Amino Acids (2011) 40:1169–1183
DOI 10.1007/s00726-010-0741-x

ORIGINAL ARTICLE

Novel amino acids indices based on quantum topological molecular similarity and their application to QSAR study of peptides

Bahram Hemmateenejad · Saeed Yousefinejad · Ahmad Reza Mehdipour

Received: 4 April 2010 / Accepted: 31 August 2010 / Published online: 16 September 2010

Abstract A new source of amino acid (AA) indices based on quantum topological molecular similarity (QTMS) descriptors has been proposed for use in QSAR study of peptides. For each bond of the chemical structure of AA, eight electronic properties were calculated using the approaches of bond critical point and theory of atom in molecule. Thus, for each molecule a data matrix of QTMS descriptors (having information from both topology and electronic features) were calculated. Using four different criterion based on principal component analysis of the QTMS data matrices, four different sets of AA indices were generated. The indices were used as the input variables for QSAR study (employing genetic algorithm-partial least squares) of three peptides' data sets, namely, angiotensin-converting enzyme inhibitors, bactericidal peptides and the peptides binding to the HLA-A*0201 molecule. The obtained models had better prediction ability or a comparable one with respect to the previously reported models. In addition, by using the proposed indices and analysis of the variable important in projection, the active site of the peptides which plays a significant role in the biological activity of interest, was identified.

Keywords Amino acid indices · QTMS · QSAR · Peptide

Electronic supplementary material The online version of this article (doi:10.1007/s00726-010-0741-x) contains supplementary material, which is available to authorized users.

B. Hemmateenejad (✉) · S. Yousefinejad
Department of Chemistry, Shiraz University, Shiraz, Iran
e-mail: hemmatb@sums.ac.ir

B. Hemmateenejad · S. Yousefinejad · A. R. Mehdipour
Medicinal & Natural Products Chemistry Research Center, Shiraz University of Medical Sciences, Shiraz, Iran

Introduction

Peptides play significant roles in biological world especially in human life. There are a great number of peptides and proteins used in therapeutics and more that are under development as pharmaceutical targets. Therefore, thousands of peptides and proteins are designed, synthesized and screened for various pharmacological effects. However, design and prediction of their activity remain one of the most challenging areas in the life sciences due to large amount of arrangement possibilities. Therefore, computational study of structure/activity relationships becomes an attractive field in peptides design (Padron-Garcia et al. 2009; Tong et al. 2008; Ramos de Armas et al. 2005; Du et al. 2008; Zhou et al. 2010).

Quantitative structure–activity relationship (QSAR) methods represent a mathematical attempt to explain the relationship between the structure of a set of compounds and their biological activities (Hansch et al. 2002; Selassie et al. 2002). Indeed, the activities are predicted by a regression analysis using some molecular descriptors. Once a correlation is established, these models would be used to predict the compounds of unknown activities and/or to study the action mechanism of chemical–biological interactions in drug design (Mehdipour et al. 2007). So far QSAR is a well established method and has become a standard tool in the drug discovery and development (Du et al. 2008).

The research on QSAR modeling of the peptides has attracted major interest despite many challenges in this field (Lin et al. 2008). Therefore, a lot of efforts were undertaken in order to model the relation between the peptide structure/sequence and its biological activity (Jenssen et al. 2006; Tian et al. 2009). On the basis of QSAR concept, functions and structures of peptides or

proteins are resolved by the information enclosed in the amino acid arrangements. Thus, there are a number of ways based on the principle of describing diverse properties of a molecule and then relating these properties to the biological activities (Jenssen et al. 2006). In the case of the peptides, since the chain is built of repeating units of similar structure, there are two different approaches that can be followed for encoding of the structure: first, using the global descriptors which are based on the whole structure of the peptides. However, there are some difficulties in calculating the characteristics of whole peptides related to the fact that all the atoms belonging to a peptide should be considered and complex descriptors cannot be calculated on big structures constituted by thousands of atoms (Lin et al. 2008; Mauri et al. 2008). Furthermore, they do not show a good predicting ability compared to other types of descriptors (Doytchinova et al. 2005). The second, using the structural properties of the individual building blocks, in this case amino acids, for modeling the biological response which has been receiving more attention. In this methodology, a set of descriptors is calculated for each amino acid and they are put together to produce a descriptor data matrix for a set of peptides. Then, in most cases, factor analysis is applied to the various properties of the 20 natural amino acids in order to obtain structure descriptors (Lin et al. 2008). In this context, definition of a new amino acid index which can describe the structure–function/activity relationship in more convenient way is the frontier of the peptide QSAR research.

Recently, merging quantum mechanical data with the topological approach, according the theory of atoms in molecules (AIM), led to a novel category of descriptors, called quantum topological molecular similarity (QTMS) indices (O'Brien and Popelier 2001, 2002). These descriptors characterize a molecule, based on the topology of its electron density (Bytheway et al. 1996). Over the last several years, QTMS methods have produced excellent results of relevance in different fields. Primarily, they were used to predict Hammett σ values of benzoic acids (O'Brien and Popelier 2002). Subsequently, their action radius was successfully explored in biological (Popelier et al. 2004), environmental (Roy and Popelier 2008a), medicinal (Popelier et al. 2002; Mohajeri et al. 2008; Roy and Popelier 2008b; Popelier and Smith 2006; Hemmateenejad et al. 2008), industrial, and physical organic chemistry (Alsberg et al. 2000, 2001; Chaudry and Popelier 2003; Hemmateenejad and Mohajeri 2007; Harding et al. 2009). In the all cases, the models revealed excellent validation statistics and also provided information about the active center of the compounds which are important for the activity studies. An interesting feature of QTMS indices, which makes them different from other sources of descriptors is that, a two-dimensional array of descriptors is produced for each molecule (quantum property in one dimension and topology or chemical bonds in the other direction).

In this work, we calculated the QTMS parameters of amino acid and then transformed them using factor analysis into novel structural indices of amino acids, using different approaches. These indices were used for modeling of three different peptides data sets and efficient models were obtained.

Materials and methods

QTMS descriptors

The theory of "Atoms in Molecules" retrieves chemical insight from electronic wave functions (Bader 1990; Rafat et al. 2006). Bond critical points (BCPs) are labeled as points in real 3D space where the gradient of the electron density, ρ, faded away ($\nabla\rho = 0$). A BCP is characterized by the sign pattern of the local principal curvature of ρ. One curvature (or eigenvalue) is positive and the remaining curvatures are negative. The sum of the three Eigen values is the Laplacian of the electron density, denoted by $\nabla^2\rho$, which measures how much ρ is concentrated or depleted in a point:

$$\nabla^2\rho = \lambda_1 + \lambda_2 + \lambda_3. \quad (1)$$

Another quantity derived from these eigenvalues is the ellipticity (at the BCP) which is denoted by ε. It is defined as:

$$\varepsilon = \lambda_1/\lambda_2 - 1. \quad (2)$$

Ellipticity provides a measure of the charge which is accumulated in a given plane and, under appropriate conditions, can be used as a quantitative index of the π-character of a bond (Bader et al. 1981). Bonds can be further characterized by evaluating two types of kinetic energy densities denoted by Lagrangian kinetic energy density, $G(\mathbf{r})$ (Bader and Preston 1969)

$$G(\mathbf{r}) = (1/2)N \int \mathrm{d}\tau' \nabla\psi * \cdot \nabla\psi \quad (3)$$

and Hamiltonian kinetic energy density, $K(\mathbf{r})$,

$$K(\mathbf{r}) = -(1/4)N \int \mathrm{d}\tau' \left[\psi * \nabla^2\psi + \psi\nabla^2\psi*\right] = -(1/2)N \int \mathrm{d}\tau' \psi\nabla^2\psi \quad (4)$$

where N is the number of electrons, $N\int \mathrm{d}\tau'$ summarizes the one-electron integration mode and ψ is the Schrödinger wave function. The quantities introduced above can be used together as electro-chemical descriptors of a bond.

Basically, there are two steps in the calculation of QTMS descriptors. First, a geometry optimization which is

performed for each molecule to obtain structural parameters and wave functions at a level of theory ranging from semi-empirical to ab initio or density functional theory (DFT) calculations with varying basis sets. Then, the calculated wave functions are used for the topological analysis of the electron density. In this step, the BCPs are located for each individual bond in all molecules. Here, we keep track of eight BCP descriptors (or QTMS descriptors) for each bond. They are λ_1, λ_2, λ_3, ρ, $\nabla\rho^2$, ε, G and K.

Calculation procedure of AA indices based on QTMS descriptors

AAs have the same structural backbone of the form of NH_2-CHR-COOH. Thus, AAs possess nine common chemical bonds (i.e., 2 N–H, N–C, C–H, C–R, C–C, C=O, C–O and O–H). For each bond, eight QTMS indices were calculated and a descriptor data matrix with the size of 9×8 was provided. By collecting the QTMS of 20 naturally occurring AAs beside each other, a three-dimensional array of descriptors with the size of $(20 \times 9 \times 8)$ was obtained. The AA indices were obtained from this three-dimensional array of QTMS descriptors employing four different approaches including bond and descriptor-based factor analysis of QTMS (CBFQ and CDFQ indices), unfolded-data-based factor analysis of QTMS (CUFQ indices) and all bonds' descriptor-based factor analysis of QTMS descriptors (ADFQ indices). Each letter of the acronyms stands for the number of bonds (C, common bonds; A, all bonds), arrangement of QTMS data matrix (B, arrangement in the direction of bonds; D, arrangement in the direction of descriptors; U, unfolding the matrix in a row vector), application of factor analysis on the data (F) and using QTMS data (Q), respectively. They are explained as follows.

