Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Structure 830 (2007) 106–115
Journal of MOLECULAR STRUCTURE
www.elsevier.com/locate/molstruc

T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides

Feifei Tian [a,b,*], Peng Zhou [a,b], Zhiliang Li [a,b,*]

[a] College of Chemistry and Chemical Engineering, Department of Applied Chemistry, Institute for Molecular Pharmacy, Chongqing University, Chongqing 400044, China
[b] Key Laboratory of Biomedical Engineering of Educational Ministry, Chongqing 400044, China

Received 13 January 2006; received in revised form 19 June 2006; accepted 5 July 2006
Available online 30 August 2006

Abstract

In this paper, a new topological descriptor T-scale is derived from principal component analysis (PCA) on the collected 67 kinds of structural and topological variables of 135 amino acids. Applying T-scale to three peptide panels as 58 angiotensin-converting enzyme (ACE) inhibitors, 20 thromboplastin inhibitors (TI) and 28 bovine lactoferricin-(17–31)-pentadecapeptides (LFB), the resulting QSAR models, constructed by partial least squares (PLS), are all superior to reference reports, with correlative coefficient r^2 and cross-validated q^2 of 0.845, 0.786; 0.996, 0.782 (0.988, 0.961); 0.760, 0.627, respectively.

Keywords: Amino acids; Peptide; Topological scale (T-scale); Quantitative structure–activity relationship (QSAR); Partial least square regression (PLS)

1. Introduction

Peptides play important roles in life status of human beings and other organisms; they function as hormone, enzyme inhibitor/substrate, growth promoter, inhibitor, neurotransmitter, immunomodulating agents as well as antibiotics, driving considerable pharmacological interest in design and application of new drugs. With the exploitation and development of new drug, quantitative structure activity relationship (QSAR) has been brought into the spotlight, involving in not only the key idea of pharmaceutical chemistry and pharmacology but also the foundations of drug design. Especially in recent years, QSAR models have been extensively used to study peptide sequence in relation with its biological activities by relevant scientists from different research fields, achieving good results [1–3]. Sneath [4] did the precedent works where amino acid descriptors, resulted from several semi-quantified experimental-based parameters, were successfully employed to predict activities of oxytocin vasopressine analogues. Then experiment/calculation-based descriptors for single amino acid were put forward one by one, e.g., Kidera et al. [5] extracted 10 orthogonal factors from factor analysis on the collected 188 kinds of properties of natural amino acids and applied them to predict high-level structures of polypeptides or proteins. Following that, Hellberge et al. [6,7] primarily proposed Z_1, Z_2, Z_3 descriptors from three dominant principal component scores which were resulted from a PCA process on 29 amino acid physicochemical properties. Hereafter, Sandberg et al. [8] extended Z-scale to 87 amino acids including 20 natural ones. Furthermore, Collantes et al. [9] established 3D-QSAR models on the base of three-dimensional structural characters of amino acid side chains, i.e. isotropic surface (ISA) together with electronic charge index (ECI) for polypeptides. However, original variables of above-mentioned amino acid descriptors are mostly experiment-restricted and the available numbers are also finite.

Nowadays there are mainly three methods on molecular structural representation in modern QSAR fields: (a)

* Corresponding authors. Tel.: +86 023 65106677.
E-mail address: ggootc@163.com (Z. Li).

0022-2860/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.molstruc.2006.07.004

Physicochemical parameters are directly served as molecular descriptors; (b) Quantum chemical indexes and structural fragments are utilized to exploit molecular structural information; (c) Topological methods. In contrast with the top two approaches, topological methods have unique merits in its simpleness and effectiveness. Dramatically dissimilar to traditional physicochemical parameters, topological descriptors, free from experimental restrictions, directly derive from atom-connecting manners in molecular two-dimensional topological map and well correlate with molecular physicochemical properties and biological activities. Therefore, topological descriptors are widely utilized in modern QSAR/QSPR studies [10–13], e.g., Winer index (W index) [14], Hosoya index (Z index) [15], Randic index (χ index) [16], Balaban index (J index) [17] and Kier–Hall link index ($^{m}\chi^{v}$ index) [18]. Herein, via performing principal component analysis (PCA) on the collected 67 common topological descriptors of 135 amino acids, a new topological scale of amino acids (T-scale) is proposed in this context. Applying T-scale to 58 angiotensin-converting enzyme (ACE) inhibitors, 20 thromboplastin inhibitors (TI) and 28 bovine lactoferricin-(17–31)-pentadecapeptides (LFB), the constructed QSAR models all obtain satisfying results.

2. Principle and methodology

2.1. Amino acid dataset

One hundred and thirty five amino acids are compiled from reference reports [8,19–22] (with their names shown in Table 1 and molecular structures as supplementary material). Amongst, numbers 1–20 are standard natural coded amino acids; numbers 21 and 22, recently discovered, are two non-standard coded ones as selenocysteine [20] and pyrrolysine [21] and the remains are common non-standard and non-coded ones, e.g., ornithine, hydroxyproline and homoserine, etc.

2.2. T-scale

Original structures of 135 amino acids, which are built up by molecular graphics software Chemoffice 8.0, are saved as the file format of MDL MolFlie (*.mol). Then, a series of descriptor calculation software as Chem3D 8.0, ChemSW 5.0, Codessa 2.7 and Dragon 5.2 are utilized to generate the 67 common topological descriptors for each single amino acid (Table 2 lists their names and the values are as supplementary material). What should be elucidated is that during this process, there is no requirement for molecular geometrical optimization due to topological descriptors, having nothing to do with molecular three-dimensional conformations, only relate to atom-connecting manners. Among different topological descriptors, there may be large information overlap due to they all directly reflect the overall molecular structural features and a direct usage of these 67 topological descriptors would also increase complexity of the QSAR models, so in conjunction with classical data information extraction technique PCA, variable dimensions are to be largely compressed. Primarily, original variable matrix $\boldsymbol{X}_{135\times67}$ is subject to autoscaling, then a PCA on that generates 5 prominent principal components (PCs), cumulatively explaining 91.14% variances with the contribution from each 66.60%, 10.29%, 7.52%, 4.33% and 2.40%, respectively. These 5 PCs, corresponding to each single amino acid, are the results of original 67 variables multiplied by each PC scoring coefficient and are then utilized to replace original variable matrix with only 8.86% information loss. Here, the 5 PCs for 135 amino acids are termed as T-scale, with their values in Table 1. Statistics software SPSS 13.0 implements PCA program.

