Amino Acids (2009) 37:583–591
DOI 10.1007/s00726-008-0177-8

ORIGINAL ARTICLE

Using multidimensional patterns of amino acid attributes for QSAR analysis of peptides

G. Liang · L. Yang · L. Kang · H. Mei · Z. Li

Received: 3 April 2008 / Accepted: 25 August 2008 / Published online: 28 September 2008

Abstract On the basis of exploratory factor analysis, six multidimensional patterns of 516 amino acid attributes, namely, factor analysis scales of generalized amino acid information (FASGAI) involving hydrophobicity, alpha and turn propensities, bulky properties, compositional characteristics, local flexibility and electronic properties, are proposed to represent structures of 48 bitter-tasting dipeptides and 58 angiotensin-converting enzyme inhibitors. Characteristic parameters related to bioactivities of the peptides studied are selected by genetic algorithm, and quantitative structure–activity relationship (QSAR) models are constructed by partial least square (PLS). Our results by a leave-one-out cross validation are compared with the previously known structure representation method and are shown to give slightly superior or comparative performance. Further, two data sets are divided into training sets and test sets to validate the characterization repertoire of FASGAI. Performance of the PLS models developed by training samples by a leave-one-out cross validation and external validation for test samples are satisfying. These results demonstrate that FASGAI is an effective representation technique of peptide structures, and that FASGAI vectors have many preponderant characteristics such as straightforward physicochemical information, high characterization competence and easy manipulation. They can be further applied to investigate the relationship between structures and functions of various peptides, even proteins.

Keywords Peptide · Factor analysis scales of generalized amino acid information · Quantitative structure–activity relationship · Partial least squares · Genetic algorithm-partial least square

Abbreviations

FASGAI	Factor analysis scales of generalized amino acid information
QSAR	Quantitative structure–activity relationship
PLS	Partial least squares
GA-PLS	Genetic algorithm-partial least square
BTD	Bitter-tasting dipeptide
ACE	Angiotensin-converting enzyme

Introduction

Known as critical elements in life science, peptides have attracted considerable pharmacological and medicinal interest in recent years (Sewald and Jakubke 2002). With the development of peptide library, thousands of peptides have been designed and synthesized. Then, a quantitative structure–activity relationship (QSAR) (Gonzalez-Díaz et al. 2007, 2008; Selassie et al. 2002) model provides a practical tool for the analysis of biological data. The main idea of QSAR is that structural features of biomolecules can be correlated with biological activities; i.e., biological activities can be modelled as functions of molecular structures. QSAR techniques can be 2D or 3D. The former uses physicochemical descriptors, while the latter also

Electronic supplementary material The online version of this article (doi:10.1007/s00726-008-0177-8) contains supplementary material, which is available to authorized users.

G. Liang (✉) · L. Yang · L. Kang · H. Mei
College of Bioengineering, Chongqing University, 400030 Chongqing, China
e-mail: gzliang@cqu.edu.cn

Z. Li
College of Chemistry and Chemical Engineering, Chongqing University, Chongqing, China

takes the spatial features of the molecules into account (Fauchere et al. 1988; Collantes and Dunn III 1995; Norinder 1991). Statistical methods, such as partial least squares (PLS), are used in QSAR fields to produce models that relate changes in activities to molecular properties (Felipe-Sotelo et al. 2003; Hasegawa and Funatsu 2000).

In the development of peptide QSARs, numerous methods have been used to describe the variation in activity as a function of structure. On the basis of predictive ability, the best models are probably those derived from parameters describing macroscopic properties of the peptides, e.g. hydrophobicity, charge and alpha-helicity, since these properties reflect the actual behaviour of the peptide. The drawback of such models is that new improved peptides will be described by the same input descriptors used in the modelling. This will again lead to difficulties in translation of these properties into amino acid sequences, and thereby construction of novel peptides (Lejon et al. 2002). One way of overcoming this problem is to use amino acid descriptors (Hellberg et al. 1987) rather than peptide descriptors. Since some amino acid descriptors were used to construct some quantitative sequence–activity modelling of oxytocin–vasopressin analogues by Sneath (1966), a number of quantitative amino acid descriptors have been put forward in the past few years (Collantes and Dunn III 1995; Kidera et al. 1985; Hellberg et al. 1987; Sandberg et al. 1998). Particularly, a recent development in QSAR field of peptides was the use of amino acid "*z*-scales" which were scales of hydrophilicity, bulk and electronic properties by principal component analysis relying on 29 physicochemical variables of the 20 coded amino acids (Hellberg et al. 1987). Subsequently, the *z*-scales were revised by adding new data to the multi-property matrix (Kawashima and Kanehisa 2000) and eventually were extended to the non-coded amino acids (Tomii and Kanehisa 1996). The *z*-scales have been proved to be powerful for describing small peptide structural characteristics related to their activities with good results obtained. Collantes and Dunn III (1995) established 3D-QSAR models on the base of 3D structural characters of amino acid side chains, i.e., isotropic surface (ISA) together with electronic charge index (ECI) for polypeptides. Recently, VHSE (principal components score vectors of hydrophobic, steric, and electronic properties) was derived from total 50 physicochemical variables including 18 hydrophobic properties, 17 steric properties, and 15 electronic properties of the 20 coded amino acids using principal component analysis by Mei et al. (2005). Then, a comparison of the results to those obtained with *z*-scales and other 2D or 3D descriptors showed that the scales are comparable for parameterizing the structural variability of some oligopeptides. The quality of amino acid descriptors is an important factor in producing QSAR models with good predictive ability (Hunt 1999). An excellent descriptor should can reasonably characterize structural characteristics of peptides and extract important structural information related to functions of peptides. Most importantly, the information extracted can give some insight into relationship between structures and functions of peptides studied.