Bond and descriptor-based factor analysis of QTMS

A schematic representation of this approach can be seen in Fig. 1a. In this approach, the QTMS data matrix of the common bonds of each amino acid (size $= 9 \times 8$) was separately subjected to PCA. The results were two compressed matrices **T** and **P** named score and loading, spanning the row (bond) and column (quantum properties) spaces of the original data matrix, respectively. The first three principal components (PCs), explained up to 99% of variances, were extracted so that the **T** and **P** matrices had the respective dimensions of 9×3 and 3×8. Each matrix was vectorized (or unfolded) into a row vector of the dimension of 27 and 24, naming **b** (in the direction of

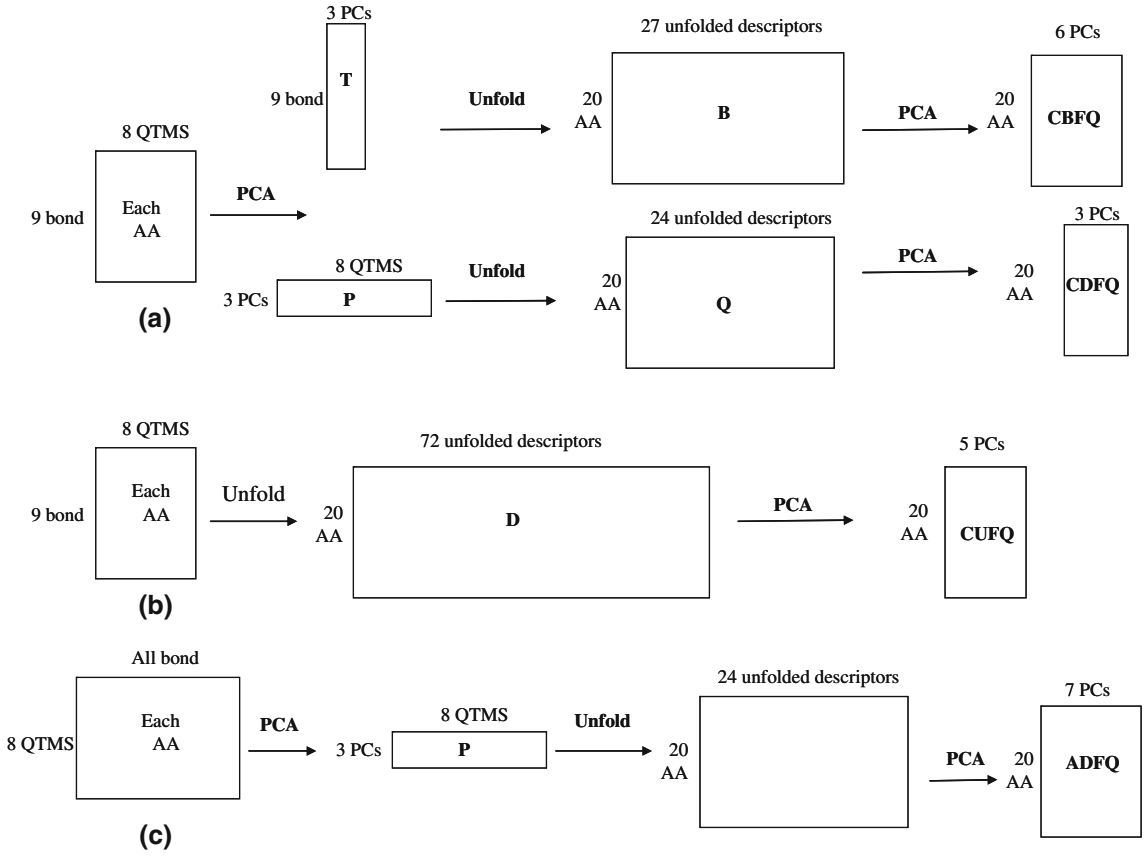

Fig. 1 Schematic representation of the procedures used for calculation of four types of AA indices: **a** CBFQ and CDFQ, **b** CUFQ and **c** ADFQ

bond) and **q** (in the direction of quantum properties), respectively.

The **b** or **q** vectors of all AAs can be arranged into two new data matrices of **B** and **Q** with the size of 20 × 27 and 20 × 24, respectively. Again PCA is applied separately on these data matrices and the significant scores explaining more than 99% of variances were chosen as final descriptors of AAs. For the bond-based data matrix (**B**), six significant scores were selected whereas for the quantum property-based matrix (**Q**), three significant scores were chosen. Thus, the elements in each row of the scores of **B** were named as common bond bond-based factor analysis of the QTMS (CBFQ) indices of the respective AA, and the elements in each row of the scores of **Q** were named as common bond descriptor-based factor analysis of the QTMS (CDFQ) indices.

Unfolded-data-based factor analysis of QTMS

A schematic representation of this approach can be seen in Fig. 1b. In this approach, the QTMS descriptor data matrix of each AA was unfolded into a row vector (**d**) having $9 \times 8 = 72$ elements. Therefore, a matrix of dimension of 20 × 72 (**D**) was provided by collecting the row vector of QTMS indices of each AA below each other. PCA was then applied on this data matrix producing five major PCs, which can explain more than 99% of variances. Each row vector of the resulting score matrix including five elements ($\mathbf{T}_D$) comprises another set of index (called CUFQ) for the corresponding AA.

All bonds' descriptor-based factor analysis of unfolded QTMS descriptors

A schematic representation of this approach is showed in Fig. 1c. In the previous approaches, the QTMS descriptors of the bonds common between all AAs were analyzed. However, in this approach, all chemical bonds presented in AA chemical structure was considered, and consequently, the QTMS data matrix of each AA had the dimension of $nb \times 8$, where nb is being the number of chemical bonds. By applying PCA on each data matrix and using f significant PCs, the score and loading matrices with the size of $nb \times f$ and $f \times 8$ was obtained, respectively. Since the scores had different dimensions, it was not possible to produce a data matrix of the scores of all AAs, similar to that obtained in CBFQ approach. However, it was possible to provide a data matrix from the unfolded loading matrix of the AAs, similar to CDFQ approach. Here, 3 PCs were enough to explain more than 99% of variances and thus the loading had a dimension of 3 × 8, and the descriptor data matrix which was obtained from the unfolded loading had a dimension of 20 × 24. Compression of this data matrix resulted in seven significant scores for each AA, which comprises the ADFQ indices of the AAs.

Data set

Three peptide data sets with known biological activity were used to investigate the performances of the suggested AA indices. They were a set of 55 angiotensin-converting enzyme (ACE) inhibitors, a set of 12 bactericidal peptides and a set of 177 nonameric peptides binding to the HLA-A*0201 molecule. The data sets were picked from the papers of Lin et al. (2008) and Doytchinova et al. (2005) (see Table 1).

Quantum chemical computations

The chemical structure of the AAs was drawn by HyperChem Software and then was transferred into GAUSSIAN98 program to optimize the three-dimensional geometry of the molecules at the RHF/6-31 + G* level of theory. Then, AIM 2000 program (AIM2000, Version 2.0, 2002, http://www.aim2000.de/) located all bond critical point (BCPs) in each of the molecules and evaluated, at each BCP position, the eight QTMS descriptors described previously. Therefore, eight descriptors were obtained for each bond of amino acids. Then, these parameters were transformed into a new set of parameters using the approaches explained in the theory section.

Modeling procedure

Statistical modeling of the relationship between the proposed indices of the AAs presented in the structure of the peptides and the biological activity of the peptides was achieved utilizing partial least square (PLS), because PLS

Table 1 Characteristics of the data sets used in this paper

	Name	No. of molecules	No. of sequences	No. of training sets	No. of test sets	References
Data set 1	ACE	55	3	45	10	Lin et al. (2008)
Data set 2	Bactericidal	12	18	12	–	Lin et al. (2008), Tong et al. (2008)
Data set 3	HLA	177	9	131	46	Doytchinova et al. (2005)

is generally used when the number of descriptors is high. As an accepted procedure of refinement process in selecting the optimum number of PLS latent variables, leave-one-out cross validation (LOO-CV) and leave-many-out cross validation (LMO-CV), using Q^2 as scoring function, were used (Wold et al. 2001):

$$Q^2 = 1.0 - \frac{\sum_{i=1} (Y_{\text{exp}} - Y_{\text{pred}})^2}{\sum_{i=1} (Y_{\text{exp}} - Y_{\text{mean}})^2}. \quad (5)$$

In order to find better structure–activity relationships, genetic algorithm (GA) was performed to select the most suitable set of input variables. In addition, to achieve more robust results, GA was run many times each time with different initial population. Then, more repeated parameters in the collection of different runs were chosen for final modeling. Moreover, correlation coefficients of prediction sets (R_{p}^2) (for ACE inhibitors data set and peptides binding to the HLA) were used to validate the prediction ability of models in the model development steps (the prediction set was the same as in original articles) (Lin et al. 2008; Doytchinova et al. 2005).

