Journal of Peptide Science
J. Pept. Sci. 2007; **13**: 549–566
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/psc.892

JOURNAL OF
PEPTIDE SCIENCE

Three-dimensional holograph vector of atomic interaction field (3D-HoVAIF): a novel rotation–translation invariant 3D structure descriptor and its applications to peptides

FEIFEI TIAN,[a,b,c] PENG ZHOU,[b] FENGLIN LV,[c] RONG SONG[c] and ZHILIANG LI[a,b]*

[a] College of Bioengineering, Chongqing University, Chongqing, China
[b] College of Chemistry and Chemical Engineering, Chongqing University, Chongqing, China
[c] Research Institute of Surgery & Daping Hospital, Third Military Medical University, Chongqing, China

Received 3 March 2007; Revised 17 May 2007; Accepted 29 May 2007

Abstract: Quantitative structure–activity relationship (QSAR) study, important in drug design, mainly involves two aspects, molecular structural characterization (MSC) and construction of a statistical model. MSC focuses on transforming molecular structural and property characteristics into a group of numerical codes, dedicated to minimizing information loss during this process. In this context, common atoms in organic compounds are classified according to their families in the periodic table, and hybridization states, and on the basis of these, three nonbonding interactions (i.e. electrostatic, van der Waals and hydrophobic) are calculated, ultimately resulting in a new rotation–translation invariant, 3D-MSC, as a three-dimensional holograph vector of atomic interaction field (3D-HoVAIF). By applying 3D-HoVAIF to QSAR studies on two classical peptides including 58 angiotensin-converting enzyme (ACE) inhibitors and 48 bitter-tasting dipeptides, we get two excellent genetic algorithm-partial least squares (GA-PLS) models, with statistics r^2, q^2, root mean square error (RMSEE), and root mean square error of cross-validation (RMSCV) of 0.857, 0.811, 0.376, and 0.432 for ACE inhibitors and 0.940, 0.892, 0.153 and 0.205 for bitter-tasting dipeptides, respectively. By equally dividing the two datasets into training and test sets by D-optimal, the 3D-HoVAIF approach undergoes rigorous statistical validation. Furthermore, the superior performance of 3D-HoVAIF is confirmed in comparison with two other peptide MSC approaches referring to z-scale and ISA-ECI. For 58 ACE inhibitors, the GA-PLS model yields two principal components, with the following statistics: $r^2 = 0.893$, $q^2 = 0.824$, RMSEE = 0.349, RMSCV = 0.425, $q^2_{\text{ext}} = 0.739$, $r^2_{\text{ext}} = 0.784$, $r^2_{0,\text{ext}} = 0.781$, $r^{'2}_{0,ext} = 0.779$, $k = 0.962$, $k' = 1.019$, and RMSEP = 0.460; for 48 bitter-tasting dipeptides, three principal components resulted, with the statistics as: $r^2 = 0.950$, $q^2 = 0.893$, RMSEE = 0.152, RMSCV = 0.222, $q^2_{ext} = 0.875$, $r^2_{ext} = 0.919$, $r^2_{0,ext} = 0.919$, $r^{'2}_{0,ext} = 0.919$, $k = 1.018$, $k' = 0.974$, and RMSEP = 0.198. In addition, the relationship of ACE-inhibiting activities with bitter-tasting thresholds has been investigated by applying the above-constructed models to predictions on 400 theoretically possible dipeptides. Through analysis, the ACE-inhibiting activities are found to be prominently related to bitter-tasting intensities. Thus, it is deemed to be difficult to find such dipeptides that simultaneously satisfy pharmacodynamic action (high ACE-inhibiting activities) and comfortable tastes, suggesting that active components of dipeptides that are served as functional food to lower blood pressure are not very ideal. Copyright © 2007 European Peptide Society and John Wiley & Sons, Ltd.

Supplementary electronic material for this paper is available in Wiley InterScience at http://www.interscience.wiley.com/jpages/1075-2617/suppmat/

Keywords: quantitative structure–activity relationship (QSAR); three-dimensional holograph vector of atomic interaction field (3D-HoVAIF); angiotensin-converting enzyme (ACE) inhibitor; bitter-tasting dipeptide; functional food

INTRODUCTION

It is understood that since molecules are the fundamental constituents of substances, the physicochemical properties of substances are mostly interpreted in terms of the constituent molecules. Once the molecular structures are known, the properties can be concomitantly determined. Molecular structural characterization (MSC), indispensable to drug design and pharmacodynamic action evaluation, is the first key involved in quantitative structure–activity relationship (QSAR) studies. With the main idea of transforming structure and properties of organic molecules into a group of characteristic codes, MSC commits itself to minimizing information loss during this process. Current MSC techniques mainly include two types: one is based on the two-dimensional (2D) molecular skeleton and the other on three-dimensions (3D) [1]. For the 2D descriptors, since the first topological approach was proposed by Wiener [2] in 1947, there have been many other singly parameterized methods based on molecular topological structures (e.g. Hosoya index [3], Randic index [4], Balaban index [5], etc.), achieving predictable performance on organic homologous physicochemical properties [6,7]. Consequently, the past few years have witnessed important developments of 2D descriptors, confirmed by the E-State index [8–9] proposed by Kier and coworkers and a series of molecular fingerprint descriptors developed by

*Correspondence to: Zhiliang Li, College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 40044, China;
e-mail: ggootc@163.com

Wild *et al.* [10–14]. However, 2D structural descriptors are essentially unable to construct valid QSAR model on drug and bio-macromolecules because they are incapable of reproducing the true spatial conformations of molecules and also overlook ligand–receptor active sites. Because of this, 3D approaches are given much importance currently in the MSC field. Comparative molecular field analysis (CoMFA) proposed by Cramer *et al.* [15] in the 1980s and a number of similar methods based on molecular spatial structures (e.g. CoMSIA [16], HASL [17], GRID [18] and COMPASS [19]) have become the mainstay of current QSAR studies. But these methods are all confronted with some insurmountable issues, exemplified by conformation alignment prior to performing structure–activities studies, division of spatial grids, control of the number of variables, selection of reasonable probes, etc. In view of that, the weighted holistic invariant molecular (WHIM) method [20–22] proposed by Todeschini's team is completely different from CoMFA methods, and resolves molecular-field energies by classical probes, behaving as a group of rotation–translation invariants derived from a weighted transformation of different physical variables against the steric coordinates of atoms. Since then, several other QSAR methods including COMMA [23], EVA [24], DiP [25], etc. have also been developed, which are independent of conformation alignments, but these methods suffer from the demerits of low resolution of molecular structural information, implicit physicochemical meanings and complicated calculations.

In early works, the molecular distance-edge (MDE) vector [26] was proposed based on of molecular 2D structures and interatomic Pauling's electronegativity interactions; in subsequent researches, MDE evolved into a series of similar descriptors (e.g. MEDV [27], VAED [28], ADEV [29], etc.). However, these MDE-derived methods all pertain to electro-topologies because they involve only molecular 2D topological structures and simple atomic charges. Inspired by MDE and the ideas of rotation–translation invariants in WHIM and molecular potential field functions in CoMFA, we propose a new 3D MSC method, three-dimensional holograph vector of atomic interaction field (3D-HoVAIF). Considering two spatial invariants (with regard to the relative atomic distance and inherited atomic properties) of the molecule, 3D-HoVAIF descriptors result from calculations of three nonbonding interactions directly relating to bioactivities, involving neither an experimental parameter nor conformation alignment. In contrast with traditional 3D-MSC methods, 3D-HoVAIF is differerent: (i) distributions of nonbonding potential fields around drug molecules are indirectly reflected in calculations of the interatomic interactions; (ii) modeling interpretabilities and resolution power on molecular structures are augmented by classifying atoms in terms of their chemical properties; (iii) such descriptors are easy to calculate, avoiding the disadvantages of conformation alignment, grid allocation and probe settings in CoMFA. By applying 3D-HoVAIF to QSAR studies on two classical data sets, 58 angiotensin-converting enzyme (ACE) inhibitors, and 48 bitter-tasting dipeptides, 3D-HoVAIF descriptors were confirmed to be competent to extract information on molecular nonbonding potential fields and to relate with bioactivities.

PRINCIPLE AND METHODOLOGY

Three-dimensional Holograph Vector of Atomic Interaction Field

Common atoms in organic molecules, including H, C, N, P, O, S, F, Cl, Br, and I, are mainly located in five groups of the periodic table (i.e. group IA, IVA, VA, VIA, and VIIA). Enlightened by the idea that 'atoms of similar chemical properties pertain to the same species', the atoms under consideration are naturally grouped into five classes according to their families in the periodic table. In further consideration of the molecular fine structures, the above five atomic types are subsequently subdivided into 10 classes in terms of their hybridization state, which is deemed to be the key to present distinct chemical properties; thereby, a molecule ultimately corresponds to 55 atomic interaction items (Table 1). Here, what should be elucidated is how one can make further classifications (i.e. a classification beyond the above-mentioned 55 items) according to practical 3D-HoVAIF applications. Therefore, considering the three common potential energies, electrostatic, van der Waals and hydrophobic interactions, which directly relate with bioactivities, the 55 interaction items are multiplied by 3, resulting in $3 \times 55 = 165$ interaction items to represent a molecule. Although not indicating the direct ligand–receptor interaction mode, the 3D-HoVAIF descriptors mostly contain abundant information about molecular potential energy distributions, even under conditions of unknown receptor structures.

Electrostatic interaction: As an important nonbonded interaction, it obeys Coulomb's law. In Eqn (1), r_{ij} denotes the interatomic Euclidean distance, with m serving as its unit; e is the elementary charge ($1.6021892 \times 10^{-19}$C); ε_0 ($8.85418782 \times 10^{-12}\mathrm{C}^2/\mathrm{J}\cdot\mathrm{m}$) represents the dielectric constant in vacuum; q is the number of Mülliken partial charges for the atoms; m and n are the atomic types.

$$E_{mn}(E) = \sum_{i \in m, j \in n} \frac{e^2}{4\pi\varepsilon_0} \cdot \frac{q_i \cdot q_j}{r_{ij}} (1 \leq m \leq 10, m \leq n \leq 10) \tag{1}$$

Van der Waals interaction: Behaving as interatomic spatial nondipole–dipole or dipole-induced interactions, it is here expressed by the Lennard–Jones Eqn (2). Here $\varepsilon_{ij} = (\varepsilon_{ii} \cdot \varepsilon_{jj})^{1/2}$ is the potential well of

Table 1 The ten atomic types and 55 interactions in 3D-HoVAIF

No.	Atomic type	1	2	3	4	5	6	7	8	9	10
1	H	1–1	1–2	1–3	1–4	1–5	1–6	1–7	1–8	1–9	1–10
2	C(sp^3)		2–2	2–3	2–4	2–5	2–6	2–7	2–8	2–9	2–10
3	C(sp^2)			3–3	3–4	3–5	3–6	3–7	3–8	3–9	3–10
4	C(sp)				4–4	4–5	4–6	4–7	4–8	4–9	4–10
5	N(sp^3), P(sp^3)					5–5	5–6	5–7	5–8	5–9	5–10
6	N(sp^2), P(sp^2)						6–6	6–7	6–8	6–9	6–10
7	N(sp), P(sp)							7–7	7–8	7–9	7–10
8	O(sp^3), S(sp^3)								8–8	8–9	8–10
9	O(sp^2), S(sp^2)									9–9	9–10
10	F, Cl, Br, I										10–10

the atom pairs, with its value taken from Ref. 30; $R_{ij}{}^* = (C_h \cdot R_{ii}{}^* + C_h \cdot R_{jj}{}^*)/2$ indicates the van der Waals radius for modified atom pairs, with the correction factor $C_h = 1.00$ in case of sp^3 hybridization, 0.95 for sp^2 hybridization and 0.90 for sp hybridization [31].

$$E_{mn}(V) = \sum_{i \in m, j \in n} \varepsilon_{ij} \times \left[\left(\frac{R_{ij}^*}{r_{ij}} \right)^{12} - 2 \cdot \left(\frac{R_{ij}^*}{r_{ij}} \right)^6 \right]$$
$$(1 \le m \le 10, m \le n \le 10) \qquad (2)$$