It should be mentioned that a number of reports (Agüero-Chapín et al. 2008; González-Díaz and Uriarte 2005; Gonzalez-Díaz et al. 2005, 2006, 2007a, b, c, d, 2008) have constructed some Markov chain scales and descriptors of peptides and proteins using various parameters of amino acids and develop QSAR models. However, these scales and descriptors are complicated to a certain extent. In this paper, six multidimensional patterns of amino acid attributes, namely, factor analysis scales of generalized amino acid information (FASGAI), reflecting hydrophobicity, alpha and turn propensities, bulky properties, compositional characteristics, local flexibility and electronic properties are obtained from multidimensional properties of the 20-coded amino acids using multivariate statistical analysis. Structures of two classical dipeptide datasets including 48 bitter-tasting dipeptides (BTDs) and 58 angiotensin-converting enzyme inhibitors are represented by FASGAI vectors, and QSAR analysis is carried out to investigate the relationship between their structures and functions. These results indicate that FASGAI is an effective characterization method of peptide structures, and it cherishes straightforward physicochemical significance and is easy to be operated.

Materials and methods

Factor analysis scales of generalized amino acid information

Functions and structures of peptides or proteins are determined by the information contained in the amino acid sequence (Anfinsen 1973). Hence, 335 significative properties (Table S1 in Supplementary material) of the 20 coded amino acids, representing alpha and turn propensities, beta propensity, hydrophobicity properties, physicochemical properties, composition properties and so on, were culled from the AAindex database (Kawashima and Kanehisa 2000; Tomii and Kanehisa 1996; Nakai et al. 1988) based on relative loading coefficients and communalities of the variables from initial factor analysis, along with relative ease of interpretation and perceived structural importance.

Exploratory factor analysis (Johnson and Wichern 2002), as a powerful statistical procedure, was used to produce a subset of numerical variables that would summarize the entire constellation of all 335 properties. Factor analysis simplifies high-dimensional data by generating a smaller number of ''factors'' that describe the structure of

highly correlated variables (Johnson and Wichern 2002). The resultant factors are linear functions of the original data, fewer in number than the original, and reflect clusters of covarying variables that describe the underlying or ''latent structure'' of the variables. Factor analysis models assume that observation i denoted as $x_i \in \mathbf{R}^p$ can be decomposed into $x_i = \wedge f_i + u_i$, where $\wedge$: $\mathbf{R}^k \rightarrow \mathbf{R}^p$ is linear, and $f_i \sim N_k(0, I_k)$, $u_i \sim N_p(0, \psi)$, where ψ is diagonal, all f_i and u_j are independent, and $k < p$. The new set of inferred variables f_i are called common or latent factors, whereas u_j are called unique factors. Factor analysis differs from principal components analysis in that the latter does not distinguish between common and unique variance; with principal components, all $u_i = 0$.

A number of different methods, including maximum likelihood, principal component, principal axis extraction, etc., can be used to extract factors (Field 2005). Here, principal component method was used to extract and obtain several clusters of highly intercorrelated physiochemical variables. The factor coefficient contained in the $\wedge$ matrix is a regression coefficient quantifying the relationship between the variable and the common factor. The interpretability of factors is improved through rotation. Such rotations maximize the loading of each variable on the extracted factors while minimizing the loading on all other. There are two major categories of rotations, orthogonal rotations, which produce uncorrelated factors, and oblique rotations, which produce correlated factors (Field 2005). There was certain correlation between diversified properties of proteins or peptides; consequently, in this paper, promax algorithm with Kaiser normalization, as an oblique solution, was used to rotate the factors to simple structure to improve their interpretation.