In order to assess the risk of chance correlation, permutation test (*y*-scrambling) was performed (Baumann 2003, 2005). In this way, the biological activity of the molecules was randomly shuffled 50 times and maximum Q^2 were reported in order to show that model was not obtained by chance. This is shown by Q_{MP}^2.

The subroutines for doing PLS were written in MATLAB (Mathwork Inc., version 7). For variable selection by GA, the PLS-toolbox developed by Eigen vector Company (Eigenvector Research, Inc.) was employed. A Pentium IV personal computer with windows XP operating system was used throughout.

Results and discussion

QTMS indices describe the quantum property of the molecules based on the molecular topology (or orientation of chemical bonds in the molecule). As it was explained, we used the QTMS indices of AAs to generate four different set of AA indices for use in QSAR study of peptides. Application of PCA on the 9 × 8 QTMS data matrix of all AA revealed that they can be abstracted and represented in a three-dimensional space, which explains more than 99% of variances. The scores and loadings of the QTMS data matrices can be used as a source of new AA indices. New data matrices can be prepared by unfolding the scores or loadings of 20 naturally occurring AAs. Since the number of elements provided for each AA was large, PCA was used again to compress the data. It was found that the score data matrix can be abstracted into a 6-dimensional space, whilst the loading data matrix was abstracted well in a 3-dimensional space. Since scores were in the direction of chemical bonds and loadings were in the direction of electronic properties, the lower dimensionality of the latter suggested that the electronic properties were more correlated than bonds, and introducing of the bond information, would increase the dimensionality of the data. It should be noted that data matrices of non-correlated variables possess more chemical information. The AA indices calculated from PCA of the score and loadings of the QTMS data matrices are listed in Table 2. The percent of variances explained by each PC are also listed in this Table. The PCs having higher explained variance possess more information about the QTMS data. However, the PCs of high explained variances do not essentially have high correlation with biological activity (Hemmateenejad et al. 2003; Hemmateenejad 2005). Thus, it was necessary to employ a variable selection method for selecting the most convenient subset of AA indices in a specified data set.

Another set of AA indices were calculated by unfolding the QTMS of each AA into a row vector and then collecting them into a data matrix with the size of 20 × 72. Application of PCA on this data matrix resulted in five significant scores for each AA, which they called CUFQ indices. The values of these indices in association with their respective percent of the explained variances are listed in Table 3.

One limitation of the QTMS descriptors for using in QSAR studies is that they can be calculated for the chemical bonds which are in common between all molecules in the data set. However, in this work, we used PCA to overcome this problem and in another approach the QTMS indices of all bonds of the AAs were used to calculate some other AA indices. The approach is similar to CDFQ and CBFQ approaches. Since each AA had a total of *nb* chemical bonds, its QTMS data matrix would have a dimension of $nb \times 8$. The score of such matrices for different AAs did not have the same dimension. However, all loading matrices had the dimension of $f \times 8$, where f was the number of significant PCs. It was found that the QTMS data matrices of all AAs could be abstracted into three significant PCs. In the same manner as CDFQ approach, the loadings of these matrices were unfolded and then subjected again to PCA to calculate the last set of AA indices. In this case, the final data matrix was compressed into seven significant scores, the element of each row comprise the ADFQ indices of AAs. The values of these indices and their explained variances are listed in Table 3

In the next sections, the application of these newly proposed AA indices in modeling of three different biological activities of peptides will be discussed. Each set of AA indices (i.e., CDFQ, CBFQ, ADFQ and CUFQ) was used to develop separate QSAR models. Consider a peptide data set, which is composed of some oligopeptides, all having the same number of residue (*nR*). As the set of AA indices used

Table 2 The AA indices obtained by application of PCA on the unfolded scores and loadings of the QTMS data matrix of the common bonds of AAs (CBFQ and CDFQ indices, respectively)

AA	CBFQ indices						CDFQ indices		
	$CBFQ_1$	$CBFQ_2$	$CBFQ_3$	$CBFQ_4$	$CBFQ_5$	$CBFQ_6$	$CDFQ_1$	$CDFQ_2$	$CDFQ_3$
Ala (A)	0.656	−3.454	−1.976	−0.150	−1.121	−0.596	−2.403	3.139	1.480
Arg (R)	−0.329	0.344	−0.233	−0.068	0.261	0.287	0.308	0.065	−0.311
Asn (N)	−0.367	0.097	−0.142	−0.488	−1.142	2.888	0.299	0.136	−0.377
Asp (D)	−0.342	−0.451	1.166	4.000	−0.549	0.034	0.622	−1.717	3.533
CyS (C)	−0.295	0.286	−0.440	−0.005	1.194	−0.587	0.319	0.064	−0.395
Gln (Q)	−0.348	0.288	−0.350	−0.063	0.128	1.472	0.307	0.081	−0.349
Glu (E)	−0.317	0.462	0.234	−0.461	−0.452	−0.196	0.307	0.081	−0.340
Gly (G)	4.065	1.171	−0.203	0.282	−0.135	0.115	−3.358	−2.465	−0.829
His (H)	−0.328	0.368	0.230	−0.239	−0.905	−1.038	0.307	0.115	−0.435
Ile (I)	−0.294	0.449	−0.138	−0.133	0.439	−1.315	0.302	0.072	−0.301
Leu (L)	−0.299	0.351	0.259	−0.425	−0.498	−1.006	0.308	0.062	−0.297
Lys (K)	−0.312	0.417	0.263	−0.389	−0.569	−0.728	0.307	0.065	−0.299
Met (M)	−0.342	0.270	−0.307	−0.073	0.013	1.121	0.304	0.078	−0.320
Pro (P)	0.598	−1.921	3.278	−1.168	1.310	0.342	0.209	−0.155	1.313
Phe (F)	−0.324	0.352	0.180	−0.404	−0.752	−0.197	0.305	0.081	−0.344
Ser (S)	−0.237	−0.116	−1.086	0.355	2.085	0.783	0.322	0.037	−0.380
Thr (T)	−0.228	−0.056	−0.964	0.293	2.125	−0.118	0.322	0.042	−0.394
Trp (W)	−0.330	0.304	0.153	−0.350	−0.947	−0.062	0.304	0.071	−0.309
Tyr (Y)	−0.326	0.361	0.175	−0.384	−0.820	−0.074	0.304	0.080	−0.337
Val (V)	−0.303	0.478	−0.100	−0.129	0.334	−1.123	0.304	0.069	−0.308
% Variance	53.2	22.9	13.0	7.2	2.7	0.6	57.4	37.3	5.0

in QSAR study of peptides had *nI* indices, the vector of descriptor for each oligopeptide was composed of the AA indices of its residue, collecting beside each other in the order of the appearances of that AA in the peptide sequence. So the number of elements of this row vector was equal to $nD = nR \times nI$. Based on the type of AA indices which was used, each element of this vector is denoted by $CBFQ_{K,L}$, $CDFQ_{K,L}$, $ADFQ_{K,L}$ or $CUFQ_{K,L}$, where the subscript K and L refer to the order of the index in its corresponding set and the order of residue in the oligopeptide sequence, respectively. For example the fifth index of the set of CBFQ indices corresponding to the third residue in the oligopeptide is represented as $CDFQ_{5,3}$. It should be noted that, the order of the index in each set of AA indices was determined by the explained percent of the variances of that index, and they were ranked based on the decreasing variances.

If the number of polypeptides in the data set was considered as *nP*, a descriptor data matrix with the number of row and columns of *nP* and *nD*, respectively, would be obtained.

QSAR models of ACE inhibitors

This data set is composed of 55 tri-peptides as inhibitors of the angiotensin-converting enzyme (ACE), among which 45 molecules were used as training set and the remainders were used as external prediction set to develop the most convenient QSAR model. The resulting models based on the use of four different sets of AA indices are summarized in Table 4.

In the case of CBFQ indices, among the 18 original input variables, GA selected 10 indices as the input of PLS. Projection of these indices into 4 latent variables produced a QSAR model of high statistical quality so that, it could explain 87.1% of variances in the ACE inhibitory data of the studied tri-peptides. In order to confirm the reliability of the model, leave-one-out and leave-many-out (leave-15-out) cross validation was applied on the training set and the respective Q^2_{LOO} and Q^2_{LMO} of 0.824 and 0.821 were obtained. The closeness of these parameters to each other and to that of training confirms that the generated model is stable. The R^2 of the test set ($R^2_p = 0.854$) was also very close to that of training and cross validation, indicating the good prediction ability of the model without the presence of significant over-fitting. Moreover, the model reported a very low value of chance correlation Q^2_{MP}, emphasizing the model was not chancy and the obtained relationship was systematic. Fig. S1 (supporting information) shows the plot of calculated activities by CBFQ-based PLS model versus observed activities for the ACE inhibitor tri-peptides.

