315100 Ningbo, People's Republic of China

amino acid descriptors have been recruited and successfully applied in QSAR studies (Kidera et al. 1985; Hellberg et al. 1986, 1987, 1991; Cocchi and Johansson 1993; Collantes and Dunn 1995; Liu et al. 2001; Zaliani and Gancia 1999; Li et al. 2001). Hellberg et al. (1986, 1987, 1991) applied orthogonal transformation to 29 physicochemical properties of individual amino acids to produce one set of *Z* scales (*Z*) including hydrophobicity (*z*1), bulk (*z*2), and electrogenicity (*z*3). Collantes and Dunn (1995), on the basis of 3D structural characteristics of amino acid side chains, set two computable 3D descriptors-isotropic surface area (ISA) and electronic charge index (ECI). The *z*-scores and ISA-ECI descriptors have proved to be powerful tools for modeling a variety of bioactivities of small peptides with good results generated. However, there still exist limitations with these descriptors, for the available descriptors are only applicable for the coded amino acids. Since non-coded amino acids exist ubiquitously in nature and have been used for structural alteration and polypeptide modification, in recent development of peptidomimetics, it is not convenient to use QSAMs on structural representation of 20 coded amino acids to meet the subject development. Sandberg et al. (1998) extended *Z* scales up to 87 amino acids including 20 coded amino acids. However, because the structures of non-natural amino acids are diverse and their property data are difficult to collect experimentally, fewer studies have reported this. Herein, based on our group previous work (Mei et al. 2004, 2005; Liang et al. 2006; Liang and Li 2007), we have reported a novel structural topological scale (ST-scale) of amino acid descriptor on 827 structural variables of 167 amino acids using principal component analysis (PCA). In the present studies, we applied ST-scale to 58 angiotensin-converting enzyme (ACE) inhibitors, 34 antimicrobial peptides (AMPs) and 89 elastase substrates (ES).

Principles and methodologies

Structure of amino acid and ST-scale

The information about 167 amino acids was collected from the literatures (Sandberg et al. 1998; Jonsson et al. 1989; Böck et al. 1991; Atkins and Gesteland 2002; Yan et al. 2000; Robert and Phillips 2004; Caligiuri et al. 2006; Liu and Kit 2001). Table 1 lists their structural names (and Molecular structures are provided in supplementary Table S1). No. 1–20 are coded amino acids, and the rest are non-coded ones. The structure of 167 amino acids was constructed and optimized with Sybyl 6.8 package (Sybyl Version 6.8 2001, Tripos Inc, USA), and Dragon1.1 software (Milano Chemometrics and QSAR Research Group, Italy, 2001) was used to generate 827 descriptor parameters for each amino acid, as described in supplementary Table S2. These parameters are mainly related to constitutional, topological, geometrical, hydrophobic, electronic, and steric properties of the amino acids (Gilvez et al.1994; Rucker and Rucker 1993; Balaban et al. 1991; Diudea et al. 1995; Randic et al. 1994; Schuur et al. 1996; Gasteiger et al. 1996; Todeschini et al. 1995; Consonni et al.2002). In order to avoid overlap in descriptor parameters and hence increase in complexity of QSAR model, PCA was applied to the descriptor matrix of 167 amino acids. First, original variable matrix $\boldsymbol{X}_{167\times827}$ was mean centered and autoscaled, then eight significant principal components (8PCs) were extracted by PCA that explained 71.5% variances with the contribution from each as 40.11, 8.33, 6.71, 4.42, 4.03, 3.07, 2.66 and 2.4%, respectively. The score matrix $\boldsymbol{X}_{167\times8}$ of eight PCs were then used to replace the corresponding original data matrix $\boldsymbol{X}_{167\times827}$ with less information loss. From the loadings of the eight PCs (see Table S3), the extracted information was mainly related to molecular constitutional, topological, geometrical, connectivity information, atomic molecular electro-topological variation, and polarization and so on. For convenience, the eight PC scores for 167 amino acids were denoted as ST-scale with their values in Table 1, and SPSS 13.0 was used to implement PCA program.

Sequence representation and variable selection

For a set of peptides and their analogues, the chemical structure would be now characterized by describing each varied amino acid position with eight ST-scale values. For example, the chemical structure of di-peptide would be described by 16 (8 × 2) variables. Thus, a set of peptides and their analogues varied in *n* positions can be described by 8*n* variables.

For a QSAM dataset, not all the structural variables are relevant to bioactivities; therefore, the redundant variables should be deleted from the model in order to promote its robustness and predictive capability, especially when the number of variables is very large. Several variable selection methods including stepwise multiple regression (SMR), genetic algorithm (GA) (Rogers and Hopfinger 1994), and simulated annealing (SA) (Sutter et al. 1995) have been widely used to exclude irrelevant variables. In case of small variables, SMR should be the optimal one because it is less time-consuming and easy to implement. SPSS 13.00 was used to implement SMR.

PLS modeling

PLS is a widely used modeling method, for it has the advantage to overcome multicollinearity issues in an effective way and especially is suitable for the occasion