Interpretation on loadings (Tables S1 and S2 in Supplementary material) demonstrates that the first six factors possess straightforward and physiochemical information, reflecting hydrophobicity, alpha and turn propensities, bulky properties, composition characteristics, local flexibility and electronic properties, respectively. The fifth and the sixth factors were still considered because they obviously represent physiochemical information, although they explained relatively little variance. Six factors account for an 83.5% variance of these 335 variables according to relationship between component number and eigenvalues (Table S3 in Supplementary material). Communality values (Table S1 in Supplementary material), the sum of the squared factor coefficients for each property, reflect the portion of the common variation in a variable by six factors. Communality value for each variable is all larger than 0.500; especially, many of the properties express high communality values (>0.9), suggesting that they have high factor coefficients on at least one factor and that the six-factor model is sufficient. Factor analysis produces a new set of synthetic traits called factor scores (Table S3 in Supplementary material) that are linear combinations of the original variables. Here, these six-factor score vectors are tentatively called factor analysis scales of generalized amino acid information (FASGAI). FASGAI vectors, summarize most information of the 335 properties, so they can be utilized to represent structural features of peptides or proteins. Each residue in a sequence is described by six FASGAI vectors according to varied amino acid position. Accordingly, the structural characteristics of a sequence with n amino acid residues are represented by the concatenation of $6 \times n$ FASGAI vectors. Illustration of the general workflow of the analysis with FASGAI descriptors is displayed in Fig. 1.

Feature selection

Here, variable selection is completed by genetic arithmetic-partial least square (GA-PLS) as an effective variable selection tool nowadays, which is a sophisticated hybrid approach that combines GA as a powerful optimization method with PLS as a robust statistical method for variables selection (Hasegawa et al. 1997; Hasegawa and Funatsu 1998). In GA-PLS, the chromosome and its fitness in the species correspond to a set of variables and internal prediction ability of the derived PLS model, respectively. The fitness of each chromosome is evaluated by internal prediction ability of the PLS model derived from a binary bit pattern. The internal predictive performance of the model is expressed in terms of a cross validation (CV)

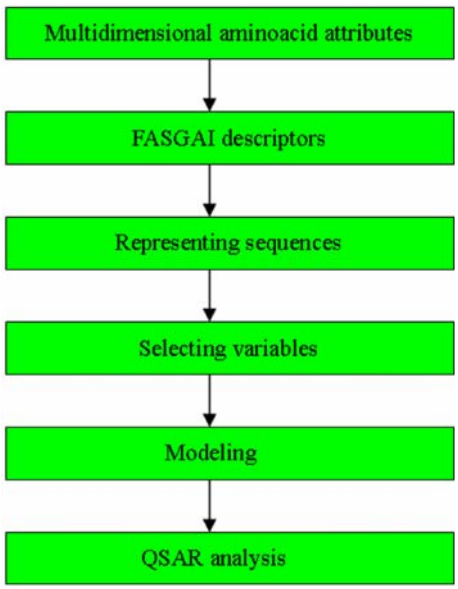

Fig. 1 The general workflow of the analysis with FASGAI descriptors

square of cumulative multiple correlation coefficient (R^2_{cum}) value (hereafter, denoted by Q^2_{cv}) by the leave-one-out (LOO) procedure as follows:

$$Q^2_{cv} = 1 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Big/ \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad (1)$$

where y_i and $\hat{y}_i$ represent the observed value and the predicted value of the dependent variable, respectively; $\bar{y}$ indicates the mean observed value of the dependent variable; n is the number of samples.

Partial least-squares modelling

The PLS also has the desirable property that the precision of the model parameters is improved with the increasing number of relevant variables and observations (Gramatica et al. 2004; Wold et al. 2001a, b; Helland 2001). The PLS regression algorithm consists of outer relations (X and Y block individually) and an inner relation linking both blocks:

$$x_{ik} = \sum_{a=1}^{A} t_{ia} p_{ak} + e_{ik} \quad (2)$$

$$x_{im} = \sum_{a=1}^{A} u_{ia} c_{am} + g_{im} \quad (3)$$

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

$$\hat{u} = bt \quad (4)$$

Model validation

In the present work, a LOO CV for internal validation criteria is used. Predictive performance of the model is assessed by the prediction values of Q^2_{cv} (Eq. 1). External validation can only be achieved by splitting the total data set into a training set for establishing a model and a test set for evaluating the performance of the model obtained. The external prediction power of a model can be evaluated by an external CV correlation coefficient (hereafter, denoted by Q^2_{ext}) as follows (Tropsha et al. 2003):

$$Q^2_{ext} = 1 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Big/ \sum_{i=1}^{n} (y_i - \bar{y}_{tr})^2 \quad (5)$$

where y_i and $\hat{y}_i$ are the observed and predicted values over the test set of the dependent variable, respectively; $\bar{y}_{tr}$ is the mean value of the dependent variable for the training set; n is the number of test set.

Software used

Factor analysis was implemented using the SPSS 13.0 statistical software. SIMCA-P (Version 11.0, Umetrics 2004) software was employed to perform the PLS analysis. GA-PLS programme was written in M-file based on software Matlab (Version 6.1.0.450 release 12.1. MathWorks, Natick 2001).

