Electronic Supporting Material on the publication entitled

In silico rational design and virtual screening of bioactive peptides based on QSAR modeling

Mehri Mahmoodi-Reihani [a], Fatemeh Abbasitabar [b,*], Vahid Zare-Shahabadi [a]

[a] **Department of Chemistry, Mahshahr Branch, Islamic Azad University, Mahshahr, Iran**

[b] **Department of Chemistry, Marvdasht Branch, Islamic Azad University, Marvdasht, Iran**

Table of Contents

Table S1. AA indices used in the construction of the ACE QSAR model.

AA index	min	max
Signal sequence helical potential (Argos et al., 1982)	0.05	3.23
Short and medium range non-bonded energy per residue (Oobatake-Ooi, 1977)	-14.42	-5.46
Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)	-0.43	0.33
Weights for coil at the window position of 1 (Qian-Sejnowski, 1988)	-0.99	2.27
Information measure for extended without H-bond (Robson-Suzuki, 1976)	-3.70	7.40
Information measure for N-terminal turn (Robson-Suzuki, 1976)	-4.30	6.50
Alpha helix propensity of position 44 in T4 lysozyme (Blaber et al., 1993)	-2.50	0.96
Hydrophobicity coefficient in RP-HPLC, C18 with 0.1% TFA/MeCN/H_2O (Wilce etal. 1995)	-2.24	4.80

Table S2. Calculated numerical descriptive matrix (NDM) for ACE data set together with the experimental and predicted biological activities.

	Sequence	Numerical descriptive vector			Experimental	Predicted
Training	VVV	-16.059	-17.474	-16.059	1.63	1.951
Training	RPG	-15.869	-18.640	-10.460	3.09	2.913
Training	GRP	-10.709	-16.374	-18.336	0.48	1.191
Training	LLL	-19.907	-21.660	-19.907	1.35	1.561
Training	GLG	-12.276	-18.156	-12.276	2.45	2.354
Training	LGL	-18.060	-14.667	-18.060	1.52	1.728
Training	FGG	-18.470	-12.909	-10.830	2.79	3.002
Training	GFG	-12.157	-19.112	-12.157	2.53	2.370
Training	GGF	-10.830	-12.909	-18.470	1.11	1.162
Training	FFG	-20.986	-21.518	-12.740	2.71	2.875
Training	FGF	-19.376	-14.717	-19.376	1.29	1.589
Training	GFF	-12.740	-21.518	-20.986	1.02	0.889
Training	GGG	-10.452	-11.373	-10.452	2.61	2.520
Training	GYG	-11.467	-18.638	-11.467	2.33	2.441
Training	GGY	-10.405	-12.211	-18.099	1.35	1.195
Training	YGY	-19.194	-13.875	-19.194	1.82	1.605
Training	GYY	-11.956	-21.384	-21.008	1.07	0.832
Training	YYY	-21.965	-23.900	-21.965	1.54	1.352
Training	FIV	-20.314	-18.730	-16.134	2.04	2.231
Training	FPP	-20.214	-22.188	-20.771	1.50	1.434
Training	FPK	-20.046	-20.149	-15.551	2.45	2.319
Training	PFP	-20.270	-21.894	-20.270	1.74	1.523
Training	RRR	-16.445	-17.894	-16.445	1.77	1.912
Training	PPP	-21.211	-23.080	-21.211	1.86	1.429
Training	FFF	-21.809	-23.731	-21.809	1.20	1.368
Training	RGP	-14.714	-11.989	-17.310	1.73	1.622
Training	PGR	-17.310	-11.989	-14.714	2.67	2.247
Training	GGV	-10.390	-11.634	-13.873	1.99	1.924
Training	GVV	-10.928	-16.035	-15.508	1.82	1.694
Training	PPG	-20.098	-20.235	-10.251	3.18	3.241
Training	PGG	-16.635	-10.795	-9.652	3.14	3.074
Training	PGP	-18.073	-11.965	-18.073	1.82	1.716
Training	GPG	-10.021	-17.042	-10.021	2.65	2.588
Training	GGP	-9.652	-10.795	-16.635	1.28	1.392
Training	PGI	-17.183	-11.993	-14.779	2.23	2.227
Training	KPK	-15.462	-18.690	-15.462	2.63	2.019
Training	ADA	-12.482	-15.568	-12.482	2.17	2.322
Training	GEG	-11.098	-15.803	-11.098	2.28	2.469
Training	LEL	-18.749	-18.799	-18.749	1.19	1.672
Training	RGP	-14.714	-11.989	-17.310	1.73	1.622
Training	PIP	-19.358	-17.903	-19.358	1.69	1.604
Training	FPF	-20.665	-21.651	-20.665	1.32	1.481
Training	KPF	-15.551	-20.149	-20.046	1.51	1.237
Training	VYP	-16.493	-21.952	-20.309	0.82	1.262
Training	YPF	-21.324	-22.213	-20.575	1.60	1.543
Test	LGG	-17.406	-13.060	-11.014	2.49	2.899
Test	GGL	-11.014	-13.060	-17.406	1.63	1.359
Test	LLG	-19.297	-19.979	-12.822	2.33	2.741
Test	GLL	-12.822	-19.979	-19.297	1.47	1.181
Test	YGG	-18.099	-12.211	-10.405	3.07	3.048
Test	YYG	-21.008	-21.384	-11.956	2.79	3.012
Test	LDL	-18.636	-18.192	-18.636	1.42	1.681
Test	VIF	-16.134	-18.730	-20.314	0.78	1.224
Test	RPF	-16.590	-20.826	-20.176	1.59	1.287
Test	PPF	-20.771	-22.188	-20.214	1.68	1.568

