Amino Acids (2015) 47:125–134
DOI 10.1007/s00726-014-1850-8

ORIGINAL ARTICLE

Quantitative sequence–activity modeling of antimicrobial hexapeptides using a segmented principal component strategy: an approach to describe and predict activities of peptide drugs containing L/D and unnatural residues

Saeed Yousefinejad · Mojtaba Bagheri · Ali Akbar Moosavi-Movahedi

Received: 18 March 2014 / Accepted: 3 October 2014 / Published online: 17 October 2014

Abstract The treatment of infections caused by multi-drugs resistant bacteria and fungi is a particular challenge. Whereas cationic antimicrobial peptides (CAPs) are considered as promising drug candidates for treatment of such superbugs, recent studies have focused on design of those peptides with increased bioavailability and stability against proteases. In between, applications of the quantitative structure–activity relationship (QSAR) studies which provide information on activities of CAPs based on descriptors for each individual amino acid are inevitable. However, the satisfactory results derived from a QSAR model depend highly on the choice of amino acid descriptors and the mathematical strategy used to relate the descriptors to the CAPs' activity. In this study, the quantitative sequence–activity modeling (QSAM) of 60 CAPs derived from O-$\underline{\text{W}}$-$\underline{\text{F}}$-I-$\underline{\text{F}}$-H(1-Bzl)-NH_2 sequence which showed excellent activities against a broad range of hazardous microorganisms: e.g., MR*SA*, MR*SE*, *E. coli* and *C. albicans*, is discussed. The peptides contained natural and non-natural amino acids (AAs) of the both isomers D and L. In this study, a segmented principal component strategy was performed on the structural descriptors of AAs to extract AA's indices. Our results showed that constructed models covered more than 82, 94, 80 and 78 % of the cross-validated variance of *C. albicans*, *MRSA*, *MRSE* and *E. coli* data sets, respectively. The results were also used to determine the important and significant AAs which are important in CAPs activities. According to the best of our knowledge, it is the first successful attempt in the QSAM studies of peptides containing both natural and non-natural AAs of the both L and D isomers.

Keywords Cationic antimicrobial peptides · Superbugs · Chemoinformatic · Quantitative sequence–activity modeling · Amino acid descriptors

Introduction

The fact of increasing use of antibiotics in immunosuppressant patients has resulted in the prevalence and drug resistance of bacterial/fungal superbugs (Guilhelmelli et al. 2013; Mayer et al. 2013). Cationic antimicrobial peptides (CAPs) as a wide category of compounds with their primitive defense mechanism could be an effective immune wall against the superbugs-associated infections (Zasloff 2002). CAPs are found in a wide range of eukaryotic organisms, which mostly show action by damaging bacterial/fungal cell membrane (Dathe and Wieprecht 1999; Bagheri et al. 2011).

Design and introduction of new CAPs with better therapeutic activity are always demanding and are in progress (Reddy et al. 2004; Junkes et al. 2011). On the other hand, because of growing emergence of bacterial/fungal superbugs some efforts have been started to introduce suitable CAPs as a new generation of antibiotics.

Some methods have been proposed to help the design of peptides and also analog peptide libraries with desired

Electronic supplementary material The online version of this article (doi:10.1007/s00726-014-1850-8) contains supplementary material, which is available to authorized users.

S. Yousefinejad (✉) · M. Bagheri · A. A. Moosavi-Movahedi
Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
e-mail: yousefinejad@ibb.ut.ac.ir; yousefinejad.s@gmail.com

A. A. Moosavi-Movahedi
Centre of Excellence in Biothermodynamics, University of Tehran, Tehran, Iran

activity for final synthesis in laboratory. For example, peptide sequence scanning (Savio et al. 2012; Jamieson et al. 2013) and quantitative sequence–activity modeling (QSAM) (Jonsson et al. 1993; Zhou et al. 2010) are two well-known techniques showing good potential in peptide design.

QSAM is an effective chemoinformatic technique, firstly proposed by (Jonsson et al. 1993), employing quantitative structure–activity relationship for biomolecules. In this approach, biosequence activities of biomolecules, e.g., therapeutic peptides, are linked to their functional/structural properties (Doytchinova et al. 2005; Hemmateenejad et al. 2011). The most important part of a structure–activity relationship in peptides is the definition and extraction of suitable descriptors or indices for the AAs in the peptide sequences. Different attempts have been done to propose new indices for AAs to use in the QSAM studies (Sneath 1966; Kidera et al. 1985; Hellberg et al. 1987; Sandberg et al. 1998; Raychaudhury et al. 1999; Zaliani and Gancia 1999; Mei et al. 2005; Tong et al. 2008; Lin et al. 2008; Yang et al. 2010; Hemmateenejad et al. 2011; Yousefinejad et al. 2012; Hemmateenejad et al. 2012). Most of these proposed AA indices used in the QSAM studies are for natural AAs; however, some of them are extended for non-natural AAs [e.g., Sandberg et al. (1998); Yang et al. (2010)].

Here, a QSAM study was performed on a series of linear hexa-CAPs with different activity profiles against bacterial/fungal superbugs using segmented principal component regression strategy (SPCR) (Hemmateenejad and Elyasi 2009). Since these CAPs contained simultaneously non-natural AAs and AAs of the both D- and L-isomers and no indices have been proposed before for some of these AAs, SPCR helped to extract potent indices for this kind of AAs and the data set of target CAPs.

Methods and materials

Data set

60 recently synthesized linear hexapeptides derived from O-W-F-I-F-H(1-Bzl)-NH$_2$ sequence which showed excellent activities against a broad range of hazardous microorganisms: e.g., *C. albicans*, MR*SA*, MR*SE* and *E. coli*, were taken from the literature (Sharma et al. 2010). All of these 60 peptides had defined antifungal activity against *C. albicans*. Among these 60 peptides, 48 had defined activity against "methicillin-resistant *Staphylococcus epidermidis*" (MR*SE*) and 40 of them were active against "methicillin-resistant *Staphylococcus aureus*" (MR*SA*), respectively. Also 28 of these 60 linear peptides showed activity against "*Escherichia coli*" (*E. coli*). All of the MR*SE*, MR*SA* and *E. coli* activities of the under-study peptides were adopted from the literature (Sharma et al. 2010). The data used in the current QSAM study consisted of activity of these linear hexapeptides (IC_{50} in mg/mL), against *C. albicans,* MR*SE,* MR*SA* and *E. coli*. The inhibition activities were converted to the logarithmic scale pIC_{50} ($-\log IC_{50}$) and then used for subsequent quantitative sequence–activity modeling as the dependent variables. The sequence of peptides and their experimental antimicrobial activities could be found in Table 1. It should be mentioned that the presentation of peptides in this table was arranged according to rational of AA replacement in the base peptide, i.e., O-W-F-I-F-H(1-Bzl)-NH$_2$.