Results and discussion

QSAR study on bitter-tasting dipeptides

Taste plays very important roles in human and other organisms, and often the gustatory sensitivities are typed for four classes including sweet, bitter, salty and acid. Amongst, bitter sensitivities protect humans and organisms from injury by toxic substances (Cocchi and Johansson 1993). As a classical sample set in QSAR studies, 48 BTDs (Table S4 in Supplementary material) have their activities expressed as negative logarithm of concentration (pT) (Collantes and Dunn III 1995). Accordingly, each residue in a dipeptide was described by six FASGAI vectors according to varied amino acid position. Each dipeptide was then characterized by the concatenation of 12 vectors.

The values of empirical parameters influencing the performance of GA-PLS feature selection were determined by the experience from the series of GA-PLS studies. As a result, parameters were set as follows: the number of populations was 200, the maximum number of generations was 100, the generation gap was 0.8, the crossover frequency was 0.5, the mutation rate was 0.005, and the fitting function was Q^2_{cv}. Finally, an optimal model containing nine variables was obtained from ten trained models. An 84.8% variable dispersion was cumulatively explained by three principal components through the LOO CV procedure, and root mean square error of modelling simulation (RMS_{ms}) of 0.198 was obtained.

The PLS model score of BTDs (Fig. 2) shows that their high-dimensional property of the independent variables may be similar to each other when the two samples relatively approach. It can be seen that other sample dots locate in *Hotelling* T^2 ellipse except the 42nd sample, which

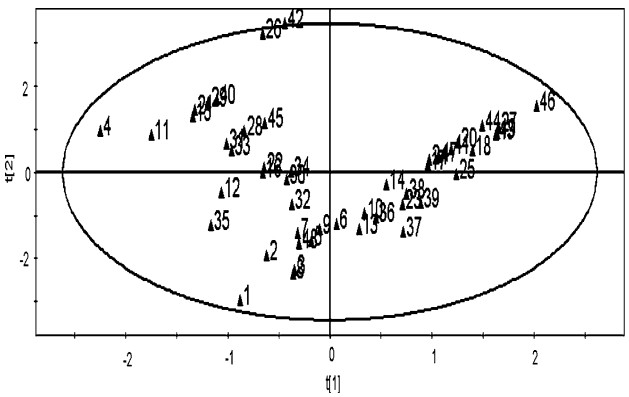

Fig. 2 Plot of the GA-PLS model score of BTDs

demonstrates that the high-dimensional properties of the independent variables of this sample may obviously be different from those of other samples. The PLS model was further validated by response permutations. The 20-random-permutation validation of the PLS model of BTDs (Fig. 3) indicates that intercepts of R^2_{cum} and Q^2_{cv} are 0.087 and −0.379, respectively. For a model to be valid, the desirable intercept limits should be $R^2_{\text{cum}} < 0.300$ and $Q^2_{\text{cv}} < 0.050$ (Andersson et al. 1998); hence, it is considered that relatively large R^2_{cum} and Q^2_{cv} are not resulted from accidental factors.

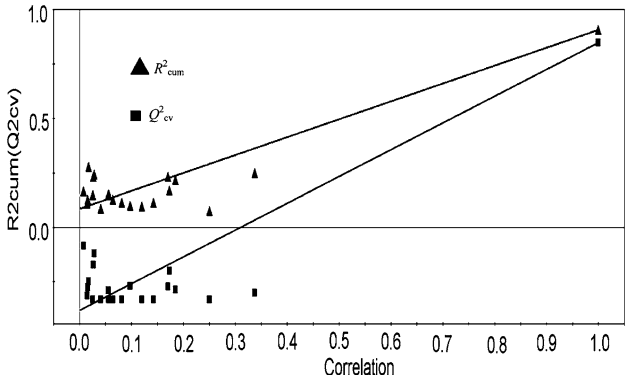

Fig. 3 Plot of the 20-random-permutation validation for the GA-PLS model of BTDs

Performance comparison among QSAR models of BTDs are summarized in Table 1. It can be seen that these models of present study achieve relatively good results. When comparing with various modelling methods listed in Table 1, we find that the model based on the procedure involving FASGAI representation, GA-PLS selection variable and PLS modelling (FASGAI-GA-PLS) produce the highest internal predictive capability, and these results based on the FASGAI-GA-PLS procedure are significantly superior to those of *z*-scales, ISA-ECI and MS-WHIM methods. Ramos de Armas et al. (2004) developed 12 QSAR models based on 12 different types of molecular descriptors and a multiple linear regression analysis. The first three upmost modelling results (Nos. 7, 8 and 9) are also summarized in Table 1. These comparative results display the present methodology is distinctly effective on QSAR study of BTDs.