When parameterized a peptide analogue, the overall compositive amino acids are orderly represented by corresponding T-scales. Here, a dipeptide is exemplified: each 5 T-scales specially correspond to every sequence site and ultimately there would be 10 T-scales ($2\times5=10$) to represent the dipeptide molecule. Accordingly, peptide sequence with the length of n generates $n\times5$ variables.

2.3. PLS regression

Nowadays, partial least square regression (PLS) is a widely used modeling method. It has merits in an effective overcoming of multicollinearity issues and especially suits for condition of sample size smaller than variable numbers. In the 1980s, PLS was proposed by Wold et al. [23]. The principle is as follows; primarily independent variable matrix $\mathbf{X}$ is subject to a bi-linear decomposition.

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{T}+\mathbf{F} \tag{1}$$

Amongst, matrix $\mathbf{T}$ contains mutually orthogonal latent variable or scoring vector $\mathbf{t}$ which is a linear combination of variables in matrix $\mathbf{X}$. Dissimilar to PCA, PLS also implements bi-linear decomposition on target matrix $\mathbf{Y}$:

$$\mathbf{Y}=\mathbf{U}\mathbf{Q}^{T}+\mathbf{E} \tag{2}$$

Of that, matrix $\mathbf{U}$ comprises latent variable $\mathbf{u}$ of $\mathbf{Y}$. Upon that, PLS requires for latent variable $\mathbf{t}$, obtained by decomposing $\mathbf{X}$, maximumly overlaps with latent variable $\mathbf{u}$ derived from $\mathbf{Y}$ decomposition. Therefore:

$$\mathbf{u}=\mathbf{v}\mathbf{t}+\mathbf{e} \tag{3}$$

In Eq. (3), $\mathbf{e}$ is error vector; coefficient $\mathbf{v}$ is determined by least square. Computational and other details are given in Refs. [24,25].

2.4. Variable selections

Among peptide sequences, the compositive amino acids are often different and variables regarding the same amino acid may also contribute distinctly to activities, so variables should be screened beforehand by such common methods as orthogonalization [26], stepwise multiple regression (SMR), genetic algorithm (GA) [27], and simulated anneal-

Table 1
135 amino acid and their T-scale values

No.	Abbreviations	Name	T_1	T_2	T_3	T_4	T_5
1	Ala	Alanine	−9.11	−1.63	0.63	1.04	2.26
2	Arg	Arginine	0.23	3.89	−1.16	−0.39	−0.06
3	Asn	Asparagine	−4.62	0.66	1.16	−0.22	0.93
4	Asp	Aspartic acid	−4.65	0.75	1.39	−0.40	1.05
5	Cys	Cysteine	−7.35	−0.86	−0.33	0.80	0.98
6	Gln	Glutamine	−3.00	1.72	0.28	−0.39	0.33
7	Glu	Glutamic acid	−3.03	1.82	0.51	−0.58	0.43
8	Gly	Glycine	−10.61	−1.21	−0.12	0.75	3.25
9	His	Histidine	−1.01	−1.31	0.01	−1.81	−0.21
10	Ile	Isoleucine	−4.25	−0.28	−0.15	1.40	−0.21
11	Leu	Leucine	−4.38	0.28	−0.49	1.45	0.02
12	Lys	Lysine	−2.59	2.34	−1.69	0.41	−0.21
13	Met	Methionine	−4.08	0.98	−2.34	1.64	−0.79
14	Phe	Phenylalanine	0.49	−0.94	−0.63	−1.27	−0.44
15	Pro	Proline	−5.11	−3.54	−0.53	−0.36	−0.29
16	Ser	Serine	−7.44	−0.65	0.68	−0.17	1.58
17	Thr	Threonine	−5.97	−0.62	1.11	0.31	0.95
18	Trp	Tryptophan	5.73	−2.67	−0.07	−1.96	−0.54
19	Tyr	Tyrosine	2.08	−0.47	0.07	−1.67	−0.35
20	Val	Valine	−5.87	−0.94	0.28	1.10	0.48
21	Acp	α-Aminocaprylic acid	−0.89	3.26	−3.07	0.91	−0.92
22	Aec	(*S*)-2-Aminoethyl-L-cysteine · HCl	−2.46	2.10	−2.56	1.55	−1.01
23	Afa	Aminophenylacetate	−1.05	−2.07	0.22	−1.58	−0.27
24	Aib	α-Aminoisobytyric acid	−7.66	−1.90	1.46	1.66	1.42
25	Ail	Alloisoleucine	−4.38	0.28	−0.49	1.45	0.02
26	Alg	L-Allylglycine	−5.93	0.06	−0.07	0.12	0.94
27	Aba	α-Aminobutyric acid	−7.34	−0.93	−0.07	0.53	1.15
28	Aph	*p*-Aminophenylalanine	2.08	−0.57	−0.21	−1.46	−0.43
29	β-Ala	β-Alanine	−8.79	−0.23	−0.98	0.11	2.06
30	Brp	*p*-Bromophenylalanine	2.14	−0.93	−1.38	−0.18	−1.23
31	Cha	Cyclohexylalanine	1.04	−1.16	−2.26	0.91	−1.86
32	Cit	Citrulline	0.22	3.95	−0.95	−0.55	−0.00
33	Cla	β-Chloroalanine	−7.41	−0.81	−0.06	0.53	1.14
34	Cle	Cycloleucine	−3.53	−3.86	0.16	0.30	−0.96
35	Clp	*p*-Chlorophenylalanine	2.07	−0.66	−0.63	−1.00	−0.74
36	Cya	Cysteic acid	−2.95	0.84	2.57	1.41	−0.04
37	Dab	2,4-Diaminobutyric acid	−5.83	0.21	−0.16	0.28	0.84
38	Dap	2,3-Diaminopropionic acid	−7.41	−0.75	0.42	0.03	1.47
39	Dhp	3,4-Dehydroproline	−5.29	−3.45	−0.00	−1.06	0.18
40	Dha	3,4-Dihydroxyphenylalanine	3.60	−0.14	1.04	−1.79	−0.24
41	Fph	*p*-Fluorophenylalanine	2.33	−0.47	0.52	−1.93	−0.52
42	Gaa	D-Glucoseaminic acid	1.69	2.87	2.77	−0.12	−0.45
43	Hag	Homoarginine	0.23	3.89	−1.16	−0.39	−0.06
44	Hly	δ-Hyroxylysine · HCl	−1.23	2.55	−0.48	0.24	−0.31
45	Hnv	DL-β-Hydroxynorvaline	−4.36	0.09	0.74	0.58	0.30
46	Hog	Homoglutamine	−1.36	2.80	−0.43	−0.08	0.02
47	Hop	Homophenylalanine	2.17	−0.08	−1.46	−1.28	−0.65
48	Hos	Homoserine	−5.84	0.30	0.09	0.08	0.95
49	Hpr	Hydroxyproline	−3.67	−2.92	0.28	−0.51	−0.24
50	Iph	*p*-Iodophenylalanine	2.21	−1.02	−1.89	0.37	−1.56
51	Ise	Isoserine	−9.24	−1.37	1.37	0.42	2.70
52	Mle	α-Methylleucine	−2.88	−0.26	0.44	2.19	−0.71
53	Msm	DL-Methionine-*s*-methylsulfoniumchloride	−2.46	0.79	−2.97	3.66	−2.12
54	1Nala	β-(1-Naphthyl)alanine	7.16	−2.66	−0.23	−1.71	−0.51
55	2Nala	β-(2-Naphthyl)alanine	7.35	−2.26	−0.51	−1.70	−0.25
56	Nle	Norleucine(or 2-aminohexanoic acid)	−4.16	1.15	−1.56	0.63	−0.11
57	Nma	*N*-Methylalanine	−7.35	−1.01	−0.23	0.54	1.14
58	Nva	Norvaline(or 2-aminopentanoic acid)	−5.77	0.03	−0.66	0.79	0.54
59	Obs	*O*-Benzylserine	3.80	1.23	−1.52	−1.82	−0.15
60	Obt	*O*-Benzyltyrosine	14.75	0.43	−2.06	−1.86	2.44
61	Oet	*O*-Ethyltyrosine	5.45	1.43	−1.41	−1.46	−0.76
62	Oms	*O*-Methylserine	−5.89	0.25	−0.02	0.03	1.01
63	Omt	*O*-Methylthreonine	−4.32	−0.05	0.53	0.61	0.27
64	Omy	*O*-Methyltyrosine	3.72	0.14	−0.60	−1.39	−0.37