Hydrophobic interaction: It is very important for drug molecules to bind to organisms. Indicating information on the systemic entropy changes, such an interaction does not have a unique expression. In 3D-HoVAIF, hydrophobic interaction is indicated by Eqn (3), which is defined in the method proposed by Kellogg *et al.* [32]. In that, S is the atomic solvent accessible surface area (SASA) [33], indicating the surface area formed by a water-molecule probe with its center on an atom surface in a circle; a is atomic hydrophobic constant, with the value taken from Ref. 34; T is the discriminant function, denoting entropy changing orientation in the case of different interatomic interactions.

$$E_{mn}(H) = \sum_{i \in m, j \in n} S_i \cdot a_i \cdot S_j \cdot a_j \cdot e^{-r_{ij}} \cdot T_{ij}$$
$$(1 \le m \le 10, m \le n \le 10) \qquad (3)$$

Partial Least Square Regression

Partial least square (PLS) regression, proposed by Wold *et al.* [35] in the 1980s to overcome multicollinearity during the modeling process, is widely used and especially suitable for the case in which the sample number is below the variable number. As in the following, the independent variable matrix **X** is subjected to a bilinear decomposition:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{F} \qquad (4)$$

In the above, matrix **T** is composed of mutually orthogonal latent variables or the scoring vector **t** which derives from a linear combination of variables in the matrix **X**. Unlike PCA, PLS simultaneously implements bilinear decomposition on the target matrix **Y**:

$$\mathbf{Y} = \mathbf{UQ}' + \mathbf{E} \qquad (5)$$

Of these, the matrix **U** comprises the latent variable u in **Y**. On the basis of that, the latent variable t, obtained by decomposing **X**, maximally overlaps with the latent variable u derived from the decomposition of **Y**. Therefore:

$$\mathbf{u} = \mathbf{vt} + \mathbf{e} \qquad (6)$$

In Eqn (6), **e** is the error vector and coefficient **v** is determined by the method of least-squares. Computational and other details are given in Refs. 36,37.

The optimal PLS principal component number is determined upon leave-one-out cross-validation.

Statistical Parameters

In common practice, the leave-one-out cross-validation correlation coefficient q^2 and root mean square error of cross-validation (RMSCV) are jointly used to evaluate modeling predictabilities, with separate expressions as in Eqns (7) and (8):

$$q^2 = 1 - \frac{PRESS}{SSQ} \qquad (7)$$

$$RMSCV = \sqrt{\frac{PRESS}{n}} \qquad (8)$$

In the above, *PRESS* is the error sum of predicted squares between Y_{obsd} (indicating the observed sample values) and leave-one-out cross-validation Y_{pred}; SSQ denotes residual sum of squares of Y_{obsd}:

$$PRESS = \sum_{i=1}^{n} (Y_{\text{obsd}}^i - Y_{\text{pred}}^i)^2 \qquad (9)$$

$$SSQ = \sum_{i=1}^{n} (Y_{\text{obsd}}^i - \overline{Y}_{\text{obsd}})^2 \qquad (10)$$

The non-cross-validated models are estimated by standard PLS parameters – explained variance r^2 and root mean square error (RMSEE) which are given by the formulas:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(Y^i_{\text{obsd}} - Y^i_{\text{est}})^2}{\sum_{i=1}^{n}(Y^i_{\text{obsd}} - \overline{Y}_{\text{obsd}})^2} \qquad (11)$$

$$RMSEE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y^i_{\text{obsd}} - Y^i_{\text{est}})^2} \qquad (12)$$

It has been recently pointed out that only q^2 underdetermines the predictabilities of a QSAR model, and therefore, an external validation is required [38–40]. The common criteria to confirm modeling predictabilities is as follows [41,42]:

$$RMSEP = \sqrt{\frac{1}{n_{\text{ext}}}\sum_{i=1}^{n_{\text{ext}}}(Y^i_{\text{obsd}} - Y^i_{\text{ext}})^2} \qquad (13)$$

$$q^2_{\text{ext}} = 1 - \frac{\sum_{i-1}^{n_{\text{ext}}}(Y^i_{\text{obsd}} - Y^i_{\text{pred}})^2}{\sum_{i-1}^{n_{\text{ext}}}(Y^i_{\text{obsd}} - \overline{Y}_{\text{tra}})^2} \qquad (14)$$

$$\frac{r^2_{\text{ext}} - r^2_{0,\text{ext}}}{r^2_{\text{ext}}} < 0.1 \text{ or } \frac{r^2_{\text{ext}} - r'^2_{0,\text{ext}}}{r^2_{\text{ext}}} < 0.1 \qquad (15)$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15 \qquad (16)$$

where, RMSEP is the root mean square error of predictions on the test set; $q_{\text{ext}}{}^2$ (external q^2) is external correlation coefficient indicating predictabilities on the test set by the model. Also, $Y^i{}_{\text{obsd}}$ denotes the observed bioactivities on test set while $Y^i{}_{\text{pred}}$ is the predicted value by the model for test samples. $Y^i{}_{\text{tra}}$ represents the average observed bioactivities over training samples; $r^2{}_{\text{ext}}$ indicates the correlation coefficient of the observed-to-predicted regression for the test set, $r^2_{0,\text{ext}}$ and $r'^2_{0,\text{ext}}$ are the correlation coefficients of the regression passing through the origin for the test set (predicted *versus* observed activities $r^2_{0,\text{ext}}$, and observed *versus* predicted activities $r'^2_{0,\text{ext}}$), with k and k' corresponding to separate slopes.

PREPARING WORK

Dataset

Angiotensin-converting enzyme (ACE) inhibitors. The rennin–angiotensin system plays an important role in regulating blood pressure in human bodies. Angiotensinogen, produced by liver, is catalyzed by rennin to disrupt the inactive angiotensin I, which is further catalyzed by the ACE to rupture into angiotensin II, an agent highly responsible for blood vessel contractions. In view of that, ACE becomes the biotarget of many important antihypertensive drugs [43]. By simulating structural characteristics of active sites in angiotensin I (an ACE substrate), ACE inhibitors competitively bind to ACE, thereby inhibiting effective ACE bioactivities. Fifty-eight ACE inhibitors, originally reported by Cushman *et al.* [44], have been extensively used as a classical QSAR sample set [45–56] to validate newly proposed MSC methods. Structures and activities for such a dataset are taken from Ref. 47, with activities expressed in the form of pIC_{50}, listed in Table 2.

Bitter-tasting dipeptides. Taste is very important to humans and other organisms, often classified into four typical types, as sweet, bitter, salty, and acid. Of that, the bitter perception protects human and many other higher animals from injury by toxic substances. In gustatory receptor cell, conduction of the gustatory signal includes a series of intricate processes mediated by the G-protein coupled receptor [57,58]. Bitter-tasting thresholds of 48 dipeptides are reported by Asao *et al.* [59]; their activities are expressed as the negative logarithm of bitter-tasting threshold concentrations (pT) (Table 3). This dataset also has wide applications in testing newly proposed descriptors [45,47–49,51,53,60,61].

Structural Optimizations

Captopril, as the first ACE inhibitor of peptide analogs, was developed by Cushman and Ondetti in 1977 [62]. Figure 1 presents crystal structure of captopril–ACE complex which was measured by Natesh *et al.* [63] using X-ray diffraction at 2 Å resolution (PDB ID: 1UZF), which shows that at the core of enzyme active site, there is a monomolecular captopril. Figure 2 reveals the spatial structure of captopril separated from the complex,

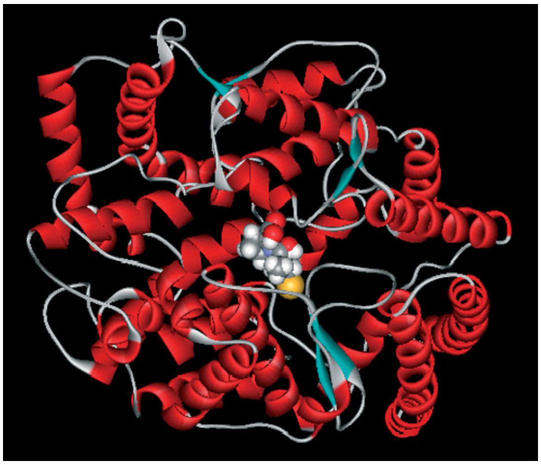

Figure 1 Three-dimensional crystal structures of captopril–ACE complex measured by X-ray diffraction.

Table 2 Sequences of 58 ACE Inhibitors and their Observed and Calculated Activities (pIC_{50})

No.	Peptide[a]	*n*	Cald1[c]	Cald2[d]	Cald3[e]	No.	Peptide[a]	Obsd[b]	Cald1[c]	Cald2[d]	Cald3[e]
1	VW	5.80	5.65	5.67	5.39	30	KG*	2.49	2.81	2.58	2.26
2	IW	5.70	5.44	5.53	5.74	31	FG*	2.43	3.01	2.59	3.19
3	IY*	5.43	4.62	4.93	4.93	32	GS	2.42	2.13	1.95	1.83
4	AW*	5.00	4.52	4.78	5.59	33	GV	2.34	2.84	2.89	2.98
5	RW	4.80	4.47	4.51	4.51	34	MG	2.32	2.54	2.53	2.40
6	VY*	4.66	4.37	4.63	4.65	35	GK	2.27	2.94	2.72	2.34
7	GW	4.52	4.16	4.34	4.36	36	GE	2.27	2.03	1.90	1.90
8	VF	4.28	4.06	3.80	4.05	37	GT*	2.24	2.23	2.39	1.82
9	AY*	4.06	3.78	4.11	4.13	38	WG	2.23	3.69	3.42	3.16
10	IP*	3.89	3.92	3.77	3.22	39	HG*	2.20	2.20	2.06	1.79
11	RP*	3.74	3.27	3.42	3.19	40	GQ*	2.15	1.81	2.32	2.06
12	AF*	3.72	3.49	3.34	3.76	41	GG*	2.14	1.87	1.84	2.17
13	GY	3.68	3.43	3.65	3.37	42	QG*	2.13	1.76	2.28	2.02
14	AP*	3.64	3.07	3.22	2.90	43	SG*	2.07	1.81	1.91	1.69
15	RF	3.64	3.69	3.56	3.78	44	LG*	2.06	3.00	2.66	2.59
16	VP*	3.38	3.65	3.66	2.79	45	GD	2.04	2.01	2.07	2.15
17	GP*	3.35	2.74	2.45	2.81	46	TG*	2.00	2.09	2.31	1.94
18	GF*	3.20	3.15	2.88	3.45	47	EG	2.00	1.91	1.81	1.83
19	IF*	3.03	4.35	3.92	3.62	48	DG	1.85	1.76	1.57	1.92
20	VG	2.96	2.75	2.76	2.77	49	PG	1.77	2.44	1.70	1.82
21	IG*	2.92	3.01	2.87	3.05	50	LA*	3.51	3.37	3.34	3.22
22	GI	2.92	3.10	2.96	3.26	51	KA*	3.42	3.23	3.10	2.55
23	GM	2.85	2.59	2.68	2.58	52	RA	3.34	2.77	3.01	2.68
24	GA*	2.70	2.24	2.38	2.42	53	YA*	3.34	3.70	3.91	3.93
25	YG	2.70	3.29	3.14	3.05	54	AA*	3.21	2.56	2.79	2.69
26	GL*	2.60	3.05	3.12	3.04	55	FR	3.04	3.67	3.45	3.53
27	AG	2.60	2.28	2.32	2.37	56	HL	2.49	2.70	2.85	2.30
28	GH	2.51	2.75	2.74	2.78	57	DA	2.42	2.87	2.75	2.68
29	GR	2.49	2.48	2.77	2.61	58	EA	2.00	2.33	2.36	2.12

[a] '*' superscript indicates that the peptide was chosen to be a member of the test set.

[b] Obsd: observed activity, pIC_{50}.

[c] Cald1: calculated activity of the PLS model.

[d] Cald2: calculated activity of the GA-PLS model.

[e] Cald3: predicted activity in case of dividing training/test set.