Table 1 167 amino acid and their ST-scale values

No.	Abbreviation	Name	ST1	ST2	ST3	ST4	ST5	ST6	ST7	ST8
1	Ala(A)	Alanine	−1.552	−0.791	−0.627	0.237	−0.461	−2.229	0.283	1.221
2	Arg(R)	Arginine	−0.059	0.731	−0.013	−0.096	−0.253	0.3	1.256	0.854
3	Asn(N)	Asparagine	−0.888	−0.057	−0.651	−0.214	0.917	0.164	−0.14	−0.166
4	Asp(D)	Aspartic acid	−0.907	−0.054	−0.781	−0.248	1.12	0.101	−0.245	−0.075
5	Cys(C)	Cysteine	−1.276	−0.401	0.134	0.859	−0.196	−0.72	0.639	−0.857
6	Gln(Q)	Glutamine	−0.622	0.228	−0.193	−0.105	0.418	0.474	0.172	0.408
7	Glu(E)	Glutamic acid	−0.629	0.39	−0.38	−0.366	0.635	0.514	0.175	0.367
8	Gly(G)	Glycine	−1.844	−0.018	−0.184	0.573	−0.728	−3.317	0.166	2.522
9	His(H)	Histidine	−0.225	0.361	0.079	−1.037	0.568	0.273	1.208	−0.001
10	Ile(I)	Isoleucine	−0.785	−1.01	−0.349	−0.097	−0.402	1.091	−0.139	−0.764
11	Leu(L)	Leucine	−0.826	−0.379	0.038	−0.059	−0.625	1.025	−0.229	−0.129
12	Lys(K)	Lysine	−0.504	0.245	0.297	−0.065	−0.387	1.011	0.525	0.553
13	Met(M)	Methionine	−0.693	0.498	0.658	0.457	−0.231	1.064	0.248	−0.778
14	Phe(P)	Phenylalanine	−0.019	0.024	1.08	−0.22	−0.937	0.57	−0.357	0.278
15	Pro(P)	Proline	−1.049	−0.407	−0.067	−0.066	−0.813	−0.89	0.021	−0.894
16	Ser(S)	Serine	−1.343	−0.311	−0.917	−0.049	0.549	−1.533	0.166	0.28
17	Thr(T)	Threonine	−1.061	−0.928	−0.911	−0.063	0.538	−0.775	−0.147	−0.717
18	Trp(W)	Tryptophan	0.853	0.039	0.26	−1.163	0.16	−0.202	1.01	0.195
19	Tyr(Y)	Tyrosine	0.308	0.569	1.1	−0.464	−0.144	−0.354	−1.099	0.162
20	Val(V)	Valine	−1.133	−0.893	−0.325	0.303	−0.561	−0.175	−0.02	−0.311
21	Acp	α−Aminocaprylic acid	−0.171	1.546	0.11	0.217	−1.067	0.92	0.802	0.943
22	Aec	(S)-2-aminoethyl-L-cysteine·HCl	−0.475	0.955	0.251	0.464	0.167	0.757	0.641	−0.283
23	Afa	Aminophenylacetate	−0.226	0.507	1.074	−0.669	−0.487	−0.009	−1.273	−0.666
24	Aib	α-Aminoisobytyric acid	−1.476	−1.115	−0.261	1.11	−1.134	−2.008	−0.609	1.471
25	Ail	Alloisoleucine	−0.826	−0.352	0.041	−0.041	−0.638	1.088	−0.216	−0.17
26	Alg	L-allylglycine	−0.987	−0.561	0.026	−0.181	−0.458	0.077	0.155	−0.197
27	Aba	α-Aminobutyric acid	−1.317	−0.367	−0.523	0.08	−0.308	−0.767	0.344	−0.045
28	Aph	*p*-Aminophenylalanine	0.332	0.666	1.085	−0.403	−0.258	−0.217	−1.305	−0.064
29	β-Ala	β-Alanine	−1.609	0.168	0.326	0.63	−0.88	−2.191	0.371	1.695
30	Brp	*p*-Bromophenylalanine	0.528	0.865	2.289	1.35	−0.151	−0.706	−0.509	0.298
31	Cha	Cyclohexylalanine	−0.099	0.303	0.604	0.088	−1.304	2.232	0.074	0.617
32	Cit	Citrulline	−0.078	1.881	−0.57	0.045	0.118	0.125	0.959	1.167
33	Cla	β-Chloroalanine	−1.285	−0.851	−0.003	0.871	0.145	−0.992	0.497	−0.411
34	Cle	Cycloleucine	−0.959	−0.812	0.239	0.749	−1.36	−0.167	−0.759	−0.64
35	Clp	*p*-chlorophenylalanine	0.403	0.276	1.707	0.15	−0.459	−0.538	−0.331	0.674
36	Cya	Cysteic acid	−0.716	−0.116	−0.902	0.734	1.646	0.843	−0.236	−0.912
37	Dab	2,4-Diaminobutyric acid	−1.084	0.092	−0.653	−0.224	0.294	0.179	0.304	−0.255
38	Dap	2,3-diaminopropionic acid	−1.349	−0.651	−0.885	0.026	0.126	−1.274	0.329	0.046
39	Dhp	3,4-Dehydroproline	−1.015	−0.709	0.08	−0.271	−0.83	−0.942	0.543	−0.929
40	Dha	3,4-Dihydroxyphenylalanine	0.656	0.114	−0.001	−0.907	1.144	−0.419	−0.287	−0.04
41	Fph	*p*-Fluorophenylalanine	0.326	0.591	1.023	−0.476	0.107	−0.212	−1.167	0.19
42	Gaa	D-Glucoseaminic acid	0.018	−0.407	−1.366	−0.331	1.855	1.496	−0.42	−0.47
43	Hag	Homoarginine	−0.051	1.502	−0.331	−0.061	−0.103	0.099	1.179	1.307
44	Hly	δ-Hydroxylysine · HCl	−0.306	0.305	−0.534	−0.166	0.463	1.286	0.522	1.146
45	Hnv	DL-β-hydroxynorvaline	−0.818	−0.959	−0.659	−0.288	0.383	0.572	−0.201	−0.807
46	Hog	Homoglutamine	−0.331	0.926	−0.253	−0.025	0.318	1.053	0.56	1.263
47	Hop	Homophenylalanine	0.358	0.912	1.062	−0.239	−0.842	−0.074	−0.638	0.733
48	Hos	Homoserine	−1.095	−0.261	−0.607	−0.246	0.387	−0.13	0.081	−0.09

Table 1 continued

No.	Abbreviation	Name	ST1	ST2	ST3	ST4	ST5	ST6	ST7	ST8
49	Hpr	Hydroxyproline	−0.766	−0.298	−0.611	−0.527	0.494	−0.066	−0.07	−0.681
50	Iph	*p*-Iodophenylalanine	0.718	0.34	3.525	2.911	0.094	−1.674	0.628	0.996
51	Ise	Isoserine	−1.59	−0.515	−1.108	−0.028	0.53	−2.653	0.045	1.564
52	Mle	α-Methylleucine	−0.708	−0.341	0.249	0.71	−1.004	1.61	−1.041	−0.191
53	Msm	DL-methionine-s-methylsulfoniumchloride	−0.472	0.178	0.943	0.72	−0.314	1.263	−0.018	−0.041
54	1Nala	*β*-(1-naphthyl)alanine	1.173	−1.17	0.856	−1.049	0.124	−0.818	−0.499	−0.299
55	2Nala	*β*-(2-naphthyl)alanine	1.205	−0.577	0.692	−1.033	−0.199	−1.319	−0.351	−0.169
56	Nle	Norleucine(or 2-aminohexanoic acid)	−0.708	0.358	0.171	−0.219	−0.533	1.164	0.097	0.436
57	Nma	N-methylalanine	−1.321	−1.039	−0.496	0.229	−0.45	−0.944	0.321	−0.374
58	Nva	Norvaline(or 2-aminopentanoic acid)	−1.031	0.089	−0.266	−0.121	−0.266	0.536	0.113	−0.314
59	Obs	O-benzylserine	0.505	0.999	0.513	−0.068	−0.75	−1.172	−0.468	0.473
60	Obt	O-benzyltyrosine	1.758	2.164	0.187	0.547	−0.925	−1.905	−2.421	−1.004
61	Oet	O-ethyltyrosine	0.722	1.552	0.453	−0.037	−0.219	−0.6	−1.716	−0.869
62	Oms	O-methylserine	−1.078	−0.117	−0.519	−0.156	0.396	−0.416	−0.09	−0.255
63	Omt	O-methylthreonine	−0.845	−0.798	−0.779	−0.121	0.402	0.274	−0.163	−1.044
64	Omy	O-methyltyrosine	0.458	0.544	0.866	−0.257	−0.371	−0.503	−1.061	0.03
65	Orn	Ornithine	−0.77	0.334	−0.236	−0.236	−0.056	0.822	0.386	0.363
66	Pen	Penicillamine	−0.925	−1.168	0.265	1.369	−0.523	0.055	−0.092	−1.167
67	Pga	Pyroglutamic acid	−0.833	−0.321	−0.324	−0.473	0.145	−0.209	0.273	−0.842
68	Pip	Pipecolic acid	−0.858	−0.268	0.102	0.063	−0.974	0.595	0.031	−0.819
69	Sar	Sarcosine	−1.588	0.394	0.287	0.513	−0.922	−1.977	0.245	1.171
70	Tfa	3,3,3-Trifluoroalanine	−1.016	−0.921	−1.395	0.214	2.09	−1.333	−1.456	0.825
71	Thp	6-Hydroxydopa	0.86	−0.246	−0.45	−1.097	1.698	−0.534	−0.466	−0.07
72	Vig	L-vinylglycine	−1.277	−0.573	−0.487	−0.118	−0.18	−1.199	0.36	−0.273
73	Aas	(-)-(2R)-2-amino-3-(2-aminoethylsulfonyl)propanoic acid dihydrochloride	−0.2	0.49	−0.366	0.574	1.289	1.596	0.023	−1.046
74	Ahd	(2S)-2-amino-9-hydroxy-4,7-dioxanonanoic acid	0.052	2.655	−1.078	0.296	0.404	−1.006	0.049	−0.744
75	Aho	(2S)-2-amino-6-hydroxy-4-oxahexanoic acid	−0.57	1.049	−0.455	−0.22	0.808	−0.052	0.144	0.292
76	Ahs	(-)-(2R)-2-amino-3-(2-hydroxyethylsulfonyl)propanoic acid	−0.211	0.5	−0.434	0.517	1.518	1.499	−0.119	−0.999
77	Ahp	(-)-(2R)-2-amino-3-(2-hydroxyethylsulfanyl)propanoic acid	−0.48	0.962	0.19	0.423	0.426	0.656	0.409	−0.228
78	Ahd	(2S)-2-amino-12-hydroxy-4,7,10-trioxadodecanoic acid	0.516	4.277	−2.015	0.819	0.27	−1.582	−0.318	−2.1
79	Dad	(2S)-2,9-diamino-4,7-dioxanonanoic acid	0.043	2.705	−1.133	0.33	0.186	−0.753	0.304	−0.835
80	Dat	(2S)-2,12-diamino-4,7,10-trioxadodecanoic acid	0.512	4.292	−2.03	0.864	0.213	−1.532	−0.222	−2.221
81	Dfn	(S)-5,5-difluoronorleucine	−0.364	0.237	−0.305	0.013	1.02	0.884	−1.008	1.725
82	Dfv	(S)-4,4-difluoronorvaline	−0.693	−0.415	−0.529	−0.074	1.271	0.14	−1.182	0.499
83	Dtc	(3R)-1-1-dioxo-[1,4]thiaziane-3-carboxylic acid	−0.481	−0.463	−0.288	0.617	0.696	1.379	0.358	−1.559
84	Hfn	(S)-4,4,5,5,6,6,6-heptafluoronorleucine	0.462	0.094	−1.571	0.756	3.958	0.825	−3.569	2.56
85	Pfn	(S)-5,5,6,6,6-pentafluoronorleucine	0.17	0.677	−1	0.497	2.984	0.679	−1.998	3.037
86	Pfv	(S)-4,4,5,5,5-pentafluoronorvaline	−0.109	−0.042	−1.201	0.282	3.106	0.724	−2.726	2.094
87	Tca	(3R)-1,4-thiazinane-3-carboxylic acid	−0.828	−0.359	0.694	0.76	−0.828	0.527	0.499	−1.373
88	Pyl	Pyrrolysine	0.858	1.385	−1.146	0.517	−1.082	0.715	1.707	−1.107
89	Ath	*β*-(9-Anthracenyl)alanine	2.049	−2.013	0.2	−0.528	−0.229	−1.617	−0.873	0.237
90	Bal	*β*-(3-Benzothienyl)alanine	0.817	−0.99	0.846	−0.399	−0.351	0.566	1.491	−0.229
91	Bip	*β*-(4,4′-Biphenyl)alanine	1.358	1.094	0.807	−0.04	−1.159	−1.141	−2.037	−1.213
92	Dip	*β*,*β*-Diphenylalanine	1.306	−0.68	1.265	−0.255	−0.828	0.532	−2.079	0.074