Also, we extract the normalized k ($k = 21$) order central moment features of FASGAI vectors of 48 peptides to develop a QSAR model using multiple linear regression (MLR). The statistical parameters (Table 1) displayed that the MLR model possessed high simulative ability (R^2_{cum})

Table 1 Performance comparison among QSAR models of BTDs

No.	Descriptors	Data size	Correlation methods	A[a]	R^2_{cum}[b]	Q^2_{cv}[c]	$\text{RMS}^{\text{d}}_{\text{sm}}$
1	*z*-Scales (Hellberg et al. 1991)	48	PLS	2	0.824	nd	0.260
2	GRID (Cocchi and Johansson 1993)	48	PLS	1	nd[e]	0.780	nd
3	ISA-ECI (Collantes and Dunn 1995)	48	PLS	2	0.847	nd	nd
4	MS-WHIM (rotameric) (Zaliani and Gancia 1999)	48	PLS	3	0.704	0.633	nd
5	MS-WHIM (extended) (Zaliani and Gancia 1999)	48	PLS	3	0.754	0.710	0.320
6	VHSE (Mei et al. 2005)	48	PLS	3	0.910	0.816	0.200
7	MARCH-INSIDE (Ramos de Armas et al. 2004)	48	PLS	3	0.858	nd	0.230
8	Topological (Ramos de Armas et al. 2004)[f]	48	MLR	4[i]	0.910	0.890	nd
9	Geometrical (Ramos de Armas et al. 2004)[f]	48	MLR	3[i]	0.909	0.895	nd
10	3D-Morse (Ramos de Armas et al. 2004)[f]	48	MLR	5[i]	0.914	0.880	nd
11	Normalized central moment features of FASGAI vectors	48	MLR	12[i]	0.923	0.701	0.183
12	FASGAI	48	PLS	3	0.886	0.723	0.220
13	FASGAI	48	GA-PLS	3	0.907	0.848	0.198
14[g]	FASGAI	24/24[h]	GA-PLS	2	0.936	0.761	0.172

[a] A is the number of principal component

[b] R^2_{cum} is cumulative multiple correlation coefficient

[c] Q^2_{cv} is a cross validation square of cumulative multiple correlation coefficient by the leave-one-out procedure

[d] RMS_{sm} is root mean square error of modelling simulation (RMS_{ms})

[e] nd shows that the correlative value is not given out

[f] The descriptors were generated with the DRAGON software (Todeschini et al. 2002)

[g] External cross validation correlation coefficient (Q^2_{ext}) and root mean square error of external validation (RMS_{ext}) are 0.797 and 0.273, respectively

[h] Two numbers separated by slashes denote the number of samples in training and test sets, respectively

[i] Number of variables

and low predictive ability (Q^2_{cv}) relative to that of the PLS model, which demonstrated that PLS is necessary for modelling of BTDs QSAR.

The description, coefficient and variable importance in the projection (VIP) of the variables for the GA-PLS model of BTDs are displayed in Table S5 in Supplementary material. VIP is the sum of the variable influence over all model dimensions, and is a measure of variable importance. Higher VIP values indicate good correlation between the variable and the data. It can be seen that corresponding VIPs of five variables, including bulky properties of the second and the first residue, hydrophobicity, electronic properties, and alpha and turn propensities of the second residue, are larger than those of the other ones. Corresponding coefficients of these variables show that bioactivities of BTDs may be markedly positive to bulky properties of the first residue, bulky properties and hydrophobicity of the second residue, and that may be markedly negative to alpha-helix and turn propensities of the second residue. Preferred amino acids at positions that are important for bioactivities of BTDs are gathered in Table 2. Peptides with high bioactivities can be obtained by the alteration of these important amino acid residues.

To further validate the characterization repertoire of FASGAI applied in QSAR analysis of BTDs, we used *k*-means cluster analysis (Kowalski and Wold 1982; Molina et al. 2004) to divide 48 BTDs into two groups. Afterwards, each group samples were sorted from low activity to high activity. Then the first, the third, the fifth sample, etc., in each group were chosen to unite into training set, and the second, the fourth, the sixth sample, etc., in each group into test set. Finally, 24 training samples were applied to construct a QSAR model and 24 test samples were applied to validate the external prediction power of the model developed. Grouped samples and predicted activities of BTDs are gathered in Table S4 in Supplementary material. Internal and external validation results (Table 1) give $Q^2_{cv} = 0.761$, $RMS_{sm} = 0.172$, $Q^2_{ext} = 0.797$ and $RMS_{ext} = 0.273$.

Table 2 Preferred amino acids at positions which are important for activities of BTDs

Amino acid positions	Properties	Preferred amino acids	Contribution to peptide activity
The first residue	Bulky properties	W, Y, F	Positive
The second residue	Bulky properties	W, Y, F	Positive
The second residue	Hydrophobicity	I, V, F, L	Positive
The second residue	Alpha and turn propensities	E, M, L	Negative

QSAR study on angiotensin-converting enzyme inhibitors

Rennin–angiotensin system plays important roles in blood pressure regulation in human bodies. Angiotensinogen, produced by the 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 (Crackower et al. 2002). A new set of dipeptide sequences of 58 ACE inhibitors (Hellberg et al. 1991) have often been utilized to test the effectiveness of diverse kinds of amino acid descriptors (Collantes and Dunn III 1995; Mei et al. 2005; Cocchi and Johansson 1993; Hellberg et al. 1991; Zaliani and Gancia 1999; Li et al. 2001; Mei et al. 2004). The log values of $1/IC_{50}$ (pIC_{50}) are used in QSAR correlations, as they are related to changes in activities of ACE inhibitors. Each dipeptide (Table S6 in Supplementary material) was represented by 12 FASGAI variables.