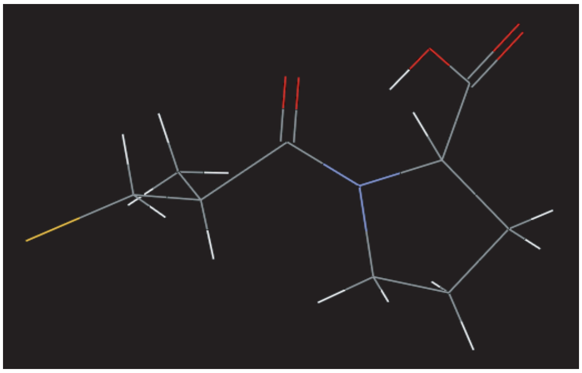

Figure 2 Steric conformation of captopril separated from the complex.

demonstrating that the main chain stretches out, while the side chains are distorted to some extent under the influences of the nearby target enzyme residues. With captopril crystal structure separated from the complex serving as the pharmacophoric conformation template, the original steric structures of the 58 dipeptides were constructed by HyperChem 7.5 [64]. To further eliminate irrationality of the structures, combinational optimization was subsequently implemented by molecular mechanics and molecular dynamics (MM + force field). First, each molecule was simulated for 1 ps at 300 K by molecular dynamics, in steps of 1 fs; it was then optimized by conjugate gradient until convergence conditions were achieved with RMS gradient <0.001 kcal mol^{-1}, resulting in the ultimate conformations.

The pharmacophoric conformations of the 48 bitter-tasting peptides, however, were taken from their low-energy conformations because there are still no utilizable crystal structures of the peptide analog–gustatory receptor complex. For the 48 bitter-tasting dipeptides, the original structures were auto-generated in the module Database of HyperChem 7.5 [64]; and considering

Table 3 Sequences of 48 bitter-tasting dipeptides and their observed and calculated activity (pT)

No.	Peptide[a]	Obsd[b]	Cald1[c]	Cald2[d]	Cald3[e]	No.	Peptide[a]	Obsd[b]	Cald1[c]	Cald2[d]	Cald3[e]
1	GV	1.13	1.26	1.28	1.17	25	II*	2.26	2.32	2.37	2.41
2	GL	1.68	1.43	1.51	1.43	26	IP	2.40	2.05	2.21	2.30
3	GI*	1.70	1.45	1.51	1.52	27	IW	3.05	3.03	3.20	2.87
4	GP*	1.35	1.25	1.44	1.76	28	IN	1.49	1.47	1.30	1.33
5	GF	1.80	1.80	1.79	1.60	29	ID	1.37	1.25	1.34	1.34
6	GW	1.89	1.76	1.99	1.79	30	IQ*	1.49	1.65	1.37	1.36
7	GY*	1.77	1.71	1.49	1.35	31	IE*	1.37	1.41	1.42	1.35
8	AV	1.16	1.50	1.46	1.42	32	IK	1.65	2.18	1.56	1.58
9	AL*	1.70	1.65	1.66	1.66	33	IS*	1.49	1.41	1.41	1.38
10	AF	1.72	2.02	1.92	1.76	34	IT	1.49	1.52	1.58	1.80
11	VG	1.19	1.26	1.26	1.14	35	PA	1.32	1.47	1.48	1.41
12	VA*	1.16	1.46	1.40	1.34	36	PL	2.22	2.53	2.41	2.29
13	VV*	1.71	1.88	1.83	1.79	37	PI	2.33	2.09	2.14	2.17
14	VL	2.00	2.06	2.08	2.11	38	PY	1.80	2.35	2.16	1.98
15	LG*	1.72	1.43	1.50	1.48	39	PF*	2.80	2.44	2.46	2.87
16	LA*	1.72	1.63	1.65	1.65	40	FG*	1.77	1.78	1.77	2.04
17	LL*	2.35	2.24	2.36	2.47	41	FL*	2.87	2.67	2.79	3.11
18	LF*	2.75	2.63	2.69	3.05	42	FP	2.70	2.41	2.59	2.56
19	LW*	3.40	2.99	3.20	3.63	43	FF	3.10	3.12	3.28	3.23
20	LY*	2.46	2.54	2.39	2.33	44	FY	3.13	3.10	3.13	2.99
21	IG*	1.68	1.45	1.48	1.43	45	WE	1.56	2.01	1.67	1.77
22	IA*	1.68	1.65	1.64	1.61	46	WW	3.60	3.61	3.52	3.69
23	IV	2.05	2.09	2.09	2.12	47	YL*	2.40	2.58	2.48	2.39
24	IL*	2.26	2.27	2.34	2.42	48	SL*	1.49	1.34	1.58	1.43

[a] '*' superscript indicates that the peptide was chosen to be a member of the test set.

[b] Obsd: observed activity, pT.

[c] Cald1: calculated activity of the PLS model.

[d] Cald2: calculated activity of the GA-PLS model.

[e] Cald3: estimated/predicted activity in case of dividing training/test set.

their high molecular flexibilities, they were subsequently optimized by utilizing a conformational search, with the related software BioMedCAChe 6.1 [65]. First, each rotational bond in the molecule is appended by a search label by the Geometry Label Wizard, and simultaneously a searching realm of −180° to +180° and a step length of 36° were defined. Then for each molecule, once the searching process is completed, an aggregation comprising many low-energy conformations would be correspondingly generated. Following that, the lowest-energy conformation is taken out to implement molecular mechanics optimization by HyperChem 7.5 [64], giving rise to the ultimate conformation (the parameters included in molecular mechanics are the same as above, and the optimization-derived dipeptide WW of the highest bitter-tasting activities is presented in its low-energy conformation in Figure 3).

Calculations of 3D-HoVAIF Descriptors

With the optimal conformations inputted into the semiexperimental quantum chemistry software MOPAC 6.0 [66], Mülliken partial charges for each atom were worked out in the single-point form at the PM3 level. Then inputting each atomic Cartesian coordinates and partial charge into program GET3D, which is edited in the C-language, we ultimately obtained the corresponding 3D-HoVAIF descriptors for all the samples. For the reason that $C_{(sp)}$, N and halogen atoms are absent in natural dipeptides, there are 81 empty items in a total of 165 3D-HoVAIF descriptors. Taking off all these empty items, 84 3D-HoVAIF descriptors remained for a molecule, of which, variables V1–V28 demote electrostatic items, V29–V56 indicate van der Waals interactions and V57 – V84 characterize the hydrophobic interactions.

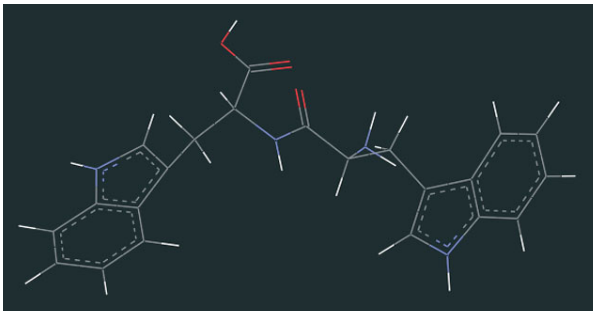

Figure 3 Theoretical low-energy conformation of WW.

DOI: 10.1002/psc

RESULTS AND DISCUSSION

QSAR Studies on ACE Inhibitors

On the basis of the chemometrics software Simca-p 10.0 [67], the PLS model is constructed to relate 3D-HoVAIF descriptors (X) with bioactivities (Y) for 58 ACE inhibitors, yielding two prominent principal components which together account for 79.2% variance of the Y variables, with cross-validation achieving 68.7%. The relative statistics are listed as $r^2 = 0.792$, $q^2 = 0.687$, RMSEE = 0.459 and RMSCV = 0.542. Table 4 lists the results from this paper and other available reference reports, of which the earlier references R1–R4 contain no variable selection, while the latter references R6–R13 show the opposite, i.e. they all implement variable selections by different methods. As seen from this table, the 3D-HoVAIF approach is demonstrated to have greatly gained in comparison with the methods of references R1 – R4, while fading into significance by comparison with those in references R6–R13 with respect to q^2 (indicating modeling stabilities). However, variable selection, which aims to seek a few factors directly relating to the dependent variable (Y), is deemed to be efficacious in improving the modeling qualities by filtering out noise and other interferences. Thus, the genetic algorithm-partial least squares (GA-PLS) [68] is utilized here, with the related programs Gaot_Toolbox [69] and PLS_Toolbox [70] based on the Matlab 6.1 [71] environment. The parameter settings included in the GA listed here are: original population size, 150; maximum genetic algebra, 200; convergent condition, 80% of individuals achieve an agreement; mutation probability, 0.5%; cross-interchange, 2 points; cross-validation, leave-1/5-out; data pretreatment, autoscaling; others are taken as defaults. The optimal variable subset is composed of the variables V3, V5, V6, V9, V15, V16, V17, V18, V22, V25, V29, V32, V34, V37, V43, V45, V49, V50, V53, V55, V60, V61, V69, V73, V75, V78, V80, and V81. Fitness curves with the evolution of the GA population are shown in Figure 4. By such an implementation, the PLS model has been largely improved, ultimately resulting in 28 variables from the whole 84 3D-HoVAIF descriptors. On further analysis by the software Simca-p 10.0 [67], the PLS model yielded three principal components, with r^2, q^2, RMSEE, and RMSCV of 0.857, 0.811, 0.376, and 0.432, respectively, and in contrast with references R6–R13, it has been found to be only slightly inferior to the MEDV model proposed by Liu *et al.* [50] utilizing GA-MLR modeling methods with respect to q^2, while surpassing other reference reports (Table 4). Figure 5 is the scoring scatter at the top two PLS principal component spaces in X, and of these, samples of $\text{pIC}_{50} > 4$ are marked by circles, $3 \leq \text{pIC}_{50} \leq 4$ by

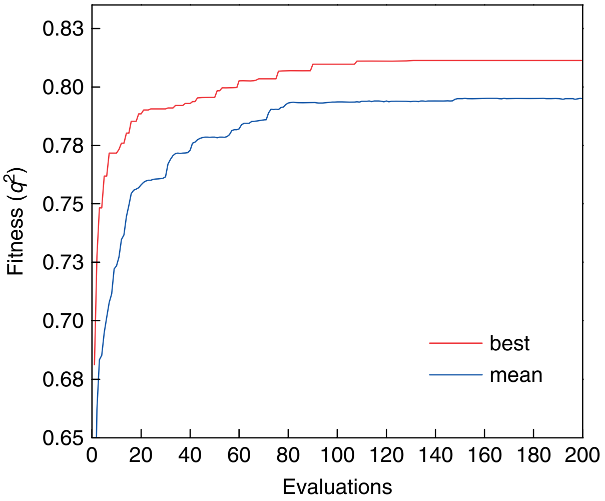

Figure 4 Fitness curve with the evolution of the GA population.

Table 4 Comparisons among different QSAR models for ACE inhibitors

No.	Descriptor	Method	PC	r^2	q^2	*RMSEE*	*RMSCV*
1	*z*-scale [45]	PLS	2	0.770	0.723	—	—
2	*t*-score [46]	PLS	1	0.744	—	0.50	—
3	ISA-ECI [47]	PLS	2	0.700	—	—	—
4	MSW-score [48]	PLS	2	0.708	0.637	—	—
5	3D-HoVAIF	PLS	2	0.792	0.687	0.459	0.542
6	VMEE [49]	SMR-MLR	2	0.741	0.711	0.504	—
7	MEDV [50]	GA-MLR	5	0.883	0.861	0.339	0.370
8	MHDV [51]	SMR-PCR	19	0.878	0.753	0.347	0.50
9	MEDV-13 [52]	SMR-PCR	19	0.895	0.783	0.32	0.47
10	VHSE [53]	SMR-PLS	1	0.770	0.745	0.48	—
11	SSIA [54]	SMR-PLS	—	0.789	0.773	0.47	—
12	T-scale [55]	SMR-PLS	2	0.845	0.786	0.39	—
13	GVSC [56]	GA-PLS	1	0.766	0.712	0.48	—
14	3D-HoVAIF	GA-PLS	3	0.857	0.811	0.376	0.432

DOI: 10.1002/psc

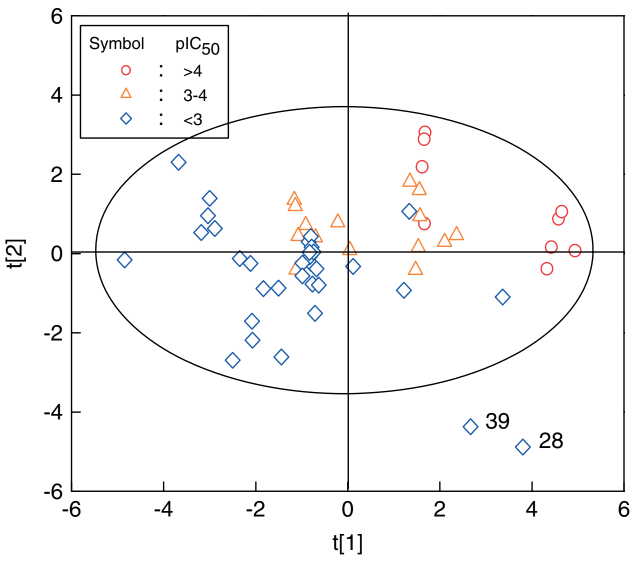

Figure 5 The GA-PLS scores $t1$ and $t2$ for ACE inhibitors.