Table 1 (*continued*)

No.	Abbreviations	Name	T_1	T_2	T_3	T_4	T_5
65	Orn	Ornithine	−4.21	1.33	−1.06	0.11	0.18
66	Pen	Penicillamine	−4.42	−1.12	0.66	2.27	−0.38
67	Pga	Pyroglutamic acid	−3.74	−2.95	0.68	−0.93	−0.06
68	Pip	Pipecolic acid	−3.62	−3.09	−0.90	0.02	−0.66
69	Sar	Sarcosine	−8.81	−0.29	−1.10	0.14	2.06
70	Tfa	3,3,3-Trifluoroalanine	−4.56	0.24	5.16	−0.76	1.50
71	Thp	6-Hydroxydopa	5.12	0.20	2.17	−1.79	−0.20
72	Vig	L-Vinylglycine	−7.51	−0.90	0.50	−0.12	1.55
73	Aas	(-)-(2*R*)-2-Amino-3-(2-aminoethylsulfonyl)propanoic acid dihydrochloride	0.79	2.08	1.87	1.04	−1.03
74	Ahd	(2*S*)-2-Amino-9-hydroxy-4,7-dioxanononanoic acid	2.22	6.21	−2.41	−1.11	0.06
75	Aho	(2*S*)-2-Amino-6-hydroxy-4-oxahexanoic acid	−2.73	2.71	−0.74	−0.64	0.42
76	Ahs	(-)-(2*R*)-2-Amino-3-(2-hydroxyethylsulfonyl)propanoic acid	0.78	2.20	2.17	0.82	−0.91
77	Ahp	(-)-(2*R*)-2-Amino-3-(2-hydroxyethylsulfanyl)propanoic acid	−2.48	2.20	−2.30	1.32	−0.90
78	Ahd	(2*S*)-2-Amino-12-hydroxy-4,7,10-trioxadodecanoic acid	7.68	9.71	−3.91	−0.93	0.68
79	Dad	(2*S*)-2,9-Diamino-4,7-dioxanonanoic acid	2.22	6.13	−2.65	−0.90	−0.01
80	Dat	(2*S*)-2,12-Diamino-4,7,10-trioxadodecanoic acid	7.70	9.63	−4.15	−0.72	0.61
81	Dfn	(*S*)-5,5-Difluoronorleucine	−1.26	2.19	1.98	0.15	−0.12
82	Dfv	(*S*)-4,4-Difluoronorvaline	−2.86	1.08	2.94	0.28	0.38
83	Dtc	(3*R*)-1-1-Dioxo-[1,4]thiaziane-3-carboxylic acid	0.64	−2.37	1.56	0.15	−1.19
84	Hfn	(*S*)-4,4,5,5,6,6,6-Heptafluoronorleucine	7.11	6.32	15.84	1.43	−2.07
85	Pfn	(*S*)-5,5,6,6,6-Pentafluoronorleucine	3.43	4.14	7.69	−0.34	−0.27
86	Pfv	(*S*)-4,4,5,5,5-Pentafluoronorvaline	1.69	3.15	8.24	−0.81	−0.00
87	Tca	(3*R*)-1,4-Thiazinane-3-carboxylic acid	−3.50	−3.41	−1.74	1.17	−1.49
88	Sec	Selenocysteine	−7.31	−1.45	−1.02	1.76	0.32
89	Pyl	Pyrrolysine	11.25	3.40	−1.91	−0.72	−0.13
90	Ath	β-(9-Anthracenyl)alanine	14.19	−3.65	−0.21	−1.74	0.06
91	Bal	β-(3-Benzothienyl)alanine	5.89	−3.15	−1.21	−0.47	−1.60
92	Bip	β-(4,4′-Biphenyl)alanine	10.77	−1.14	−1.30	−1.77	0.76
93	Dip	β,β-Diphenylalanine	10.02	−2.62	−0.06	−1.47	−0.18
94	Tbt	β-[3-(2,5,7-Tri-tert-butyl-indolyl)]alanine	24.18	−1.01	0.76	13.40	−2.14
95	Tpc	β-{3-[2-(2,2,5,7,8-Pentamethyl-chroman-6-sulfonyl)-indolyl]}alanine	36.90	−4.22	−0.72	5.46	8.53
96	Asu	Aminosuberic acide	2.03	5.00	−1.72	0.1	−0.40
97	Hcy	Homocysteine	−5.77	0.11	−0.91	1.07	0.36
98	Sta	Statine	0.24	2.51	−0.67	1.60	−0.97
99	Thi	β-(2-Thienyl)alanine	−2.35	−2.75	−0.52	0.04	−1.04
100	γ-Abu	L-γ-Aminobutyric acid	−7.20	0.89	−1.63	0.36	1.40
101	Aca	ε-Aminocaproic acid	−3.95	3.02	−3.14	0.46	0.22
102	Ach	1-Aminocyclohexane-1-carboxylic acid	−2.04	−3.43	−0.22	0.70	−1.28
103	Afb	β-Amino-β-phenyl-*p*-nitro-L-butyric acid	8.47	0.86	2.47	−1.57	−0.17
104	Aoq	α-Amino-β-[4-(1,2-dihydro-2-oxo-quinolinyl)]propionic acid	7.16	−2.34	0.28	−2.26	−0.21
105	Bpa	4′-Benzoylphenylalanine	15.14	−0.69	−0.58	−3.65	3.19
106	Mas	β-Methyl aspartic aicd	−3.09	0.41	2.11	0.15	0.34
107	Ceg	2-Chloroethylglycine	−5.83	0.16	−0.65	0.79	0.50
108	Cha	β-Cyclohexyl(*p*-methoxyl)-L-alanine	4.29	−0.07	−2.20	0.78	−1.79
109	Dty	α,β-Divinyltyrosine	8.27	−0.13	1.89	−0.83	−0.80
110	Chg	2-L-Cyclohexylglycine	−0.51	−2.32	−1.39	0.56	−1.68
111	Cpa	4-Chlorophenylalanine	2.07	−0.66	−0.63	−1.00	−0.74
112	Deg	α,α-Diethyl glycine	−4.28	−1.30	0.72	1.42	−0.38
113	Dmt	2′,6′-Dimethyltyrosine	5.26	−0.56	0.59	−0.32	−1.04
114	Dvg	Divinyl glycine	−4.50	1.18	−0.44	−0.70	0.69
115	Gav	2-Guanidine-5-amino-L − *n*-valeric acid	0.02	2.83	−0.10	0.08	−0.41
116	Hat	2-Amino-6-hydroxytetralin-2-carboxylic acid	6.03	−3.75	0.94	−1.03	−1.10
117	Hai	2-Amino-5-hydroxyindan-2-carboxylic acid	4.40	−3.99	1.43	−1.16	−1.02
118	Hpp	3-(4′-Hyroxyphenyl)proline	5.88	−3.31	−0.01	−1.38	−0.88
119	Ing	1-Indanylglycine	4.34	−3.82	−0.1	−0.80	−1.40
120	Mhp	*p*-Methoxyhomophenylalanine	5.48	1.01	−1.43	−1.29	−0.39
121	Oct	*n*-Octylglycine	2.52	5.39	−4.60	1.36	−1.35
122	Oic	Octahydroindole-2-carboxylic acid	1.66	−4.70	−1.30	0.80	−2.21
123	Pal	β-Pyridylalanine	0.45	−0.86	−0.33	−1.59	−0.28
124	Tic	1,2,3,4-Tetrahydroisoquinoline-3-carboxylic acid	2.69	−4.12	−0.21	−1.28	−0.90
125	Thz	L-4-Thiazolidine carboxylic acid	−5.01	−3.85	−1.33	0.72	−1.07
126	Tle	L-*tert*-Butylglycine	−4.43	−1.15	0.89	1.88	−0.13
127	Dpg	Diphenylglycine	8.69	−3.44	0.94	−2.00	−0.38
128	Dbz	Dibenzylglycine	11.79	−2.00	−0.54	−1.27	−0.16
129	β-Phe	β-Phenylalanine	0.49	−1.07	−0.62	−1.27	−0.49