Determined parameters of the GA-PLS operation were set as follows: the number of populations, the maximum number of generations, the generation gap, the crossover frequency, the mutation rate and the fitting function were 200, 100, 0.8, 0.5, 0.005 and Q^2_{cv}, respectively. Ten models were trained to select an optimal model containing nine variables. The model used one principal component to capture 79.6% of variance, and variance cumulatively explained by the LOO CV procedure was 77.5% and RMS_{ms} was 0.456.

Dimensional distributing characteristics of 58 samples in the principal component space are depicted in Fig. 4. It can be seen that the first and the second samples are outliers according to two multiple standard deviations. It was worthy of noticing that the samples with high activities

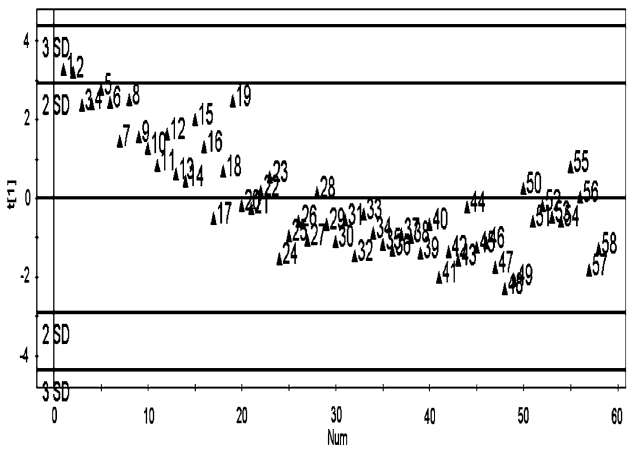

Fig. 4 Plot of the GA-PLS model score of ACE inhibitors

distribute the above area of score plot, while the samples with low activities locate in the below area of score plot. Figure 5 shows that the performance of Y random permutations through 20 times validations. It can be concluded that relatively high R^2_{cum} and Q^2_{cv} value is not caused by accidental factors according to intercepts of $R^2_{cum} = 0.024$ and $Q^2_{cv} = -0.175$.

As such, the normalized central moment features of FASGAI vectors were used as input descriptors to develop a QSAR model using MLR. Performance comparison among QSAR models of ACE inhibitors (Table 3) indicates that the performance by the LOO CV procedure in the present study shows significant superiority to that of other modelling methods listed.

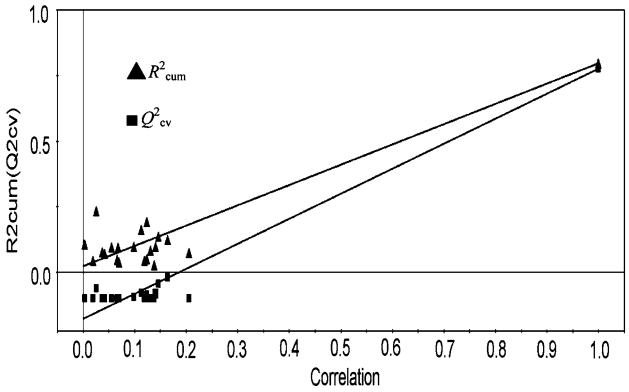

Fig. 5 Plot of the 20-random-permutation validation for the GA-PLS model of ACE inhibitors

The description, coefficient and VIP of the variables for the GA-PLS model of ACE inhibitors are summarized in Table S7 in Supplementary material. The VIP values corresponding to these variables including bulky properties, compositional characteristics and hydrophobicity of the second residue, and electronic properties of the first residue are higher than 1.000. Corresponding coefficients show that the improvements of these property parameters which involve bulky properties and hydrophobicity of the second residue may be conducive to enhancing bioactivities of ACE inhibitors, and that there is an inverse relationship between electronic properties of the first residue and bioactivities of ACE inhibitors. Table 4 presents preferred residues which produce relatively great impact on activities of ACE inhibitors. We can alter the corresponding residues to acquire ACE inhibitors with high activities according to corresponding characteristics of these amino acids listed.

We used the same method that is used to group BTD samples to divide 58 ACE inhibitors into 29 training samples and 29 test samples (Table S6 in Supplementary material) to further validate the characterization repertoire of FASGAI vectors. Twenty-nine test samples were applied to validate the external prediction power of the model developed by 29 training samples. Through general internal and external validation, $Q^2_{cv} = 0.835$, $RMS_{sm} = 0.357$, $Q^2_{ext} = 0.706$ and $RMS_{ext} = 0.558$ were obtained, respectively (Table 3).