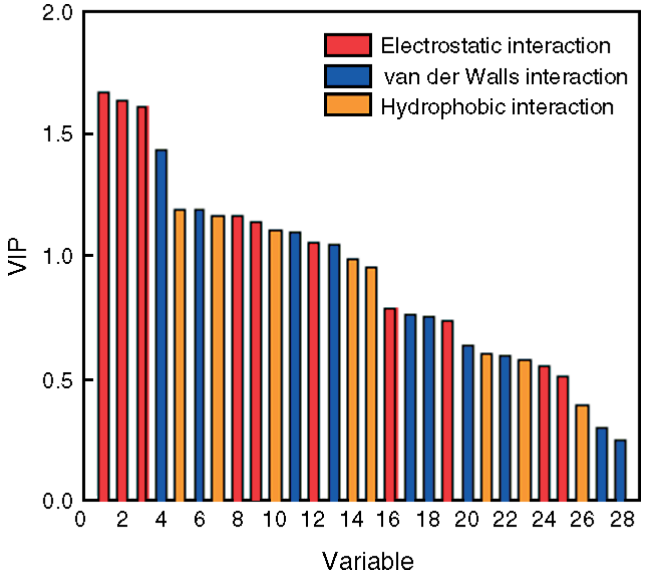

Figure 6 Variable importance in projection (VIP) in the GA-PLS model.

triangles and $\text{pIC}_{50} < 3$ by squares. In this figure, the ACE inhibitors are inclined to be increasingly distributed from the bottom left to the top right corner according to the activity values, with samples of similar bioactivities favorably assembled together, thereby suggesting that the top two GA-derived principal components are already sufficient to characterize activity distributions. Besides, out of all the samples, only samples #28 and #39 exceed Hotelling's T^2 ellipse of 95% confidence. By analysis, these two dipeptides are found to have the same compositions (GH and HG, respectively), with one of their constituent residues as H (histidine) which is very rare in other peptides, and therefore behaving a little differently, to be separated at the bottom right corner of this scoring scatter plot. Figure 6 delineates the variable importance in projection (VIP) [37] for the 28 3D-HoVAIF descriptors, the different colors corresponding to the electrostatic interactions, the van der Waals interactions and the hydrophobicities separately, indicating approximately the same contribution to the model but relatively more contribution by electrostatic interaction, than van der Waals and hydrophobicities in turn. To further investigate the reliability of this GA-PLS model, *Y* random permutations test [72], accompanied by modeling over-fitness validation, was carried out. Here, the *Y* variables are subject to 100 times random permutations, and then plots of r^2 and q^2 of the permutated model against correlation coefficients of original and permutated *Y* variables were separately given out. Results of the *Y* random permutations test are demonstrated in Figure 7, in which the slopes of r^2 and q^2 regression lines are 0.156 and -0.385, respectively. The high values of r^2 and q^2 are not deemed to occur by accident. Figures 8 and 9 are plots of GA-PLS-calculated and cross-validation-predicted *versus* observed activities for 58 ACE inhibitors, respectively, wherein most samples are uniformly dispersed along a line passing through the origin and forming an angle of 45°, except for

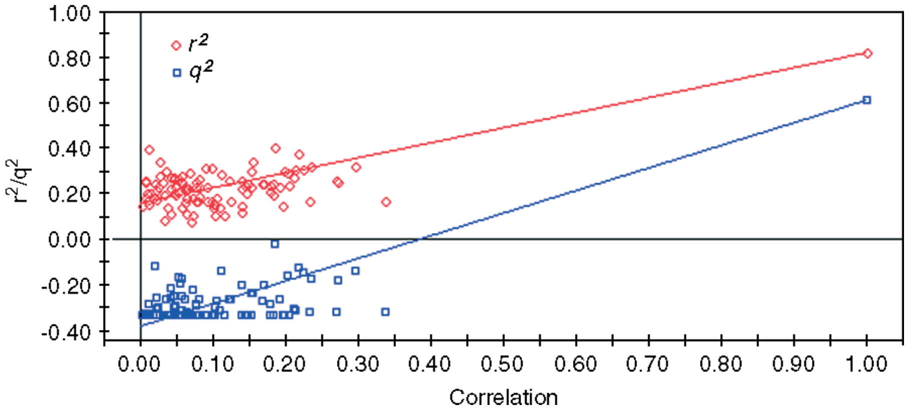

Figure 7 *Y* random permutations test in the GA-PLS model.

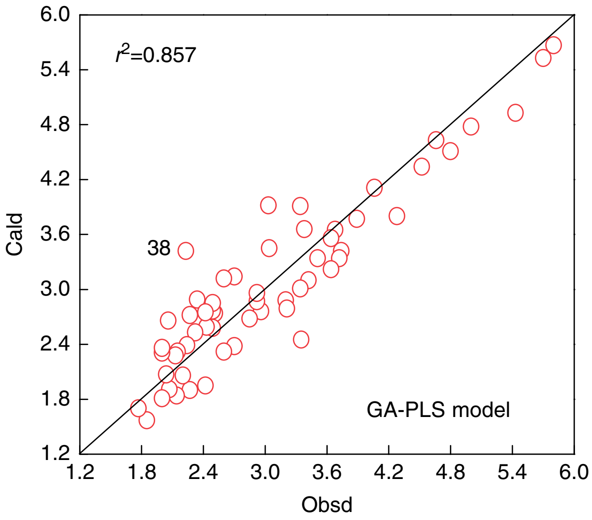

Figure 8 Plot of GA-PLS-calculated *vs* observed activities for 58 ACE inhibitors. This figure is available in colour online at www.interscience.wiley.com/journal/jpepsci.

peptide #38 which has large positive residue errors. By structural analysis, dipeptide #38 comprises residue W (tryptophan), which possesses a conjugated dicyclic side chain of large volume. In Table 2, compounds of high activities ($pIC_{50} > 4$) are often found to contain residue W or structurally similar residues, such as Y (tyrosine) and F (phenylalanine), being opposite to the case of sample #38, which has a relative low observed activity ($pIC_{50} = 2.23$). This predicted large positive error is ascribed to one of the following: (i) the experimental value is a bit low; (ii) sample #38 is special by itself; or (iii) the model selected is irrational. In many cases, outliers are concealed from the model, risking the loss of some valuable information. So, sample #38 is carefully reserved here.

QSAR Studies On Bitter-Tasting Dipeptides

First, we directly employed 84 3D-HoVAIF descriptors to construct the QSAR model for activities (pT) of 48 dipeptides, yielding a PLS model with three prominent principal components, of which r^2, q^2, *RMSEE* and *RMSCV* were 0.876, 0.798, 0.219, and 0.280, respectively. Such results have already been satisfactory in spite of no variable selection, but to further improve the modeling quality, GA is utilized to select variables (with the GA parameter settings as above). The optimal variable subset is composed of 25 3D-HoVAIF descriptors, of which 8 electrostatic interactions are indicated by variables V1, V2, V9, V10, V14, V19, V22, and V27; 7 van der Waals interactions are as variables V32, V37, V45, V48, V52, V54, and V55, and 10 hydrophobic interactions as variables V57, V 58, V60, V64, V74, V76, V78, V80, V83, and V84. Statistics of this GA-PLS model are $r^2 = 0.940$, $q^2 = 0.892$, $RMSEE = 0.153$ and $RMSCV = 0.205$, indicating both largely advanced fitting abilities (r^2) and stabilities (q^2) (Figures 10 and 11 are plots of PLS and GA-PLS cross-validation-predicted *vs* observed activities for 48 bitter-tasting dipeptides respectively). Table 5 gives the available reference reports on this dataset; by comparison, it is found that irrespective of whether implementing variable selection or not, the 3D-HoVAIF model is superior, especially with its predictabilities q^2 remarkably improved. To validate normal hypothesis, we then implement the normal probability of the standardized residual [37] for the regression model; in Figure 12 most of the residue errors follow a normal

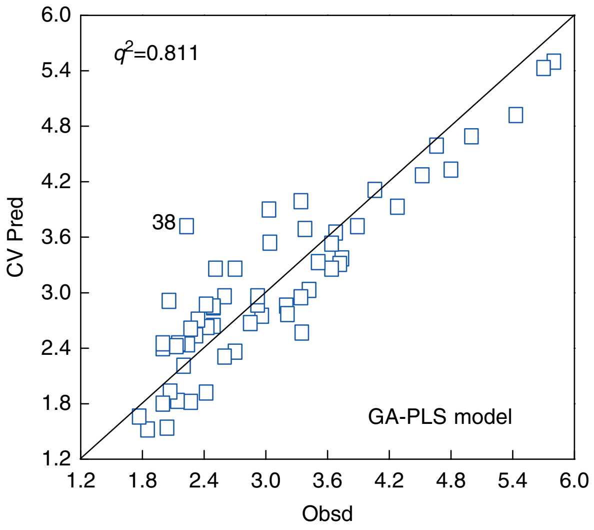

Figure 9 Plot of GA-PLS cross-validation-predicted *vs* observed activities for 58 ACE inhibitors. This figure is available in colour online at www.interscience.wiley.com/journal/jpepsci.

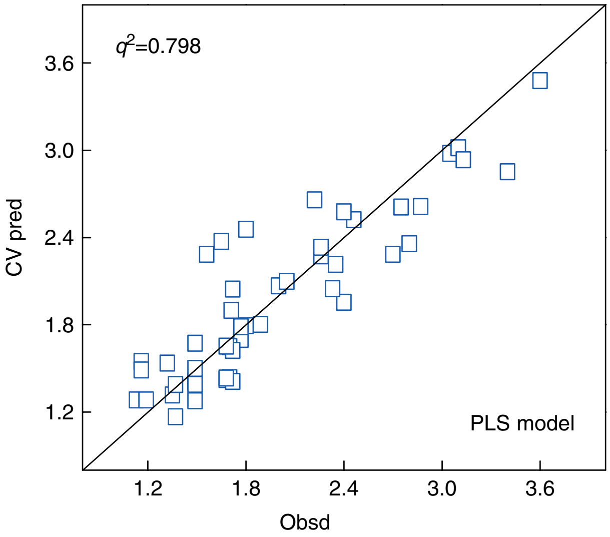

Figure 10 Plot of PLS cross-validation-predicted *vs* observed activities for 48 bitter-tasting dipeptides. This figure is available in colour online at www.interscience.wiley.com/journal/jpepsci.

Table 5 Comparison among different QSAR models for bitter-tasting dipeptides

No.	Descriptor	Method	PC	r^2	q^2	RMSEE	RMSCV
1	*z*-Scale [45]	PLS	2	0.824	—	0.26	—
2	Extended *z*-scale [60]	PLS	1	0.780	—	—	—
3	ISA-ECI [47]	PLS	2	0.847	—	0.24	—
4	MSW-score [48]	PLS	3	0.754	0.710	—	—
5	MARCH-INSIDE [61]	PLS	3	0.858	—	0.226	—
6	**3D-HoVAIF**	**PLS**	**3**	**0.876**	**0.798**	**0.219**	**0.280**
7[a]	VMEE [49]	SMR-MLR	3	0.735	0.677	0.323	—
8	MHDV [51]	SMR-PCA	10	0.919	0.857	0.178	0.232
9	VHSE [53]	SMR-PLS	3	0.910	0.816	0.20	—
10	**3D-HoVAIF**	**GA-PLS**	**3**	**0.940**	**0.892**	**0.153**	**0.205**

[a] There is an outlier for this model.