(*continued on next page*)

Table 1 (*continued*)

No.	Abbreviations	Name	T_1	T_2	T_3	T_4	T_5
130	α-Abu	α-Aminobutyric acid	−7.34	−0.93	−0.07	0.52	1.15
131	Mpr	3-Methyproline	−3.56	−3.67	−0.36	0.17	−0.85
132	Hva	3-Hydroxyvaline	−4.52	−0.77	1.81	1.05	0.34
133	Dcp	3,5-Dihydroxy-4-chloro-phenylalanine	5.17	−0.03	1.42	−1.16	−0.58
134	Car	β-Carbonylarginine	1.41	3.64	1.32	−1.50	0.09
135	Has	β-Hydroxyaspartate	−3.21	0.74	2.96	−0.58	0.82

Table 2
67 structural and topological variables used to characterize the amino acids

No.	Structural and topological variables	Software
1	Balaban Index	Chem3D 8.0
2	Cluster Count	Chem3D 8.0
3	Diameter of side chain	Chem3D 8.0
4	Molecular Topological Index	Chem3D 8.0
5	Radius	Chem3D 8.0
6	Shape Attribute	Chem3D 8.0
7	2D Petitjean Shape Coefficient	Chem3D 8.0
8	Sum of Degrees	Chem3D 8.0
9	Sum of Valence Degrees	Chem3D 8.0
10	Wiener Index	Chem3D 8.0
11	KAPPA Shape Index 2	ChemSW 5.0
12	Connectivity Index 0	ChemSW 5.0
13	Connectivity Index 1	ChemSW 5.0
14	Connectivity Index 2	ChemSW 5.0
15	Connectivity Index 3	ChemSW 5.0
16	Connectivity Index 4	ChemSW 5.0
17	Valence Connectivity Index 0	ChemSW 5.0
18	Valence Connectivity Index 1	ChemSW 5.0
19	Valence Connectivity Index 2	ChemSW 5.0
20	Valence Connectivity Index 3	ChemSW 5.0
21	Valence Connectivity Index 4	ChemSW 5.0
22	Difference Index 0	ChemSW 5.0
23	Difference Index 1	ChemSW 5.0
24	Difference Index 2	ChemSW 5.0
25	Difference Index 3	ChemSW 5.0
26	Difference Index 4	ChemSW 5.0
27	Randic Index 0	Codessa 2.7
28	Randic Index 1	Codessa 2.7
29	Randic Index 2	Codessa 2.7
30	Randic Index 3	Codessa 2.7
31	Kier & Hall Index 0	Codessa 2.7
32	Kier & Hall Index 1	Codessa 2.7
33	Kier & Hall Index 2	Codessa 2.7
34	Kier & Hall Index 3	Codessa 2.7
35	Kier Shape Index 1	Codessa 2.7
36	Kier Shape Index 2	Codessa 2.7
37	Kier Shape Index 3	Codessa 2.7
38	Kier Flexibility Index	Codessa 2.7
39	Zagreb Index	Dragon 5.2
40	Quadratic Index	Dragon 5.2
41	Narumi Simple Topological Index (log)	Dragon 5.2
42	Narumi Harmonic Topological Indx	Dragon 5.2
43	Narumi Harmonic Geometric Indx	Dragon 5.2
44	Total Structure Connectivity Index	Dragon 5.2
45	Pogliani Index	Dragon 5.2
46	Log of Product of Row Sum (PRS)	Dragon 5.2
47	Average Vertex Distance Degree	Dragon 5.2
48	Mean Square Distance Index	Dragon 5.2
49	Schultz Molecular Topological Index	Dragon 5.2
50	Gutman Molecular Topological Index	Dragon 5.2
51	Xu Index	Dragon 5.2
52	Superpendentic Index	Dragon 5.2
53	Harary H Index	Dragon 5.2
54	Square Reciprocal Distance Sum Index	Dragon 5.2