Table S3. Physicochemical indices of amino acids used to calculate numerical descriptive vector for bitter peptide

AA index	Min	Max
STERIMOL maximum width of the side chain (Fauchere et al., 1988)	1.00	6.72
Retention coefficient in HPLC, pH 2.1 (Meek, 1980)	-7.50	18.10
Optimized side chain interaction parameter (Oobatake et al., 1985)	-0.78	10.68
Surrounding hydrophobicity in alpha-helix (Ponnuswamy et al., 1980)	10.98	15.36
Average relative fractional occurrence in AL(i-1) (Rackovsky-Scheraga, 1982)	0.00	6.49
RF rank (Zimmerman et al., 1968)	2.80	18.80
Surface composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)	0.42	9.67
Diameter (maximum eccentricity) (Karkbara-Knisley, 2016)	0.00	14.00

Table S4. Calculated NDM for BBT data set together with the experimental and predicted biological activities.

	Sequence	NDV		Experimental	Predicted
Training	GV	-1.456	-1.468	1.13	1.31
Training	GL	-1.536	-1.872	1.68	1.68
Training	GI	-1.493	-1.650	1.7	1.68
Training	GP	-1.516	-1.706	1.35	1.47
Training	GF	-1.572	-2.099	1.8	2.01
Training	GW	-1.600	-2.223	1.89	1.97
Training	AV	-1.461	-1.473	1.16	1.17
Training	AL	-1.561	-1.890	1.7	1.58
Training	AF	-1.606	-2.120	1.72	1.91
Training	VG	-1.468	-1.456	1.19	1.30
Training	VA	-1.473	-1.461	1.16	1.18
Training	VV	-1.481	-1.481	1.71	1.39
Training	VL	-1.602	-1.912	2	1.86
Training	LG	-1.872	-1.536	1.72	1.72
Training	LA	-1.890	-1.561	1.72	1.65
Training	LL	-2.085	-2.085	2.35	2.42
Training	LF	-2.164	-2.330	2.75	2.81
Training	LY	-1.951	-1.704	2.46	2.38
Training	IG	-1.650	-1.493	1.68	1.73
Training	IA	-1.663	-1.508	1.68	1.67
Training	IV	-1.681	-1.540	2.05	1.93
Training	II	-1.759	-1.759	2.26	2.50
Training	IP	-1.775	-1.808	2.4	2.11
Training	IW	-1.947	-2.367	3.05	2.93
Training	ID	-1.489	-1.547	1.37	1.28
Training	IQ	-1.444	-1.725	1.49	1.57
Training	IE	-1.436	-1.782	1.37	1.36
Training	IK	-1.351	-2.187	1.65	1.65
Training	IS	-1.555	-1.439	1.49	1.46
Training	IT	-1.513	-1.455	1.49	1.82