Table 3 The AA indices obtained by application of PCA on the unfolded original QTMS data matrices and the unfolded loadings of the QTMS data matrices of all bonds of AAs (CUFQ and ADFQ indices, respectively)

AA	CUFQ indices					ADFQ indices						
	$CUFQ_1$	$CUFQ_2$	$CUFQ_3$	$CUFQ_4$	$CUFQ_5$	$ADFQ_1$	$ADFQ_2$	$ADFQ_3$	$ADFQ_4$	$ADFQ_5$	$ADFQ_6$	$ADFQ_7$
Ala (A)	2.862	3.132	−0.085	−0.032	0.156	1.113	−3.053	1.864	1.477	−1.097	0.177	0.356
Arg (R)	−0.329	0.008	−0.197	0.281	0.026	0.334	1.765	0.604	1.205	−0.521	−0.293	0.619
Asn (N)	−0.360	0.056	−0.220	0.229	−1.972	−0.437	−0.136	0.171	−1.459	−0.784	0.614	0.179
Asp (D)	−0.396	0.000	−0.410	−4.209	0.099	0.478	1.747	1.076	0.042	−1.812	−0.258	−0.103
CyS (C)	−0.337	−0.015	−0.248	0.270	0.510	−0.416	−0.519	−0.106	0.293	−0.050	0.312	−0.586
Gln (Q)	−0.332	−0.020	−0.155	0.271	0.238	−0.311	−0.234	0.080	−1.429	−0.871	0.755	0.125
Glu (E)	−0.342	0.003	−0.199	0.267	−0.242	0.265	0.498	2.638	−0.922	3.096	0.329	0.322
Gly (G)	2.977	−3.020	−0.191	−0.040	−0.160	3.761	0.039	−1.725	−0.657	0.631	0.119	−0.079
His (H)	−0.353	0.077	−0.402	0.278	−2.254	−0.205	−0.418	0.138	−0.948	−0.076	−3.969	−0.896
Ile (I)	−0.305	0.024	−0.232	0.292	0.368	−0.527	−0.176	−0.569	0.279	0.376	0.536	−1.264
Leu (L)	−0.318	0.007	−0.182	0.276	0.304	−0.527	−0.178	−0.565	0.282	0.381	0.548	−1.268
Lys (K)	−0.321	−0.005	−0.184	0.282	0.209	−0.585	−0.093	−0.558	0.301	0.476	0.307	−1.177
Met (M)	−0.333	0.019	−0.204	0.271	−0.309	−0.510	−0.328	−0.464	0.932	0.539	0.047	−0.923
Pro (P)	−0.329	0.039	−0.228	0.249	−0.523	−0.484	−0.148	−0.571	0.177	0.326	0.321	−0.748
Phe (F)	−0.127	−0.072	4.238	−0.176	−0.092	−0.560	0.008	−0.984	0.802	0.591	−0.183	1.388
Ser (S)	−0.353	−0.158	−0.222	0.341	1.961	−0.318	−0.074	0.055	−1.778	−0.909	0.594	0.758
Thr (T)	−0.343	−0.155	−0.221	0.341	2.151	−0.393	−0.127	−0.141	−1.091	−0.460	0.522	0.054
Trp (W)	−0.323	0.035	−0.215	0.260	−0.369	−0.593	−0.112	−0.770	0.777	0.476	−0.604	2.217
Tyr (Y)	−0.328	0.033	−0.221	0.254	−0.457	−0.481	−0.043	−0.761	0.263	0.233	−0.140	1.782
Val (V)	−0.310	0.011	−0.222	0.295	0.358	0.394	1.584	0.586	1.453	−0.546	0.267	−0.759
% Variance	72.9	18	5.5	2.4	0.8	37.6	25.8	14.1	12.2	6.3	2.8	0.8

Table 4 QSAR models of the peptides as ACE inhibitors obtained using different sets of the proposed AA indices

AA indices	n_{OV}[a]	n_{SV}[b]	n_{LV}[c]	R_{cal}^2[d]	Q_{LOO}^2[e]	Q_{LMO}^2[f]	R_P^2[g]	RMSE cal[h]	Q_{MP}^2[i]
CBFQ	18	10	4	0.871	0.824	0.821	0.845	0.361	0.084
CDFQ	9	6	2	0.760	0.722	0.682	0.713	0.486	0.105
CUFQ	15	7	2	0.855	0.803	0.803	0.924	0.381	0.190
ADFQ	21	10	6	0.867	0.820	0.807	0.861	0.366	0.307

[a] Number of initial variables (overall variable)

[b] Number of selected variables

[c] Number of PLS latent variables

[d] Calibration correlation coefficient

[e] Leave-one-out cross-validation correlation coefficient

[f] Leave-many-out cross-validation correlation coefficient

[g] Prediction correlation coefficient

[h] Calibration root mean square of errors

[i] Maximum cross-validation correlation coefficient for *Y*-randomization test

Another QSAR model was built using CDFQ descriptors. As it is shown in Table 4, the GA-PLS regression selected a subset of six variables, among nine original CDFQ descriptors, and projected them into a 2-dimensional space of latent variables. The statistical quality of this model was lower than that of the previous one, suggesting that the bond space of the QTMS indices possess more useful information about the ACE inhibitory capacity of the study peptides than the space of quantum property. Figure S2 shows the plot of calculated activity of training set and test set and leave-one-out cross validation by CDFQ-based PLS model versus actual activity.

In the case of CDFQ and CBFQ indices, the bond and quantum information were investigated separately.

However, CUFQ indices have mixed information from both bond (topology) and quantum properties of the AAs. As it is reported in Table 4, the descriptor data matrix of CUFQ indices for the tri-peptides of ACE inhibitors contained 15 variables, among which 7 variables were selected by GA-PLS to model the ACE inhibitory of the studied peptides. The resulting model revealed calibration and cross-validation statistics lower than those of the models obtained by CDFQ and CBFQ descriptors. However, it possessed much better prediction quality.

The last model developed for the ACE data set was based on ADFQ descriptors (21 input variables). The GA-PLS modeling of this data set produced a 10-variable QSAR model with the training quality similar to and prediction ability slightly better than CBFQ model. Both ADFQ and CDFQ descriptors were calculated from the loading of the QTMS data matrices. However, the latter used the information of all bonds in the chemical structure of AAs. As can be observed in Table 4, by incorporation of this extra information, the resulting indices could produce more appropriate model for the ACE inhibitory of the peptides.

A comparison between the QSAR models obtained from different types of the proposed AA indices revealed that the model which was made based on CUFQ indices not only used a low number of input variables, but also possessed more prediction ability. Thus, it is preferred over the other models. This model was compared with the previously reported model for ACE data set in Table 5. Obviously, the QSAR model of CUFQ indices possesses much better quality than that obtained by using z-scale indices (Lin et al. 2008) and is comparable with that obtained from Lin-scale indices (Lin et al. 2008). But according to the RMSE values of prediction set (RMSEP), as a brief indicator of accuracy, it could be said that CUFQ-based model (RMSEP = 0.275) is not as accurate as that of Lin-scale (RMSEP = 0.152).

QSAR models of bactericidal peptides

This data set is composed of 12 peptides of known bactericidal activity. Each peptide is composed of 18 AA residues. Since the number of molecules in this data set was very small, we did not select a separate external test set and the models were validated by cross validations. It should be noted that the QSAR models generated from small number of molecules are not highly reliable; however, this data set has been used by many researchers to validate their AA indices. Thus, we used this data set to compare the potentiality of our proposed AA indices with the previously reported ones.

The obtained QSAR models for this data set using different subsets of the proposed AA indices are summarized in Table 6. Obviously, all subsets of indices resulted in QSAR models of high statistical qualities. In the same manner as the ACE data set, the models obtained from CUFQ and ADFQ indices were of the highest internal prediction ability (Q^2_{LMO} of 0.958 and 0.975, respectively),

Table 6 QSAR models of the Bactericidal peptides obtained using different sets of the proposed AA indices

AA indices	n_{OV}	n_{SV}	n_{LV}	R^2_{cal}	Q^2_{LOO}	Q^2_{LMO}	$RMSE_{cal}$	Q^2_{MP}
CBFQ	108	11	3	0.998	0.991	0.943	0.048	0.312
CDFQ	54	17	3	0.978	0.918	0.923	0.148	0.470
CUFQ	90	9	4	0.992	0.977	0.958	0.089	0.458
ADFQ	126	17	3	0.997	0.986	0.975	0.056	0.580