Table 1 continued

No.	Abbreviation	Name	ST1	ST2	ST3	ST4	ST5	ST6	ST7	ST8
93	Tbt	*β*-[3-(2,5,7-Tri-tert-butyl-indolyl)]alanine	3.025	−1.714	−3.315	3.518	−3.605	2.459	0.863	5.092
94	Tpc	*β*-{3-[2-(2,2,5,7,8-Pentamethyl-chroman-6-sulfonyl)-indolyl]}alanine	4.497	−2.785	−5.946	4.66	−2.052	−1.404	−1.133	−2.607
95	Asu	Aminosuberic acide	0.207	2.374	−0.711	0.463	−0.787	0.412	0.933	1.264
96	Hcy	Homocysteine	−0.947	0.007	0.642	0.678	−0.52	1.203	0.152	−0.936
97	Sta	Statine	−0.104	−0.371	−0.154	0.1	−0.162	1.672	0.052	0.546
98	Thi	*β*-(2-Thienyl)alanine	−0.517	−0.684	0.68	0.141	−0.371	0.527	0.563	−1.803
99	*γ*-Abu	L-*γ*-Aminobutyric acid	−1.265	0.63	0.558	0.201	−0.922	−0.144	0.127	1.15
100	Aca	*ε*-Aminocaproic acid	−0.625	1.814	0.827	0.302	−1.091	0.89	0.329	1.851
101	Ach	1-Aminocyclohexane-1-carboxylic acid	−0.825	−0.735	0.51	0.909	−2.161	1.442	−0.831	−0.044
102	Afb	*β*-Amino-*β*-phenyl-*p*-nitro-L-butyric acid	0.847	0.047	0.183	0.205	0.311	0.771	−2.003	0.609
103	Aoq	*α*-Amino-*β*-[4-(1,2-dihydro-2-oxo-quinolinyl)] propionic acid	0.964	−0.796	0.371	−0.97	0.303	−0.375	−0.027	0.07
104	Bpa	4′-Benzoylphenylalanine	1.5	1.588	−0.639	0.889	−1.456	0.781	−0.358	−1.763
105	Mas	*β*-Methyl aspartic aicd	−0.678	−1.02	−0.875	−0.219	0.913	0.566	−0.356	−0.74
106	Ceg	2-Chloroethylglycine	−1	0.201	0.076	0.378	0.274	0.804	0.21	−0.543
107	Cha	*β*-Cyclohexyl(*p*-methoxyl)-L-alanine	0.059	0.725	−0.23	0.342	−0.648	1.78	0.267	0.173
108	Dty	*α*,*β*-divinyltyrosine	1	−0.706	0.336	−0.172	−0.142	1.237	−1.373	−0.206
109	Chg	2-L-cyclohexylglycine	−0.374	−0.161	0.315	0.069	−1.12	2.281	−0.187	−0.027
110	Cpa	4-chlorophenylalanine	0.379	0.371	1.655	0.172	−0.654	−0.326	−0.187	0.758
111	Deg	*α*,*α*-Diethyl glycine	−1.011	−1.13	−0.006	0.768	−0.992	0.4	−0.667	−0.636
112	Dmt	2′,6′-Dimethyltyrosine	0.704	−0.675	0.365	−0.581	0.282	0.896	−0.623	0.136
113	Dvg	Divinyl glycine	−0.659	0.542	0.108	−0.572	−0.074	0.42	0.302	−0.292
114	Gav	2-Guanidine-5-amino-L-n-valeric acid	−0.163	0.266	−0.406	−0.308	0.397	1.666	0.425	0.631
115	Hat	2-Amino-6-hydroxytetralin-2-carboxylic acid	0.743	−0.814	0.342	−0.451	−0.099	0.766	−0.171	−0.687
116	Hai	2-Amino-5-hydroxyindan-2-carboxylic acid	0.512	−0.871	0.283	−1.017	0.462	0.271	−0.252	−0.643
117	Hpp	3-(4′-hyroxyphenyl)proline	0.716	−0.232	0.548	−0.677	−0.039	0.226	−1.271	−0.764
118	Ing	1-Indanylglycine	0.419	−0.482	0.14	−0.629	−0.556	1.61	0.414	−1.101
119	Mhp	*p*-Methoxyhomophenylalanine	0.698	1.123	0.51	−0.018	−0.476	−0.699	−1.065	0.456
120	Oct	n-Octylglycine	0.182	2.918	−0.655	0.703	−1.481	0.923	0.622	−0.044
121	Oic	Octahydroindole-2-carboxylic acid	−0.001	−0.499	−0.049	−0.548	−0.65	2.018	−0.092	−0.988
122	Pal	*β*-Pyridylalanine	0.026	−0.322	0.783	−0.813	−0.078	0.148	−0.02	0.31
123	Tic	1,2,3,4-Tetrahydroisoquinoline-3-carboxylic acid	0.345	−0.315	0.4	−0.847	0.08	0.414	−0.436	−1.33
124	Thz	L-4-Thiazolidine carboxylic acid	−1.036	−0.522	0.513	0.722	−0.722	−0.629	0.754	−1.442
125	Tle	L-tert-butylglycine	−0.945	−1.28	−0.325	0.682	−0.784	0.377	−0.389	−0.578
126	Dpg	Diphenylglycine	1.007	−0.898	1.356	−0.238	−0.628	0.35	−3.165	−0.331
127	Dbz	Dibenzylglycine	1.511	−0.489	1.556	−0.11	−1.827	−0.461	−2.083	1.596
128	*β*-Phe	*β*-Phenylalanine	0.059	0.09	1.271	−0.575	−0.416	0.263	−0.963	−0.411
129	*α*-Abu	*α*-Aminobutyric acid	−1.31	−0.422	−0.483	0.069	−0.329	−0.822	0.361	−0.168
130	Mpr	3-Methyproline	−0.812	−1.058	−0.53	−0.362	−0.267	0.032	0.138	−1.546
131	Hva	3-Hydroxyvaline	−0.976	−1.107	−0.768	0.503	0.057	−0.338	−0.598	−0.422
132	Dcp	3,5-Dihydroxy-4-chloro-phenylalanine	0.745	0.152	0.313	−0.04	1.275	−0.533	−0.161	0.227
133	Car	*β*-Carbonylarginine	0.05	1.189	−0.734	−0.023	0.583	0.41	1.167	0.696
134	Has	*β*-Hydroxyaspartate	−0.711	−0.836	−1.174	−0.325	1.648	0.169	−0.629	−0.611
135	1Nag	1-Naphthylglycine	0.902	−1.657	0.817	−1.316	0.567	−0.587	−1.376	−1.074
136	2Nag	1-Naphthylglycine	0.987	−0.845	0.83	−1.268	0.429	−1.461	−0.705	−0.625
137	Cpc	1-Aminocyclopropanecarboxylic acid	−1.515	−0.912	0.101	1.005	−2.001	−2.041	0.163	1.894
138	Hyp	4-Hydroxypyrrolidine-2-carboxylic acid	−0.781	−0.541	−0.57	−0.566	0.221	−0.038	0.024	−0.64
139	Aad	2-Aminohexanedioic acid	−0.343	1.195	−0.416	−0.102	0.676	0.93	0.585	1.067