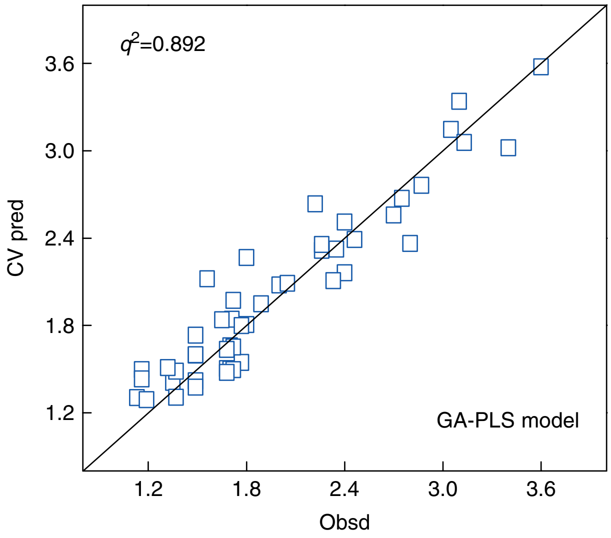

Figure 11 Plot of GA-PLS cross-validation-predicted *vs* observed activities for 48 bitter-tasting dipeptides. This figure is available in colour online at www.interscience.wiley.com /journal/jpepsci.

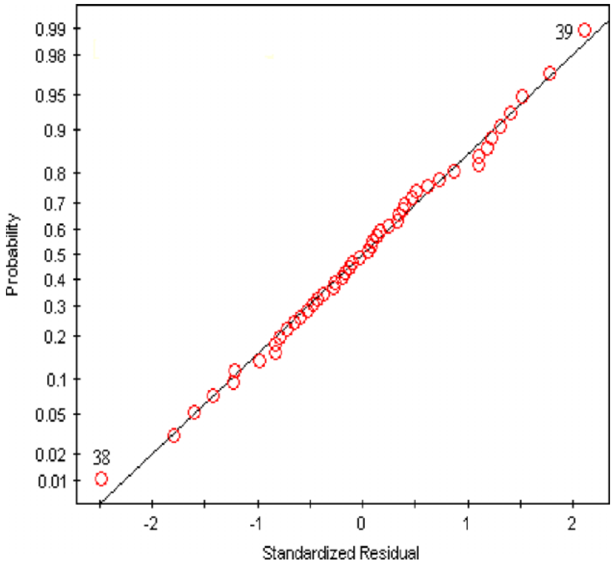

Figure 12 The normal probability plot of the *Y*-standardized residuals for the bitter-tasting dipeptides. This figure is available in colour online at www.interscience.wiley.com/journal /jpepsci.

distribution, with the only exceptions of samples #38 and 39, for which the standardized residues (SD) are beyond the range of ±2 (Wold *et al.* [37] has pointed out that it would be permissible for SD to be within the range of ±3). So the normal hypotheses are confirmed to be true. Figure 13 delineates the scoring scatter of the top two principal components in the *X*-space, wherein the sample points are demonstrated to present an increasing distribution, with activities increasing in the first principal component space (different activity ranks are marked by different symbols in Figure 13), and dipeptides of similar activities are obviously assembled together. At the second principal component space, however, the distribution of sample points is less regular. Besides, in Figure 13 there is an extreme outlier caused by sample #46 of which the scoring point remarkably deviates from the other 47 samples. By analysis, sample #46 comprises two large W (tryptophan) residues in addition to possessing the largest observed activity (pT = 3.60), so its abnormality in the scoring distribution is ascribed to specificities in molecular structure and observed activity. Such a phenomenon also occurred for sample #45, for which also the scoring point slightly deviates from other ones because its *N*-terminus is occupied by the residue W (only samples #45 and #46 contain a residue W at the *N*-terminus). Therefore, it is concluded 3D-HoVAIF descriptors are efficacious in mapping molecular structural characteristic onto an independent variable space, and to favorably reproduce molecular fine structures in the statistical model. Figure 14 is the plot of *u*1 against *t*1 in the GA-PLS model (*t*1 and *u*1 indicate the first principal component in the *X* and *Y* scoring space, respectively).

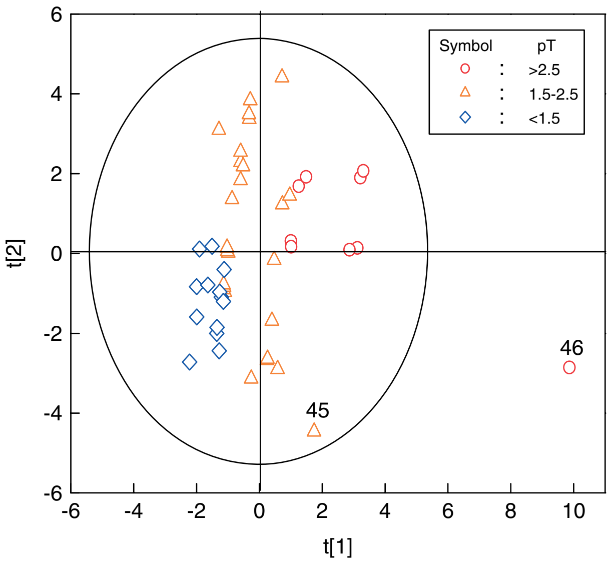

Figure 13 The GA-PLS scores $t1$ and $t2$ for the bitter-tasting dipeptides. This figure is available in colour online at www.interscience.wiley.com/journal/jpepsci.

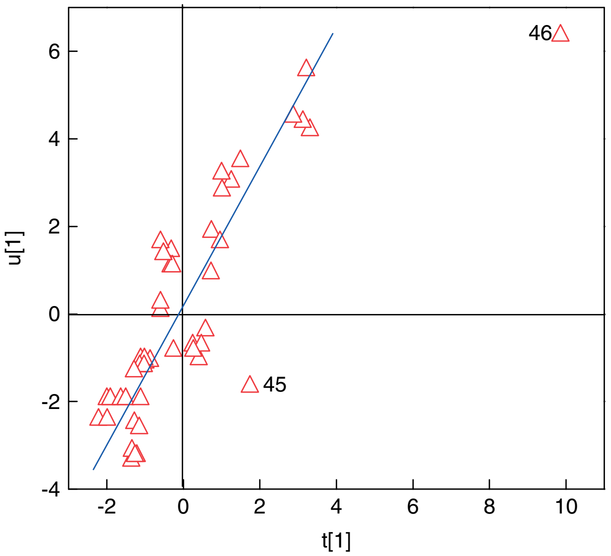

Figure 14 Plot of the GA-PLS scores $u1$ against $t1$ for the bitter-tasting dipeptides. This figure is available in colour online at www.interscience.wiley.com/journal/jpepsci.

In Figure 14, a strong linear relation is displayed between $t1$ and $u1$, and except for #45 and #46, most samples are dispersed nearly along this line, which is consistent with conclusions of Figure 13. Figure 15 shows the loading contribution of the GA-PLS model, wherein hydrophobic and van der Waals interactions positively contribute to Y at the first principal component, while electrostatic interaction has negative contributions. Out of this, large loading contributions ($|\text{loading}| > 0.3$) to the first principal component are provided by electrostatic interactions of H–H, C(sp^3)–C(sp^2) and C(sp^2)–C(sp^2), van der Waals interactions of C(sp^3)–C(sp^2) and hydrophobic interactions of C(sp^3)–N(sp^3). Generally, prominent variables to the model are mostly C–H interactions, just in agreement with the true case because dipeptide skeletons are mainly made up of these two atomic types.

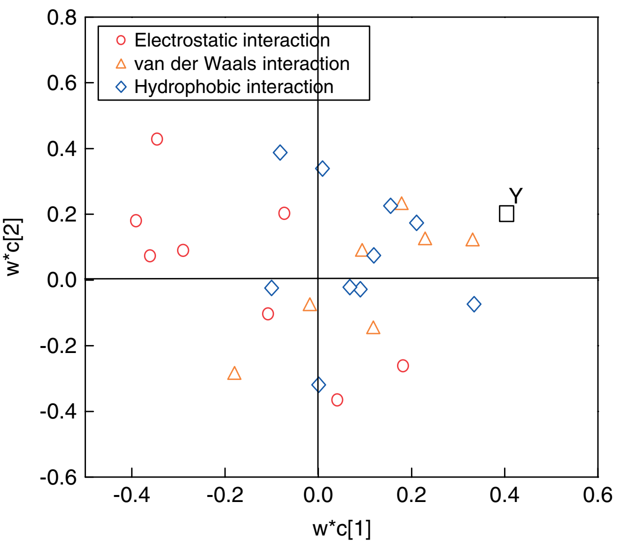

Figure 15 Loading plot of the GA-PLS model for the bitter-tasting dipeptides. This figure is available in colour online at www.interscience.wiley.com/journal/jpepsci.

Model Validation

It has been recently recognized that only leave-one-out cross-validation correlation coefficient q^2 underdetermines reliabilities of a QSAR model, and therefore a rigorous validation by an external test set is required. For three important statistical parameters referring to non-cross-validation, correlation coefficient r^2, cross-validation correlation coefficient q^2 and external correlation coefficient q^2_{ext} on the test set, the former serves as the necessary and sufficient condition for the latter in turns, i.e. a QSAR model of excellent predictabilities (q^2_{ext}) must have high stabilities (q^2) which is ensured by strong fitting abilities (r^2). In the above discussions, it has already been confirmed that the 3D-HoVAIF model possesses favorable fitting abilities and stabilities on 58 ACE inhibitors and 48 bitter-tasting dipeptides, and following that, a further validation would be implemented by an external sample set. Meanwhile, a parallel has been drawn between 3D-HoVAIF and two classical amino acid descriptors including z-scale and ISA-ECI. Among these, the z-scale indicates an amino acid scoring vector that is extracted from 29 physicochemical properties of natural amino acids by principal component analysis by Hellberg *et al.* [73], and has wide applications in peptide activity predictions [74–77], protein design [78], analysis of peptide–protein binding affinities [79] and protein stability discriminations [80]. ISA-ECI (referring to isotropic surface area and electronic charge index) consists of two fundamental

DOI: 10.1002/psc

parameters proposed by Collantes *et al.* [47] to indicate information on steric characteristics and local dipole properties of amino acid side chains, also contributing much to the fields of combinational peptide library design [81], HLA-A*0201-restricted CTL epitope recognition [82], charge distribution simulation for peptide skeletons [83,] etc. The z-scale and ISA-ECI values are listed in Table 6.