Descriptor extraction and model development

One of the essential steps in structure–activity relationship studies is extraction of some numerical codes from the desired compounds which could describe the structural properties. Because the values of many descriptors are dependent on the bond lengths, bond angles and other geometrical characteristics, each AA structure was firstly drawn using Hyperchem (Version 8, Hyper Cube Inc.) and subsequently optimized by the semi-empirical AM1 method. This process obtained the most stable state of AAs according to the interactions between the chemical groups in a molecule, before determination of its molecular descriptors.

Then, the optimized structure of the AAs was transferred to the Dragon software (Version 2.1, Milano Chemometrics and QSAR research group, http://michem.disat.unimib.it/chm/) to extract the molecular descriptors. A total of 531 descriptors (Charge, Randic molecular profiles, Geometrical, Radial Distribution function, 3D-MoRSE, WHIM, GETAWAY) were calculated for each AA molecule. These descriptors were calculated after deleting the redundant and collinear descriptors (i.e., $R^2 > 0.95$). Another criterion which should be included in the generated descriptor was the ability of differentiating between the two isomers of an AA (D and L). It was necessary because in some of the peptides in the under-study data set, D and L conformers of some AAs were existed. So, the descriptors with high similarity in the D and L forms were deleted and finally 297 descriptors were remained for each AA molecule.

It is clear that the large number of descriptors for each AAs could not be logical for constructing the independent (descriptor) matrix for the peptides that consist of six AAs. In such case, we have a matrix containing a large number of descriptors ($6 \times 297 = 1{,}782$) and apparently the risk of obtaining chance models is increased if the number of original descriptors is increased significantly with respect to the number of molecules (Topliss and Costello 1972; Livingstone and Salt 2005). One of the strategies to decrease

Table 1 Sequences of 60 antimicrobial linear peptides used in this study which are based on the O–$\underline{W}$–$\underline{F}$–I-$\underline{F}$–H(1-Bzl)-NH_2 with their observed and predicted activities against *C. albicans*, Gram-positive MR*SA* and MR*SE* and Gram-negative *E. coli* bacteria obtained by the models based on 16 segments' SPCR indices

Peptide denotation	Sequence[a]	Activity								Rationale
		C. albicans		MR*SA*		MR*SE*		*E. coli*		
		PIC_{50} (Obs.)	PIC_{50} (Pred.)	PIC_{50} (Obs.)	pIC_{50} (Pred.)	pIC_{50} (Obs.)	pIC_{50} (Pred.)	pIC_{50} (Obs.)	pIC_{50} (Pred.)	
Diastereomers of original peptide sequence										
P1[b]	O-W-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	2.10	2.22	2.96	2.96	2.21	2.58	1.63	1.76	Trp → D-Trp
P2[c]	O-$\underline{W}$-F-I-$\underline{F}$-H(1-Bzl)-NH_2	1.87	1.77	1.64	1.99	2.42	2.37			Phe_1 → D-Phe_1
P3[b]	O-$\underline{W}$-$\underline{F}$-I-F-H(1-Bzl)-NH_2	1.69	1.78							Phe_2 → D-Phe_2
Substitution of the amino acid residue at 1st position										
P4[c,d]	R-$\underline{W}$-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	1.73	1.75	3.12	2.82	2.17	2.26	1.88	1.88	Arg → Orn
P5	K-$\underline{W}$-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	1.93	1.84	2.32	2.32	2.43	2.21	1.44	1.50	Lys → Orn
P6	H-$\underline{W}$-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	1.68	1.6							His → Orn
P7	H(1-Bzl)-$\underline{W}$-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	1.75	1.76							H(1-Bzl) → Orn
P8[d]	$\underline{K}$-$\underline{W}$-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	2.00	2.00	1.76	1.76	2.10	2.26	1.37	1.38	D-Lys → Orn
Substitution of the amino acid residue at 2nd position										
P9	O-$\underline{P}$-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	1.636	1.662							D-Pro → D-Trp
P10	O-Bal-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	2.027	2.019							Bal → D-Trp
P11	O-Nal-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	2.959	3.010	2.43	2.45	2.42	2.36			Nal → D-Trp
P12	O-F(4-CF_3)-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	2.143	2.127	1.79	1.78	1.47	1.56			F(4-CF_3) → D-Trp
P13[b,c]	O-F(4-F)-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	2.398	2.459	2.47	2.68	1.66	1.62			F(4-F) → D-Trp
P14	O-F(4-Me)-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	2.824	2.802	1.57	1.58	1.49	1.61	1.49	1.33	F(4-Me) → D-Trp
P15	O-F-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	1.827	1.791							Phe → D-Trp
P16[d]	O-Phg-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	1.951	1.957			1.55	1.91			Phg → D-Trp
P17	O-Cha-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	2.538	2.524	2.14	2.14	2.44	2.46			Cha → D-Trp
P18	O-I-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	1.684	1.780	1.64	1.64	1.94	2.12			Ile → D-Trp
P19[c]	O-Y-$\underline{F}$-I-$\underline{F}$-H(1-Bzl)-NH_2	2.347	2.295	2.07	2.17	2.70	2.22	2.85	2.81	Tyr → D-Trp
Substitution of the amino acid residue at 3rd position										
P20[c]	O-$\underline{W}$-$\underline{W}$-I-$\underline{F}$-H(1-Bzl)-NH_2	1.951	1.879	2.05	2.08	3.74	3.84			D-Trp → D-Phe_1
P21	O-$\underline{W}$-$\underline{I}$-I-$\underline{F}$-H(1-Bzl)-NH_2	2.018	2.104	1.77	1.78	2.07	2.08			D-Ile → D-Phe_1
P22	O-$\underline{W}$-$\underline{L}$-I-$\underline{F}$-H(1-Bzl)-NH_2	2.092	2.135	1.21	1.33	2.20	2.24			D-Leu → D-Phe_1
P23[c,d]	O-$\underline{W}$-$\underline{C}$-I-$\underline{F}$-H(1-Bzl)-NH_2	1.547	1.522	2.25	2.45	2.85	2.54	1.39	1.51	D-Cys → D-Phe_1
P24[d]	O-$\underline{W}$-F(4-CF_3)-I-$\underline{F}$-H(1-Bzl)-NH_2	2.398	2.304	3.03	3.05	3.09	2.77	1.67	1.71	F(4-CF_3) → D-Phe_1
P25[b]	O-$\underline{W}$-F(4-*t*Bu)-I-$\underline{F}$-H(1-Bzl)-NH_2	2.143	2.030	2.96	2.98	3.25	3.22	1.10	1.11	F(4-*t*Bu) → D-Phe_1
P26[d]	O-$\underline{W}$-F(4-F)-I-$\underline{F}$-H(1-Bzl)-NH_2	1.886	1.984	3.60	3.60	2.62	2.41	1.54	1.69	F(4-F) → D-Phe_1
P27	O-$\underline{W}$-F(4-Me)-I-$\underline{F}$-H(1-Bzl)-NH_2	2.180	2.125	2.74	2.67	3.70	3.65	1.79	1.50	F(4-Me) → D-Phe_1
P28[d]	O-$\underline{W}$-Bal-I-$\underline{F}$-H(1-Bzl)-NH_2	1.752	1.811	2.74	2.69	3.00	2.99	1.28	1.32	Bal → D-Phe_1
P29[b,d]	O-$\underline{W}$-W-I-$\underline{F}$-H(1-Bzl)-NH_2	2.292	2.069	1.74	1.61	2.96	2.69			Trp → D-Phe_1
P30	O-$\underline{W}$-Dip-I-$\underline{F}$-H(1-Bzl)-NH_2	2.194	2.191	2.92	2.93	3.24	3.25	1.45	1.56	Dip → D-Phe_1
P31[b,c]	O-$\underline{W}$-Phg-I-$\underline{F}$-H(1-Bzl)-NH_2	1.551	1.489	2.06	2.50	2.16	2.01	2.23	2.19	Phg → D-Phe_1
P32[b]	O-$\underline{W}$-I-I-$\underline{F}$-H(1-Bzl)-NH_2	1.680	1.544			1.11	1.12			Ile → D-Phe_1
P33	O-$\underline{W}$-L-I-$\underline{F}$-H(1-Bzl)-NH_2	1.914	1.837	1.68	1.54	1.94	1.84			Leu → D-Phe_1
P34[b]	O-$\underline{W}$-Nle-I-$\underline{F}$-H(1-Bzl)-NH_2	1.830	1.544	2.17	2.16	2.80	2.90			Nle → D-Phe_1
P35	O-$\underline{W}$-Cha-I-$\underline{F}$-H(1-Bzl)-NH_2	3.000	3.007	3.96	3.93	3.07	2.86	1.72	1.65	Cha → D-Phe_1
P36	O-$\underline{W}$-Y-I-$\underline{F}$-H(1-Bzl)-NH_2	1.646	1.778			1.33	1.40			Tyr → D-Phe_1
Substitution of the amino acid residue at 4th position										
P37[c]	O-$\underline{W}$-$\underline{F}$-Pfp-$\underline{F}$-H(1-Bzl)-NH_2	1.623	1.634	3.20	2.98	1.71	1.68			Pfp → Ile
P38	O-$\underline{W}$-$\underline{F}$-A(9-anth)-$\underline{F}$-H(1-Bzl)-NH_2	1.752	1.685			1.41	1.39			A(9-anth) → Ile
P39[b]	O-$\underline{W}$-$\underline{F}$-F(4-Me)-$\underline{F}$-H(1-Bzl)-NH_2	1.815	1.634	3.38	3.35	4.15	4.19			F(4-Me) → Ile