Table 3 Performance comparison between QSAR models of ACE inhibitors

No.	Descriptors	Data size	Correlation methods	A[a]	R^2_{cum}[b]	Q^2_{cv}[c]	RMS_{sm}[d]
1	z-Scales (Hellberg et al. 1991)	58	PLS	2	0.770	nd[e]	nd
2	GRID (Cocchi and Johansson 1993)	58	PLS	1	0.744	nd	0.500
3	ISA-ECI (Collantes and Dunn 1995)	58	PLS	2	0.700	nd	nd
4	MS-WHIM (rotameric) (Zaliani and Gancia 1999)	58	PLS	6	0.657	0.541	nd
5	MS-WHIM (extended) (Zaliani and Gancia 1999)	58	PLS	2	0.708	0.637	0.540
6	VHSE (Mei et al. 2005)	58	PLS	1	0.770	0.745	0.480
7	Normalized central moment features of FASGAI vectors	58	MLR	12[h]	0.881	0.709	0.301
8	FASGAI	58	PLS	1	0.760	0.728	0.495
9	FASGAI	58	GA-PLS	1	0.796	0.775	0.456
10[f]	FASGAI	29/29[g]	GA-PLS	1	0.869	0.835	0.357

[a] A is the number of principal component

[b] R^2_{cum} is cumulative multiple correlation coefficient

[c] Q^2_{cv} is a cross validation square of cumulative multiple correlation coefficient by the leave-one-out procedure

[d] RMS_{sm} is root mean square error of modelling simulation (RMS_{ms})

[e] nd shows that the correlative value is not given out

[f] External cross validation correlation coefficient (Q^2_{ext}) and root mean square error of external validation (RMS_{ext}) are 0.706 and 0.558, respectively

[g] Two numbers separated by slashes denote the number of samples in training and test sets, respectively

Table 4 Preferred amino acids at positions which are important for activities of ACE inhibitors

Amino acid positions	Properties	Preferred amino acids	Contribution to peptide activity
The first residue	Electronic properties	R, K, V, T, C	Positive
The second residue	Bulky properties	W, Y, F	Positive
The second residue	Hydrophobicity	I, V, F, L	Positive
The second residue	Compositional characteristics	L, A, G, V	Negative

Conclusions

With the exploitation and development of new drug, QSAR has been brought into the spotlight, involving not only the key idea of pharmaceutical chemistry and pharmacology but also the foundations of drug design. Molecular structural characterization is the key to success of QSAR study. Concerning with the culled 335 physicochemical parameters of the coded 20 amino acids, FASGAI vectors by multivariate statistical analysis are proposed to represent peptide structures. QSAR analysis of BTDs and ACE inhibitors is performed to validate the effectiveness of FASGAI vectors in structural representation of peptides. Satisfying results indicate that FASGAI vectors are advantageous of rapid calculation, easy operation, excellent performance, definite physicochemical meaning, etc. and that there is a wide prospect for applications of FASGAI vectors in relationship between structures and activities of various peptides or proteins.

Acknowledgments We thank the reviewers for the constructive comments. This work was supported by the National high-tech Research Program (The "863" Program) (2006AA02Z312), National 111 Programme of Introducing Talents of Discipline to Universities (0507111106) and Innovative Group Program for Graduates of Chongqing University, Science and Innovation Fund (200711C1A0010260).

References

Agüero-Chapín G, Gonzalez-Díaz H, de la Riva G, Rodríguez E, Sanchez-Rodríguez A, Podda G, Vazquez-Padrón RI (2008) MMM-QSAR recognition of ribonucleases without alignment: comparison with an HMM model and isolation from Schizosaccharomyces pombe, prediction, and experimental assay of a new sequence. J Chem Inf Model 48:434–448

Agüero-Chapin G, González-Díaz H, Molina R, Varona-Santos J, Uriarte E, González-Díaz Y (2006) Novel 2D maps and coupling numbers for protein sequences: the first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L. FEBS Lett 580:723–730

Andersson PM, Sjostrom M, Lundstedt T (1998) Preprocessing peptide sequences for multivariate sequence-property analysis. Chemom Intell Lab Syst 42:41–50

Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230

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

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

Crackower MA, Sarao R, Oudit GY, Yagil C, Kozieradzki I, Scanga SE, Oliveira-dos-Santos AJ, da Costa J, Zhang L, Pei Y, Scholey J, Ferrario CM, Manoukian AS, Chappell MC, Backx PH, Yagil Y, Penninger JM (2002) Angiotensin-converting enzyme 2 is an essential regulator of heart function. Nature 417:822–828

Fauchere JL, Charton M, Kier LB, Verloop A, Pliska V (1988) Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res 32:269–278