Table 2 (*continued*)

No.	Structural and topological variables	Software
55	First Mohar Index TI1	Dragon 5.2
56	Second Mohar Index TI2	Dragon 5.2
57	Hyper-Distance-Path Index	Dragon 5.2
58	Detour Index	Dragon 5.2
59	Balaban Distance Connectivity Index (J)	Dragon 5.2
60	Maximal Electrotopological Negative Variation	Dragon 5.2
61	Maximal Electrotopological Positive Variation	Dragon 5.2
62	Molecular Electrotopological Variation	Dragon 5.2
63	E-state Topological Parameter	Dragon 5.2
64	Kier Symmetry Index	Dragon 5.2
65	Mean Distance Dregee Deviation	Dragon 5.2
66	Balaban Centric Index	Dragon 5.2
67	Lopping Centric Index	Dragon 5.2

ing (SA) [28]. In case of small number of variables, SMR is deemed to be excellent, ascribed to its quick and easy calculations. Variables are introduced in terms of prominence by Fisher test in SMR, and PLS model is constructed to determine the number of the most optimal variables by correlative coefficient q^2 of leave-one-out cross-validation (LOO CV). It should be elucidated that the number of PLS latent variables is in compliance with the default of Simca-P 10.0, that is to say, variables are firstly iterative by non-linear iterative partial least square (NIPALS) algorithms, and then principal component is calculated orderly according to the value of variance explaining original variables. Simultaneously, the contributions of each principal component to the correlative coefficient q^2 of models by cross-validation are verified, and calculation of more principal components is stopped in case q^2 of some introduced principal component indicating model correlativeness is no more than 0.097, which is suggestive of indistinction.

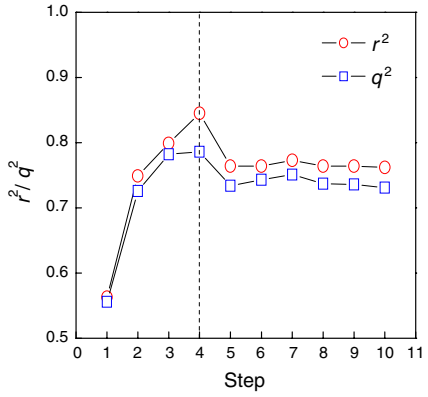

Fig. 1. Value of fitness r^2 and cross-validation q^2 change with SMR-introduced variables in the PLS model.

3. Results and analysis

3.1. QSAR model for angiotensin converting enzyme inhibitors

Rennin–angiotensin system plays an important role in blood pressure regulation in human bodies. Angiotensinogen, produced by liver, is catalyzed by rennin to disrupt into inactive angiotensin I which is further catalyzed by angiotensin converting enzyme (ACE) to rupture into angiotensin II, an extremely responsible agent for blood vessel contractions. Thus ACE drives considerable interests in developing antihypertension drugs [29]. Besides, dipeptide sequences of 58 ACE inhibitors, a classical sample set for QSAR studies [9,30–35], are often utilized to test effectiveness of diverse kinds of amino acid descriptors. Structural representation of each ACE inhibitor by T-scale yields 10 variables. Fig. 1 presents variable selection process, where it shows fitness r^2 and cross-validation q^2 change consistently with SMR-introduced variables, both achieving the maximum at the fourth step. For that, we ultimately obtain a 4-variable PLS model which, by only two PLS components, accounts for 84.5% variances of Y variables with cross-validation achieving 78.6% and RMSEE of 0.39. Table 3 presents experimental data by observation and calculated value (*cald_1*) by model. Compared with reference reports (Table 4), it is indicated that T-scale, with regard to modeling qualities, distinctly outperforms many other traditional amino acid descriptors that mainly focus on experimental or physicochemical parameters, especially with a remarkable q^2 of 0.786 in contrast with the normal averaged level of 0.637–0.767. Fig. 2 presents the plot of calculated value against experimental data for ACE inhibitors. It is shown most samples are uniformly dispersed around the origin-passed diagonal except for the 19th which deviates more from its calculated value than others. By structural analysis, it is shown the 19th sample is composed of two bulky amino acid residues (i.e. isoleucine I and phenylalanine F), which means its topological structure would be dramatically larger than the other ones at the same activity rank. Fig. 3 presents scoring distribution scatter of 58 samples at the first and second PLS principal component spaces, of which, circle marks samples with the pIC_{50} bigger than 4, triangle marks samples with the pIC_{50} between 3 and 4 and rhombus marks samples with the pIC_{50} smaller than 3. It can be seen

Table 3
The sequences of angiotensin-converting enzyme inhibitors with observed and calculated activity