Table 1 continued

Peptide denotation	Sequence[a]	Activity								Rationale
		C. albicans		MR*SA*		MR*SE*		*E. coli*		
		PIC_{50} (Obs.)	PIC_{50} (Pred.)	PIC_{50} (Obs.)	pIC_{50} (Pred.)	pIC_{50} (Obs.)	pIC_{50} (Pred.)	pIC_{50} (Obs.)	pIC_{50} (Pred.)	
P40	O-$\underline{W}$-$\underline{F}$-Dip-$\underline{F}$-H(1-Bzl)-NH_2	1.646	1.754	1.91	1.88	3.15	3.20			Dip → Ile
P41[b]	O-$\underline{W}$-$\underline{F}$-Phg-$\underline{F}$-H(1-Bzl)-NH_2	1.572	1.522	2.03	2.01	2.92	2.76	1.44	1.43	Phg → Ile
P42	O-$\underline{W}$-$\underline{F}$-Cha-$\underline{F}$-H(1-Bzl)-NH_2	1.947	1.950					1.47	1.44	Cha → Ile
P43	O-$\underline{W}$-$\underline{F}$-Nle-$\underline{F}$-H(1-Bzl)-NH_2	2.155	2.148							Nle → Ile
P44	O-$\underline{W}$-$\underline{F}$-L-$\underline{F}$-H(1-Bzl)-NH_2	1.533	1.538							L → Ile
P45	O-$\underline{W}$-$\underline{F}$-$\underline{V}$-$\underline{F}$-H(1-Bzl)-NH_2	1.903	1.954	1.57	1.56	1.69	1.84			D-Val → Ile
P46[c,d]	O-$\underline{W}$-$\underline{F}$-$\underline{F}$-$\underline{F}$-H(1-Bzl)-NH_2	2.357	2.336	2.68	2.73	2.40	2.50	1.16	1.17	D-Phe → Ile
Substitution of the amino acid residue at 5th position										
P47	O-$\underline{W}$-$\underline{F}$-I-$\underline{W}$-H(1-Bzl)-NH_2	1.947	1.921			1.83	1.88	1.22	1.22	D-Trp → D-Phe_2
P48	O-$\underline{W}$-$\underline{F}$-I-$\underline{Y}$-H(1-Bzl)-NH_2	1.559	1.623							D-Tyr → D-Phe_2
P49	O-$\underline{W}$-$\underline{F}$-I-Dip-H(1-Bzl)-NH_2	1.762	1.890	1.45	1.44	2.02	2.23			Dip → D-Phe_2
P50	O-$\underline{W}$-$\underline{F}$-I-F(4-*t*Bu)-H(1-Bzl)-NH_2	2.237	2.220			1.56	1.54	1.57	1.55	F(4-*t*Bu) → D-Phe_2
P51	O-$\underline{W}$-$\underline{F}$-I-Cha-H(1-Bzl)-NH_2	1.793	1.780	1.84	1.85			1.47	1.49	Cha → D-Phe_2
P52	O-$\underline{W}$-$\underline{F}$-I-A-H(1-Bzl)-NH_2	1.717	1.703							Ala → D-Phe_2
Substitution of the amino acid residue at 6th position										
P53	O-$\underline{W}$-$\underline{F}$-I-$\underline{F}$-F(4-NH_2)-NH_2	1.815	1.803	2.39	2.39	2.48	2.32	2.19	2.16	F(4-NH_2) → H(1-Bzl)
P54[b]	O-$\underline{W}$-$\underline{F}$-I-$\underline{F}$-R-NH_2	1.544	1.692	2.80	2.70	3.92	3.92	1.78	1.87	Arg → H(1-Bzl)
P55	O-$\underline{W}$-$\underline{F}$-I-$\underline{F}$-K-NH_2	1.821	1.862			1.96	1.96			Lys → H(1-Bzl)
P56[d]	O-$\underline{W}$-$\underline{F}$-I-$\underline{F}$-O-NH_2	1.967	1.989	2.17	2.51	2.92	2.57	1.91	1.77	Orn → H(1-Bzl)
P57[b]	O-$\underline{W}$-$\underline{F}$-I-$\underline{F}$-Dab-NH_2	2.585	2.653	1.82	1.89	2.41	2.48	1.98	2.14	Dab → H(1-Bzl)
P58	O-$\underline{W}$-$\underline{F}$-I-$\underline{F}$-H-NH_2	1.896	1.800			2.33	2.36	1.88	1.83	His → H(1-Bzl)
P59	O-$\underline{W}$-$\underline{F}$-I-$\underline{F}$-$\underline{K}$-NH_2	1.928	1.920			2.02	2.05	2.82	2.78	D-Lys → H(1-Bzl)
P60[c]	O-$\underline{W}$-$\underline{F}$-I-$\underline{F}$-$\underline{H}$-NH_2	1.730	1.683	1.92	2.35	2.21	2.27	2.09	2.08	D-His → H(1-Bzl)