Felipe-Sotelo M, Andrade JM, Carlosena A, Prada D (2003) Partial least squares multivariate regression as an alternative to handle interferences of Fe on the determination of trace Cr in water by electrothermal atomic absorption spectrometry. Anal Chem 75:5254–5261

Field A (2005) Discovering statistics using SPSS, 2nd edn. Sage, London

González-Díaz H, Uriarte E (2005) Proteins QSAR with Markov average electrostatic potentials. Bioorg Med Chem Lett 15:5088–5094

González-Díaz H, Molina R, Uriarte E (2005) Recognition of stable protein mutants with 3D stochastic average electrostatic potentials. FEBS Lett 579:4297–4301

González-Díaz H, Sánchez-González A, González-Díaz Y (2006) 3D-QSAR study for DNA cleavage proteins with a potential anti-tumor ATCUN-like motif. J Inorg Biochem 100:1290–1297

González-Díaz H, Pérez-Castillo Y, Podda G, Uriarte E (2007a) Computational chemistry comparison of stable/nonstable protein mutants classification models based on 3D and topological indices. J Comput Chem 28:1990–1995

González-Díaz H, Saíz-Urra L, Molina R, González-Díaz Y, Sánchez-González A (2007b) Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments. J Comput Chem 28:1042–1048

González-Díaz H, Vilar S, Santana L, Uriarte E (2007c) Medicinal chemistry and bioinformatics: current trends in drugs discovery with networks topological indices. Curr Top Med Chem 7:1015–1029

González-Díaz H, Saiz-Urra L, Molina R, Santana L, Uriarte E (2007d) A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions. J Proteome Res 6:904–908

González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E (2008) Proteomics, networks and connectivity indices. Proteomics 8:750–778

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

Hasegawa K, Funatsu K (1998) GA strategy for variable selection in QSAR studies: GAPLS and D-optimal designs for predictive QSAR model. J Mol Struct (Theochem) 425:255–262

Hasegawa K, Funatsu K (2000) Partial least squares modeling and genetic algorithm optimization in quantitative structure–activity relationships. SAR QSAR Environ Res 11:189–209

Hasegawa K, Miyashita Y, Funatsu K (1997) GA strategy for variable selection in QSAR Studies: GA based PLS analysis of calcium channel antagonists. J Chem Inf Comput Sci 37:306–310

Helland IS (2001) Some theoretical aspects of partial least squares regression. Chemom Intell Lab Syst 58:97–107

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

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

Hunt PA (1999) QSAR using 2D descriptors and Tripos' SIMCA. J Comput Aided Mol Des 13:453–467

Johnson RA, Wichern DW (2002) Applied multivariate statistical analysis. Prentice-Hall, Upper Saddle River

Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. Nucleic Acids Res 28:374

Kidera A, Konishi Y, Poka M, Ooi T, Scheraga HA (1985) Statistical analysis of the physical properties of the 10 naturally occurring amino acids. J Protein Chem 4:23–55

Kowalski RB, Wold S (1982) Pattern recognition in chemistry. In: Krishnaiah PR, Kanal LN (eds) Handbook of statistics. North-Holland, Amsterdam

Lejon T, Svendsen JS, Haug BE (2002) Simple parameterization of non-proteinogenic amino acids for QSAR of antibacterial peptides. J Peptide Sci 8:302–306

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

Mei H, Liao Z, Zhou Y, Li SZ (2005) A new set of amino acid descriptors and its application in peptide QSARs. Biopolymers (Pept Sci) 80:775–786

Molina E, Diaz HG, Gonzalez MP, Rodriguez E, Uriarte E (2004) Designing antibacterial compounds through a topological substructural approach. J Chem Inf Comp Sci 44: 515–521

Nakai K, Kidera A, Kanehisa M (1988) Cluster analysis of amino acid indices for prediction of protein structure and function. Protein Eng 2:93–100

Norinder U (1991) Theoretical amino acid descriptors: application to bradykinin potentiating peptides. Peptides 12:1223–1227

Ramos de Armas R, González-Díaz H, Molina R, Pérez-González M, Uriarte E (2004) Stochastic-based descriptors studying peptides biological properties: modeling the bitter tasting threshold of dipeptides. Bioorg Med Chem 12:4815–4822

Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S (1998) New chemical descriptors relevant for the design of biologically active peptides: a multivariate characterization of 87 amino acids. J Med Chem 412:481–2491

Selassie CD, Mekapati SB, Verma RP (2002) QSAR: then and now. Cur Top Med Chem 2:1357–1379

Sewald N, Jakubke HD (2002) Peptides: chemistry and biology. Wiley-VCH Verlag GmbH, Weinheim

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

Todeschini R, Consonni, V, Pavan M (2002) DRAGON software version 2.1. http://www.talete.mi.it/main_exp.htm

Tomii K, Kanehisa M (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng 9:27–36

Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and inerpretation of QSPR models. QSAR Comb Sci 22:69–77

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

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

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