First, we employ D-optimal [84,85] to divide the datasets for the 58 ACE inhibitors and 48 bitter-tasting dipeptides separately, resulting in 29 and 24 training/test samples. As an optimal algorithm for the determinant, D-optimal ensures maximization of information matrix (X^TX) determinant value of training set. Here, for the reason that QSAR studies on sample sets are performed by three different characteristic methods, z-scale, ISA-ECI and 3D-HoVAIF, orthogonal coding is taken to achieve equalization when implementing D-optimal calculations on each dipeptide (i.e. each dipeptide location is characterized by 20 binary variables, and each variable indicates a type of amino acid residue with occurrence or not, corresponding to 1 or 0, respectively). Test samples resulting by D-optimal algorithm are marked by the symbol '*' in Tables 2 and 3. D-optimal algorithm is implemented by Matlab 6.1 [71]. Based upon that and by using the training set, GA-PLS models are created for the z-scale, ISA-ECI and 3D-HoVAIF, with separate statistics referred to in Tables 7 and 8. It is revealed that that the 3D-HoVAIF model is remarkably superior to the z-scale and ISA-ECI models with respect to both fitting abilities for the training set and predictabilities for the test set, with external validation statistics as $\frac{r^2_{\text{ext}} - r^2_{0,\text{ext}}}{r^2_{\text{ext}}} = 0.004$, $\frac{r^2_{\text{ext}} - r'^2_{0,\text{ext}}}{r^2_{\text{ext}}} = 0.006$, $k = 0.962$ and $k' = 1.019$ for ACE inhibitors, and $\frac{r^2_{\text{ext}} - r^2_{0,\text{ext}}}{r^2_{\text{ext}}} = 0.000$, $\frac{r^2_{\text{ext}} - r'^2_{0,\text{ext}}}{r^2_{\text{ext}}} = 0.000$, $k = 1.018$

Table 6 z-Scale and ISA-ECI descriptors for 20 natural amino acids (AAs)

AAs	z-Scale			ISA-ECI	
	z1	z2	z3	ISA	ECI
Ala, A	0.07	−1.73	0.09	62.90	0.05
Arg, R	2.88	2.52	−3.44	52.98	1.69
Asn, N	3.22	1.45	0.84	17.87	1.31
Asp, D	3.64	1.13	2.36	18.46	1.25
Cys, C	0.71	−0.97	4.13	78.51	0.15
Gln, Q	2.18	0.53	−1.14	19.53	1.36
Glu, E	3.08	0.39	−0.07	30.19	1.31
Gly, G	2.23	−5.36	0.30	19.93	0.02
His, H	2.41	1.74	1.11	87.38	0.56
Ile, I	−4.44	−1.68	−1.03	149.77	0.09
Leu, L	−4.19	−1.03	−0.98	154.35	0.10
Lys, K	2.84	1.41	−3.14	102.78	0.53
Met, M	−2.49	−0.27	−0.41	132.22	0.34
Phe, F	−4.92	1.30	0.45	189.42	0.14
Pro, P	−1.22	0.88	2.23	122.35	0.16
Ser, S	1.96	−1.63	0.57	19.75	0.56
Thr, T	0.92	−2.09	−1.40	59.44	0.65
Trp, W	−4.75	3.65	0.85	179.16	1.08
Tyr, Y	−1.39	2.32	0.01	132.16	0.72
Val, V	−2.69	−2.53	−1.29	120.91	0.07

Table 7 Comparisons among different QSAR models separately constructed by z-scale, ISA-ECI and 3D-HoVAIF descriptors on ACE inhibitors

No.	Descriptor	Method	PC	Training set				Test set						
				r^2	q^2	RMSEE	RMSCV	q^2_{ext}	r^2_{ext}	$r^2_{0,\text{ext}}$	$r'^2_{0,\text{ext}}$	k	k'	RMSEP
1	z-Scale	GA-PLS	2	0.864	0.794	0.394	0.487	0.671	0.759	0.755	0.753	0.957	1.019	0.516
2	ISA-ECI	GA-PLS	1	0.721	0.633	0.566	0.649	0.572	0.652	0.650	0.650	0.941	1.028	0.589
3	**3D-HoVAIF**	**GA-PLS**	**2**	**0.893**	**0.824**	**0.349**	**0.425**	**0.739**	**0.784**	**0.781**	**0.779**	**0.962**	**1.019**	**0.460**

Table 8 Comparisons among different QSAR models separately constructed by z-scale, ISA-ECI and 3D-HoVAIF descriptors on bitter-tasting dipeptides

No.	Descriptor	Method	PC	Training set				Test set						
				r^2	q^2	*RMSEE*	*RMSCV*	q^2_{ext}	r^2_{ext}	$r^2_{0,\text{ext}}$	$r'^2_{0,\text{ext}}$	k	k'	RMSEP
1	z-scale	GA-PLS	3	0.897	0.837	0.217	0.274	0.721	0.793	0.790	0.790	1.011	0.969	0.297
2	ISA-ECI	GA-PLS	2	0.884	0.842	0.231	0.269	0.802	0.844	0.843	0.843	1.021	0.975	0.250
3	**3D-HoVAIF**	**GA-PLS**	**3**	**0.950**	**0.893**	**0.152**	**0.222**	**0.875**	**0.919**	**0.919**	**0.919**	**1.018**	**0.974**	**0.198**

DOI: 10.1002/psc

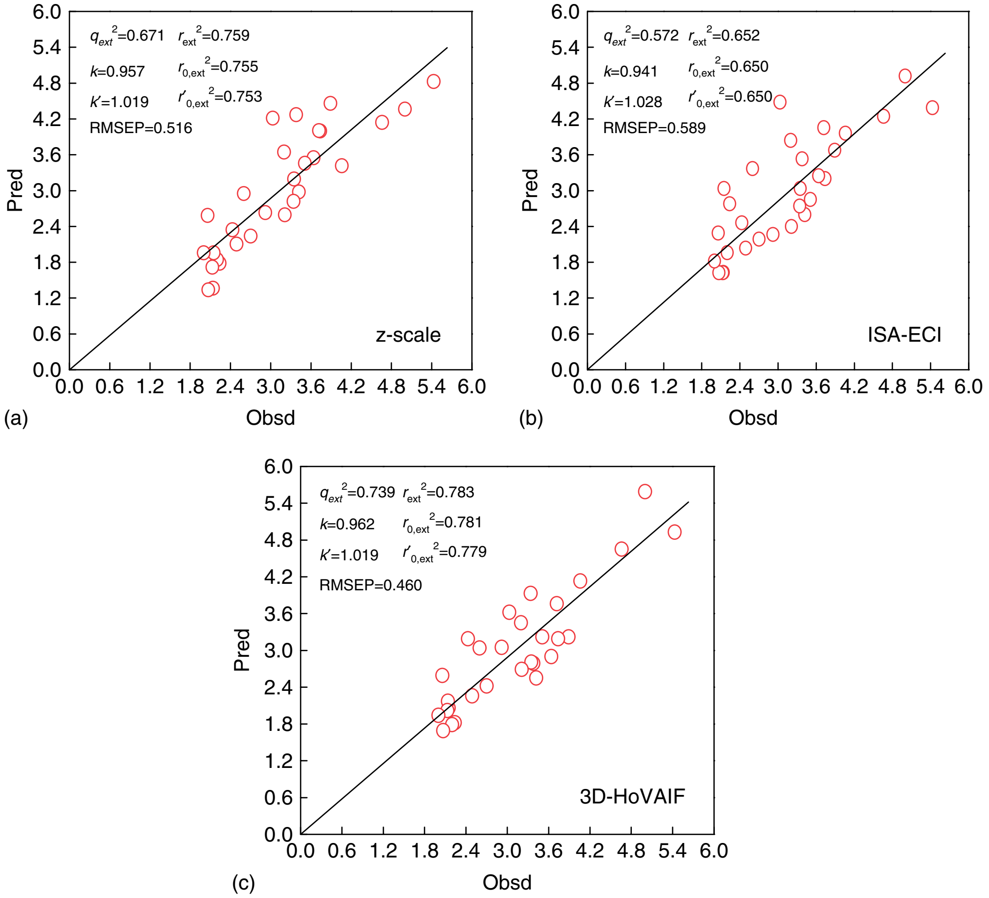

Figure 16 Plot of predicted *vs* observed activities for 29 ACE inhibitors in test set: (a) z-scale; (b) ISA-ECI; (c) 3D-HoVAIF. This figure is available in colour online at www.interscience.wiley.com/journal/jpepsci.

and $k' = 0.974$ for bitter-tasting dipeptides, satisfying Eqns (13) and (14). In contrast to ACE inhibitors, bitter-tasting dipeptides obtain better modeling qualities. This can be ascribed to the fact that ACE-inhibiting activities are predominently related to the two-dimensional topological structures of the dipeptides, with large inhibiting activities corresponding to the large two-dimensional topological structures, while for bitter-tasting dipeptides, such a rule is impermissible and activities are first related to the information on three-dimensional potential fields, and second to two-dimensional topological structures. Figures 16 and 17 are plots of predicted *versus* observed activities for 29 ACE inhibitors and 24 bitter-tasting dipeptides in the test set: (a) z-scale; (b) ISA-ECI; (c) 3D-HoVAIF, respectively, wherein 3D-HoVAIF-predicted sample points are uniformly dispersed along a regression line passing through the origin, while with inferior uniformities by z-scale and ISA-ECI. In Figure 16(b), the ISA-ECI model overestimates the ACE-inhibiting activity of sample #19, thereby largely undermining predictabilities. In Figure 17(a), the z-scale model, however, yields large calculated errors over all the 24 bitter-tasting dipeptides in the test set, so this model is also deemed to be poor in predictabilities.

Generally, z-scale and ISA-ECI, serving as two-dimensional structural descriptors based on peptide primary-order sequence characteristics, are both unable to provide insight into three-dimensional potential field information on peptide–receptor interactions. Besides, z-scale and ISA-ECI, characterizing intricate peptide analogs by only taking a few principal properties or side-chain parameters, suffer considerably in comparison with multidimensional 3D-HoVAIF vectors with respect to resolution capabilities on intricate pharmaceutical properties, so the resulting models are inferior to the 3D-HoVAIF model to different extents for ACE inhibitors and bitter-tasting dipeptides. But it should be remarked that 3D-HoVAIF would also be infeasible in practical applications in cases of more difficult structural optimizations caused by high structural flexibilities with the peptide chains becoming longer.

Correlation Analysis between Dipeptide Bitter-tasting Intensities and ACE-inhibiting Activities

Numerous proteins and polypeptide analogs are absorbed by the human body via foods, and consequently they are hydrolyzed by proteases *in vivo* into

DOI: 10.1002/psc

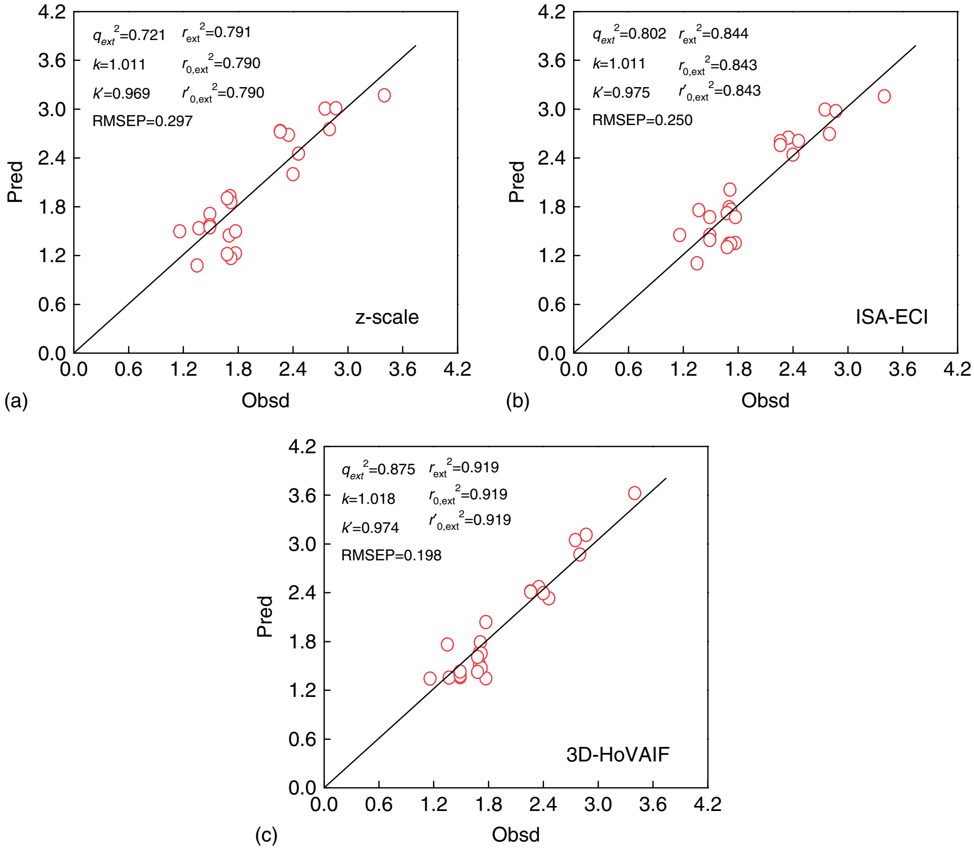

Figure 17 Plot of predicted *vs* observed activities for 24 bitter-tasting dipeptides in test set: (a) z-scale; (b) ISA-ECI; (c) 3D-HoVAIF. This figure is available in colour online at www.interscience.wiley.com/journal/jpepsci.