[a] The single letters representing D-amino acids residues were underlined. Chemical formulae of schematic

[b] The test set for *C. albicans*

[c] The test set for MR*SA*

[d] The test set for MR*SE*

the number of initial descriptors before variable selection is using principle component analysis (PCA). This is a good way to compress data in some principal components (PCs) without losing information. Then, these PCs could be utilized in the regression step after variable selection. In the current study, a segmented principal component regression (SPCR) was used which was proposed by Hemmateenejad and Elyasi (2009). The SPCR showed also good ability in the QSAR study of peptides (Hemmateenejad et al. 2012). According to the SPCR method, the original descriptors for all AAs were segmented into different groups based on their information similarity and then each segment was subjected to PCA. Segmentation of original descriptors could be done by different strategies like PCA, self-organization mapping (SOM) based on Kohonen network and classification, and fuzzy c-means clustering and K-means clustering (Hemmateenejad et al. 2013). In this study, PCA based on singular value decomposition algorithm was utilized as a simple segmentation strategy. The extracted PCs of each segment were then subjected to PC selection and the most relevant subset of PCs was used as the AA indices. Finally, the obtained indices of AAs in each peptide were arranged beside each other to construct the independent-variable matrix of linear hexapeptide data set.

To test of the models performance, in addition to cross-validation, about 20 % of the data (12 peptides out of 60 in *C. albicans* model, 10 peptides out of 48 and 40 in MR*SE* and MR*SA* models respectively) were selected as the external test set. These peptides were selected randomly based on descriptors spaces. Because of the limited number of peptides in the *E. coli* data set, none of the peptides were used as the external test set and the goodness of the model was checked only by cross-validation. To further evaluate the model performance, cross-validation, $\overline{r_{\mathrm{m}}^{2}}$ metrics,

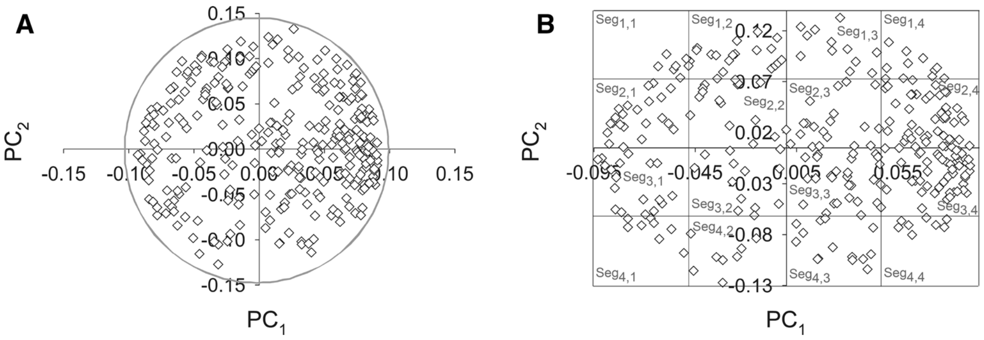

Fig. 1 **a** Distribution of the calculated descriptors for amino acids in the 2D-space of the first and second PCs in the descriptors' direction (loading 1 and loading 2). **b** Partitioning of the descriptors into 16 segments by vertical and *horizontal* lines to extract amino acid indices

external validation and *y* scrambling test were used. All the statistics and modeling procedures were done in MATLAB environment (version 7, Math work, Inc., http://www.mathworks.com, USA) and necessary routines were written in our laboratory.

Results and discussion

Model construction for *C. albicans* activity

It is worth mentioning that one of the important aspects of the peptides data set with antimicrobial activity, studied in the current work, is the presence of non-natural AAs as well as D and L conformers of some other AAs. So it was necessary to define a new source of indices for these non-formal kinds of AAs. First, the model development was performed for the *C. albicans* data set which was the largest one. As it was noted, we used segmented principal component regression (SPCR) to extend SPCR index (Hemmateenejad et al. 2012) for the AAs used in the synthesis of the 60 hexa peptides. The name and structure of AAs used in this work are represented in Table S1 (Supplementary materials).

After extraction of AA structural codes, they are classified into groups with similarity in their informational contents using principal component analysis. The distribution of descriptors of AAs in the two-dimensional space of loadings (loading 1 and loading 2) is classified into different groups. Then each group was separately subjected to principal component analysis (PCA). The extracted PCs were used as the indices of AAs in the hexapeptides data set.

The distribution of descriptors in the factor space of the first and the second loadings (PCs in the direction of descriptors) is presented in Fig. 1a. These two PCs covered about 70 % of variances of total 297 extracted descriptors. This pattern of distribution was used to classify the descriptors into different number of segments. According to Fig. 1a, it is clear that the distribution of variables related to the AAs was almost homogenous in both x- and y-directions of space. So both axes of distribution space (between minimum and maximum values) were divided into 1, 2, 4 and 5 portions and finally 1, 4, 16 and 25 segments were obtained, respectively. For instance the clustering of descriptors in 16 segments is shown in Fig. 1b. After segmentation (1, 4, 9, 16 and 25), the descriptors in each segment were separately subjected to singular value decomposition and the PCs with eigenvalues bigger than one were extracted from each segment. Total number of PCs in each segmentation step and number of PCs in each segment are abstracted in Table 2. It is clear that by increasing the number of segments, number of extracted PCs as the descriptive indices of AAs was increased. SPCR is a way to input the relevant information into the model and by increasing the number of segments, the probability of separating irrelevant and relevant information increased (Hemmateenejad and Elyasi 2009).

After arranging AA indices obtained with different number of indices for 60 linear hexapeptides in *C. albicans* data set, five matrices of independent variables with the size of 60 × *n* were constructed based on *n* extracted indices in five different segmentation strategies. Each of these matrices was subjected to stepwise variable selection to make QSAM models to relate peptides' independent variables to their activities against *C. albicans*. During the modeling procedure, 80 % of peptides (48 of 60) were randomly selected as the training set for model development and remaining 20 % were used as the test set to check the prediction ability of models. In the current work stepwise multiple linear regressions (SMLR) were used for variable selection and model development.