Table 5 Comparison between the QSAR models of three data sets

Data set	Descriptors	Model	LVs	R^2_{cal}	Q^2_{LOO}	$Rpred^2$	RMSE $_{cal}$	Reference
ACE	z-scales	PLS	NR	0.500	0.426	NR	0.404	Lin et al. (2008)
	Lin scale	MLR	–	0.970	0.943	0.976	0.154	Lin et al. (2008)
	QTMS-CUFQ	GA-PLS	2	0.855	0.803	0.924	0.381	This work
Bactericidal	z-scale	PLS	4	0.985	0.525	NR	0.34	Lin et al. (2008), Mei et al. (2004)
	VSTV	PLS	4	0.996	0.879	NR	0.17	Lin et al. (2008), Mei et al. (2004)
	Lin scale	MLR		0.926	0.773	NR	0.41	Lin et al. (2008)
	VSW	PLS	3	0.997	0.954	NR	0.13	Tong et al. (2008)
	QTMS-ADFQ	GA-PLS	3	0.997	0.986	NR	0.056	This work
HLA	Additive	GA-MLR	–	0.97	NR	0.24	0.25	Doytchinova et al. (2005)
	Additive	PLS	3	0.85	0.54	0.64	NR	Doytchinova et al. (2005)
	Global	GA-MLR	–	0.43	NR	0.42	0.75	Doytchinova et al. (2005)
	z-scales	GA-MLR	–	0.67	NR	0.50	0.59	Doytchinova et al. (2005)
	QTMS-ADFQ	GA-PLS	3	0.648	0.563	0.498	0.593	This work

NR not reported

and that of CDFQ represented the lowest internal prediction results (Q^2_{LMO} of 0.923). Since the number of AA residues in the peptides was large, the number of variables in the original descriptor data matrix was also high. However, GA-PLS selected a small number of them as predictor variables. In addition, the selected variables were projected in a low-dimensional space of latent variables (3 or 4 variables), which were an acceptable dimension with respect to the number of molecules. This shows the unique property of GA-PLS to produce a low-dimensional model from a data set of extremely high dimension of input variables with respect to the number of samples. However, the permutation test indicated that still there was a sign of chance correlation in the models as a large value of Q^2_{MP} was obtained for them.

The ADFQ-based QSAR model of the bactericidal peptides (which represented the most appropriate quality) is compared in Table 5 with the previously reported QSAR model for the same data set. Clearly, the ADFQ-based QSAR model would be preferred over the models obtained by *z*-scale, Lin scale and VSTV indices (Mei et al. 2004; Lin et al. 2008) with respect to both calibration and cross-validation statistics. Whilst our model and that obtained based on VSW indices (Tong et al. 2008) have the same correlation coefficient of calibrations, the former exhibited a slightly better cross-validation result. Thus, the ADFQ-based QSAR model of the bactericidal peptides can be preferred over four previously reported models.

QSAR models of HLA data set

In this data set, the binding of 177 nonameric peptides to the HLA-A*0201 molecule was modeled by the proposed AA indices based on QTMS descriptors. This is an example of a peptide data set of structurally diverse so that, the previous attempts for creating a QSAR model of appropriate prediction ability have been failed. As it is reported in Table 5, the QSAR models based on additive descriptors possessed high calibration quality (i.e., R^2 of 0.85 or 0.92); however, their prediction ability was poor. These models can be considered as over-fitted models for a large difference between calibration and prediction qualities. The QSAR models produced from global and *z*-scale indices possessed moderate qualities whereas the latter represented better results for both calibration and prediction (R^2 and R^2_P of 0.67 and 0.50, respectively).

The QSAR models based on the AA indices of QTMS descriptors are summarized in Table 7. Evidently, these models are not more accurate than the previously reported models. Their calibration correlation coefficients were in the range of 0.266 (for CDFQ-based model) and 0.648 (for ADFQ-based model). Similar trends were observed for the cross-validation and prediction data. It is interesting to note that all models represented a balance between prediction ability and calibration quality and thus they did not suffer from serious over-fitting. In the same manner as ACE and bactericidal data sets, the QSAR model based on ADFQ indices would be the most appropriate one for HLA data set and it has comparable performances with the previously reported models (Table 5) for this data set.

Active site analysis

The term "active site" is commonly used as the binding cavity in an enzyme, but here it refers to a general highlighted zone in the molecule whose structure represents the highest impact on the activity. A convenient way for determining the active part of a bio-molecule in a QSAR study is using the so called variables importance in the projection (VIP) of the variables presented in the QSAR model. VIP values reflect the importance of the terms used in the PLS model with respect to response variable (Erikson et al. 2001). Variables with higher VIP scores are more relevant in explaining the activity. Variables with VIP values more than 1 were indicated as highly influential, while those with VIP values less than 0.8 were considered as the least influential. VIP values lying between 0.8 and 1 indicate moderate influences. VIP values are typically shown as histogram plots, which are given in many figures in this article. Ideally, after the first few variables with the highest VIP values, the profile of the VIP histogram displays a sharp decline. In this case, the active site is well localized, whereas if the VIP profile drops off slowly, the active site of molecule is poorly localized or diffuses (Popelier and Smith 2006).

Table 7 QSAR models of the peptides binding to the HLA obtained using different sets of the proposed AA indices

AA indices	n_{OV}	n_{SV}	n_{LV}	R^2_{cal}	Q^2_{LOO}	Q^2_{LMO}	R^2_P	$RMSE_{cal}$	Q^2_{MP}
CBFQ	54	14	4	0.485	0.396	0.302	0.419	0.717	−0.077
CDFQ	27	5	2	0.266	0.228	0.253	0.213	0.857	0.081
CUFQ	45	14	3	0.379	0.204	0.201	0.266	0.788	0.021
ADFQ	63	25	3	0.648	0.563	0.502	0.498	0.593	0.0018

Active site of the ACE inhibitors

CBFQ descriptors describe the topological properties of AAs; as mentioned previously, the CBFQs were the scores of PCA in the direction of amino acid bonds. The VIP of the variables appeared in the CBFQ-based QSAR models of ACE inhibitors (Table 4) are presented in Fig. 2a. Among the 10 CBFQ-based indices selected by GA for the tri-peptide ACE data set, $CBFQ_{3,3}$ and $CBFQ_{4,3}$ (i.e., the third and fourth indices of the third AA in the tri-peptide sequence) possessed the highest VIP and it could be guessed that the third sequence plays the most significant role in the ACEs activity from the topological point of view. The next three indices having VIP values between 0.8 and 1 are $CBFQ_{3,2}$, $CBFQ_{1,2}$ and $CBFQ_{5,3}$. It was observed that the fifth CBFQ index of the third residue was among the variables of moderate influence. In addition, two indices of the second residue appeared in the list of the variables of moderate influence. Since among the highly and moderately influential variable, three belonged to the third residue and a relatively sharp drop-off in the VIP histogram of CBFQ-based model of ACEs was observed after the first two VIP (belonging to the third amino acid of peptide), from the topological point of view, the third sequence of the ACE inhibitors would be considered as highly significant active site.

The CDFQ indices include the quantum information of the amino acids. The VIP values of the variables that appeared in the corresponding models (Table 4) are shown in Fig. 2b. As it can be seen, among the six selected CDFQ-PCs by GA-PLS for modeling the relation between structure and activity of the ACE inhibitors, the index $CDFQ_{1,3}$ (the first CDFQ of the third residue) represented the highest influences in peptide activity (VIP >1) and the two indices $CDFQ_{3,1}$ and $CDFQ_{2,3}$ were in the second order of importance. We observed that the CDFQ indices of the third residue have been selected as important variables in the electronic point of view too. However, contrary to the CBFQ descriptors, the CDFQ indices of the second residue were not selected as influential variables, and instead, the electronic index of the first residue was selected.

Figure 2c shows that among the variables that appeared in the CUFQ-based QSAR model of the ACE data set, the index $CUFQ_{1,3}$ possessed the highest VIP value with an exactly sharp difference relative to the next VIPs values. This analysis confirmed the significant role of the third residue on the inhibitory capacity of the studied tri-peptides against ACE. The next three CUFQ indices ($CUFQ_{5,3}$, $CUFQ_{3,3}$ and $CUFQ_{5,1}$) can be considered as variables of moderate influence. Thus, according to the CUFQ indices (having information from both electronic and topology of