Training	PA	-1.718	-1.531	1.32	1.34
Training	PL	-1.882	-2.013	2.22	2.05
Training	PI	-1.808	-1.775	2.33	2.07
Training	FG	-2.099	-1.572	1.77	2.10
Training	FL	-2.330	-2.164	2.87	2.85
Training	FP	-2.249	-1.948	2.7	2.51
Training	FF	-2.415	-2.415	3.1	3.27
Training	WW	-2.586	-2.586	3.6	3.50
Training	YL	-1.704	-1.951	2.4	2.38
Training	SL	-1.424	-1.756	1.49	1.39
Test	GY	-1.473	-1.545	1.77	1.65
Test	LW	-2.206	-2.457	3.4	2.87
Test	IL	-1.837	-2.000	2.26	2.46
Test	IN	-1.489	-1.547	1.49	1.55
Test	PY	-1.767	-1.654	1.8	2.02
Test	PF	-1.948	-2.249	2.8	2.44
Test	FY	-2.187	-1.767	3.13	2.85
Test	WE	-1.885	-1.526	1.56	1.79

Table S5. Statistical analysis of the selected QSAR model for BTT data set.

Regression coefficient	SE	*t-value*	*P*–value
1.92	0.024	78.60	8.91×10^{-43}
-0.33	0.031	-10.74	6.26×10^{-13}
-0.32	0.031	-10.40	1.57×10^{-12}

Table S6. QSAR models for HLA data set obtained using different sets of the AA indices.

Number of used AA indices	$R^2_{training}$	R^2_{test}	q^2_{loo}	q^2_{lmo}	$RSME_{training}$	$RSME_{test}$	R^2_{MP}[a]
2	0.504	0.478	0.415	0.416	0.690	0.679	0.034
3	0.518	0.505	0.397	0.434	0.680	0.638	0.030
4	0.511	0.507	0.392	0.416	0.685	0.661	0.063
5	0.517	0.533	0.377	0.435	0.681	0.627	0.020
6	***0.540***	***0.535***	***0.407***	***0.461***	***0.664***	***0.623***	***0.048***
7	0.527	0.530	0.396	0.450	0.673	0.627	0.055
8	0.541	0.535	0.406	0.456	0.663	0.625	0.058
9	0.549	0.534	0.397	0.460	0.657	0.623	0.041
10	0.540	0.537	0.399	0.454	0.664	0.622	0.032
11	0.540	0.535	0.395	0.456	0.664	0.629	0.037
12	0.543	0.536	0.413	0.466	0.662	0.625	0.061
13	0.514	0.521	0.373	0.413	0.683	0.652	0.054
14	0.524	0.534	0.386	0.423	0.676	0.645	0.059
15	0.537	0.531	0.427	0.453	0.666	0.647	0.026
16	0.522	0.519	0.390	0.426	0.677	0.651	0.030
17	0.521	0.522	0.387	0.422	0.678	0.654	0.029
18	0.522	0.520	0.387	0.423	0.677	0.654	0.057
19	0.533	0.511	0.411	0.445	0.669	0.659	0.045
20	0.490	0.517	0.331	0.361	0.699	0.642	0.047