All statistics of five constructed models like training correlation coefficient (R^2_{cal}), root mean square error of calibration (RMSEC), correlation coefficient and root mean square error of cross-validation (Q^2_{cv} and RMSEcv, respectively) related to the *C. albicans* model are presented in Table S2. Correlation coefficient and root mean square error of the external test set (R^2_{test} and $\mathrm{RMSE}_{\mathrm{test}}$, respectively) are also summarized for each model in Table S2.

Validation of *C. albicans* models

According to the criteria noted in the literature, all the models had acceptable values of R^2_{cal} and Q^2 ($R^2_{\mathrm{cal}} > 0.6$

Table 2 Initial number of indices extracted by SPCR for each amino acid

Segments' no.	N^{a}_{EPCs}	$N^{b}_{EPCs\text{-}seg}$
1	24	$Seg_{1,1} = 24$
4	45	$Seg_{1,1} = 13$; $Seg_{2,1} = 12$; $Seg_{1,2} = 13$; $Seg_{2,2} = 7$
9	54	$Seg_{1,1} = 5$; $Seg_{2,1} = 6$; $Seg_{3,1} = 6$; $Seg_{1,2} = 12$; $Seg_{2,2} = 7$; $Seg_{3,2} = 5$; $Seg_{1,3} = 4$; $Seg_{2,3} = 6$; $Seg_{3,3} = 3$
16	62	$Seg_{1,1} = 2$; $Seg_{2,1} = 3$; $Seg_{3,1} = 5$; $Seg_{4,1} = 1$;$Seg_{1,2} = 7$; $Seg_{2,2} = 7$; $Seg_{3,2} = 4$; $Seg_{4,2} = 4$; $Seg_{1,3} = 7$; $Seg_{2,3} = 7$; $Seg_{3,3} = 4$; $Seg_{4,3} = 3$; $Seg_{1,4} = 1$; $Seg_{2,4} = 4$; $Seg_{3,4} = 2$; $Seg_{4,4} = 1$
25	64	$Seg_{1,1} = 1$; $Seg_{2,1} = 2$; $Seg_{3,1} = 3$;$Seg_{4,1} = 2$; $Seg_{5,1} = 1$; $Seg_{1,2} = 3$; $Seg_{2,2} = 4$; $Seg_{3,2} = 2$; $Seg_{4,2} = 4$; $Seg_{5,2} = 2$; $Seg_{1,3} = 6$; $Seg_{2,3} = 5$; $Seg_{3,3} = 5$; $Seg_{4,3} = 4$; $Seg_{5,3} = 2$; $Seg_{1,4} = 3$; $Seg_{2,4} = 5$; $Seg_{3,4} = 3$; $Seg_{4,4} = 2$;$Seg_{5,4} = 2$; $Seg_{1,5} = 0$;$Seg_{2,5} = 1$; $Seg_{3,5} = 1$; $Seg_{4,5} = 1$; $Seg_{5,5} = 0$

[a] Total number of extracted PCs for each amino acid

[b] Number of extracted PCs from each segment before variable selection; $Seg_{i,j}$ shows segments of the ith row and jth column (Fig. 1b)

and $Q^2 > 0.5$ (Golbraikh and Tropsha 2002a); however, the models based on segments = 4 and segments = 16 showed better statistics. It is also clear from Table S2 that all the models also show good predictability for the external test set. The squared regression coefficient between observed and predicted values of the train, cross-validation or test set peptides do not necessarily mean that the predicted values are very near to observed activity values (Roy et al. 2012). So an average modified term ($\overline{r^2_{m(cv)}}$), that was recently proposed (Roy et al. 2012), was calculated for cross-validation. According to this parameter, only models #2 and #4 (based on AA indices from 4 segments and 16 segments, respectively) were acceptable and model #4 was the best one. In model #4, $\overline{r^2_{m(cv)}}$ shows that 71 % of predicted values are very near to observed activity values of the hexapeptides. So, the number of segments equal to 16 was chosen as the optimum case for AA index extraction and QSAM of *C. albicans* hexapeptide data set. As noted previously, prediction ability of this model was also good and R^2_{test} and $RMSE_{test}$ had values of 0.87 and 0.42, respectively. Slopes k (slope of observed peptide activity vs. predicted activity) and k' (slope of observed peptide activity vs. predicted activity) (Golbraikh and Tropsha 2002b) related to regression line of model #4 through the origin were also calculated which were 0.97 and 1.0, respectively. Both values were in the range of 0.85–1.15 (Golbraikh and Tropsha 2002b) and already were near to middle of this range, showing another proof for the predictability of the proposed QSAM model.

In addition, to assess the risk of chance correlation, permutation test (*y* scrambling) was performed (Gramatica et al. 2007). In this way, the dependent variable of model (pIC_{50} of different peptides) was randomly shuffled 50 times and maximum correlation coefficient of cross-validation in permutation test (Q^2_{MP}) was calculated which was 0.14. It is clear that this value is significantly lower than Q^2 of original model and proves that the model was not chancy.

To have a comparison of the statistics of model #4 (based on 16 segments) with the modes constructed for other kinds of antimicrobial activities, which discussed future, the parameters of model #4 are abstracted in Table 2. A complete list of the peptide sequences in the training and test sets, their experimental and predicted $-\log(IC_{50})$ values by stepwise MLR and their residual values of prediction are listed in Table 1. To have a visual presentation, the plot of predicted vs. observed activity of peptides against *C. albicans* is presented in Fig. 2 (up-left). The standardized residual distribution plot of Fig. 2 related to *C. albicans* is also shown in Fig S2 (up-left). The propagation of the standardized residuals in both sides of zero line indicates that no systematic error exists in the constructed model for *C. albicans* activity. On the other hand, all the standardized residual values of training and test set were located in the acceptable range of $\pm 3\sigma$, where σ is the standard deviation.

Highlighted zone in linear antifungal peptides

One of the goals of this research was discovering "which residue(s) in the peptide sequence has (have) more effect on the bioactivity?" To do so, we look at the indices (PCs obtained by SPCR) that were selected after variable selection to include in the model. First of all, the percent of existence of variables (PCs) related to each residue in the peptide sequence for the selected model (model #4) was calculated which is shown graphically in Fig. 3 (up-left). As it is clear from this figure, the percent of existence in model (PEM) related to residue 2 and 3 was bigger than remaining AAs. However, the role of AA2 was bigger than AA3. On the other hand, AA4 also has the next order of importance in activity against *C. albicans*. So it could be concluded that the manipulation in this position of these kinds of antimicrobial linear hexapeptides might have a significant effect on their activity against *C. albicans* superbugs.