Table 1 continued

No.	Abbreviation	Name	ST1	ST2	ST3	ST4	ST5	ST6	ST7	ST8
140	Ppa	(S)-2-amino-3-(pyridin-4-yl)propanoic acid	−0.069	0.398	0.791	−0.501	−0.274	0.686	−0.04	0.115
141	Pra	2-Aminopent-4-ynoic acid	−1.051	0.135	−0.151	−0.222	−0.081	0.106	0.434	−0.94
142	Apa	(S)-2-amino-3-(4-cyanopheyl)propanoic acid	0.55	1.03	1.003	−0.375	−0.465	−0.499	−0.811	−0.55
143	Npa	(S)-2-amino-3-(4-nitropheyl)propanoic acid	0.61	0.412	0.726	−0.435	−0.031	−0.311	−1.087	0.227
144	Cpp	(S)-2-amino-3-*p*-tolylpropanoic acid	0.303	0.211	1.316	−0.295	−0.977	−0.091	−0.609	0.477
145	Npp	(S)-2-amino-3-(4-azidophenyl)propanoic acid	0.784	0.739	0.723	−0.547	−0.165	−0.532	−0.959	−0.22
146	Lpa	(S)-2-amino-3-(3-chlorophenyl)propanoic acid	0.413	−0.045	0.919	−0.307	0.634	0.23	0.455	−0.163
147	Bpp	(S)-2-amino-3-(3,5-dibromo-4-hydroxyphenyl)propanoic acid	1.073	−0.398	2.077	3.272	2.652	0.181	0.769	0.015
148	Ipp	(S)-2-amino-3-(4-hydroxy-3,5-diiodophenyl)propanoic acid	1.382	−1.054	4.314	7.008	4.495	−0.535	3.205	−1.168
149	1Fpp	(S)-2-amino-3-(5-fluoro-1H-indol-3-yl)propanoic acid	1.055	−0.363	−0.182	−1.356	0.497	−0.408	1.516	0.575
150	2Fpp	(S)-2-amino-3-(6-fluoro-1H-indol-3-yl)propanoic acid	1.058	−0.5	0.01	−1.349	0.468	−0.526	1.323	0.073
151	Opp	(S)-2-amino-3-(5-hydroxy-1H-indol-3-yl)propanoic acid	1.066	−0.33	−0.159	−1.325	0.409	−0.474	1.286	0.364
152	1Mpp	(S)-2-amino-3-(2-methyl-1H-indol-3-yl)propanoic acid	0.97	−0.989	0.143	−1.102	−0.195	0.354	1.357	0.476
153	2Mpp	(S)-2-amino-3-(4-methyl-1H-indol-3-yl)propanoic acid	1.026	−0.629	−0.163	−1.123	0.259	0.075	1.796	0.17
154	3Mpp	(S)-2-amino-3-(5-methyl-1H-indol-3-yl)propanoic acid	1.054	−0.334	−0.05	−1.208	−0.118	−0.1	1.676	0.403
155	4Mpp	S)-2-amino-3-(6-methyl-1H-indol-3-yl)propanoic acid	1.029	−0.801	0.132	−1.133	−0.202	−0.118	1.63	0.221
156	5Mpp	(S)-2-amino-3-(7-methyl-1H-indol-3-yl)propanoic acid	1.066	−0.231	−0.165	−0.973	−0.015	−0.12	1.451	−0.119
157	1Mop	(S)-2-amino-3-(5-methoxy-1H-indol-3-yl)propanoic acid	1.244	−0.295	−0.407	−0.997	0.139	−0.647	1.321	0.504
158	2Mop	(S)-2-amino-3-(6-methoxy-1H-indol-3-yl)propanoic acid	1.23	0.483	−0.516	−0.961	0.43	−0.634	0.881	−0.87
159	1Npr	(S)-2-amino-3-(4-azido-1H-indol-3-yl)propanoic acid	1.456	−0.738	−0.739	−1.135	0.149	−0.607	1.323	0.958
160	2Npr	(S)-2-amino-3-(5-azido-1H-indol-3-yl)propanoic acid	1.471	−0.396	−0.619	−1.153	0.227	−0.898	1.184	0.493
161	3Npr	(S)-2-amino-3-(6-azido-1H-indol-3-yl)propanoic acid	1.535	0.378	−0.7	−1.082	0.4	−1.081	0.831	−0.46
162	4Npr	(S)-2-amino-3-(7-azido-1H-indol-3-yl)propanoic acid	1.523	0.161	−0.621	−1.102	0.27	−1.098	1.108	−0.336
163	1Cpr	(S)-2-amino-3-(4-chloro-1H-indol-3-yl)propanoic acid	1.106	−0.624	0.212	−0.759	0.765	−0.045	1.799	0.262
164	2Cpr	(S)-2-amino-3-(5-chloro-1H-indol-3-yl)propanoic acid	1.139	−0.361	0.229	−0.835	0.235	−0.261	2.157	0.488
165	3Cpr	(S)-2-amino-3-(6-chloro-1H-indol-3-yl)propanoic acid	1.137	0.04	0.089	−0.857	0.339	−0.339	1.612	−0.239
166	4Cpr	(S)-2-amino-3-(7-chloro-1H-indol-3-yl)propanoic acid	1.122	0.334	0.125	−0.685	0.256	−0.508	1.777	0.037
167	Fpr	(S)-2-amino-3-(2-(difluoromethyl)-1H-indol-3-yl)propanoic acid	1.318	−1.243	−0.239	−0.792	0.529	0.439	0.938	1.657

that sample number is smaller than the variable number. In addition, the desirable property of PLS is that the precision of the model parameters will be improved with increasing the number of relevant variable and observation (Wold et al. 2001a, b). PLS was implemented by Simca-P 10.0 software (Umetrics AB, Sweden, 2004).