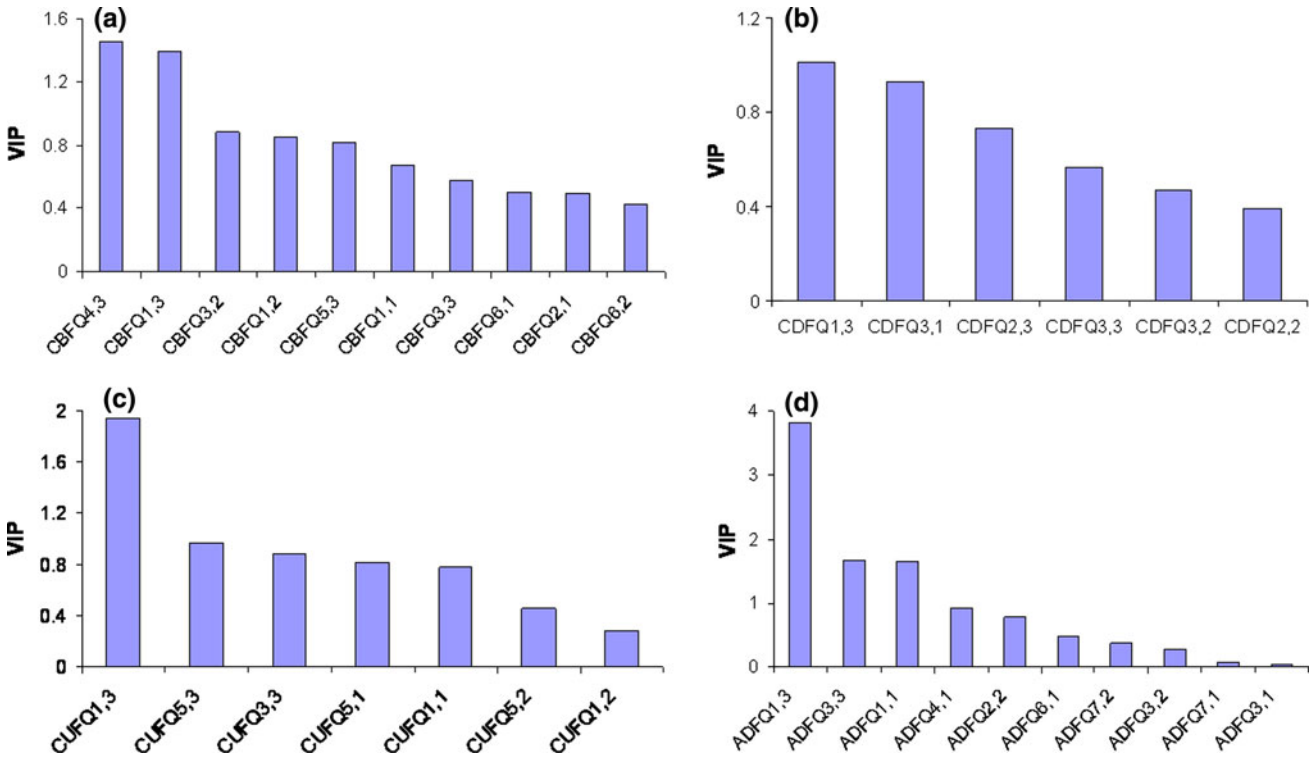

Fig. 2 VIP plot for the variables selected by GA-PLS as input of the QSAR models of ACE data set using **a** CBFQ descriptors, **b** CDFQ descriptors, **c** CUFQ descriptors and **d** ADFQ descriptors

the AAs) the residues in the two ends of the tri-peptides would have more impact on the ACE inhibition.

The last section of Fig. 2 shows the VIP plot for the variables of the ADFQ-based QSAR model of ACE data set. It is clear from the Fig. 2d that four indices including $ADFQ_{1,3}$, $ADFQ_{3,3}$, $ADFQ_{1,1}$ and $ADFQ_{4,1}$ had VIP values about 1 or higher. Again, it could be concluded that the residues of the two ends of tri-peptides play a significant role in the peptides activity; however, the prominent role of the third sequence based on the results could not be neglected.

The discussions given above showed that for a given residue in the peptide's sequence, different indices were selected as influential variables. This makes difficult to decide which sequence would be more important (or would be active site of the peptide). To get a quick snapshot on the role of residues in the studied biological activities, for a given sequence, the VIP values of all of the indices which were selected as highly influential variables in all models were summed. Thus for each sequence, a VIP value which was the representative of all VIP values and assigned to the selected indices of that sequence (denoted by RVIP), was calculated. Thus, RVIP shows the overall importance of each sequence in the studied biological activity.

Figure 3 exhibits the calculated RVIP for the three residues in the chemical structure of the ACE inhibitors. Obviously, the RVIP value of the third residue was much higher than the other residues, suggesting the more significant role of this sequence on the ACE inhibitory of the studied tri-peptides. This implies that the AAs in this sequence are more involved in the ACE inhibitory activity and thus can be considered as the active site of the peptides. This finding is in direct agreement with the previous QSAR study on this data set (Lin et al. 2008), in which the authors found that the amino acid in position 3 is the main factor influencing the biological activity of ACE inhibitors.

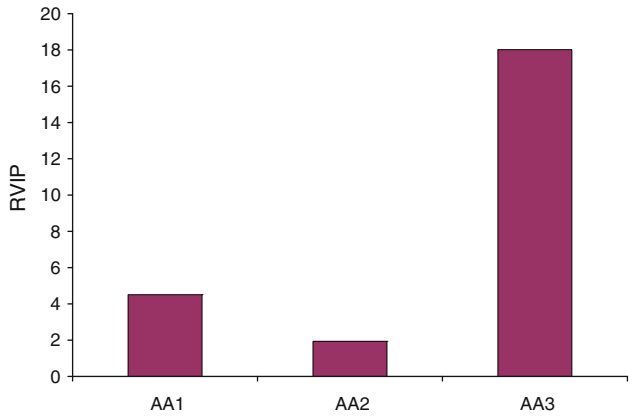

Fig. 3 Plot of RVIP for 3 sequences of the set of ACE inhibitors' peptides

Active site of the bactericidal peptides

Research on bactericidal peptides has been an active field in medicine exploitation in recent years. As it is seen in Table 6, in each descriptor groups of CBFQ, CDFQ, CUFQ and ADFQ, respectively, 11, 17, 9 and 17 descriptors were selected by GA and were used in the modeling of the relation between activity and structure of bactericidal 18 peptides.

Figure 4a shows that among the three selected indices of considerable VIP value, the most important one belonged to the 11th residue (i.e., $CBFQ_{3,11}$) for its VIP value higher than 1. The next high VIP belonged to $CBFQ_{4,9}$ variable and the next was that of $CBFQ_{5,3}$ variable. The histogram has a sharp drop-off after these three factors. So it could be suggested that from the topological point of view, the residues at the middle of peptide chain (i.e. 11th and 9th residues) might be a specific position in activity definition of bactericidal peptides of interest; however, the third amino acid that exists in the beginning of the chain showed a moderate importance.

Figure 4b shows that among the variables selected for the CDFQ-based QSAR model, 10 variables had VIP values greater than 1. However, it was observed that two variables ($CDFQ_{1,11}$ and $CDFQ_{1,18}$) had remarkably much higher VIP values. The former, which had the highest VIP value, showed again (in the electronic point of view) the importance of the 11th residue (being present in the middle of peptide chain) and the latter showed the importance of the 18th amino acid (being present in the end of the chain). The remaining variables of VIP >1 in the order of decreasing importance were $CDFQ_{3,4}$, $CDFQ_{2,1}$, $CDFQ_{1,4}$, $CDFQ_{2,4}$, $CDFQ_{1,1}$, $CDFQ_{3,2}$, $CDFQ_{2,2}$, $CDFQ_{1,2}$, which belonged to the first, second and fourth residues (the beginning of the peptide chain).

Based on the results collected in Fig. 4c, the first three variables, selected for the CUFQ-based QSAR model of the bactericidal peptides, possessed VIP values between 0.8 and 1, which were distinct from that of the remainder. These were $CUFQ_{3,2}$, $CUFQ_{4,1}$, and $CUFQ_{1,11}$, suggesting that from both topology and quantum points of view, the AAs in the first, second and 11th residues would play a significant role in the bactericidal activity of the peptides.

Figure 4d demonstrates that none of the variables used in the ADFQ-based QSAR model had VIP >0.8. Thus, this QSAR model could not reveal the parts of the peptides playing role as active site. However, it was observed that variables of the highest VIP values mostly belonged to the 11th residue.

The plot of RVIP values for all residues in the octadecamer peptides is represented in Fig. 5. Clearly, the 11th sequence of the peptide chain might play the most highlighted role in the activity of peptide. From another view it