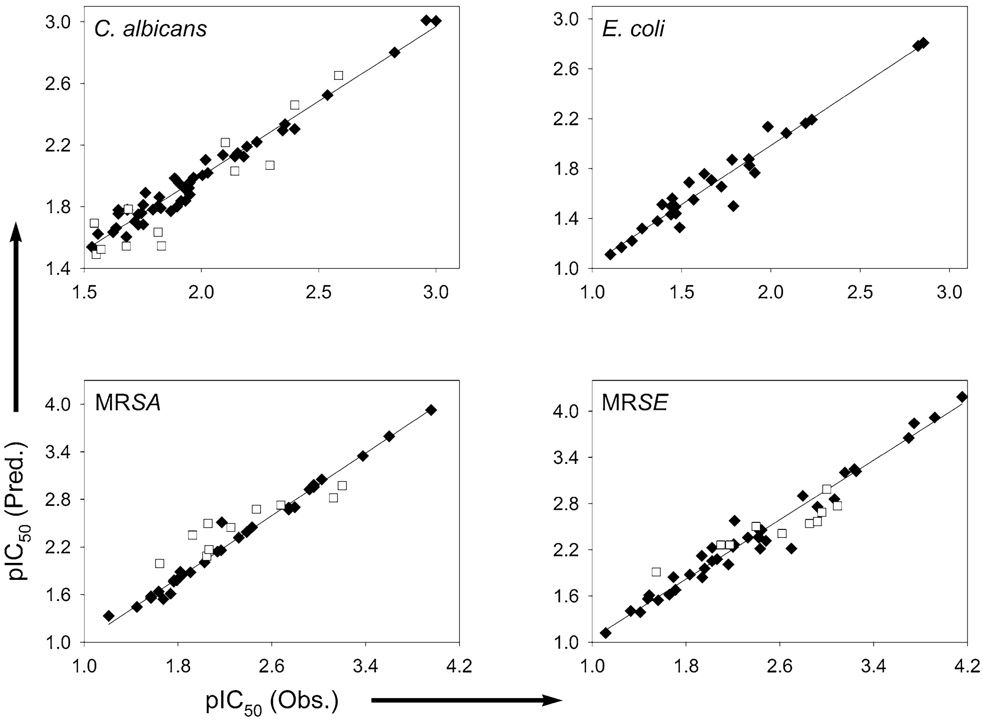

Fig. 2 Plot of different calculated antimicrobial activity against observed values

Model development for activity against MR*SE*, MR*SA* and *E. coli*

After showing the effectiveness of indices based on 16 segments (4 × 4) in modeling the anti-*C. albicans* activity, the potential of proposed indices for constructing the QSAMs for the activity of these hexapeptides against MR*SE*, MR*SA* and *E. coli* was also checked and in all cases good results were achieved. The statistics of calibration, cross-validation, external test set and y-randomization test are available in Table 3. According to the results, all models showed good statistics in both calibration and prediction points of view. However, the model of *MRSA* was the best one especially in cross-validation parameters. The average r_{m}^{2} metrics of cross-validation $(\overline{r_{\mathrm{m(cv)}}^{2}})$ for the model of *MRSA* activity was 0.88 which shows stability of the model (Roy et al. 2012).

By applying the extracted indices based on segmentation of PCs space, it was shown that this kind of parameters has enough potential in prediction of other antimicrobial activities of the linear hexapeptides studied in the current work. The plot of observed vs. predicted activity of MR*SE*, MR*SA* and *E. coli* model is represented in Fig. 2. The standardized residual distribution plots of these models is also represented in Fig. S2. In all models (MR*SE*, MR*SA* and *E. coli*), the standardized residual values of test and/or training set were located in the acceptable range of ±3σ. The propagation of the standardized residuals in both sides of zero line indicates that no serous systematic error exists in the developed models for the under-study antimicrobial activities. The observed (experimental) and predicted values of *MRSE*, *MRSA* and *E. Coli* activities are included in Table 1. It is worth mentioning that the indices based on other number of segments in SPCR (1, 4, 9 and 25) were also checked to model the *MRSE*, *MRSA* and *E. Coli* activities but similar to *C. albicans* case; it was observed that the index #4 (based on 16 segments) was the best one (data are not shown).

Highlighted zone of antimicrobial peptides with activity against MR*SE*, MR*SA* and *E. coli*

By similar manner which was noted previously for the highlighted zone of anti-*C. albicans* peptides, the PEM of each AA of peptides of three data sets (MR*SE*, MR*SA* and *E. coli*) was calculated which is plotted in Fig. 3. As it is clear from Fig. 3 (down-left and down-right), in MRSE and MRSA data sets, similar to anti-*C. albicans* peptides AA2, AA3 and A4 were the highlighted residues in antimicrobial activities. However in MRSE and MRSA models AA3 was the most important one.

The highlighted zone in the *E. coli* data set which is active against *E. coli* as Gram-negative bacteria was slightly different with the others. According to which is shown in Fig. 3 (up-right), in the anti-*E. coli* peptides, AA3 and AA6 were the most important and effective residues in their activities. However, here again the AA3 was the most important residue. It is worth mentioning that in almost all data sets, the AA3 was the most effective residue on the antimicrobial activities. So the focus on this residue could

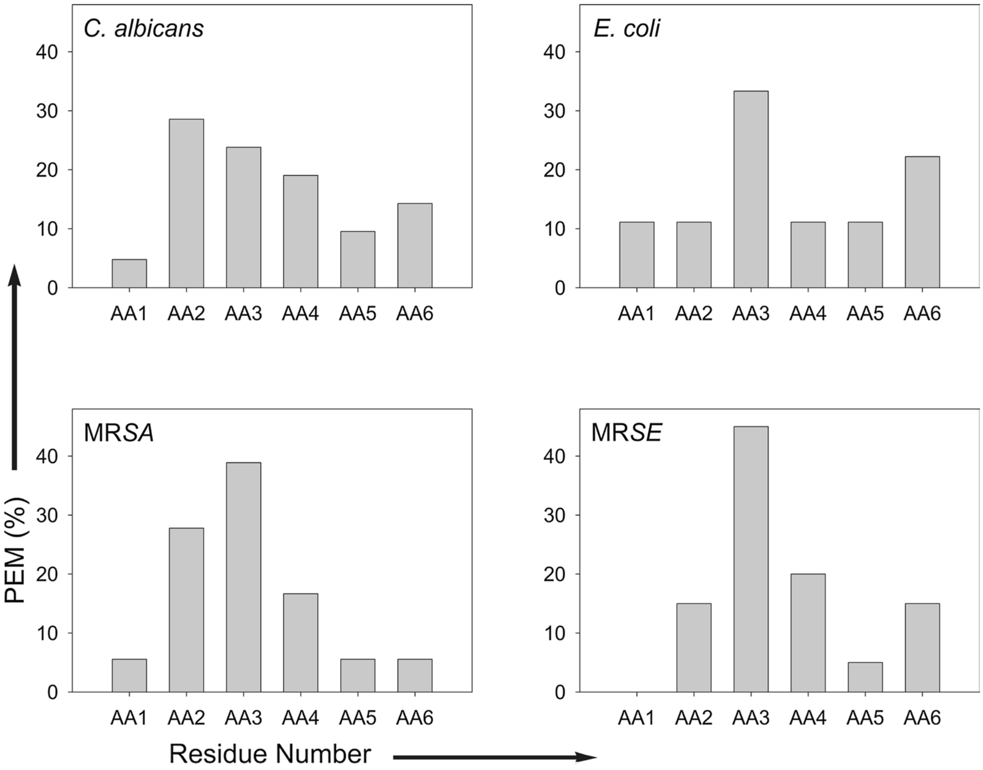

Fig. 3 Percent of existence of indices related to each residue of CAPs in the final models built for different antimicrobial activities