Results and discussions

QSAM for 58 ACE inhibitors

A set of 58 dipeptides of ACE inhibitors, as a classical data set for QSAMs studies (Hellberg et al. 1991; Collantes and Dunn 1995; Cocchi and Johansson 1993; Zaliani and Gancia 1999; Li et al. 2001; Mei et al. 2004, 2005), were used to test the effectiveness of diverse amino acid descriptors. The bioactivity of ACE inhibitors was expressed in pIC_{50}, and the structural information of each dipeptide was quantified by 16 ST-scale variables. SMR was then utilized to screen redundant variables according to the significance of its Fisher test. The optimal variable number was determined by the cross-validation q^2 of the constructed PLS model (Supplementary Fig. 1). According to q^2, the seventh step was reached the optimal model. Finally, we obtained a 7-variable PLS model in which five PLS components were enough to account for 85.5% variances of Y variables with cross-validation up to 77.4% and RMSE up to 0.4. Table 2 presents experimental data and calculated values (cald_1) for the 7-variable PLS model. Figure 1 presents the plot of calculated values against experimental data for ACE inhibitors. Most samples were uniformly scattered around the diagonal except for the 34th one, whose calculated value deviated far from its experimental value than the others; this may be because the sample consist two bulky amino acid residues (i.e. isoleucine I and phenylalanine F), which means the capacity of its topological structure would be dramatically larger than that of the others with the same activity rank. The model was further validated by response permutations (Eriksson et al. 1997), and as described in Fig. 2. For a valid model, the desirable interception limits should be $r^2 < 0.3$ and $q^2 < 0.05$. If both limits were to exceed, the model should be treated with caution. The high r^2 values for the permutations suggested that a high r^2 could be obtained with a random y-vector. However, a high q^2 value cannot be obtained with a random y-vector as shown, and thus the model could be regarded as valid.

Table 2 Sequences of ACE with observed and calculated activity

No.	Peptide	Obsd	Cal_1	Cal_2	No.	Peptide	Obsd	Cal_1	Cal_2
1	PG	1.77	2.83	2.61	30	GM*	2.85	2.15	2.14
2	DG*	1.85	1.94	2.06	31	GI*	2.92	2.65	2.5
3	EA*	2	2.3	2.05	32	IG*	2.92	2.9	2.37
4	EG	2	2	2.6	33	VG	2.96	2.6	2.43
5	TG	2	2.67	2.31	34	IF	3.03	4.16	3.91
6	GD*	2.04	2.22	2.11	35	FR	3.04	3.29	3.2
7	LG*	2.06	2.58	2.4	36	GF	3.2	3.47	3.22
8	SG	2.07	2.05	2.09	37	AA*	3.21	2.86	2.61
9	QG	2.13	2.07	2.08	38	RA*	3.34	2.87	2.65
10	GG*	2.14	1.96	1.82	39	YA*	3.34	3	2.79
11	GQ	2.15	2.26	2.05	40	GP	3.35	3.01	3.19
12	HG	2.2	2.12	2.15	41	VP*	3.38	3.65	3.8
13	WG*	2.23	2.21	2.18	42	KA*	3.42	3	2.75
14	GT*	2.24	3.12	3.07	43	LA*	3.51	3.25	2.95
15	GE*	2.27	2.26	1.86	44	AP	3.64	3.24	3.43
16	GK*	2.27	2.09	1.94	45	RF	3.64	3.7	3.51
17	MG*	2.32	2.59	2.47	46	GY*	3.68	3.91	4.21
18	GV	2.34	2.79	2.44	47	AF*	3.72	3.69	3.47
19	DA	2.42	2.4	2.45	48	RP	3.74	3.25	3.48
20	GS	2.42	2.61	2.61	49	IP*	3.89	3.7	3.87
21	FG	2.43	2.59	2.37	50	AY*	4.06	4.13	4.45
22	GR	2.49	2.66	2.65	51	VF	4.28	4.1	3.83
23	HL*	2.49	2.32	2.2	52	GW	4.52	4.77	4.98
24	KG	2.49	2.52	2.2	53	VY*	4.66	4.54	4.81
25	GH	2.51	2.91	2.97	54	RW*	4.8	5.01	5.27
26	AG*	2.6	2.37	2.06	55	AW*	5	4.99	5.22
27	GL	2.6	2.19	1.88	56	IY	5.43	4.6	4.89
28	GA	2.7	2.63	2.37	57	IW	5.7	5.46	5.66
29	YG*	2.7	2.33	2.24	58	VW	5.8	5.41	5.59

$calcd_1$ Calculated value by seven ST-scales variables PLS model, *$calcd_2$* Calculated value by D-optimal seven ST-scales variables PLS model

The validity and stability of the model were also reached by the external model validation (Golbraikh and Tropsha 2002). D-optimal algorithm (Gramatica et al. 2004) was used to divide the sample set into training and test set. 29 samples were treated as training set and used to construct QSAM model while the rest 29 samples were taken as the test set (highlighted with "*" in Table 2). Algorithm D-optimal was implemented by software MatLab 7.0 (Version 7.0.1.450 release 14.1, MathWorks, Natick, MA, 2004). A 7-ST-scale variables PLS model was constructed for the training set with its $r^2 = 0.892$, $q^2 = 0.734$ and RMSEE = 0.41. The constructed model was then used to predict the test set with the results of $r^2_{\text{ext}} = 0.815$ and $\text{RMSE}_{\text{ext}} = 0.46$ (details as cald_2 in Table 2) confirming that ST-scale model was stable and reliable.

The results from literatures and the models using ST-scale descriptors are given in Table 3. $r^{\ 2}$ values obtained with ST-scale (0.855) are far superior to those cited in literature. Similarly, q^2 of ST-scale model was also found to be superior to those descriptors models, thereby suggesting a good stability of the ST-scale model.

QSAM for 34 antimicrobial peptides

Table 4 presents the sequences of 34 antimicrobial peptides in different lengths and their corresponding bioactivities as expressed with the logarithmic value, i.e., the number of *Staphylococcus aureus* killed within 2 h (Patel et al. 1998). The peptide sequence was parameterized with ST-scale descriptors, and the peptides were composed of different amounts of amino acids obtained from different amounts of ST-scale variables. To achieve equal variable number over all the peptides in a sample set, auto cross-covariance (ACCs) was carried out as described in the literature (Sjöstrom et al. 1995). By ACCs, 832 cross-variables were generated for each peptide sequence of the sample set. A PLS model with three significant latent variables was obtained with $r^2 = 0.790$. In order to improve the predictive ability and quality of the model, the PLS model was constructed after orthogonal signal correction (OSC) (Wold et al. 2001a, b) (filtering out the overlapped information from X matrix) on the 832 cross-variables. Three latent variables were obtained. The resultant model was found to be extremely advanced, with $r^2 = 0.995$ and q^2 up to 0.848. Figure 3 shows the plot of calculated value against experimental data for AMPs. All the samples were uniformly scattered around the diagonal for OSC-PLS model. The final OSC-PLS model was further validated by random permutation test (Eriksson et al. 1997). Figure 4 summarizes the 20 permutation and cross-validation rounds for the response: intercepts, $r^2 = 0.66$ and $q^2 = -0.44$. Therefore, the ST-scale descriptor model can be considered valid and stable.