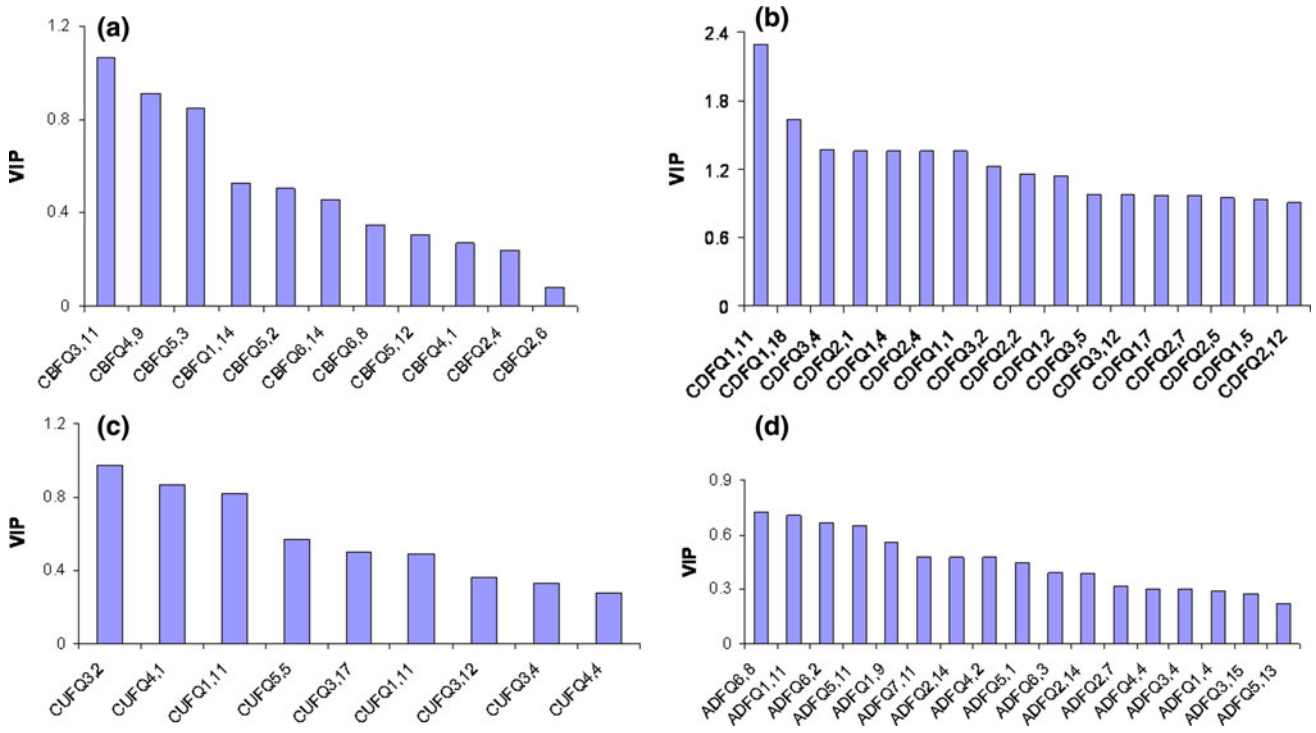

Fig. 4 VIP plot for the variables selected by GA-PLS as input of the QSAR models of Bactericidal data set using **a** CBFQ descriptors, **b** CDFQ descriptors, **c** CUFQ descriptors and **d** ADFQ descriptors

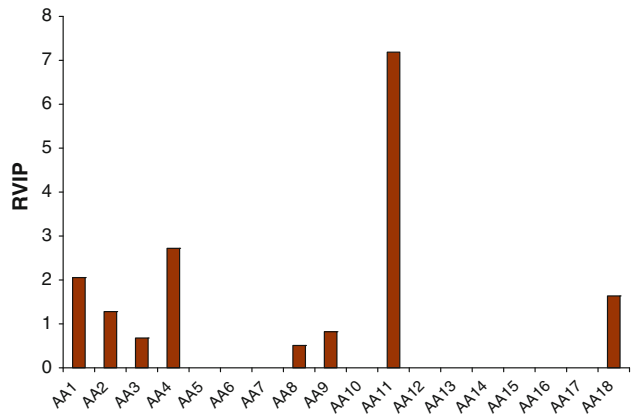

Fig. 5 Plot of RVIP for 18 sequences of the set of the bactericidal peptides

seemed that the electrotopological properties of the sequences located at the middle and two ends of the peptides could be controlling factors for the bactericidal activity. In the previous QSAR study on this kind of bactericidal peptides (Lin et al. 2008) it had been concluded from the MLR coefficient of the model that 9th, 11th and 12th residues of the bactericidal peptides were important and among them the 11th and 12th sequence were denoted as the most and least important sequences, respectively.

Active site of the HLA peptides

HLA peptides are some of the major histocompatibility complex proteins (MHCs) that have critical role in cellular immunity. Foreign proteins originating from pathogenic organisms (parasites, fungi, bacteria, or viruses) are cleaved into short peptides of 8–11 amino acids in the cell and bind to major histocompatibility complex proteins (MHCs). Peptide–MHC complexes are presented on the cell surface, where they are recognized by T cells. So designing active MHCs could be very important in immunotherapeutics and vaccine recognition (Doytchinova et al. 2005). The HLA peptides selected for this study include nine AAs. By analysis of the VIP value variables entered in the QSAR models of different AA indices, we could discuss which residue plays a significant role in the HLA activity.

The VIP values of the variables selected for different QSAR models of the HLA data set are plotted in Fig. 6. For the variables of the CBFQ-based QSAR model, it is observed that there is a significant drop-off between the VIP bars of the third and fourth variables. The first three variables, whose VIP was larger than 0.9, were $CBFQ_{3,6}$, $CBFQ_{3,7}$, $CBFQ_{6,9}$ belong to the sixth, seventh and ninth residues in the peptide sequence (Fig. 6a). It can be suggested that, from the topological point of view, the amino acids of one half of these peptides play a more significant

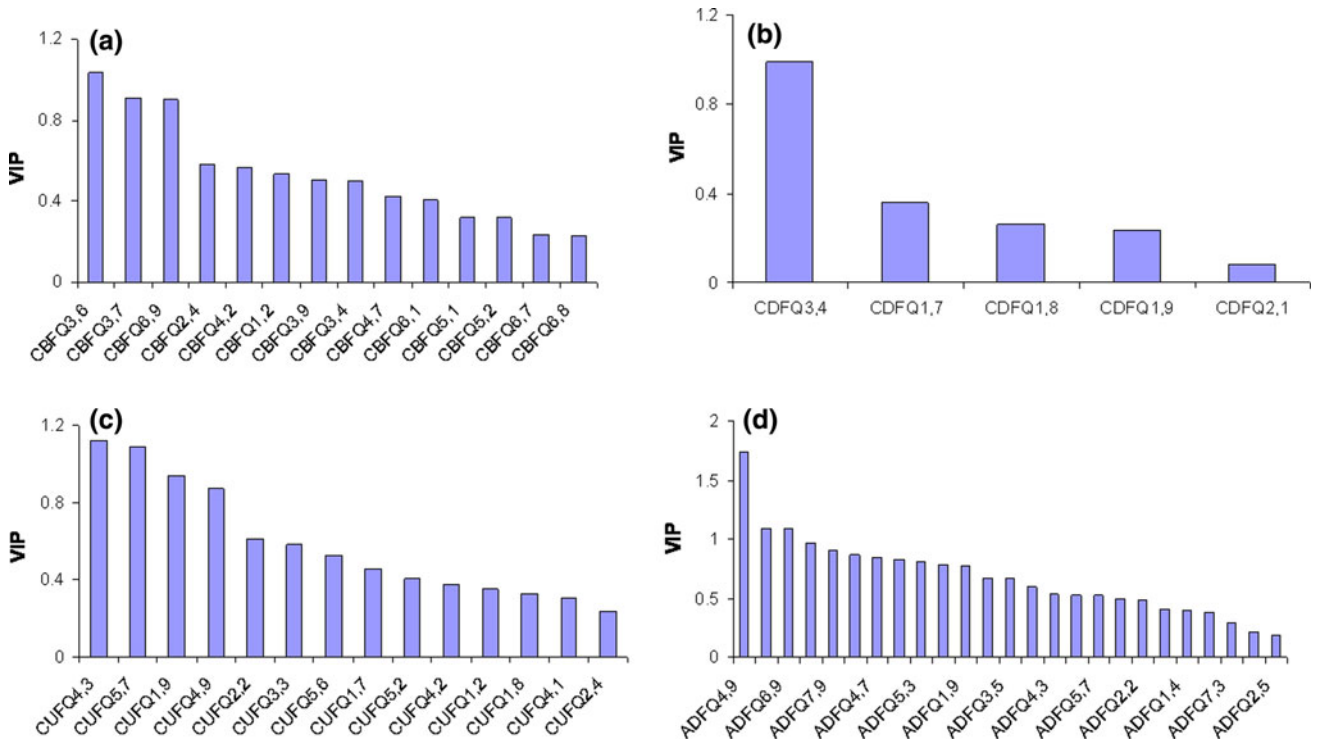

Fig. 6 VIP plot for the variables selected by GA-PLS as input of the QSAR models of HLA data set using **a** CBFQ descriptors, **b** CDFQ descriptors, **c** CUFQ descriptors and **d** ADFQ descriptors

role in pBL_{50} of the peptides, whereas the role of sixth amino acid might be more important.

The VIP values of the coefficients from the CDFQ-based model (Fig. 6b) indicated that only one variable (i.e., $CDFQ_{3,4}$) represents significant contribution in the resulting model. Thus, the electronic property of the fourth residue plays the most significant impact in the HLA activity of the studied peptides. However, the VIP values of the input variables of QSAR model obtained from the CUFQ descriptors (having mixed information from both topological and electronic features of the AAs) revealed that among the 14 selected variables, the variables $CUFQ_{4,3}$, $CUFQ_{5,7}$, $CUFQ_{1,9}$ and $CUFQ_{4,9}$ represent the highest impact(Fig. 6b). Again, it was observed that the indices of the AAs of the one half of the peptides play more significant role in the HLA activity. Also, the VIP values of the ADFQ-based model (Fig. 6d) confirmed that the indices of the AAs in the seventh and ninth sequences (the one half of the peptide) are highly influential variables. Moreover, six variables were moderate influential (VIP > 0.8). These variables were $ADFQ_{7,2}$, $ADFQ_{7,9}$, $ADFQ_{2,3}$, $ADFQ_{4,7}$, $ADFQ_{1,2}$ and $ADFQ_{5,3}$ which belonged to the ninth, seventh, third and second sequences. This suggested that in addition to the high importance of one half the peptides, the AAs in the opposite of these residues are in the second order of importance.