No.	Peptide	$pIC_{50}(obsd)$	$pIC_{50}(cald\_1)$	$pIC_{50}(cald\_2)$
1	VW	5.80	5.26	5.10
2	IW	5.70	5.46	5.29
3	IY	5.43	4.91	4.62
4	AW	5.00	5.19	4.76
5	RW	4.80	5.07	5.38
6	VY^{Δ}	4.66	4.21	4.43
7	GW	4.52	4.63	4.56
8	VF^{Δ}	4.28	4.10	4.23
9	AY^{Δ}	4.06	3.84	4.09
10	IP	3.89	4.16	4.17
11	RP	3.74	3.78	4.26
12	AF^{Δ}	3.72	3.73	3.88
13	GY^{Δ}	3.68	3.58	3.89
14	AP	3.64	3.59	3.63
15	RF	3.64	3.92	4.51
16	VP^{Δ}	3.38	3.96	3.97
17	GP	3.35	3.33	3.44
18	GF^{Δ}	3.20	3.47	3.69
19	IF^{Δ}	3.03	4.30	3.42
20	VG^{Δ}	2.96	2.62	2.60
21	IG	2.92	2.82	2.59
22	GI^{Δ}	2.92	2.65	2.66
23	GM	2.85	2.36	2.09
24	GA^{Δ}	2.70	2.30	2.16
25	YG	2.70	2.24	2.38
26	GL^{Δ}	2.60	2.49	2.28
27	AG^{Δ}	2.60	2.25	2.06
28	GH^{Δ}	2.51	2.35	2.85
29	GR	2.49	2.25	2.55
30	KG	2.49	2.63	2.62
31	FG^{Δ}	2.43	2.34	2.35
32	GS^{Δ}	2.42	2.30	2.61
33	GV^{Δ}	2.34	2.59	2.48
34	MG	2.32	2.99	2.71
35	GK	2.27	2.24	2.33
36	GE	2.27	2.30	2.90
37	GT^{Δ}	2.24	2.49	2.68
38	WG	2.23	2.22	2.32
39	HG^{Δ}	2.20	2.19	2.22
40	GQ^{Δ}	2.15	2.33	2.45
41	GG	2.14	1.99	1.86
42	QG^{Δ}	2.13	2.36	2.47
43	SG^{Δ}	2.07	2.15	2.10
44	LG^{Δ}	2.06	2.78	2.59
45	GD	2.04	2.34	2.79
46	TG^{Δ}	2.00	2.37	2.29
47	EG^{Δ}	2.00	2.30	2.45
48	DG	1.85	2.21	2.31
49	PG	1.77	2.49	2.15
50	LA^{Δ}	3.51	3.10	3.48
51	KA	3.42	2.94	2.92
52	RA	3.34	2.75	2.98
53	YA	3.34	2.55	2.68
54	AA^{Δ}	3.21	2.56	2.95
55	FR	3.04	2.60	3.04
56	HL	2.49	2.69	2.63
57	DA^{Δ}	2.42	2.52	2.60
58	EA^{Δ}	2.00	2.61	2.34

that distribution of dipeptide sequences exhibits an increasing trend from the root left corner to the top right corner in terms of their activities; compounds with similar pharmaceutical activities are close together, suggesting that the top two PCs of the 4-variable model are just enough to reflect distribution characters of this group of samples. In addition, most samples are falling into Hotelling T^2 confidence ellipse with 95% confidence, only excluding #2 and #5. By analysis, it is found these two dipeptide sequences, of which, both C-terminals comprise very bulky tryptophan and separate N-terminal is bulky isoleucine or arginine, possess the biggest molecular topological structures over the whole sample set, thus being special. To further verify reliabilities of T-scale model, algorithm D-optimal [36] is employed to averagely divide the sample set into training and test set (there are separately 29 samples in training and test set; samples in test set are highlighted with "Δ" in Table 3). Algorithm D-optimal is implemented by software Matlab 7.0. Upon that, a 4-T-scale PLS model is constructed for the training set with its fitting correlative coefficient $r^2 = 0.839$, cross-validation $q^2 = 0.778$ and $RMS = 0.48$, then such a model is utilized to predict test set with results of $r^2_{ext} = 0.798$ and $RMS_{ext} = 0.33$ (details are as *cald_2* in Table 3). Therefore, it is confirmed T-scale model is stable and generalized.

3.2. QSAR model for thromboplastin inhibitors

Table 5 presents 20 thromboplastin inhibitors (TI) sequences of different length and their corresponding activities expressed by 50% inhibitory concentration [37]. Parameterized with T-scale descriptors, peptide sequences composed of different amounts of amino acids produce different amounts of T-scale variables. To achieve agreements on variable amounts over all the peptides in sample set, auto cross-covariance (ACC), proposed by Wold, is implemented. Details about ACC are given in Ref. [38]. By ACC, 125 cross-variables are generated for each peptide sequence of the sample set, thus fulfilling the purpose that different sequence length have the same amounts of descriptors. Then performing SMR-PLS on that, an optimal subset comprising 37 variables is obtained to create the QSAR model which cumulatively accounts for 99.6% variances of Y variable by 3 prominent PCs, with its cross-validation of 78.2%, superior to reference reports (Table 6). Convenient to make a further comparison, the PLS model is constructed after the orthogonal signal correction (OSC) [39] (i.e. filtering out the overlapped information from X matrix to Y) on the 37 variables. Still 3 PCs are obtained and the resulted model is extremely advanced, with the r^2 of 98.8%, q^2 increasing to 96.1%. In references, Z-scale and VSTV descriptors, proposed by Andersson [37] and Mei et al. [35], respectively, were ever employed to construct QSAR models on this sample set with good results. Table 6 presents comparisons among these different models. It is confirmed T-scale model distinctly outperforms the Z-scale and VSTV models with respect to predictabilities, especially implementing OSC on variables before modeling.

Table 4
Comparison among different QSAR models for angiotensin-converting enzyme inhibitors

No.	Descriptor	Sum of descriptors	A	r^2	q^2	RMS
1	ISA-ECI [9]	4	2	0.700	–	–
2	Z-scale [30]	6	2	0.770	0.723	–
3	t-score [31]	14	1	0.744	–	0.50
4	MS-WHIM [32]	6	2	0.708	0.637	0.54
5	VMEE [33]	10	2	0.741	0.711	0.50
6	VHSE [34]	5	1	0.770	0.745	0.48
7	VSTV [35]	6	1	0.789	0.767	0.46
10	T-scale	4	2	0.845	0.786	0.39

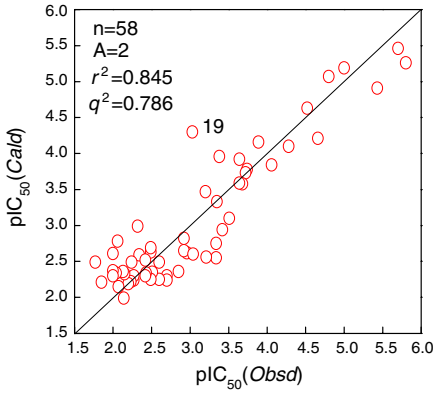

Fig. 2. Plot of calculated vs. observed values for 58 angiotensin converting enzyme inhibitors by the PLS model.