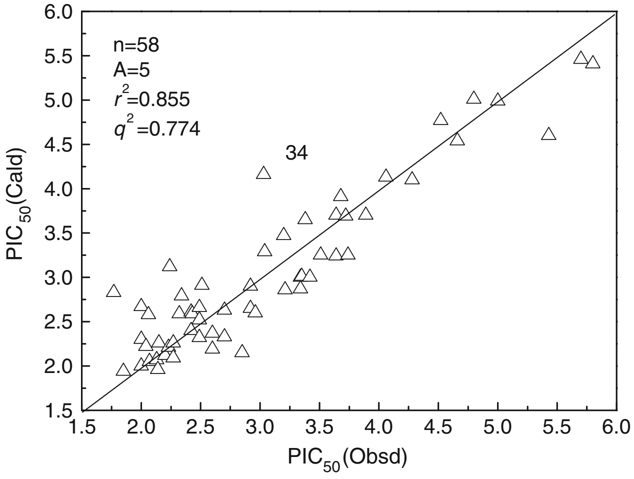

Fig. 1 Plots of calculated versus observed values for 58 ACE inhibitors by the PLS model

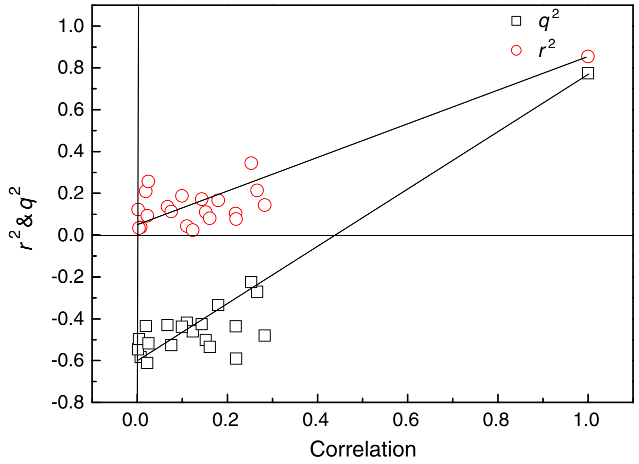

Fig. 2 Permutation validation of the PLS model; the figure summarized the 20 permutation and cross-validation rounds for two responses: intercepts, $r^2 = 0.04$ and $q^2 = -0.63$

The results of ST-scales model and other models are given in Table 5. A comparison shows r^2 of ST-scale model is superior to the other descriptors models; however, q^2 of ST-scale model is slightly inferior to that of the z-scale model.

QSAM for 89 elastase substrates

Elastase is a serine protease that participates in the pathogenesis of some diseases, e.g., emphysema. Nomizu et al. (1993) have reported 89 synthetic peptide substrates of porcine pancreatic elastase. For the 89 simulating elastase substrates, their sequences consist of Suc-A-B-Ala-pNa (Suc: succinyl; pNa: p-nitroanilide), in which A and B are varying residuals. Using spectrophotometry, logkcat/Km of reaction kinetics was determined by tracing the yield of p-nitroaniline that was catalyzed by its products (Table 6). The peptide structure was first quantified by 16 ST-scale descriptors, and GA-PLS was used to select variables (Parameters were set as follows: the number of population was 200, the maximum number of generations was 200, the

Table 3 Statistical parameters of PLS models for ACE inhibitors

No.	Descriptor	Number of descriptors	A	r^2	q^2	RMSEE
1	ISA-ECI (Collantes and Dunn 1995)	4	2	0.7	nd	nd
2	Z-scale (Hellberg et al. 1991)	6	2	0.77	0.723	nd
3	t-score (Cocchi and Johansson 1993)	14	1	0.744	nd	0.5
4	MS-WHIM (Zaliani and Gancia 1999)	6	2	0.708	0.637	0.54
5	VMEE (Li et al. 2001)	10	2	0.741	0.711	0.5
6	VHSE (Mei et al. 2005)	5	1	0.77	0.745	0.48
7	VSTV (Mei et al. 2004)	6	1	0.789	0.767	0.46
8	ST-scale	7	5	0.855	0.774	0.40

r^2 cumulative squared multiple correlation coefficient, q^2: cumulative cross-validated r^2, A the number of principal components of PLS model, *RMSEE* root mean square error of estimation for the training set, *nd* not determined

Table 4 Calculated value against observed value for 34 antimicrobial peptides

No.	Peptide sequence	Obsd	Calcd1	Calcd2
1	ESKAAKAAKKAAKAKASE	0.24	1.37	0.48
2	EKTLARTAAKTALKK	0.43	0.36	0.32
3	ETELAKKALKALKLKKLA	0.47	1.51	0.10
4	LSSALSALSSALSSK	0.54	1.08	0.50
5	GWLLLEYIPVIAAL	0.54	−0.95	0.48
6	ERSAAKSAARSLARR	0.67	−1.37	0.34
7	GESLASKAAKAAER	0.78	1.56	0.78
8	LLAILLLALLALRKKVLA	0.99	3.46	1.26
9	QKALAKLAKKALKALAKQ	1.03	2.76	1.14
10	ARLAKKALRRLAKKD	1.46	2.58	1.65
11	ESSLKKKALSKLSKLLKKG	2.57	2.16	2.45
12	ELAKKALRALKKALKSAK	2.75	4.07	2.92
13	LALLLKILLLKKLKA	3.38	4.78	3.34
14	ELAKKALKALKKALKSAR	3.64	3.81	3.61
15	QKAASRLLRALSKLLEAF	5.65	5.36	5.54
16	LALLKVLLRKIKKAL	5.68	5.23	5.57
17	FASLLGKLAKKLAKKALK	5.74	5.05	5.53
18	FASLLGKALKALLAKLAKQ	5.74	4.64	5.72
19	AASKALRTASRLARSLLT	5.85	4.99	5.80
20	LLKKLLRAASKALSLL	5.9	5.62	5.95
21	AASKAAKTLAKLLSSLLKL	5.96	5.14	5.99
22	FASLLGKALKALAKQ	6	5.54	6.25
23	LKALKKLAKKLKKLA	7	5.77	7.08
24	EKAAAKSAAAKTLARR	0.43	0.91	0.66
25	VSSKYLSKALVKAGR	0.54	0.44	0.86
26	EAALKAALDLAAKLA	0.75	1.64	0.57
27	ESLAKALSKEALKALK	1	1.24	0.92
28	ESLKARSLKKSLKLKKLL	1.58	2.70	1.81
29	ETFAKKALKALEKLLKKG	2.77	2.87	2.80
30	VSSKYLSKVKVKAGK	4	1.03	3.63
31	ALKAALLAILKIVRVIKK	5.68	5.53	5.61
32	AAKKLSKLLKTLLKLL	5.76	5.92	5.84
33	KALKKLLKLASSLLTAL	5.9	5.37	6.07
34	LKLLKKLLKKLKKLL	6.94	6.19	6.82

calcd1 Calculated value by ST-scales PLS model, *calcd2* Calculated value by ST-scales OSC-PLS model

generation gaps was 0.8, the crossover frequency was 0.5, the mutation rate was 0.005, and the fitting function was q^2, genetic algorithm was implemented by software MatLab 7.0). We obtained an optimal 12-variable PLS model, in which only three PLS components were enough to account for 84.6% variances of Y variables with cross-validation up to 74.7% and RMSE to 0.229. Figure 5 shows the plot of calculated value against experimental data for Elastase substrates, and all samples were uniformly scattered around the diagonal for the model. Figure 6 presents the 20 permutations and cross-validation rounds for the response:

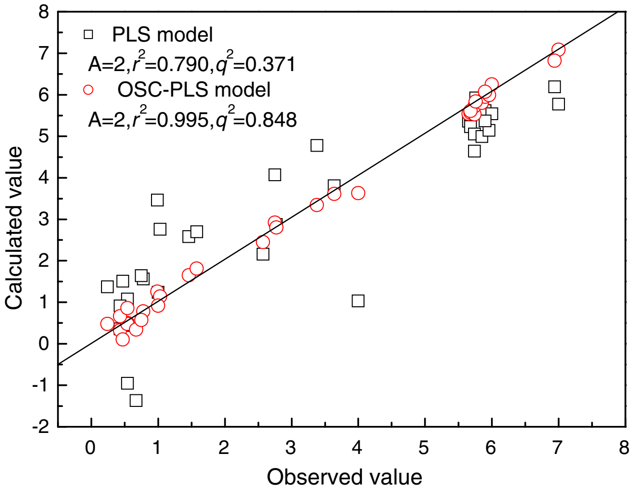

Fig. 3 Plots of calculated versus observed values for 34 AMPs by the PLS model

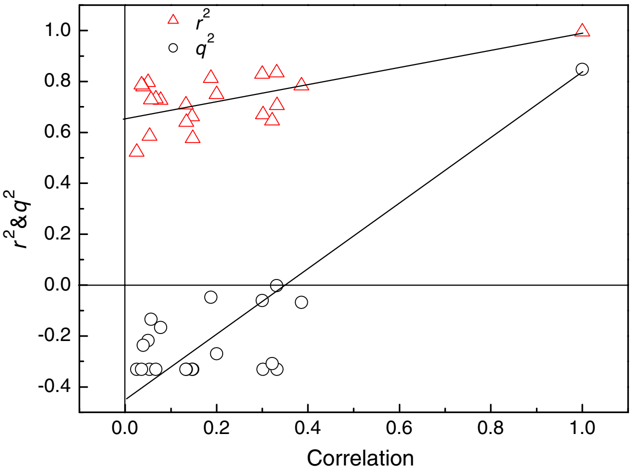

Fig 4 Permutation validation of the PLS model; the figure summarized the 20 permutation and cross-validation rounds for two responses: intercepts, $r^2 = 0.66$ and $q^2 = -0.44$