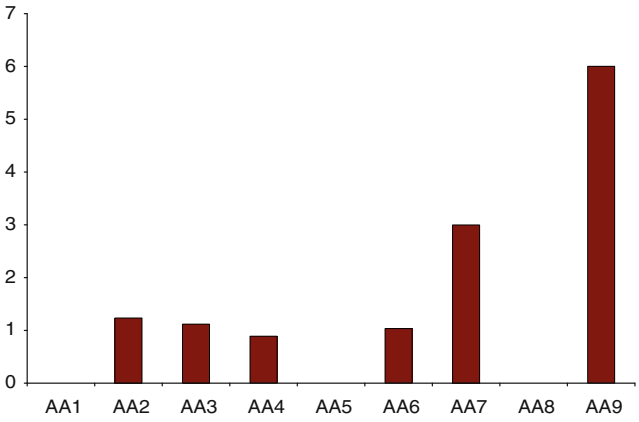

Fig. 7 Plot of RVIP for 9 sequences of the set of the HLA data set

Figure 7 shows the RVIP of the residues in the nonapeptide chain. Based on the data shown in this figure, the last sequence of peptide (i.e., the AAs in ninth residue) has the most important influence on the peptide activity and the seventh amino acid would be in the next order of importance. The influence of sixth, second, third and approximately fourth sequence also could be detected. The active site analysis of the HLA peptides was in agreement with the previous QSAR study on the data set as Doytchinova et al. (2005) determined the biggest role of the ninth sequence.

Conclusion

A new series of AA indices based on QTMS descriptors were proposed for peptide QSAR studies. The results showed that this new set of descriptors is a useful structure characterization method for peptide QSAR analysis, which has multiple advantages, such as definite physical and chemical meaning, good structural characterization ability and produces statistically significant QSAR models.

Most of the QSAR models proposed here not only exhibited good self-prediction power (interval validation), but also had a sufficient ability to predict the activity of the test set samples. Therefore, the descriptors proposed in this study could be useful in structure characterization and activity prediction of biological peptides, and would become a group of general parameters for QSAR analyses on polypeptides and proteins.

Since the four kinds of proposed QTMS-based AA indices are a powerful source of topological and quantum information, it seems that we could use these descriptors to obtain an active site analysis on the different kind of biologically and pharmaceutically important peptides. According to the results obtained in this work for short chain peptides, in the case of small chain peptides, the AAs located at the end peptide were very important in peptide activity. However, in longer chain peptides, like HLA peptides, the AA of the central sequences possessed the biggest role and those of the end sequences were in the second degree of importance.

Acknowledgments Financial support of this project by Research councils of Shiraz University and Shiraz University of Medical Sciences is appreciated.

References

Alsberg BK, Marchand-Geneste N, King RD (2000) A new 3D molecular structure representation using quantum topology with application to structure–property relationships. Chemom Intell Lab Syst 54:75–91

Alsberg BK, Marchand-Geneste N, King RD (2001) Modeling quantitative structure–property relationships in calculated reaction pathways using a new 3D quantum topological representation. Anal Chim Acta 446:3–13

Bader RFW (1990) Atoms in molecules: a quantum theory. Oxford University Press, Oxford

Bader RFW, Preston HJT (1969) The kinetic energy of molecular charge distributions and molecular stability. Int J Quantum Chem 3:327–347

Bader RFW, Nguyen-Dang TT, Tal Y (1981) A topological theory of molecular structure. Rep Prog Phys 44:893–948

Baumann K (2003) Cross-validation as the objective function for variable-selection techniques. Trends Anal Chem 22:395–406

Baumann K (2005) Chance correlation in variable subset regression: influence of the objective function, the selection mechanism, and ensemble averaging. QSAR Comb Sci 24:1033–1046

Bytheway I, Popelier PLA, Gillespie RJ (1996) Topological studies of the charge density of some group 2 metallocenes M(η5–C5H5)2 (M = Mg or Ca). Can J Chem 74:1059–1071

Chaudry UA, Popelier PLA (2003) Ester hydrolysis rate constant prediction from quantum topological molecular similarity (QTMS) descriptors. J Phys Chem A 107:4578–4582

Doytchinova IA, Walshe V, Borrow P, Flower DR (2005) Towards the chemometric dissection of peptide–HLA-A*0201 binding affinity: comparison of local and global QSAR models. J Comput-Aided Mol Des 19:203–212

Du QS, Huang RB, Chou KC (2008) Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. Curr Protein Pept Sci 9:248–260

Erikson L, Johansson E, Kettaneh-Wold N, Wold S (2001) Multi- and mega-variate data analysis. Principle and applications. Umetrics Academy, Umea

Hansch C, Hoekman D, Leo A, Weininger D, Selassie CD (2002) Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology. Chem Rev 102:783–812

Harding AP, Wedge DC, Popelier PLA (2009) pKa prediction from "Quantum Chemical Topology" descriptors. J. Chem Inf Model 49:1914–1924

Hemmateenejad B (2005) Correlation ranking procedure for factor selection in PC-ANN modeling and application to ADMETox evaluation. Chemom Intell Lab Syst 75:231–245

Hemmateenejad B, Mohajeri A (2007) Application of quantum topological molecular similarity descriptors in QSPR study of the *O*-methylation of substituted phenols. J Comput Chem 29:266–274

Hemmateenejad B, Akhond M, Miri R, Shamsipur M (2003) Genetic algorithm applied to the selection of factors in principal component-artificial neural networks: application to QSAR study of calcium channel antagonist activity of 1,4-dihydropyridines (nifedipine analogous). J Chem Inf Comput Sci 43:1328–1334

Hemmateenejad B, Mehdipour AR, Popelier PLA (2008) Quantum topological QSAR models based on the MOLMAP approach. Chem Biol Drug Des 72:551–563

Jenssen H, Gutteberg TJ, Rekdal Ø, Lejon T (2006) Prediction of activity, synthesis and biological testing of anti-HSV active peptides. Chem Biol Drug Des 68:58–66

Lin ZH, Long HX, Bo Z, Wang YQ, Wu YZ (2008) New descriptors of amino acids and their application to peptide QSAR study. Peptides 29:1798–1805

Mauri A, Ballabio D, Consonni V, Managanaro A, Todeschini R (2008) Peptides multivariate characterization using a molecular based approach. MATCH Commun Math Comput Chem 60:671–690

Mehdipour AR, Hemmateenejad B, Miri R (2007) QSAR studies on the anesthetic action of some polyhalogenated ethers. Chem Biol Drug Des 69:362–368

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

Mohajeri A, Hemmateenejad B, Mehdipour A, Miri R (2008) Modeling calcium channel antagonistic activity of dihydropyridine derivatives using QTMS indices analyzed by GA-PLS and PC-GA-PLS. J Mol Graph Model 26:1057–1065

O'Brien SE, Popelier PLA (2001) Quantum molecular similarity 3. QTMS descriptors. J Chem Inf Comput Sci 41:764–775

O'Brien SE, Popelier PLA (2002) Quantum topological molecular similarity. Part 4. A QSAR study of cell growth inhibitory properties of substituted (*E*)-1-phenylbut-1-en-3-ones. J Chem Soc Perkin Trans 2:478–483

Padron-Garcia JA, Alonso-Tarajano M, Alonso-Becerra E, Winterburn TJ, Yasser R, Kay J, Berry C (2009) Quantitative structure

activity relationship of IA3-like peptides as aspartic proteinase inhibitors. Proteins 75:859–869

Popelier PLA, Smith PJ (2006) QSAR models based on quantum topological molecular similarity. Eur J Med Chem 41:862–873

Popelier PLA, Chaudry UA, Smith PJ (2002) Quantum topological molecular similarity. Part 5: further development with an application to the toxicity of polychlorinated dibenzo-*p*-dioxins (PCDDs). J Chem Soc Perkin II:1231–1237

Popelier PLA, Chaudry UA, Smith PJ (2004) Quantitative structure–activity relationships of mutagenic activity from quantum topological descriptors: triazenes and halogenated hydroxyfuranones (mutagen-X) derivatives. J Comput-Aided Mol Des 18: 709–718

Rafat M, Shaik M, Popelier PLA (2006) Transferability of quantum topological atoms in terms of electrostatic interaction energy. J Phys Chem A 110:13578–13583

Ramos de Armas R, González Díaz H, Molina R, Uriarte E (2005) Stochastic-based descriptors studying biopolymers biological properties: extended MARCH-INSIDE methodology describing antibacterial activity of lactoferricin derivatives. Biopolymers 77:247–256

Roy K, Popelier PLA (2008a) Exploring predictive QSAR models using quantum topological molecular similarity (QTMS) descriptors for toxicity of nitroaromatics to *Saccharomyces cerevisiae*. QSAR Comb Sci 27:1006–1012

Roy K, Popelier PLA (2008b) Exploring predictive QSAR models for hepatocyte toxicity of phenols using QTMS descriptors. Bioorg Med Chem Lett 18:2604–2609

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

Tian F, Yang L, Lv F, Yang Q, Zhou P (2009) In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure–activity relationship approach. Amino Acids 36:535–554

Tong J, Liu S, Zhou P, Wu B, Li Z (2008) A novel descriptor of amino acids and its application in peptide QSAR. J Theor Biol 253:90–97

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

Zhou P, Chen X, Wu Y, Shang Z (2010) Gaussian process: an alternative approach for QSAM modeling of peptides. Amino Acids 38:199–212