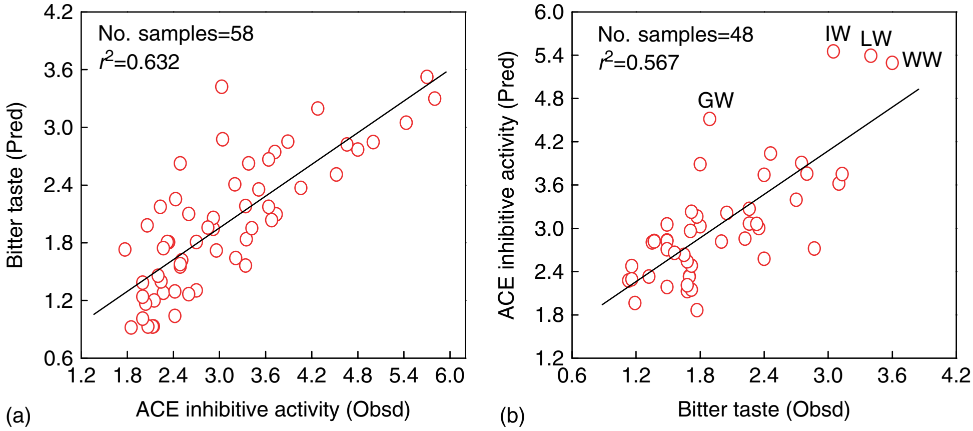

Figure 18 Relation between bitter taste and inhibiting activity for dipeptides in Tables 2 and 3: (a) GA-PLS-model-predicted bitter taste *vs* observed ACE inhibitive activity; (b) GA-PLS-model-predicted ACE-inhibiting activity *versus* the observed bitter taste. This figure is available in colour online at www.interscience.wiley.com/journal/jpepsci.

more fundamental amino acids and oligopeptide segments to be reutilized by organisms as metabolic energy resource and cell construction motifs. Often, oligopeptide compounds transformed by or contained in foods are of certain physiological activities, impacting on human cardiovascular, nervous, immune and nutritional systems [86], and therefore methods to introduce some active peptide analogs into foods are expected to assist in disease prevention and adjunctive therapy. For the reason that antihypertensive ACE inhibitors are often of bitter tastes, seeking oligopeptide sequences high in inhibiting activities but low in the bitter taste becomes an important task for the development of functional foods of comfortable tastes [87]. To this end, 3D-HoVAIF makes an attempt to perform correlation analysis on the bioactivities of two dipeptides (i.e.

ACE-inhibiting and bitter-tasting activities), aiming at providing valuable information for related applications and researches.

Among 58 ACE inhibitors and 48 bitter-tasting dipeptides with known bioactivities, 15 samples are grouped, with $r^2 = 0.573$ (indicating correlation of inhibiting activities with bitter tastes). Figure 18(a) is the plot of the GA-PLS-model-predicted bitter tastes *versus* the observed ACE-inhibiting activities ($r^2 = 0.632$) for the 58 samples in Table 2, suggesting inhibiting activities are positively related with bitter tastes. Figure 18(b) presents the plot of GA-PLS-model-predicted ACE-inhibiting activities *versus* the observed bitter tastes for the 48 samples in Table 3, suggesting that samples of high inhibiting activities, with all the *C*-termini occupied by a bulky tryptophan residue, are simultaneously of strong bitter tastes. It is revealed that among these samples, there is no perfect dipeptide compound of both high inhibiting activity and low bitter taste. Besides, it is also displayed that the GA-PLS-model-predicted correlations (referring to $r^2 = 0.567$ and 0.632) are near the observed one ($r^2 = 0.535$). Figure 19 is the plot of the GA-PLS-model-predicted ACE-inhibiting activities *versus* bitter tastes for all 400 theoretically possible dipeptides, in which three-dimensional steric structures are transformed from two-dimensional structures by CORINA 3.2 and optimized by MM+ force field. In this figure, two areas A and B are marked out to serve as distribution regions of postselecting samples, wherein area A indicating 'tenderness' has all the constituent dipeptides terminated by a histidine (H) residue at *C*-termini, possessing relative weak antihypertensive abilities but comfortable tastes. Contrary to A, area B behaving as 'stimulation', has all its constituent dipeptides terminated by a tryptophan (W) residue at *C*-termini, having good pharmaceutical performance but bad tastes. Besides, a comparison between these two areas suggests large similarities: they are both composed of seven dipeptides and behave similarly with respect to *N*-termini, wherein seven polar/neutral residue types S, Q, G (neutrality), N, D, R and E occur, but differently at the *C*-termini wherein distinctly terminated residues occur. Such a case is not an accident. Via analysis of available molecular structures of ACE inhibitors in marks, the appearance of aromatic series and hydrophobic amino acids at the *C*-termini is confirmed to actually benefit pharmacodynamic enhancements for peptide analog drugs [88]. However, in Figure 15, which presents the GA-PLS loading plot for bitter-tasting dipeptides, it is indicated that the entire peptide hydrophobicities are inclined to positively relate with bitter-tasting intensities. Obviously, only polar residues at the *N*-termini are insufficient to sustain weak hydrophobicity of the whole dipeptides, and thus hydrophobicity of the *C*-termini is confronted with a conflict between high ACE-inhibiting activities and low bitter tastes. For that, we have to take a neutral attitude. For samples in area A, the *C*-termini are occupied by histidine (H) residues, of which the side chain comprises not only π-conjugation electron systems similar to aromatic series but also the polar atom N, thereby resulting in comfortable tastes but low pharmacodynamic activities. For area B, however, slightly polar tryptophan (W) residue is chosen at the *C*-termini, which intensifies bitter tastes, although remarkably enhancing pharmacodynamic actions. So, it is actually difficult to find a perfect dipeptide compound which is expected to serve as an active component of functional foods, confirmed by the conclusion that ACE-inhibiting activities of antihypertension peptide analogs are positively related with bitter-tasting intensities. But what can be speculated is that with peptide sequence becoming longer and structural diversities increasing, the possibilities to find such 'perfect' peptide analogs would be improved.

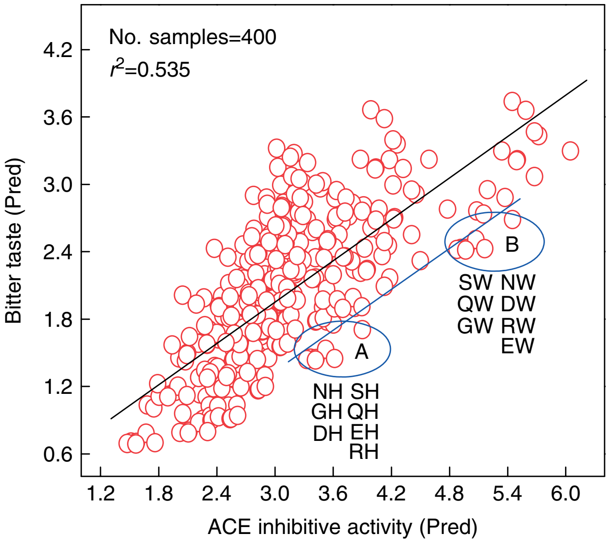

Figure 19 The plot of bitter tastes to GA-PLS-model-predicted ACE-inhibiting activities for 400 theoretically possible dipeptides. This figure is available in colour online at www.interscience.wiley.com/journal/jpepsci.

CONCLUSIONS

By defining 10 common atomic types and their 55 interactions, a novel rotation–translation invariant 3D structure descriptor, 3D-HoVAIF, is derived from calculations of three nonbonding interactions of electrostatic, van der Waals and hydrophobic interactions, which directly impact on drug activities. Such an approach has merits in easy calculation and explicit physical meanings, and is free of experimental parameters, and moreover overcomes the disadvantages inherent in most 3D-QSAR approaches, such as conformation alignment and arbitrary grid division, etc. Besides, typing atoms in terms of families in the periodic table

DOI: 10.1002/psc

and hybridization states in 3D-HoVAIF methods benefits not only amenable physicochemical meanings, but also further extensions into intricate molecular systems rich in hetero atoms, and so this method is promising in future researches. In the present work, the 3D-HoVAIF approach has been utilized to perform systematic QSAR studies on 58 ACE inhibitors and 48 bitter-tasting dipeptides, and by virtue of rigorous internal–external validations, the resulting models are confirmed to be stable and predictable. Moreover, these models are subsequently employed to seek a correlation between inhibiting activities and bitter tastes in more detail for dipeptides. Besides, some available reference reports on these datasets have been collected, aiming to benefit a comparison with the 3D-HoVAIF approach in this paper.

Supplementary Material

Supplementary electronic material for this paper is available in Wiley InterScience at: http://www.interscience.wiley.com/jpages/1075-2617/suppmat/

Parameters involved in 3D-HoVAIF and 3D-HoVAIF descriptors for 58 ACE inhibitors and 48 bitter-tasting dipeptides are provided as supporting materials.

Acknowledgements

The authors would like to acknowledge funding from National Natural Science Foundation of China (grant number 30371339) and National Project 863 Fund (grant number 2006AA02Z312).

REFERENCES

1. Tdeschini R, Consonni V. *Handbook of Molecular Descriptors*. Wiley-VCH: Weinheim, 2000.
2. Wiener H. Correlation of heats of isomerization and differences in heats of vaporiza-tion of isomers, among the paraffin hydrocarbons. *J. Phys. Chem.* 1948; **52**: 2636–2638.
3. Hosoya H. Topological index. A new proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc.* 1971; **44**: 2332–2339.
4. Randic M. On characterization of molecular branching. *J. Am. Chem. Soc.* 1975; **97**: 6609–6615.
5. Balaban AT. High discrimination distance-based topological index. *Chem. Phys. Lett.* 1982; **89**: 399–404.
6. Katritzky AR, Lobanov VS, Karelson M. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* 1995; **24**: 279–287.
7. Katritzky AR, Maran U, Lobanov VS, Karelson M. Structurally diverse quantitative structure-property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.* 2000; **40**: 1–18.
8. Kier LB, Hall LH. An electrotopological state index for atoms in molecules. *Pharm. Res.* 1990; **7**: 801–807.
9. Hall LH, Kier LB. The electrotopological state: structure information at the atomic level for molecular graph. *J. Chem. Inf. Comput. Sci.* 1991; **31**: 76–83.
10. Wild DJ, Blankley CJ. Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. *J. Chem. Inf. Comput. Sci.* 2000; **40**: 155–162.
11. Flower DR. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* 1998; **38**: 379–386.
12. Xue L, Godden JW, Bajorath J. Database searching for compounds with similar biological activities using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.* 1999; **39**: 881–886.
13. González MP, Terán C. A TOPS-MODE approach to predict adenosine kinase inhibition. *Bioorg. Med. Chem. Lett.* 2004; **14**: 3077–3079.
14. González MP, Terán C. A TOPS-MODE approach to predict affinity for A1 adenosine receptors. 2-(Arylamino) adenosine analogues. *Bioorg. Med. Chem.* 2004; **12**: 2985–2993.
15. Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* 1988; **110**: 5959–5967.
16. Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* 1994; **37**: 4130–4146.
17. Doweyko AM. The hypothetical active site lattice. An approach to Modeling active sites from data on inhibitor molecules. *J. Med. Chem.* 1988; **31**: 1396–1406.
18. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important molecules. *J. Med. Chem.* 1985; **28**: 849–857.
19. Jain AN, Dietterich TG, Lathrop RH, Chapman D, Critchlow RE, Webster TA, Lozaoperez T. COMPASS–a shape-based machine learning tool for drug design. *J. Comput. Aided Mol. Des.* 1994; **8**: 635–652.
20. Todeschini R, Lasagni M, Marengo E. New molecular descriptors for 2D and 3D structures. Theory. *J. Chemom.* 1994; **8**: 263–272.
21. Todeschini R, Gramatice P, Provenzani R. Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polycyclic aromatic hydrocarbons. *Chemom. Intell. Lab. Syst.* 1995; **27**: 221–229.
22. Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R, Zaliani A. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D-QSAR study in a series of steroids. *J. Comput. Aided Mol. Des.* 1997; **11**: 79–92.
23. Silverman BD, Platt DE. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.* 1996; **39**: 2129–2140.
24. Ferguson AM, Heritage T, Jonathon P, Pack SE, Phillips L, Rogan J, Snaith PJ. EVA: a new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *J. Comput. Aided Mol. Des.* 1997; **11**: 143–147.
25. Baumann K. Distance profiles (DiP): a translationally and rotationally invariant 3D structure descriptor capturing steric properties of molecules. *Quant. Struct.-Act. Relat.* 2002; **21**: 507–519.
26. Liu S, Cao C, Li Z. Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance edge (MDE) vector λ. *J. Chem. Inf. Comput. Sci.* 1998; **38**: 387–394.
27. Liu S, Cai S, Cao C, Li Z. Molecular electronegative distance vector (MEDV) relating to 15 properties of alkanes. *J. Chem. Inf. Comput. Sci.* 2000; **40**: 1337–1348.
28. Liu S, Liu H, Yu B, Cao C, Li Z. Investigation on quantitative relationship between chemical shift of carbon-13 nuclear magnetic resonance spectra and molecular topological structure based on a novel atomic distance-edge vector (ADEV). *J. Chemom.* 2001; **15**: 427–438.
29. Liao C, Chen Z, Yin Z, Li Z. Preliminary approach to estimation and prediction of infrared spectroscopy for Mannich bases by atomic electronegativity distance vector (VAED). *Comput. Biol. Chem.* 2003; **27**: 229–239.
30. Levitt M. Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.* 1983; **170**: 723–764.