Table 5 Comparison of the different descriptor models for 34 antimicrobial peptides

No.	Descriptor	Model	*A*	r^2	q^2	RMSEE
1	ST-scale	PLS	2	0.790	0.371	1.153
2	*z*-Scale	PLS	1	0.673	0.542	1.415
3	ISA-ECI	PLS	1	0.384	0.254	1.934
4	ST-scale	OSC-PLS	3	0.995	0.848	0.188
5	*z*-scale	OSC-PLS	2	0.996	0.966	0.155
6	ISA-ECI	OSC-PLS	1	0.659	0.642	1.446

r^2 cumulative squared multiple correlation coefficient, q^2 cumulative cross-validated r^2, *A* the number of components in a PLS model, *RMSEE* root mean square error of estimation for the training set

Table 6 The sequences of elastase substrates with observed and calculated activity

No.	Code	log kcat/Km	Calcd	No.	Code	Log kcat/Km	Calcd
1	GlyAla	2.695	2.957	46	IlePhe	2.971	2.946
2	GlyVal	2.703	3.029	47	IleAbu	3.982	4.078
3	GlyLeu	2.328	2.489	48	IleNva	3.833	4.090
4	GlyIle	2.704	2.851	49	IleNle	3.992	4.081
5	GlyPro	2.746	3.586	50	PheGly	2.225	2.587
6	GlyPhe	1.881	2.134	51	PheAla	3.465	3.447
7	GlyAbu	2.832	3.266	52	PheVal	3.303	3.519
8	GlyNva	2.882	3.278	53	PheLeu	2.965	2.979
9	GlyNle	2.994	3.269	54	PheIle	3.236	3.341
10	AlaGly	2.422	2.645	55	PhePro	3.893	4.076
11	AlaAla	3.711	3.505	56	PhePhe	2.605	2.624
12	AlaVal	3.93	3.578	57	PheAbu	3.61	3.756
13	AlaLeu	3.322	3.038	58	PheNva	3.579	3.768
14	AlaIle	3.869	3.400	59	PheNle	3.66	3.759
15	AlaPro	4.23	4.135	60	AbuGly	2.658	2.715
16	AlaPhe	3.303	2.683	61	AbuAla	3.931	3.575
17	AlaAbu	3.92	3.815	62	AbuVal	3.834	3.648
18	AlaNva	3.985	3.827	63	AbuLeu	3.435	3.107
19	AlaNle	4.265	3.818	64	AbuIle	3.876	3.470
20	ValGly	2.974	2.920	65	AbuPro	4.22	4.205
21	ValAla	3.97	3.780	66	AbuPhe	3.179	2.753
22	ValVal	3.92	3.852	67	AbuAbu	4.107	3.885
23	ValLeu	3.318	3.312	68	AbuNva	4.029	3.897
24	ValIle	3.71	3.674	69	AbuNle	4.124	3.888
25	ValPro	4.509	4.409	70	NvaGly	2.822	2.974
26	ValPhe	3.225	2.957	71	NvaAla	3.699	3.834
27	ValAbu	4.14	4.089	72	NvaVal	3.863	3.906
28	ValNva	4.025	4.101	73	NvaLeu	3.226	3.366
29	ValNle	4.09	4.092	74	NvaIle	3.631	3.728
30	LeuGly	2.628	2.653	75	NvaPro	4.238	4.463
31	LeuAla	3.781	3.513	76	NvaPhe	3.124	3.011
32	LeuVal	3.595	3.586	77	NvaAbu	4.061	4.143
33	LeuLeu	3.072	3.046	78	NvaNva	4.057	4.155
34	LeuIle	3.539	3.408	79	NvaNle	4.057	4.146
35	LeuPro	4.14	4.143	80	NleGly	2.73	2.840
36	LeuPhe	2.881	2.691	81	NleAla	3.814	3.700
37	LeuAbu	3.895	3.823	82	NleVal	3.653	3.772
38	LeuNva	3.88	3.835	83	NleLeu	2.947	3.232
39	LeuNle	4.041	3.826	84	NleIle	3.662	3.594
40	IleGly	2.9	2.908	85	NlePro	4.185	4.330
41	IleAla	3.838	3.768	86	NlePhe	3.158	2.877
42	IleVal	3.674	3.841	87	NleAbu	4.025	4.009
43	IleLeu	3.045	3.301	88	NleNva	3.914	4.021
44	IleIle	3.585	3.663	89	NleNle	4.173	4.012
45	IlePro	4.346	4.398				

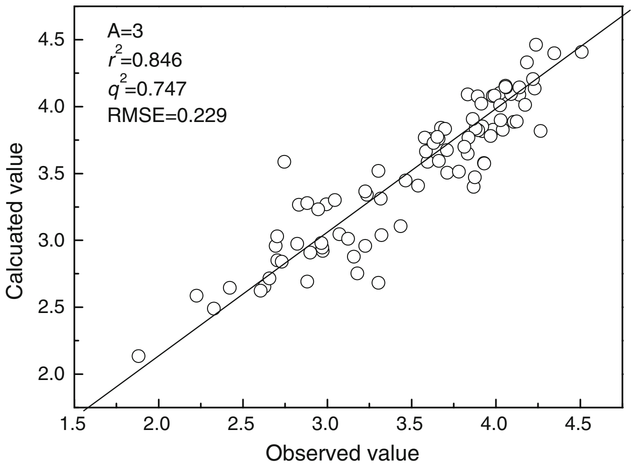

Fig. 5 Plots of calculated versus observed values for 89 elastase substrates by the PLS model

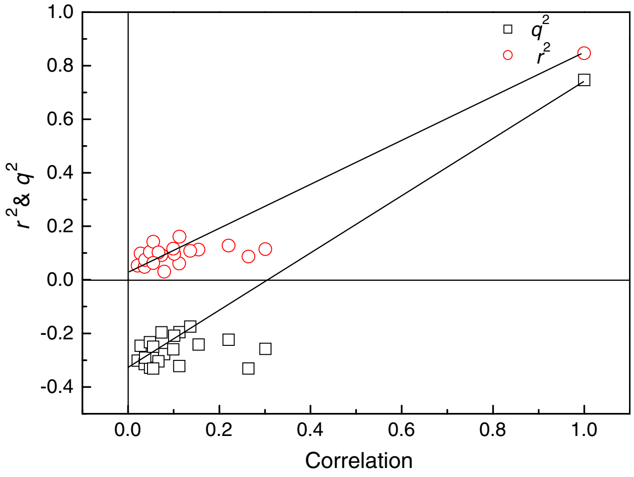

Fig. 6 Permutation validation of the PLS model, the figure summarized the 20 permutation and cross-validation rounds for two responses: intercepts, $r^2 = 0.01$ and $q^2 = -0.37$

intercepts, $r^2 = 0.01$ and $q^2 = -0.37$. From the results the ST-scale descriptor model could be considered valid and stable.