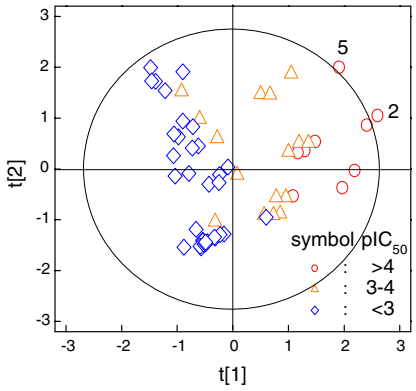

Fig. 3. Scoring distribution scatters of 58 samples at the first and second PLS principal components (different biological activities with different marks).

3.3. QSAR model for bovine lactoferricin-(17–31)-pentadecapeptides

It is indicated bovine lactoferricin-(17–31)-pentadecapeptide FKCRRWQWRMKKLGA (LFB) possesses a potential antibacterial activity by Rekdal et al. [40]. In investigations by Haug [41–43], site-directed substitution at positions 6 and/or 8 is assayed and the resulting mutants are tested for their minimal growth inhibitory concentration (MHC) against *S. aureus*. Amongst, 28 samples are selected by Lejon [44] to perform a QSAR study (Table 7). Here, they are deemed to allow for a verification of the T-scale with respect to its application to predict activities of unnatural peptides which comprise non-coded amino acids at the substitution positions on reports by Haug [41–43]. Parameterized 6- and 8-positions by T-scale, there are 10 descriptors, SMR-PLS analysis of this dataset generates a 7-variable model which accounts for 76.0% variances of Y variables by only 2 prominent PCs, with cross-valida-

Table 5
The amino acid sequences of thromboplastin inhibitors with observed and calculated activities

No.	Peptide	IC_{50}(APPT)		
		Obsd	Calcd (PLS)	Calcd (OSC-PLS)
1	PKPRPDR	5.52	5.50	5.31
2	SWKHYW	0.58	0.47	0.35
3	SWKYYW	0.79	0.69	0.82
4	SWVDAW	1.56	1.60	1.40
5	RQGRYWL	1.05	1.03	0.82
6	PPGEMD	2.66	2.69	2.69
7	EGEGGM	1.58	1.60	1.55
8	RHWNIEGRPWWS	0.66	0.57	0.72
9	SEWAIEGRPHGW	1.21	1.17	1.09
10	FLRGEV	2.32	2.24	2.69
11	FMHLST	2.26	2.30	2.26
12	FMRPQM	4.14	4.01	4.08
13	FGWGQN	4.87	5.17	5.01
14	CWPMTRGC	1.09	1.02	0.98
15	KPRWWMWK	0.05	0.03	−0.09
16	KSWQVWVK	0.80	0.81	0.84
17	KSWKYYWK	0.04	0.06	0.02
18	SWKYYWK	0.03	0.17	0.38
19	KSWKYYW	0.03	0.08	0.30
20	KMMSWKGK	0.70	0.65	0.72

Table 6
Comparisons among several QSAR models for thromboplastin inhibitors

No.	Descriptor	Method	A	r^2	q^2
1	Z-scale [37]	PLS	3	0.886	0.490
2	$Z^{1/2}$-scale [37]	PLS	1	0.658	0.332
3	VSTV [35]	PLS	4	0.981	0.480
4	T-scale	PLS	3	0.996	0.782
5	Z-scale [37]	OSC-PLS	2	0.881	0.706
6	VSTV [35]	OSC-PLS	2	0.989	0.864
7	T-scale	OSC-PLS	3	0.988	0.961

Table 7
Mutant sequences at 6- and 8-positions and the observed and calculated MIC values of LBF to *S. aureus*

No.	Position 6	Position 8	*S. aureus*	
			$\log MIC_{Obsd}$	$\log MIC_{Cald}$
LFB	Trp	Trp	1.68	1.47
1	Phe	Trp	2.17	1.66
2	Trp	Phe	2.17	1.87
3	Phe	Phe	2.18	2.06
4	Bal	Trp	1.56	1.81
5	Trp	Bal	1.23	1.06
6	Bal	Bal	1.08	1.29
7	1-Nal	Trp	1.86	1.39
8	Trp	1-Nal	1.68	1.36
9	1-Nal	1-Nal	1.38	1.27
10	2-Nal	Trp	1.56	1.33
11	Trp	2-Nal	1.38	1.30
12	2-Nal	2-Nal	0.98	1.16
13	Bip	Trp	0.85	0.98
14	Trp	Bip	0.85	0.96
15	Bip	Bip	0.15	0.28
16	Dip	Trp	1.23	1.87
17	Trp	Dip	1.23	1.45
18	Dip	Dip	0.85	0.97
19	Ath	Trp	1.20	1.06
20	Trp	Ath	0.85	0.99
21	Ath	Ath	0.53	0.59
22	Tbt	Trp	0.65	0.61
23	Trp	Tbt	0.34	0.44
24	Tbt	Tbt	0.49	0.39
25	Tpc	Trp	0.63	0.80
26	Trp	Tpc	0.51	0.59
27	Tpc	Tpc	0.46	−0.28

tion achieving to 62.7%. On the base of several molecular geometrical parameters as side-chain volume, residue length and surface area, etc., the model was constructed by Lejon [44] with its correlative coefficient $r^2 = 0.74$ and cross-validation $q^2 = 0.48$. Via a comparison, it is shown the whole modeling qualities of this paper, especially the q^2 indicting stabilities and generalized abilities of the model, are notably superior to that of reference reports. From another side, all calculations, for this sample set, have not obtained a high performance, either by T-scale in this paper or geometrical parameters on reference reports. Several reasons may contribute to this: (1) both T-scale and geometrical parameters focus on representing of amino acid structural features, overlooking some other prominently influencing factors as physicochemical properties and electron structural parameters, etc.; (2) descriptors concerning with primary structures of sequences are difficult to reflect steric information of peptides; (3) antibacterial mechanism of LFB is very intricate, structure–activity relationship can not be expressed by simple linear relation; (4) system errors introduced by different experiments. Considering above-mentioned problems that statistical model may be influenced by many external factors, and in conjunction with that q^2 of the T-scale-derived model is superior to a common value of 0.5, it is regarded T-scale model is valuable and promising.