31. Hahn M. Receptor surface models. 1. Definition and construction. *J. Med. Chem.* 1995; **38**: 2080–2090.
32. Kellogg GE, Semus SF, Abraham DJ. HINT–a new method of empirical hydrophobic field calculation for CoMFA. *J. Comput. Aided Mol. Des.* 1991; **5**: 545–552.
33. Hasel W, Hendrikson TF, Still WC. A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahed. Comp. Method.* 1988; **1**: 103–116.
34. Pei J, Wang Q, Zhou J, Lai L. Estimating protein-ligand binding free energy: atomic solvation parameters for partition coefficient and solvation free energy calculation. *Proteins* 2004; **57**: 651–664.
35. Wold S, Ruhe A, Wold H, Dunn WJ. The collinearity problem in linear regression–the partial least squares (PLS) approach to generalized inverses. *Siam J. Sci. Stat. Comput.* 1984; **5**: 735–743.
36. Hoskuldsson P. PLS regression methods. *J. Chemom.* 1988; **2**: 211–228.
37. Wold S, Sjöström M, Eriksson L. PLS regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* 2001; **58**: 109–130.
38. Leonard JT, Roy K. On selection of training and test sets for the development of predictive QSAR models. *QSAR Comb. Sci.* 2006; **25**: 235–251.
39. SYantai-Kis C, Kövesdi I, Kéri G, Orfi L. Validation subset selections for extrapolation oriented QSPAR models. *Mol. Divers.* 2003; **7**: 37–43.
40. Gramatica P, Pilutti P, Papa E. Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus Modeling. *J. Chem. Inf. Comput. Sci.* 2004; **44**: 1794–1802.
41. Golbraikh A, Tropsha A. Beware of q2!. *J. Mol. Graphics Modell.* 2002; **20**: 269–276.
42. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* 2003; **22**: 69–77.
43. Vincent M, Marchand B, Remond G, Jaguelin-Guinamant S, Damien G, Portevin B, Baumal JY, Volland J, Bouchet J, Lambert P, Serkiz B, Luitjen W, Lauibie M, Schiavi P. Synthesis and ACE inhibitory activity of the stereoisomers of perindopril (S9490) and perindoprilate (S9780). *Drug Des. Discov.* 1992; **9**: 11–28.
44. Cushman DW, Cheung HS, Sabo EF, Ondetti MA. Angiotensin-converting enzyme inhibitors: evolution of a new class of antihypertensive drugs. In *Angiotensin-Converting Enzyme Inhibitors: Mechanisms of Action and Clinical Implications*, Horovitz ZP (ed.). Urban and Schwarzenberg: Baltimore, 1981; 3–25.
45. Hellberg S, Eriksson L, Jonsson J, Lindgren F, Sjöström M, Skagerberg B, Wold S, Andrews P. Minimum analogue peptide sets (MAPS) for quantitative structure-activity relationships. *Int. J. Pept. Protein Res.* 1991; **37**: 414–424.
46. Cocchi M, Johansson E. Amino acids characterization by GRID and multivariate data analysis. *Quant. Struct.-Act. Relat.* 1993; **12**: 1–8.
47. Collantes ER, Dunn WJ. Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogues. *J. Med. Chem.* 1995; **38**: 2705–2713.
48. Zaliani A, Gancia E. MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* 1999; **39**: 525–533.
49. Li Z, Fu B, Wang Y, Liu S. On structural parameterization and molecular modeling of peptide analogues by molecular electronegativity edge vector (VMEE): estimation and prediction for biological activity of pentapeptides. *J. Chin. Chem. Soc.* 2001; **48**: 937–944.
50. Liu S, Yin C, Wang L. Combined MEDV-GA-MLR method for QSAR of three panels of steroids, dipeptides, and COX-2 inhibitors. *J. Chem. Inf. Comput. Sci.* 2002; **42**: 749–756.
51. Liu S, Yin C, Cai S, Li Z. A novel MHDV descriptor for dipeptide QSAR studies. *J. Chin. Chem. Soc.* 2001; **48**: 253–260.
52. Liu S, Yin C, Cai S, Li Z. QSAR study of steroid benchmark and dipeptides based on MEDV-13. *J. Chem. Inf. Comput. Sci.* 2001; **41**: 321–329.
53. Mei H, Liao Z, Zhou Y, Li Z. A new set of amino acid descriptors and its application in peptide QSARs. *Biopolym. Pept. Sci.* 2005; **80**: 775–786.
54. Zhou P, Zhou Y, Wu S, Li B, Tian F, Li Z. A new descriptor of amino acids based on the atomic interaction field. *Chin. Sci. Bull.* 2006; **51**: 524–529.
55. Tian F, Zhou P, Li Z. T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. *J. Mol. Struct.* 2007; **830**: 106–115.
56. Zhou P, Tian F, Li B, Wu S, Li Z. Genetic algorithm-based virtual screening of combinative mode for peptide/protein. *Acta Chim. Sinica* 2006; **64**: 691–697.
57. Adler E, Hoon MA, Mueller KL, Chandrashekar J, Ryba NJP, Zuker CS. A novel family of mammalian taste receptors. *Cell* 2000; **100**: 693–702.
58. Matsunami H, Montmayeur J, Buck LB. A family of candidate taste receptors in human and mouse. *Nature* 2000; **404**: 601–604.
59. Asao M, Iwamara H, Akamatsu M, Fujita T. Quantitative structure-activity relationships of bitter thresholds of amino acids, peptides and their derivatives. *J. Med. Chem.* 1987; **30**: 1873–1879.
60. Jonsson J, Eriksson L, Hellberg S, Sjostroem M, Wold S. Multivariate parametrization of 55 coded and non-coded amino acids. *Quant. Struct.-Act. Relat.* 1989; **8**: 203–209.
61. de Armas RR, Díaz HG, Molin R, González MP, Uriarte E. Stochastic-based descriptors studying peptides biological properties: modeling the bitter tasting threshold of dipeptides. *Bioorg. Med. Chem.* 2004; **12**: 4815–4822.
62. Cushman DW, Ondetti MA. Design of potent competitive inhibition of angiotensin-converting enzyme. *Biochemistry* 1977; **16**: 5484–5488.
63. Natesh R, Schwager SL, Evans HR, Sturrock ED, Acharya KR. Structural details on the binding of antihypertensive drugs captopril andenalaprilat to human testicular angiotensin I-converting enzyme. *Biochemistry* 2004; **43**: 8718–8724.
64. Hypercube Inc., HyperChem 7.5, 2004; www.hyper.com.
65. Fujitsu Inc., BioMedCAChe 6.1, 2004; http:///www.cachesoftware.com.
66. Stewart JJP. MOPAC: A semiepirical molecular orbital program. *J. Comput.-Aided Mol. Des.* 1990; **4**: 1–105.
67. Umetrics Inc., SIMCA-P 10.0, 2002; http://www.umetrics.com.
68. Leardi R, González AL. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemom. Intell. Lab. Syst.* 1998; **41**: 195–207.
69. Houck CR, Joines J, Kay M. A genetic algorithm for function optimization: a matlab implementation. *ACM Trans. Math. Softw.* 1996; **22**: 1–14.
70. Eigenvector Research Inc., PLS Toolbox 3.0, 2003; http://www.eigenvector.com.
71. MathWorks Inc., Matlab 6.1, 2001; http://www.mathworks.com.
72. Wold S, Eriksson L. Statistical validation of QSAR results. In *Chemometrics Methods in Molecular Design*. van de Waterbeemd H (ed.). Wiley-VCH: Weinheim, 1995; 309–318.
73. Hellberg S, Sjostrom M, Skagerberg B, Wold S. Peptide quantitative structure-activity relationships, a multivariate approach. *J. Med. Chem.* 1987; **30**: 1126–1135.
74. Andersson PM, Sjöström M, Lundstedt T. Preprocessing peptide sequences for multivariate sequence-property analysis. *Chemom. Intell. Lab. Syst.* 1998; **42**: 41–50.
75. Lejon T, Strøm MB, Svendsen JS. Antibiotic activity of pentadecapeptides modelled from amino acid descriptors. *J. Pept. Sci.* 2001; **7**: 74–81.
76. Jenssen H, Gutteberg TJ, Lejon T. Modeling of anti-HSV activity of lactoferricin analogues using amino acid descriptors. *J. Pept. Sci.* 2005; **11**: 97–103.

77. Wu J, Aluko RE, Nakai S. Structural requirements of angiotensin I-converting enzyme inhibitory peptides: quantitative structure-activity relationship modeling of peptides containing 4-10 amino acid residues. *QSAR Comb. Sci.* 2006; **25**: 873–880.
78. Genst ED, Areskoug D, Decanniere K, Muyldermans S, Andersson K. Kinetic and affinity predictions of a protein-protein interaction using multivariate experimental design. *J. Biol. Chem.* 2002; **277**: 29897–29907.
79. Guan P, Doytchinova IA, Walshe VA, Borrow P, Flower DR. Analysis of peptide-protein binding using amino acid descriptors: prediction and experimental verification for human histocompatibility complex HLA-A*0201. *J. Med. Chem.* 2005; **48**: 7418–7425.
80. Ponce YM, Marrero RM, Castro EA, de Armas RR, Díaz HG, Zaldivar VR, Torrens F. Protein quadratic indices of the "macromolecular pseudograph's α-carbon atom adjacency matrix". 1. Prediction of arc repressor alanine-mutant's stability. *Molecules* 2004; **9**: 1124–1147.
81. Cho SJ, Zheng W, Tropsha A. Rational combinatorial library design. 2. rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches. *J. Chem. Inf. Comput. Sci.* 1998; **38**: 259–268.
82. Lin Z, Wu Y, Zhu B, Ni B, Wang L. Toward the quantitative prediction of T-Cell epitopes: QSAR studies on peptides having affinity with the class I MHC molecular HLA-A*0201. *J. Comput. Biol.* 2004; **11**: 683–694.
83. de Armas RR, Díaz HG, Molina R, Uriarte E. Stochastic-based descriptors studying biopolymers biological properties: extended MARCH-INSIDE methodology describing antibacterial activity of lactoferricin derivatives. *Biopolymers* 2005; **77**: 247–256.
84. Baroni M, Clement S, Cruciani G, Kettaneh-Wold S, Wold S. D-optimal designs in QSAR. *Quant. Struct.-Act. Relat.* 1993; **12**: 225–231.
85. de Aguiar PF, Bourguignon B, Khots MS, Massart DL, Phan-Than-Luu R. D-optimal designs. *Chemom. Intell. Lab. Syst.* 1995; **30**: 199–210.
86. Silva SS, Malcata FX. Caseins as source of bioactive peptides. *Int. Dairy J.* 2005; **15**: 1–15.
87. Pripp AH, Ardö Y. Modeling relationship between angiotensin-(I)-converting enzyme inhibition and the bitter taste of peptides. *Food Chemistry* 2007; **102**: 880–888.
88. Cheung HS, Wang FL, Ondetti MA, Sabo EF, Cushman DW. Binding of peptide substrates and inhibitors of angiotensin-coverting enzyme. *J. Biol. Chem.* 1980; **255**: 401–407.

DOI: 10.1002/psc