Table 3 Statistics of models developed for the peptides activities against *C.albicans*, Gram-positive *MRSA* and *MRSE* and Gram-negative *E. coli* bacteria based on amino acids indices obtained by SPCR in 16 segments

Entry no.	Data set	R^2_{cal} [a]	RMSEC[b]	Q[2c]	$\mathrm{RMSE}^{\mathrm{d}}_{\mathrm{cv}}$	$\overline{r^2_{\mathrm{m(cv)}}}$ [e]	R^2_{test} [f]	$\mathrm{RMSE}^{\mathrm{g}}_{\mathrm{test}}$	Q^2_{MP} [h]
1	*C. albicans*	0.97	0.17	0.82	0.43	0.71	0.87	0.42	0.14
2	*MRSA*	0.98	0.07	0.94	0.26	0.88	0.90	0.53	0.25
3	*MRSE*	0.97	0.19	0.80	0.56	0.66	0.86	0.52	0.21
4	*E. coli*	0.95	0.22	0.78	0.48	0.64	ND	ND	0.19

ND not determined

[a] Correlation coefficient of calibration

[b] Root mean square error of calibration

[c] Correlation coefficient of cross-validation

[d] Root mean square error of cross-validation

[e] Average r^2_{m} metrics of cross-validation

[f] Correlation coefficient of prediction (test set)

[g] Root mean square error of prediction (test set)

[h] Maximum cross-validation correlation coefficient for Y-randomization (permutation) test

be useful in designing linear peptides with general antimicrobial properties.

Brief description of model

Another finding from the obtained model was looking at original descriptors more weighted in PCs (indices) related to the highlighted AA(s) which is more important in peptide activity changes. For this purpose, the loadings of the selected PCs were analyzed. For each selected PC of the target residue in each segment of SPCR step, the elements of its loading indicate the importance of the descriptors of that segment in construction of selected PC (AA's indices). So the original descriptors related to maximum value of loadings were extracted. Table S2 (supplementary information) contains descriptors of the highlighted zone which had the highest loading value and their definitions for all data sets (*C. albicans,* MR*SE*, MR*SA* and *E. coli*).

Table 4 Important parameters of the most important AA of peptides and their relative effect on peptide activity

Amino acids' no.	Bacteria/yeast			
	C. albicans	MR*SA*	MR*SE*	*E. coli*
AA_2	V_w (±)[a], m_a (+), X (−)			
AA_3		V_w (±)[a], m_a (−), e (+)	V_w (−), α (+), σ (−)	V_w (−), σ (−)

The effect of each property (negative, positive or both) on the antimicrobial activity is represented in the parenthesis

V_w atomic van der Waals volume, m_a atomic mass, X atomic Sanderson electronegativity, α atomic polarizability, σ molecular symmetry, e molecular eccentricity

[a] Negative is more optimum

It should be noted that almost all properties extracted as the most important features of the most highlighted AA in these peptide data sets are weighted with different properties like atomic mass, atomic van der Waals volumes and atomic Sanderson electronegativity. For example, "Mor30 m" and "H8 m" show that AAs in the 2nd position of peptide sequence with elements of the higher atomic mass have a relatively positive effect on pIC_{50} (negative effect on peptide *C. albicans* activity). Negative sign of the score of "E3e", which is a descriptor weighted by atomic van der Waals volume, probably shows that AAs with elements of lower atomic van der Waals volumes result in higher pIC_{50} (lower activity against *C. albicans*). On the other hand by looking at the descriptors weighted by atomic van der Waals volume and sign of coefficients of their original scores, it is clear that R3v+ shows negative effect on pIC_{50} while RDF030v shows a positive effect. So, it could be concluded that the van der Waals volumes might be of an optimum value (i.e., not high and not low) in this situation (See Table S3). Other descriptions of important parameters of the most highlighted AAs (i.e., 3rd residue) of other data sets are also presented in Table S3. The important properties implied by the important properties with their effect (positive, negative or optimum of both) for all four models are abstracted in Table 4. As another example from results it seems that the van der Waals volume of the AA's Atoms in the 3rd position of peptide sequence represents a relative negative effect on pIC_{50} (positive effect on peptide's antimicrobial activity).

Conclusions

A multiparameter quantitative sequence–activity relationship model was proposed for a series of linear hexapeptide antibiotics to predict their antimicrobial activity against *C. albicans,* MR*SA,* MR*SE* and *E. coli*. However, the structure of these peptides contained non-natural and very structurally similar AAs (D and L-isomers); the obtained model had enough ability to predict and model the activities. This quantitative sequence–antifungal activity model was validated with different statistical and chemometrics approaches. Also, information on the highlighted zone of the hexapeptides (representing a general part of peptide structure with highest impact upon antifungal activity) was obtained. It was shown that the 2nd and 3rd residues in the peptide sequence might highlight the importance in controlling the total antimicrobial activity. However, the role of 2nd residue was more significant in *C. albica*s activity and 3rd residue might be more important in other kinds of antimictobial properties (against MR*SA*, MR*SE* and *E. coli*). So if we want to have a peptide with multifunctionality against different kinds of microbes, focus on the 3rd AA is important. The applied strategy in the current work, which imports the properties of each peptide residue separately in calculation, makes it possible to show the importance of different parts of sequence in peptide activities. However, the data obtained by our model are well supported by the experimental results (Sharma et al. 2010), suggesting that the cationic N-terminus of the linear hexapeptides has an important role for peptide interaction with both the negatively charged components of outer membrane of Gram-negative and Gram-positive bacteria, i.e., lipopolysaccharide and peptidoglycan, as well as the fungal membrane composed of zwitterionic phospholipids. This effect was more pronounced at 2nd and 3rd residues for *C. albicans* and the bacteria, respectively. Most likely, the difference in the selectivity of CAPs against bacteria and fungi is related to the charge and physiochemical properties of the microorganisms' cell membranes (Zasloff 2002). On the other hand, some properties like atomic mass, electronegativity and van der Waals volume of the highlight zone of these peptides might have a significant role in their antifungal activity. In this work, not only the potential of SPCR in the predictive QSAM of linear peptides with relatively similar structures was shown, but also the results of the work could be used to direct the synthesis of this kind of important antibiotic peptides toward representing better activities.