4. Conclusions

Concerning with the collected 67 kinds of structural and topological variables of 135 amino acids, a novel descriptor topological scale of amino acids (T-scale) is proposed in this context. Applying T-scale to different polypeptide systems, the constructed QSAR models are satisfying. Results show that T-scale is adaptable in representation of peptide sequences not only for those comprised natural amino acids as angiotensin-converting enzyme (ACE) inhibitors and thromboplastin inhibitors (TI), but also those composed of non-coded amino acids as bovine lactoferricin-(17–31)-pentadecapeptide (LFB). Thus, T-scale has a great prospect in QSAR studies for polypeptide and its analogues.

Acknowledgements

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.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.molstruc.2006.07.004.

References

[1] S. Wold, L. Eriksson, S. Hellberg, J. Jonsson, M. Sjostrom, B. Skagerberg, C. Wikstrom, Can. J. Chem. 65 (1987) 1814–1820.
[2] C. Raychaudhury, A. Banerjee, P. Bag, S. Roy, J. Chem. Inf. Comput. Sci. 39 (1999) 248–254.
[3] S.S. Liu, C.S. Yin, S.X. Cai, Z.L. LI, J. Chin. Chem. Soc. 48 (2001) 253–260.
[4] P.H. Sneath, Theor. Biol. 12 (1966) 157–195.
[5] A. Kidera, Y. Konishi, M. Oka, M.T. Ooi, J. Protein. Chem. 4 (1985) 23–55.
[6] S. Hellberg, M. Sjöström, S. Wold, Acta Chem. Scand. B 40 (1986) 135–140.
[7] S. Hellberg, M. Sjöström, B. Skagerberg, S. Wold, J. Med. Chem. 30 (1987) 1126–1135.
[8] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, S. Wold, J. Med. Chem. 41 (1998) 2481–2491.
[9] E.R. Collantes, W.J. Dunn, J. Med. Chem. 38 (1995) 2705–2713.
[10] D. Bonchev, O. Mekenjan, G. Protic, J. Chromatogr. 176 (1979) 149–156.
[11] L. Buydens, D.L. Massart, P. Geerlings, Anal. Chem. 55 (1983) 738–744.
[12] A.C. Basak, B.D. Gute, G.D. Grunwald, J. Chem. Inf. Comput. Sci. 36 (1996) 1054–1060.
[13] L. Sun, Y. Zhou, L. Genrong, S.Z. Li, J. Mol. Struct. (Theochem) 679 (2004) 107–113.
[14] H. Winer, J. Am. Chem. Soc. 69 (1947) 2636–2641.
[15] H. Hosoya, Bull. Chem. Soc. 44 (1971) 2332–2339.
[16] M. Randic, J. Am. Chem. Soc. 97 (1975) 6609–6615.
[17] A.T. Balaban, Chem. Phys. Lett. 89 (1982) 399–404.
[18] L.B. Kier, L.H. Hall, Molecular Connectivity in Structure–Activity Analysis, Wiley, New York, 1986.

[19] J. Jonsson, L. Eriksson, S. Hellberg, Quant. Strut. –Act. Relat. 8 (1989) 204–209.
[20] A. Bock, K. Forchhammer, J. Heider, Mol. Microbiol. 5 (1991) 515–520.
[21] J.F. Atkins, R. Gesteland, Science 296 (2002) 1409–1410.
[22] A.X. Lan, G.L. Tian, Y.H. Ye, Chin. J. Org. Chem. 20 (3) (2000) 299–305.
[23] S. Wold, A. Ruhe, H. Wold, Siam. J. Sci. Statist. Comput. 5 (1984) 735–743.
[24] A. Höskuldsson, J. Chemom. 2 (1988) 211–228.
[25] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[26] M. Šoškić, J. Chem. Inf. Comput. Sci. 36 (1996) 829–832.
[27] D. Rogers, A.J. Hopfinger, J. Chem. Inf. Comput. Sci. 34 (1994) 854–866.
[28] J.M. Sutter, S.L. Dixon, P.C. Jurs, J. Chem. Inf. Comput. Sci. 35 (1995) 77–84.
[29] C.H. Hassell, J. Chem. Soc. Perkin. Trans. I 23 (1984) 155–162.
[30] S. Hellberg, L. Eriksson, J. Jonsson, Int. J. Pept. Protein Res. 37 (1991) 414–424.
[31] M. Cocchi, E. Johansson, Quant. Struct. -Act. Relat. 12 (1993) 1–8.
[32] A. Zaliani, E. Gancia, J. Chem. Inf. Comput. Sci. 39 (1999) 525–533.
[33] S. Li, B. Fu, Y. Wang, J. Chin. Chem. Soc. 48 (2001) 937–944.
[34] H. Mei, Z. Liao, Y. Zhou, S.Z. Li, Peptide Sci. 80 (2005) 775–786.
[35] H. Mei, Y. Zhou, L.L. Sun, S.Z. Li, Acta Phys. -Chim. Sin. 20 (8) (2004) 821–825.
[36] M. Baroni, S. Clement, G. Cruciani, Quant. Struct. -Act. Relat. 12 (1993) 225–231.
[37] P.M. Andersson, M. Sjöström, T. Lundstedt, Chemom. Intell. Lab. Syst. 42 (1998) 41–50.
[38] M. Sjöström, S. Rännar, Å. Wieslander, Chemometr. Intell. Lab. Syst. 29 (1995) 295–305.
[39] S. Wold, H. Antti, F. Lindgren, Chemom. Intell. Lab. Syst. 44 (1998) 175–185.
[40] Ø. Rekdal, J. Andersen, L.H. Vorland, J.S. Svendsen, J. Peptide Sci. 5 (1999) 32–45.
[41] B.E. Haug, J.S. Svendsen, J. Peptide Sci. 7 (2001) 190–196.
[42] B.E. Haug, M.L. Skar, J.S. Svendsen, J. Peptide Sci. 7 (2001) 425–432.
[43] B.E. Haug, J. Andersen, Ø. Rekdal, J.S. Svendsen, J. Peptide Sci. 8 (2002) 307–313.
[44] T. Lejon, J.S. Svendsen, B.E. Haug, J. Peptide Sci. 8 (2002) 302–306.