The *z*-scale model was constructed by Kimura et al. (1996) with its correlation coefficient $r^2 = 0.754$ and $q^2 = 0.629$. By comparison, it shows that the whole modeling qualities of ST-scale model, especially the q^2 indicting stabilities and the generalized abilities of the model, was superior to that reported in previous literatures.

Conclusion

Structural description is critical to the success of QSAMs. As it is well known, a good descriptor should contain as much chemical information relating to bioactivities as possible. In this study, a novel ST-scale was proposed by PCA on the 827 structural and topological variables of 167 amino acids. By applying ST-scale to different peptide datasets, the constructed QSAMs were stable and reliable. The results showed that ST-scale was capable of representing peptide for those composed not only of coded amino acids as ACE inhibitors and AMPs, but also of non-coded amino acids as elastase substrates. Thus, ST-scale had a promising prospect in QSAMs studies for peptide analogues.

Acknowledgments This work was supported by the "863" Program (No.2006AA02Z312, China), Innovative Group Program for postgraduates of Chongqing University (Science and Innovation Fund, No.200711C1A0010260) and Innovation Ability Training Foundation of Chongqing University (No.CDCX008). Special thanks to Dr. Gurinder K Singh, for her useful suggestions and generous assistance in the preparation process of this article.

References

Atkins JF, Gesteland R (2002) The 22nd amino acid. Science 296:1409–1410

Balaban AT, Ciubotariu D, Medeleanu M (1991) Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors. J Chem Inf Comput Sci 31:517–523

Böck A, Forchhammer K, Heider J, Leinfelder W, Sawers G, Veprek B, Zinoni F (1991) Selenocysteine: the 21st amino acid. Mol Microbiol 5:515–520

Caligiuri A, D'Arrigo P, Rosini E, Tessaro D, Molla G, Servi S, Pollegioni L (2006) Enzymatic conversion of unnatural amino acids by yeast d-amino acid oxidase. Adv Synth Catal 348:2183–2190

Cocchi M, Johansson E (1993) Amino acids characterization by GRID and multivariate data analysis. Quant Struct Act Relat 12:1–8

Collantes ER, Dunn WJ (1995) Amino acid side chain descriptors for quantitative structure activity relationship studies of peptide analogues. J Med Chem 38:2705–2713

Consonni V, Todeschini R, Pavan M (2002) Structure/response correla-tions and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci 42:682–692

Diudea MV, Horvath D, Graovac A (1995) Molecular topology.15. 3D distance matrices and related topological indices. J Chem Inf Comput Sci 35:129–135

Eriksson L, Johansson E, Miiller M, Wold S (1997) Cluster-based design in environmental QSAR. Quant Struct Act Relat 16:383–390

Gasteiger J, Sadowski J, Schuur J, Selzer P, Steinhauer L, Steinhauer V (1996) Chemical information in 3D space. J Chem Inf Comput Sci 36:1030–1037

Gilvez J, Garcia R, Salabert MT, Soler R (1994) Charge indexes: new topological descriptors. J Chem Inf Comput Sci 34:520–525

Golbraikh A, Tropsha A (2002) Beware of q^2!. J Mol Graphics Mod 20:269–276

Gramatica P, Pilutti P, Papa E (2004) Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling. J Chem Inf Comput Sci 44:1794–1802

Hellberg S, Sjöström M, Wold S (1986) The prediction of bradykinin potentiating potency of peptapetides. An example of peptide quantitave structure-activity relationship. Acta Chem Scand B 40:135–140

Hellberg S, Sjöström M, Skagerberg B (1987) Peptide quantitative structure-activity relationships, a multivariate approach. J Med Chem 30:1126–1135

Hellberg S, Eriksson L, Jonsson J, Lindgren F, Sjöström M, Skagerberg B, Wold S, Andrews P (1991) Minimum analogue peptide sets (MAPS) for quantitative structure-activity relationships. Int J Pept Protein Res 37:414–424

Jonsson J, Eriksson L, Hellberg S (1989) Multivariate parametrization of 55 coded and non-coded amino acid. Quant Struct Act Relat 8:204–209

Kidera A, Konishi Y, Oka MA (1985) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J Protein Chem 4:23–55

Kimura T, Miyashita Y, Funatsu K, Sasaki S (1996) Quantitative structure-activity relationships of the synthetic substrates for elastase enzyme using nonlinear partial least squares regression. J Chem Inf Comput Sci 36:185–189

Li SZ, Fu BH, Wang YQ (2001) On structural parameterization and molecular modeling of peptide analogues by molecular electronegativity edge vector (VMEE): estimation and prediction for biological activity of dipeptides. J Chin Chem Soc 48:937–944

Liang GZ, Li ZL (2007) Factor analysis scale of generalized amino acid information as the source of a new set of descriptors for elucidating the structure and activity relationships of cationic antimicrobial peptides. QSAR Comb Sci 26(6):754–763

Liang GZ, Zhou P, Zhou Y, Zhang QX, Li ZL (2006) New descriptors of aminoacids and their applications to peptide quantitative structure-activity relationship. Acta Chim Sin 64:393–596

Liu R, Kit S (2001) Lam automatic edman microsequencing of peptides containing multiple unnatural amino acids. Anal Biochem 295:9–16

Liu SS, Yin CS, Cai SX (2001) QSAR study of steroid benchmark and dipeptides based on MEDV-13. J Chem Inf Comput Sci 41(2):321–329

Mei H, Zhou Y, Sun LL, Li ZL (2004) A new of descriptor of amino acids and its application in peptide QSAR. Acta Phys-Chim Sin 20:821–825

Mei H, Liao ZH, Zhou Y, Li SS (2005) A new set of descriptors of amino acids and its application in peptide QSARs. Biopolymers 80:775–786

Nomizu M, Iwaki T, Yamashita T, Inagaki Y, Asano K, Akamatsu M, Fujita T (1993) Quantitative structure-sctivity relationshipm(QSAR) study of elastase substrates and Inhibitors. Int J Pept Protein Res 42:216–226

Patel S, Stott IP, Bhakoo M, Elliott P (1998) Patenting computer-designed peptides. J Comput Aided Mol Des 12:543–556

Randic M, Kleiner AF, DeAlba LM (1994) Distance/distance matrices. J Chem Inf Comput Sci 34:277–286

Robert S, Phillips RS (2004) Synthetic applications of tryptophan synthase. Tetrahedron Asymmetry 15:2787–2792

Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative structure-activity relationships and structure-property relationships. J Chem Inf Comput Sci 34:854–866

Rucker G, Rucker C (1993) Counts of all walks as atomic and molecular descriptors. J Chem Inf Comput Sci 33:683–695

Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S (1998) New chemical descriptors relevamt for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem 41:2481–2491

Schuur JH, Selzer P, Gasteiger J (1996) The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. J Chem Inf Comput Sci 36:334–344

Sjöstrom M, Rännar S, Wieslander Å (1995) Polypeptide sequence property relationships in *Escherichia coli* based on auto cross covariances. Chemometr Intell Lab Syst 29:295–305

Sneath PH (1966) Relations between chemical structure and biological activity in peptides. J Theor Biol 12:157–195

Sutter JM, Dixon SL, Jurs PC (1995) Automated descriptor selection for quantitative structure-activity relationships using generalized simulated annealing. J Chem Inf Comput Sci 35:77–84

Sybyl Version 6.8 (2001) Tripos Associates, Inc., St. Louis

Todeschini R, Gramatica P, Provenzani R, Marengo E (1995) Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaro-matic hydrocarbons. Chemon Intell Lab Syst 27:221–229

Wold S, Sjöström M, Eriksson L (2001a) PLS regression: a basic tool of chemometrics. Chemometr Intell Lab Syst 58:109–130

Wold S, Trygg J, Berglund A, Antti H (2001b) Some recent developments in PLS modeling. Chemometr Intell Lab Syst 58:131–150

Yan AX, Tian GL, Ye YH (2000) Progress in modification of bioactive peptides with non-protein amino acids and their application in the studies of structure-activity relationship. Chin J Org Chem 20(3):299–305

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