Acknowledgments The supports of Iran National Elites Foundation (INEF), Iran National Science Foundation (INSF), Center of

Excellence in Biothermodynamics (CEBiotherm), and University of Tehran are gratefully acknowledged.

Conflict of interest The authors state that they have no conflict of interest.

References

Bagheri M, Keller S, Dathe M (2011) Interaction of W-substituted analogs of cyclo-RRRWFW with bacterial lipopolysaccharides: the role of the aromatic cluster in antimicrobial activity. Antimicrob Agents Chemother 55:788–797. doi:10.1128/AAC.01098-10

Dathe M, Wieprecht T (1999) Structural features of helical antimicrobial peptides: their potential to modulate activity on model membranes and biological cells. Biochim Biophys Acta 1462:71–87. doi:10.1016/S0005-2736(99)00201-1

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. doi:10.1007/s10822-005-3993-x

Golbraikh A, Tropsha A (2002a) Beware of q2! J Mol Graph Model 20:269–276. doi:10.1016/S1093-3263(01)00123-1

Golbraikh A, Tropsha A (2002b) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Mol Divers 5:231–243. doi:10.1023/A:1021372108686

Gramatica P, Giani E, Papa E (2007) Statistical external validation and consensus modeling: a QSPR case study for Koc prediction. J Mol Graph Model 25:755–766. doi:10.1016/j.jmgm.2006.06.005

Guilhelmelli F, Vilela N, Albuquerque P et al (2013) Antibiotic development challenges: the various mechanisms of action of antimicrobial peptides and of bacterial resistance. Front Microbiol 4:353. doi:10.3389/fmicb.2013.00353

Hellberg S, Sjoestroem M, Skagerberg B, Wold S (1987) Peptide quantitative structure–activity relationships, a multivariate approach. J Med Chem 30:1126–1135. doi:10.1021/jm00390a003

Hemmateenejad B, Elyasi M (2009) A segmented principal component analysis–regression approach to quantitative structure–activity relationship modeling. Anal Chim Acta 646:30–38. doi:10.1016/j.aca.2009.05.003

Hemmateenejad B, Yousefinejad S, Mehdipour AR (2011) Novel amino acids indices based on quantum topological molecular similarity and their application to QSAR study of peptides. Amino Acids 40:1169–1183. doi:10.1007/s00726-010-0741-x

Hemmateenejad B, Miri R, Elyasi M (2012) A segmented principal component analysis–regression approach to QSAR study of peptides. J Theor Biol 305:37–44. doi:10.1016/j.jtbi.2012.03.028

Hemmateenejad B, Karimi S, Mobaraki N (2013) Clustering of variables in regression analysis: a comparative study between different algorithms. J Chemom 27:306–317. doi:10.1002/cem.2513

Jamieson AG, Boutard N, Sabatino D, Lubell WD (2013) Peptide scanning for studying structure–activity relationships in drug discovery. Chem Biol Drug Des 81:148–165. doi:10.1111/cbdd.12042

Jonsson J, Norberg T, Carlsson L et al (1993) Quantitative sequence–activity models (QSAM)—tools for sequence design. Nucleic Acids Res 21:733–739. doi:10.1093/nar/21.3.733

Junkes C, Harvey RD, Bruce KD et al (2011) Cyclic antimicrobial R-, W-rich peptides: the role of peptide structure and E. coli outer and inner membranes in activity and the mode of action. Eur Biophys J 40:515–528. doi:10.1007/s00249-011-0671-x

Kidera A, Konishi Y, Oka M et al (1985) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J Protein Chem 4:23–55. doi:10.1007/BF01025492

Lin Z, Long H, Bo Z et al (2008) New descriptors of amino acids and their application to peptide QSAR study. Peptides 29:1798–1805. doi:10.1016/j.peptides.2008.06.004

Livingstone DJ, Salt DW (2005) Judging the significance of multiple linear regression models. J Med Chem 48:661–663. doi:10.1021/jm049111p

Mayer FL, Wilson D, Hube B (2013) Candida albicans pathogenicity mechanisms. Virulence 4:119–128. doi:10.4161/viru.22913

Mei H, Liao ZH, Zhou Y, Li SZ (2005) A new set of amino acid descriptors and its application in peptide QSARs. Biopolymers 80:775–786. doi:10.1002/bip.20296

Raychaudhury C, Banerjee A, Bag P, Roy S (1999) Topological shape and size of peptides: identification of potential allele specific helper T cell antigenic sites. J Chem Inf Model 39:248–254. doi:10.1021/ci980052w

Reddy KVR, Yedery RD, Aranha C (2004) Antimicrobial peptides: premises and promises. Int J Antimicrob Agents 24:536–547. doi:10.1016/j.ijantimicag.2004.09.005

Roy K, Mitra I, Kar S et al (2012) Comparative studies on some metrics for external validation of QSPR models. J Chem Inf Model 52:396–408. doi:10.1021/ci200520g

Sandberg M, Eriksson L, Jonsson J et al (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem 41:2481–2491. doi:10.1021/jm9700575

Savio AS, Acosta OR, Pérez HG et al (2012) Enhancement of the inhibitory effect of an IL-15 antagonist peptide by alanine scanning. J Pept Sci 18:25–29. doi:10.1002/psc.1411

Sharma RK, Sundriyal S, Wangoo N et al (2010) New antimicrobial hexapeptides: synthesis, antimicrobial activities, cytotoxicity, and mechanistic studies. Chem Med Chem 5:86–95. doi:10.1002/cmdc.200900330

Sneath PHA (1966) Relations between chemical structure and biological activity in peptides. J Theor Biol 12:157–195. doi:10.1016/0022-5193(66)90112-3

Tong J, Liu S, Zhou P et al (2008) A novel descriptor of amino acids and its application in peptide QSAR. J Theor Biol 253:90–97. doi:10.1016/j.jtbi.2008.02.030

Topliss JG, Costello RJ (1972) Chance correlations in structure–activity studies using multiple regression analysis. J Med Chem 15:1066–1068. doi:10.1021/jm00280a017

Yang L, Shu M, Ma K et al (2010) ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. Amino Acids 38:805–816. doi:10.1007/s00726-009-0287-y

Yousefinejad S, Hemmateenejad B, Mehdipour AR (2012) New autocorrelation QTMS-based descriptors for use in QSAM of peptides. J Iran Chem Soc 9:569–577. doi:10.1007/s13738-012-0070-y

Zaliani A, Gancia E (1999) MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies. J Chem Inf Model 39:525–533. doi:10.1021/ci980211b

Zasloff M (2002) Antimicrobial peptides of multicellular organisms. Nature 415:389–395. doi:10.1038/415389a

Zhou P, Chen X, Wu Y, Shang Z (2010) Gaussian process: an alternative approach for QSAM modeling of peptides. Amino Acids 38:199–212. doi:10.1007/s00726-008-0228-1